Yearbook of Corpus Linguistics and Pragmatics

Jesús Romero-Trillo *Editor*

Yearbook of Corpus Linguistics and Pragmatics 2015

Current Approaches to Discourse and Translation Studies

Springer

Yearbook of Corpus Linguistics and Pragmatics

2015

More information about this series at http://www.springer.com/series/11559

Jesús Romero-Trillo
Editor

Yearbook of Corpus Linguistics and Pragmatics 2015

Current Approaches to Discourse and Translation Studies

Springer

Editor
Jesús Romero-Trillo
Departamento de Filología Inglesa
Universidad Autónoma de Madrid
Cantoblanco, Madrid, Spain

ISSN 2213-6819 ISSN 2213-6827 (electronic)
Yearbook of Corpus Linguistics and Pragmatics
ISBN 978-3-319-17947-6 ISBN 978-3-319-17948-3 (eBook)
DOI 10.1007/978-3-319-17948-3

Library of Congress Control Number: 2015946312

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Contents

Contributors

Vaclav Brezina ESRC Centre for Corpus Approaches to Social Science, Lancaster University, Lancaster, UK

Silvia Cacchiani Department of Language and Cultural Studies, University of Modena and Reggio Emilia, Modena, MO, Italy

Bart Defrancq EQTIS, Department of Translation, Interpreting and Communication, Ghent University, Ghent, Belgium

Francisco Javier Díaz-Pérez Departamento de Filología Inglesa, Facultad de Humanidades y Ciencias de la Educación, Universidad de Jaén, Jaén, Spain

Dmitrij Dobrovol'skij Russian Language Institute, Russian Academy of Sciences, Moscow, Russia

Ilka Flöck Institut für Anglistik-Amerikanistik, Fakultät III: Sprach- und Kulturwissenschaften, Universität Oldenburg, Oldenburg, Germany

Georgia Fragaki Department of Linguistics, University of Athens, Athens, Greece

Dana Gablasova ESRC Centre for Corpus Approaches to Social Science, Lancaster University, Lancaster, UK

Ronald Geluykens Institut für Anglistik-Amerikanistik, Fakultät III: Sprach- und Kulturwissenschaften, Universität Oldenburg, Oldenburg, Germany

Dionysis Goutsos Department of Linguistics, University of Athens, Athens, Greece

Raquel Lázaro Gutiérrez Departamento de Filología Moderna, Facultad de Filosofía y Letras, University of Alcalá/Group FITISPos-UAH, Alcalá de Henares (Madrid), Spain

Cédric Magnifico EQTIS, Department of Translation, Interpreting and Communication, Ghent University, Ghent, Belgium

Georg Marko Department of English Studies, Karl-Franzens-University of Graz, Graz, Austria

Silvia Molina Plaza Applied Linguistics Department UPM, Technical University of Madrid, Madrid, Spain

Koen Plevoets EQTIS, Department of Translation, Interpreting and Communication, Ghent University, Ghent, Belgium

Ludmila Pöppel Department of Slavic and Baltic Studies, Finnish, Dutch and German, Stockholm University, Stockholm, Sweden

María del Mar Sánchez Ramos Departamento de Filología Moderna, Facultad de Filosofía y Letras, University of Alcalá/Group FITISPos-UAH, Alcalá de Henares (Madrid), Spain

Jesús Romero-Trillo Departamento de Filología Inglesa, Facultad de Filosofía y Letras, Universidad Autónoma de Madrid, Madrid, Spain

Keiko Tsuchiya Foreign Language Centre, Tokai University, Tokyo, Japan

Elaine Vaughan School of Modern Languages & Applied Linguistics, University of Limerick, Ireland

Rachelle Vessey School of Education, Communication and Language Sciences, Newcastle University, Newcastle upon Tyne, UK

Richard Xiao Department of Linguistics and English Language, Lancaster University, Lancaster, UK

Current Approaches of Corpus Pragmatics on Discourse and Translation Studies, an Introduction

Jesús Romero-Trillo

The third volume of the series *Yearbook of Corpus Linguistics and Pragmatics* describes current approaches to discourse and translation studies. The chapters in the volume will undoubtedly be useful to scholars interested in translation and discourse analysis, but also to linguists who want to investigate new ways of applying pragmatic theories to the interpretation of new textual domains. For this purpose, the authors have employed a great variety of theories, registers, topics and corpus collection methodologies. The chapters include corpus data and analyses of several languages: English, Spanish, Greek, Chinese, Dutch, Russian, Galician, Swedish and Italian.

The volume is divided into three sections: Current Approaches to Discourse Studies, Current Approaches to Translation Studies, and a third section devoted to the review of recent relevant publications.

The first section, *Current Approaches to Discourse Studies*, opens with a chapter by Ilka Flöck and Ronald Geluykens entitled 'Speech Acts in Corpus Pragmatics: A Quantitative Contrastive Study of Directives in Spontaneous and Elicited Discourse'. The study compares the use of directives in three English corpora compiled through different methods: spontaneous spoken data (from the British component of the International Corpus of English), spontaneous written data of business letters, and elicited written data via Discourse Completion Tasks. The results show the existence of significant differences between the three types of corpus data. These differences lie not only in the directive act itself, but also in the accompanying modification strategies, i.e. downgrading and upgrading. Consequently, the resulting

J. Romero-Trillo (✉)
Departamento de Filología Inglesa, Facultad de Filosofía y Letras,
Universidad Autónoma de Madrid, 28049-Madrid, Spain
e-mail: jesus.romero@uam.es; www.jesusromerotrillo.es

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_1

speech acts convey different levels of directness. The implications for further research imply the revision of current methodologies, as corpus comparability is influenced by discourse genres and data collection methodologies.

The second chapter, authored by Silvia Molina Plaza, is called 'Black and White Metaphors and Metonymies in English and Spanish: A Cross-Cultural and Corpus Comparison'. In it, the author presents the metaphoric conceptualizations of black and white, in English and Spanish, based on data drawn from the British National Corpus and the *Corpus de Referencia del Español Actual*. The research focuses on the figurative meanings by relating multiword units to their various cultural contexts. Based on Piraiinen's taxonomy, the author analyses the implied cultural phenomena of the use of these terms qualitatively. The results show that the uses of *black/negro* representing 'bad, unhappy' and of *white/blanco* representing 'good, innocent' seem to be cultural facts in both cultures.

The next chapter, by Georg Marko, is entitled 'Making Informed Healthy Lifestyle Choices: Analysing Aspects of Patient-Centred and Doctor-Centred Healthcare in Self-Help Books on Cardiovascular Diseases'. The study offers a corpus-based Critical Discourse Analytical approach that examines the relation between doctor-centred and patient-centred elements in self-help books on cardiovascular diseases. The study investigates the speech act of advice as realized by acronyms and imperatives, concluding that self-help health promotion often tends to be doctor-centred rather than patient-centred, which can be considered a *contradictio in terminis*.

The fourth chapter, authored by Georgia Fragaki and Dionysis Goutsos, has the title 'Women and Men Talking About Men and Women in Greek'. The study explores the frequency and meaning distinctions of gender-related nouns in Greek for man and woman, and boy and girl. The data for the analysis is drawn from the spontaneous conversations in the Corpus of Greek Texts. The results show that speakers tend to talk about their group members, classified in terms of age and gender. Also, the study proves that Greek women tend to talk about men and women as specific persons, rather than as epitomes of their gender.

The last chapter in this section, by Dana Gablasova and Vaclav Brezina, is entitled 'Does Speaker Role Affect the Choice of Epistemic Adverbials in L2 Speech? Evidence from the Trinity Lancaster Corpus'. In their work, the authors investigate stance-taking strategies in the context of an examination of spoken English. Specifically, they present the use of epistemic adverbial markers like 'maybe', 'certainly' and 'surely'. The authors' contention is that these markers are not only employed to express speakers' degree of certainty towards a statement, but also to express speakers' position towards the addressees. Through the comparison of the expression of epistemic stance by candidates and examiners, the study underlines the importance of considering the pragmatic choices involved in this type of interaction, beyond the labels of 'native' or 'non-native' speakers of a certain language.

The second section of the volume, *Current Approaches to Translation Studies*, starts with Richard Xiao's chapter 'Source Language Interference in English-to-Chinese Translation'. The author departs from the notion of translational language

as a "third code" that differs from both source and target languages. Xiao's study investigates the "source language shining through" hypothesis by exploring source language interference in translations, at both lexical and grammatical levels. The texts supporting the study are English-to-Chinese translations from comparable and parallel corpora of the two languages. The results of the study of the two genetically distant languages are of critical importance in advocating the source language interference as a translation universal.

The next chapter, by Francisco Javier Díaz-Pérez, is entitled 'From the Other Side of the Looking Glass: A Cognitive-Pragmatic Account of Translating Lewis Carroll'. The author's intention is to analyse wordplay in the books by Lewis Carroll 'Alice's Adventures in Wonderland, and 'Through the Looking Glass and What Alice Found There'. The author avers that the 137 puns identified for the study pose a real challenge for translators, from a cognitive-pragmatic perspective. The study, which follows Sperber and Wilson's Relevance Theory, investigates the techniques used to translate wordplay in six different Spanish versions and in one Galician version. The results of the study show the importance of Relevance Theory in the analysis of translation alternatives of wordplay, and as a tool that opens avenues for future research in pragmatics.

Bart Defranq, Koen Plevoets and Cédric Magnifico contribute to the volume with a chapter entitled 'Connective Items in Interpreting and Translation: Where Do They Come From?'. Their study presents corpus-based research into the use of connective items by English and Dutch translators and interpreters. The aim of the investigation is to compare the role of connective items in translations and interpretations in relation to source texts. The corpus data is drawn from a corpus of interpretations and translations of the European Parliament. The results show, in the first place, that interpreters and translators differ in their strategies and, secondly, that interpreters omit more connective elements than translators. However, the data shows that interpreters use connective items to make clausal relations explicit and to connect clauses after omissions and to face processing difficulties.

The next chapter is authored by Dmitrij Dobrovol'skij and Ludmila Pöppel, and is entitled 'Corpus Perspectives on Russian Discursive Units: Semantics, Pragmatics, and Contrastive Analysis'. It analyses a group of Russian discursive units with focus-sensitive semantics such as *imenno* (just/precisely), *kak raz* (just/precisely), *to-to i ono* (that's just it/the point/problem), *to-to i est'* (that's just it/the point/problem) and *to-to i delo* (that's just it/the point/problem). The pragmatic functions of this group of units depend on the dialogic situation and can express agreement, disagreement, doubt, etc. Using relevant lexicographic information and text corpora, including parallel corpora and works of fiction, the authors attempt to clarify the semantic and pragmatic properties of these elements and usage peculiarities of the focus-sensitive discursive units. The authors also illustrate their position with the analysis of the systemic and translational equivalents of these expressions in English and Swedish.

The penultimate chapter, by Silvia Cacchiani, is entitled 'On Concluders and Other Discourse Markers in the Concluding Moves of English and Italian Historical Research Articles'. The author's aim is to study genre variation across English and

Italian research articles in history with a corpus-assisted approach. In particular, the study concentrates on conclu* and its lemmatizations, i.e., second-level summarizers and concluders, and how they interact with other discourse markers and with metadiscourse across moves. The results indicate that second-level discourse markers add extra meaning to their more general and fewer specific counterparts. In the author's opinion, variation across English and Italian in this regard can be accounted for following an interpersonal model of metadiscourse characterized by different strategies at the interactional level.

The last chapter of the second section is authored by Raquel Lázaro Gutiérrez and María del Mar Sánchez Ramos. Its title is 'Corpus-Based Interpreting Studies and Public Service Interpreting and Training: The Case of Interpreters Working in Gender Violence Contexts'. The chapter touches upon a very sensitive issue in our societies, that of violence against women. In order to tackle communication difficulties with foreign victims new mechanisms have been established, like Public Service Interpreting and Translation or Community Interpreting and Translation. The chapter presents a corpus with legal texts and real case interactions that will be used to train interpreters in gender violence contexts in Spain. The authors' contention is that this specific and delicate type of interpretation demands the accurate understanding of assistance protocols, as well as the key applicable legal terminology and procedures. The corpus data will be essential to understand the language of the victims from an intercultural pragmatic perspective.

The last section of the volume reviews relevant recent books of great interest to pragmaticians and corpus linguists. The first is written by Rachelle Vessey and reviews Zappavigna's volume (2012) entitled 'Discourse of Twitter and Social Media: How We Create Affiliation on the Web'. The second review is Elaine Vaughan's on Aijmer and Altenberg (2013) 'Advances in Corpus-Based Contrastive Linguistics. Studies in Honour of Stig Johansson'. The last review, written by Keiko Tsuchiya, is on Adolphs and Carter (2013) 'Spoken Corpus Linguistics: From Monomodal to Multimodal'.

Part I
Current Approaches to Discourse Studies

Speech Acts in Corpus Pragmatics: A Quantitative Contrastive Study of Directives in Spontaneous and Elicited Discourse

Ilka Flöck and Ronald Geluykens

Abstract This study compares directives in three different language corpora collected under different conditions: (1) spontaneous spoken data (taken from the British component of the *International Corpus of English*); (2) spontaneous written data (viz. business letters), and (3) elicited written data (collected through Discourse Completion Tasks). It is shown that there are significant differences between spontaneous and elicited data sets as well as between spoken and written natural data. These differences occur both in the so-called directive head act as well as in the modification strategies accompanying the head act (downgrading and upgrading), resulting in various levels of directness in the realization of directives in all three data sets. These results show the importance of quantitative comparative research not just across data collection methods, but also across discourse genres, based on corpora of authentic speech.

Keywords Directive speech acts • Methodology • Directness • Spontaneous spoken data • Spontaneous written data • DCT data

1 Introduction

The current paper is a comparative investigation into the production of directive speech acts in three types of data. By contrasting directives in three different data sets we will attempt to show that the type of corpus used for analyzing speech acts can greatly influence the obtained results. Put differently: language users' speech output, at least with regard to the realization of speech acts, varies depending on the contextual conditions under which these speech acts have to be produced. In examining a variety of data, we will attempt to address two significant research gaps in the current literature.

I. Flöck (✉) • R. Geluykens
Institut für Anglistik-Amerikanistik, Fakultät III: Sprach- und Kulturwissenschaften, Universität Oldenburg, Oldenburg, Germany
e-mail: ilka.floeck@uni-oldenburg.de; ronald.geluykens@uni-oldenburg.de

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_2

Empirical research into speech act production has hitherto focused to a large degree on the analysis of elicited speech, collected under controlled conditions. This is especially true in the area of contrastive pragmatics, where (following Blum-Kulka's et al. 1989 influential study) the use of Discourse Completion Tasks (DCTs) has long been the dominant data collection paradigm for quantitative studies. The analysis of uncontrolled, spontaneous data sets has so far been the exception rather than the norm, at least in the speech act based literature. To the extent that such studies exist, they tend to be qualitative rather than quantitative in nature, more often than not carried out within a conversation analysis framework.

Large scale quantitative analyses based on corpora of spontaneous discourse have been few and far between. The current paper attempts to address this first research gap by investigating speech acts in two separate spontaneous data sets: (informal) conversations and (formal) business letters. Analyzing not one but two corpora of spontaneous discourse allows us to compare quantitatively the effect of discourse factors (formal vs. informal) and production modes (spoken vs. written). Our starting point is the assumption that these contextual factors play a significant role in the production choices language users make, more particularly in their choices with regard to the directness level and politeness strategies involved.

The second research gap concerns the extent to which the data collection method influences speech act production. By comparing the two types of spontaneous speech mentioned above (conversations and letters) to speech acts produced in DCTs, we will be able to assess to what extent such elicited speech acts can be considered "natural". While there have been previous attempts at comparing spontaneous speech and DCTs in the literature (cf. the discussion in Sect. 3 below), such attempts have been fairly limited in nature, and at best concerned with a two-way comparison of DCTs and one other discourse type. By offering a three-way quantitative comparison, this paper hopes to break new methodological ground.

To sum up, then, we will take two hypotheses as our starting point: (i) speech act realization is dependent on the data collection method (elicited vs. non-elicited, spontaneous language use); and (ii) speech act production in spontaneous discourse is dependent on the contextual conditions of the discourse genre (e.g. written vs. spoken). Only a three-way comparison of different types of language corpora allows for the testing of these hypotheses.

The focus on directives as a testing ground is justified by the fact that first of all, directives have received a lot of attention in the contrastive pragmatic literature, thus making previous claims easier to verify or falsify. Moreover, directive speech acts occur with great regularity in the spontaneous data sets examined here, thereby facilitating quantitative comparisons. We loosely define directives here as attempts by the speaker/writer to get the hearer/reader to do something, without attempting a more sophisticated subclassification into requests, suggestions, and the like as there appears to be no objective basis for distinguishing between illocutionary subtypes.

The remainder of this paper is organized as follows. In Sect. 2, we will deal with some essential theoretical preliminaries to our analysis. First of all (Sect. 2.1), we will discuss the inherent methodological difficulties involved when trying to use automatic searches for the study of speech acts in existing language corpora.

Secondly (Sect. 2.2), we will examine the current state of empirical speech act research, which has been based on a variety of data collection methods. We will argue that the effect of such collection methods on the realization of speech acts has not been addressed adequately. In Sect. 2.3, finally, we will attempt to justify our choice of directives as a prime example of speech act to be analyzed. Section 3 discusses the methodology and data sets employed in our study. Section 4 contains the quantitative analysis; we will examine the realization of so-called head acts (the directive proper) as well as modification devices used for downgrading and upgrading the head act, and will also pay attention to the possible correlation between head act and modification, an area that has also been largely neglected in the literature. Section 5 will attempt to interpret these results from a more qualitative point of view.

2 Methodology in Speech Act Research

2.1 Speech Acts and Corpus Linguistics

Corpora have usually been developed with the aim of electronically accessing linguistic forms in large language data bases. It is therefore not surprising that corpus linguistic research has traditionally taken a form-to-function approach where linguistic forms (lexical items or morphosyntactic structures) constitute the basic starting points for corpus searches. Pragmatics, or more particularly speech act research, has traditionally taken the opposite route (function-to-form approach) in that the point of departure is a language function (e.g. a certain illocution) and the objective is to investigate its formal realizations. Language functions, however, do not easily lend themselves as a starting point for electronic searches in language corpora. While many corpora available today are tagged for parts of speech or even parsed for sentence structures, there are no corpora available that are tagged for individual illocutions or even illocutionary types. Consequently, in their study on compliments in the *British National Corpus* (BNC), Jucker et al. claim that speech acts "are not readily amenable to corpus-linguistic investigations" (2008: 273). The authors explain that speech acts are defined by their illocutionary force or their perlocutionary effect, neither of which can be searched for directly in a corpus. Speech acts can therefore only be identified electronically in language corpora when they appear in routinized forms or in conventionalized combinations with illocutionary force indicating devices (IFIDs).

By translating Manes and Wolfson's (1981) compliment formulae into abstract search strings, Jucker et al. (2008) partially overcome this problem and are able to retrieve compliments from the BNC automatically. The authors note however, that almost every query fails to have complete precision (searches for relevant patterns may generate forms that are functionally not equivalent to the speech act in question) and recall (searches may fail to find all instances of the speech act in the corpus).

Other pragmatic features that have been studied using automated corpus searches include, e.g. the speech act of thanking in the *Wellington Spoken Corpus* (Jautz 2008), discourse particles in the *London Lund Corpus of Spoken English* (Aijmer 2002), hedging in the *Limerick Corpus of Irish English* (Farr and O'Keeffe 2002), and non-minimal response tokens in the *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE, McCarthy 2003) and in CANCODE and the *Limerick Corpus of Irish English* (O'Keeffe and Adolphs 2008).

When language functions do not appear in routinized forms or in reliable combination with IFIDs (as is the case for directive speech acts), retrieving them with automated searches is either impossible or causes severe problems of precision and recall. The only remaining option to retrieve speech acts from corpora are manual searches of the corpus material (or what Kohnen 2008 in more elaborate terms calls a "genre-specific micro-analytic bottom-up" approach).

In order to distinguish the two kinds of corpus-driven research, Jucker (2009) differentiates terminologically between a "corpus approach" that is based on automated corpus searches and a "conversation analytic approach" that includes manual searches of (published) corpus material. Although such a "conversation analytic" approach to published corpus material might not be considered to be a "classic" corpus linguistic approach, it takes up the traditional corpus linguistic desideratum of increasing representativeness and reliability in linguistic research, in that it invites research to be done on the same material by a large number of researchers. In speech act research in general, however, data collection has usually been carried out using other methods.

2.2 Speech Acts and Different Methods of Data Collection

In empirical pragmatics, a variety of methods of data collection is available. In her classic article on data collection in pragmatics research, Kasper (2000) gives an overview of methods that ranges from the observation of naturally occurring discourse to eliciting language by different experimental procedures. In applying Clark and Bangerter's (2004) categories of methods of data collection to speech act research, Jucker (2009) distinguishes between three fundamental types of data collection tools: field, laboratory and armchair. While armchair approaches investigate participants' intuitions and attitudes about language use, field and laboratory approaches aim at studying actual language use. They differ, however, in the way language data are produced. While in laboratory approaches, language use is elicited by researchers (by employing role-plays or administering discourse completion tasks [DCTs]), field data are defined by the absence of such elicitation techniques. Field methods are therefore observational in nature; i.e. they require an authentic communicative intent by participants to produce language. The non-authentic character of language produced in DCTs also expands to the medium in which they are produced. Discourse completion tasks are predominantly administered in writing although the scenarios they contain are simulations of spoken language. It can

therefore be argued that DCTs are simulations involving two different dimensions: written vs. spoken language and elicited vs. authentic speech.

The respective (dis)advantages of laboratory and field data are straightforward. Laboratory methods give full variable control to the researcher, allow for exact comparability of different data sets, and can generate large amounts of data; however, participants use language without their own intrinsic communicative intent in fictional scenarios (cf. Jucker 2009: 1618). Field data, on the other hand, may pose the problem of the observer's paradox (Labov 1972) and are often difficult to compare to other data sets since variable control is low; however, (apart from possible but reducible observer effects) they do not interfere with participants' authentic communicative intents. The comparability problem for field data was at least partially solved by the advent of parallel corpora compiled with the specific aim of comparing different varieties of English (e.g. the *International Corpus of English*, or ICE) Nelson et al. (2002). The conversational parts of ICE are comparable in length, number of transcripts and informant demographics and therefore serve the purpose of comparability well.

Despite the wealth of data collection methods available, most studies concerned with the production of speech acts have made use of laboratory data, such as DCTs. Discourse completion tasks are production questionnaires traditionally administered in writing in which participants are asked to engage verbally in fictional scenarios. The situational context and interlocutor roles are usually described briefly and often a turn following the participant's response is also provided (the so-called rejoinder, cf. Kasper 2000). Since variable control and comparability of informants and data sets are important issues in contrastive pragmatics (both in cross-cultural and interlanguage contexts), the omnipresence of the DCT is understandable. However, a number of studies have indicated that speech acts elicited by DCTs differ from speech acts collected in authentic conditions (e.g. Wolfson 1981; Kasper 2000; Jucker 2009; Economidou-Kogetsidis 2013). Unfortunately, these studies do not provide us with conclusive results as to how the methodological influence manifests itself in the data. Beebe and Cummings (1996), for instance, comparing refusals in DCTs and telephone conversations, find that DCT data are less complex and more direct than naturally occurring conversations but elicit the same range of semantic formulae. Bodman and Eisenstein (1988) report that expressions of gratitude are less complex and shorter in DCTs than in field notes of naturally occurring conversations. Similarly, Schauer and Adolphs (2006) show that expressions of gratitude are sequentially more complex in naturally occurring situations than the speech acts elicited in experimental conditions (DCTs). In contrast, Yuan (2001) and Golato (2003) describe that compliment responses elicited by DCTs exhibit more turns while at the same time containing fewer markers of interaction than naturally occurring face to face conversations. The authors explain their findings by the absence of an interlocutor in questionnaire settings, which causes speakers to self-select if no response comes from an interlocutor.

When it comes to the linguistic strategies employed for encoding speech acts, the studies available also offer inconsistent results. While Beebe and Cummings (1996) report that DCTs elicit the same number of semantic formulae than can be found in authentic data, Hartford and Bardovi-Harlig (1992) find, in their study on rejections,

that DCT data contain fewer semantic formulae and status preserving strategies, and lack the extended negotiation strategies found in natural data.

Many authors argue that the responses elicited by DCTs differ structurally from authentic speech acts because the laboratory setting elicits social expectations rather than language forms that participants would actually use in natural conversation. Along these lines, Beebe and Cummings (1996: 80–81) argue that DCT data provide the researcher with "a good idea of the stereotypical shape of the speech act". They therefore claim that questionnaires do not only give the researcher control over situational and social variables, but also give an insight into the metapragmatic knowledge of informants. This claim is supported by Kasper, who argues that that production questionnaires are useful to reveal information about speakers' pragmalinguistic knowledge of the strategies and linguistic forms by which communicative acts can be implemented, and about their sociopragmatic knowledge of the context factors under which particular strategic and linguistic choices are appropriate (Kasper 2000: 329).

While several studies have been able to show that DCT data differ qualitatively from natural data, there is an unfortunate lack of research about whether, and if so to what extent, frequency distributions of data sets elicited in different methodological conditions diverge. To our knowledge, there are only a few studies that compare the influence of data collection methods from a quantitative point of view. Focusing on formulaic sequences in expressions of gratitude, Schauer and Adolphs (2006) compare expressions of gratitude elicited in DCT to those retrieved from the CANCODE corpus, and find differences in turn length and complexity and sequential patterns, with naturally occurring conversations being sequentially more complex than DCT data. It is open to debate, however, whether Schauer and Adolphs would have found different qualitative and quantitative differences between corpus and DCT data, had they not relied on the DCT formulae for their automated corpus searches but had searched a sample of the corpus manually. The procedure chosen serves the purpose of their study (i.e. the pedagogical application of corpus and DCT data) well but only provides limited evidence of how methodology influences the surface realization patterns of speech acts.

In a study on the development of a corpus consisting of task-based interactions of advanced EFL students, Pfingsthorn and Flöck (2014) raise the question of how directives elicited in task-based interactions compare to learner and native speaker DCT and conversational data. The authors compare the directness levels of the speech act in four data sets (learner task-based conversations, learner DCTs, native speaker conversations and native DCTs) and find that there are statistically significant differences in the distribution of head act strategies in the native speaker DCTs and the native speaker conversations. While participants in the naturally occurring conversations show a strong preference for direct strategies, the vast majority of directive head acts follows the patterns of conventional indirectness.

While Pfingsthorn and Flöck (2014) report on directives produced in naturally occurring conversations being significantly more direct than those elicited by DCTs, Economidou-Kogetsidis (2013) comes to the opposite conclusion. In her comparison of business telephone encounters and written DCT directives, she finds that

directives produced under naturally occurring conditions are significantly more indirect than under experimental conditions. She also finds quantitative differences in the usage of modifiers (lexical and syntactic). Despite those differences, the author (in line with Beebee and Cummings' 1996 and Schauer and Adolphs' 2006 findings) reports that the general distributional trends are similar if not identical for almost all pragmalinguistic strategies analyzed. She therefore concludes that the "WDCT data can indeed approximate natural data *to a certain extent*" (Economidou-Kogetsidis 2013: 34; original emphasis). However, the author also stresses that there are no conclusive results about the representativeness of DCT data and cautions that some of the quantitative differences between DCT and natural data are caused by the inherently different nature of the two data types.

2.3 Directive Speech Acts

The speech act class of directives comprises all illocutions that are "attempts (of varying degree (…)) by the speaker to get the hearer to do something" (Searle 1976: 13). As such, directive speech acts are the prototypical example of a so-called face threatening act or FTA (cf. Brown and Levinson 1987), as they interfere with the hearer's desire for freedom of action (i.e. their negative face). This alone explains why directive speech acts (or more specifically one subtype, requests) have been of high interest to researchers in cross-cultural and interlanguage pragmatics. The face threat inherent in directive speech acts makes them a good candidate for face work (i.e. the usage of politeness strategies in order to downplay the potential social dissonances associated with FTAs). It is therefore not surprising that there is a vast amount of studies in cross-cultural and interlanguage pragmatics dealing with the production of directive speech acts. We will not attempt to provide a concise overview of the literature on directives here, but will limit ourselves to the discussion of the most general patterns in this line of research.

The bulk of research on directive speech acts has either a cross-cultural or interlanguage pragmatic focus and is therefore contrastive in nature (cf. e.g. House and Kasper 1981; Blum-Kulka et al. 1989; Trosborg 1994; Breuer and Geluykens 2007; Barron 2008). This need for comparability of data sets has led to the somewhat unfortunate situation that most of the studies published in this research tradition make use of laboratory data exclusively (i.e. are based on DCTs).

Despite the fact that DCTs have been used for the study of many different varieties of English, the general distributional head act realization patterns are surprisingly similar. Following Blum-Kulka et al.'s (1989) influential coding scheme, the most frequently employed head act strategies in DCT-based studies is the so-called "query preparatory" strategy, in which speakers make reference to the preparatory condition of directive speech acts (i.e. they refer to the hearer's ability or willingness to comply with the directive). Prototypical head act strategies follow the pattern *Can you do X?* or *Could you do X?*. Direct strategies (such as imperatives) and indirect strategies (hints) are used to a much lesser degree in almost all DCT-based

studies. Frequency levels for direct realization strategies elicited by DCTs range from 42 % (Economidou-Kogetsidis 2013) to levels well below 10 % of all directive speech acts (e.g. Blum-Kulka 1989; Breuer and Geluykens 2007; Barron 2008).

The few exceptions of studies that do not rely on laboratory methods but use field directives produced by native speakers of English are set in business contexts (e.g. Vine 2009; Economidou-Kogetsidis 2013) and are often based on written data such as letters or emails (e.g. Geluykens 2011). Despite using different coding schemes, Geluykens (2011) and Economidou-Kogetsidis (2013) report on roughly comparable directness levels in their naturally occurring data sets. The most direct realization forms (imperatives, performatives, obligation, need and want statements) account for 42 % in Economidou-Kogetsidis' (2013) and for 49 % in Geluykens' (2011) business directives. For conventionally indirect and non-conventionally indirect strategies, differences in coding taxonomies prohibit comparisons.

3 Methodology

3.1 Data Sets and Data Collection

In order to analyze the effect of the data collection method on the frequency distribution of directive head act strategies, three data sets with 235 directive speech acts each were analyzed. The data sets were selected with a view to investigating the two dimensions of difference between naturally occurring discourse and DCTs: (1) elicited vs. non-elicited language use and (2) spoken vs. written medium (cf. Fig. 1). The corpora of informal conversations and written DCTs, respectively, are the two data sets that differ maximally with regard to those two dimensions. The third data set (business letters) is a hybrid category, in that the directives were produced with a genuine communicative intent (i.e. they are not elicited) but are produced in the

Fig. 1 Methodological properties of data sets

written medium. They share one feature with each of the other data sets and are therefore an excellent control group for testing whether elicitation method or the medium has a higher influence on the realizations of directive speech acts.

The directive speech acts in the conversational data set were retrieved from the conversational part of the British component to ICE in manual searches. The conversational part of each ICE corpus (s1a) consists of 100 transcripts of 2,000 words each and includes informal face-to-face and telephone conversations between participants of predominantly low social distance. The DCT data (a selection of the Breuer and Geluykens 2007 data set) were also elicited in scenarios of low social distance (i.e. the fictional characters knew each other fairly well) and low power relation (i.e. characters had equal status). The conversational and DCT data are therefore maximally comparable in terms of genre included and micro-social set up. The business letters were originally collected at the University of Antwerp, Belgium, as a part of the *Antwerp Corpus of Institutional Discourse* (cf. Geluykens and Van Rillaer 1995; see also Geluykens 2011). Due to confidentiality issues, there are no demographic information available for this data set. All three data sets include native speaker British English only and were collected in the same time span (i.e. the 1990s).

For analysis, 235 directive speech acts were selected randomly from each data set and categorized according to the coding scheme presented in Sect. 3.2 below.

3.2 Coding Scheme

The coding scheme employed here is an adapted version of Blum-Kulka et al.'s (1989) influential coding scheme, which is used in the majority of empirical studies on directive speech acts. It differentiates between the head act (i.e. "the minimal unit which can realize a request", Blum-Kulka et al. 1989: 275) and modification strategies that can occur head act internally or externally and serve either the function of downgrading or upgrading the illocution. Both the choice of directness level in the head acts and the modification devices can function as (im)politeness strategies. The more indirectly the head act is realized, the more options of non-compliance it leaves to the hearer and thus appeals to his/her negative face wants (cf. Brown and Levinson 1987). Reversely, the more directly a head act is realized, the more it threatens the hearer's negative face. Modification strategies function in a similar way from a facework perspective. The vast majority of modifiers appeal to the negative face wants of the hearer and therefore downgrade the face-threat involved in asking somebody to do something. However, there are also some strategies that refer either to the hearer's positive face (e.g. complimenting or positive assessments) or threaten the speaker's (own) positive face (e.g. by apologizing) and serve a downgrading function. All of the modifiers with a face-threatening or upgrading function are directed towards the hearer's negative face in that they reduce his/her options on non-compliance with the directive.

Following Blum-Kulka et al. (1989), three directness levels are distinguished on the head act level. They differ in the degree to which the illocutionary point is apparent from the locution: directly, conventionally indirectly or non-conventionally indirectly (referred to as 'indirect' in the following). The first category, direct strategies, include imperative or elliptical forms ('mood derivable'), performative utterances and utterances where the speaker's intention is directly derivable from the semantic meaning of the locution ('locution derivable'), for instance by employing modal expressions of obligation. Secondly, in conventionally indirect strategies, the illocutionary point is only derivable from the locution by means of conventionalization. Searle (1975: 76) suggests that some linguistic forms become "conventionally established as the standard idiomatic forms for indirect speech acts". While they keep their literal meanings "they will acquire conventional uses as, e.g. polite forms for requests" (Searle 1975: 76). Linguistically, this can be achieved by questioning or referring to the hearer's ability or willingness to comply with a directive ('preparatory' strategy) or by using a routine formula closely associated with a specific genre or different illocutions ('conventionalized formula'). Indirect head acts, finally, are defined by the absence of a direct reference to either the action desired by the speaker or the person whom the speaker wishes to carry out the action. Table 1 provides an overview of and actual examples for the head act strategies and superstrategies identified in the present study (bold letters indicate the characteristics of the head act strategies).

On the modification level, we focus on the functional differences between downgrading and upgrading modifiers and do not take into account structural differences (i.e. whether a modification device occurs head act internally or externally). The modifiers found in our three data sets are listed with examples in Table 2 (bold letters indicate the modifier in question).

Table 1 Head act strategies and superstrategies

	Strategy	Example
Direct	Mood derivable	Hey Tom, **get me some bread** would you. (DCTs_006)
	Performative	We **do ask** that (…) we receive a copy of the literature for our own records. (Letters_063)
	Locution derivable	We **should** maybe just leave a message here saying head over (Convers_090)
Convent. indirect	Conventionalized formula	**Why don't you** uhm replace one of the back doors here (…) (Convers_014)
	Preparatory	I know it's a pain, but **can you come in and water my plants?** (DCTs_159)
Indirect	Hint	You (…) wouldn't happen to leave it lying around on the table would you (Convers_179)

Table 2 Modification strategies (cf. also Blum-Kulka et al. 1989)

	Strategy	Function	Example
Downgrading	Apologizing	Speaker (S) apologizes for the imposition involved in the directive	**Sorry to be a pain**, but would you please water my plants while I'm away. (DCTs_118)
	Condition	S limits the validity of the directive to a specific condition to be met	**If you are in agreement with the above balance**, please use the reply slip to confirm this. (Letters_028)
	Downtoner	S invokes an irrealis state and thus modulates the impact of the directive on H	Mel, can I **possibly** borrow a pen? (DCTs_200)
	Modal past	S invokes an irrealis state and thus increases H's chances to opt out	We **could** go out at ten o'clock or something for a drink. (Convers_134)
	Pre-grounder	S gives reasons for the directive (strategy appears before head act)	**It's freezing**. Can we go upstairs? (Convers_146)
	Post-grounder	S gives reasons for the directive (strategy appears after head act)	Cathy, could I borrow your car **coz I haven't got any transport.** (DCTs_076)
	Positive evaluation	S positively evaluates either H or H's (future) actions	Could you water my plants? (…) **you're the only person I trust with my spare key**. (DCTs_133)
	Politeness marker	S uses the politeness marker *please*	**Please** could you give me details of account no. NMBR deposit provisions. (Letters_025)
Upgrading	Consequences	S specifies possible consequences that non-compliance could have	Don't tell me to keep talking **or I'm going to keep quiet** (Convers_098)
	Intensifier	S intensifies certain elements of the proposition	In which case you ought to be doing some phonetics **surely** (Convers_018)
	Time intensifier	S intensifies the temporal dimension of the proposition	Please call me **as soon as you can**. (Letters_054)

4 Quantitative Analysis

4.1 Comparison of Head Acts

Our first contrastive analysis concerns the form of the head act (Blum-Kulka et al. 1989; see Table 1 above). Direct directive speech acts include –at least in our data– three different types of head acts. The first are mood derivables, in which "the grammatical mood of the locution determines its illocutionary force as a request" (Blum-Kulka et al. 1989: 278). The prototype examples for this category are imperatives, which occur frequently in the spontaneous data, both in the letters and the conversations (see below). The second category consists of performatives (including both explicit and hedged performatives), which for the purposes of this paper we have defined relatively broadly: we have not only counted the verbs *request* and *ask* as marking the illocutionary force of a request, but also all other performative verbs that may contain a directive force (e.g. *suggest, require, urge, propose*). Locution derivables, the third category, are formulations in which the "illocutionary intent is directly derivable from the semantic meaning of the locution" (Blum-Kulka et al. 1989: 279). Examples include the use of modal verbs expressing obligation or necessity (e.g. *you must/have to/ought to…*, *we need you to do* x…). As can be seen in Fig. 2 below, direct requests are significantly more frequent in both the naturally occurring data sets (conversations and letters) than in the non-natural data (DCTs). In fact, in the DCTs, a mere 5 % of all request head acts are realized as a direct request.

Conventionally indirect request, the second main category, include two main types of realization, the first being 'conventionalized formulas', in which the "illocutionary intent is phrased as a suggestion by means of a framing routine formula"

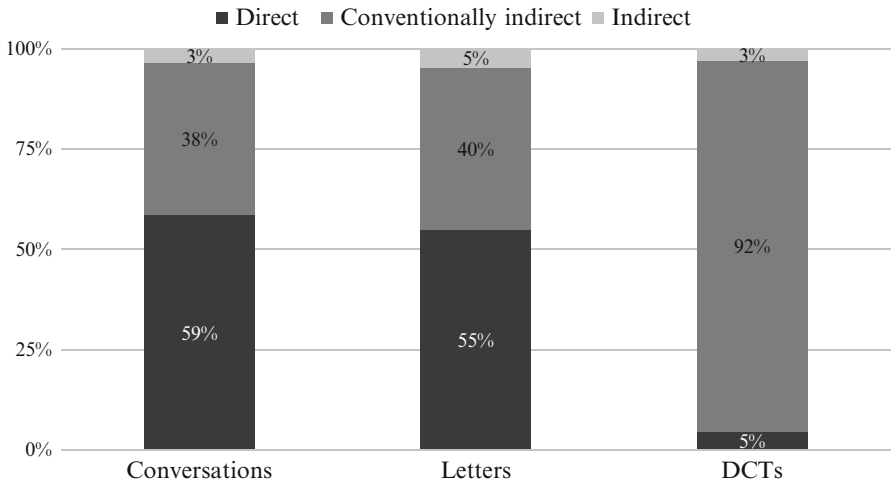


Fig. 2 Distribution of directness levels among the three data sets

(Blum-Kulka et al. 1989: 280 call these 'suggestory formulas') and genre-specific formulas. What are referred to here are routinized expressions such as *let's* and *how about* as well as some more genre specific formulas that appear predominantly in the letters, such as *we await your* X and *I/we (would) appreciate* X. Such formulas are virtually absent from the DCTs, but occur with some regularity in the other two data sets (see Table 3 below). The second category of conventionally indirect directives consists of the preparatory strategies, which are "utterances contain[ing] reference to a preparatory condition of the Request, typically one of ability, willingness, or possibility" (Blum-Kulka et al. 1989: 280), in particular query preparatories such as *can you/could you*. The category 'preparatory' also contains utterances with declarative sentence type of the structure *you can/could*. While in all data sets the interrogative form is prevalent, there are clear differences in the distribution of sentence type among preparatory strategies. The declarative form is much more frequent in the naturally occurring data sets (letters and conversations; with numbers ranging from 39 to 42 % of all preparatory strategies) than in the DCTs (11 %). This pattern becomes even more striking when one considers that 22 out of the 23 declarative preparatories in the DCTs occur in negative declarative structures with directive tag (cf. Carter and McCarthy 2006) of the type *You couldn't do* X*, could you?* or *You wouldn't do* X*, would you?*. This structure only appears once in the conversational data set and is not used at all in the business letters.

Since all preparatory strategies are conventionalized to a high degree, they cannot be considered off-record. Such preparatories are by far the most common directive strategy on the DCT data, but occur significantly less often in the natural data sets (see Table 3 below). This is also true for conventional indirectness is general: 92 % of all requests are realized in this manner in the DCTs (see Fig. 2 above).

The third main directive strategy consists of indirect strategies, or hints. Blum-Kulka et al. (1989: 280) make a distinction between 'strong' and 'mild' hints, depending on the degree to which the locution "refers to relevant elements of the intended illocutionary and/or propositional act" or not. We do not think such a distinction is tenable, however, for (at least) two reasons. First of all, the criteria whether or not certain locutionary elements are "relevant" [sic] to the illocution appears quite subjective, if not arbitrary. Secondly, even if tenable, this distinction appears to be gradable rather than binary, in that (assuming relevant elements can be

Table 3 Distribution of head act strategies across data sets

		Conversations		Letters		DCTs	
	Head act strategy	n	%	n	%	n	%
Direct	Mood derivable	96	40.9	72	30.6	8	3.4
	Performative	2	0.9	26	11.1	0	0.0
	Locution derivable	40	17.0	31	13.2	3	1.3
Convent. indirect	Convent. formula	22	9.4	50	21.3	0	0.0
	Preparatory	67	28.5	45	19.1	217	92.3
Indirect	Hint	8	3.4	11	4.7	7	3.0
TOTAL		***235***	***100***	***235***	***100***	***235***	***100***

identified in the first place) such elements are probably not all relevant to the same degree. What all hints have in common, however, is that they are off-record, i.e. the speaker/writer does not provide any formal clues as to what the illocutionary force might be, but the latter needs to be worked out by the hearer/reader through a conversational implicature. On this basis, one might even argue that the concept of 'strong hints' is a contradiction in terms, for whenever an utterance contains "relevant" clues as to the intended speech act status, the speaker may be considered to have gone on-record (and therefore not to have hinted). For the present study we have used an operational definition, in that for a directive to count as a hint, neither the action desired by the speaker nor the agent of the action may be mentioned explicitly in the utterance. As was shown in Fig. 2 above, hints are very infrequent anyway, compared to the other main types of strategies (direct and conventionally indirect).

As an interim conclusion, then, we can say that both types of naturally occurring data prefer direct head act strategies (59 % in the conversations, 55 % in the letters), while the controlled-elicitation data (DCTs) show a very strong preference for conventionally indirect strategies (92 %). While the slight difference in head act directness levels within the natural data sets (conversations and business letters) are not statistically significant ($\chi^2 = 0.97$, dF = 2, $p > 0.1$), the differences between natural and controlled data sets are highly significant ($\chi^2 = 161.86$, dF = 2, $p < 0.001$ for the comparison between conversations and DCTs and $\chi^2 = 148.05$, dF = 2, $p < 0.001$ for the comparison between letters and DCTs). In short, while all three data sets exhibit different request realizations, the two sets of natural data are remarkably similar as to the type of head act used.

However, the figures for the three main categories of head act as shown in Fig. 2 obscure some striking differences with regard to the actual strategies employed, as can be confirmed with even a quick glance at Table 3. Within the category of direct strategies (which, as already stated, occur only very rarely in the DCTs), imperatives are slightly more frequent in the conversations (40.9 % of all directives) than in the letters (30.6 %). Conversely, performatives occur more frequently in the letters (11.1 %) than in the conversations (0.9 %). Locution derivables do not occur in the DCTs at all, but they do show up, with similar frequencies, in the conversations (17.0 %) and the letters (13.2 %).

The differences are even more outspoken when turning to specific strategies within the conventionally indirect category. As we have mentioned already, this is by far the most frequent strategy in the DCTs (92.3 %, cf. Fig. 2); what is more, all such head acts in the DCTs are realized through a so-called 'preparatory' strategy as in Examples 1 and 2:

(1) Can you give me a lift to my friend's birthday party today? (DCT_071)
(2) John, can I possibly borrow your car this evening? (DCT_065)

In short, head acts in the DCTs tend to be realized uniformly in the same manner: fewer than 8 % of directives in all is produced through some other type of head act. The naturally produced data show a lot more variation with regard to the type of conventionally indirect strategy used. Preparatory strategies account for about half

of the conventionally indirect directives in the letters, and about 3 out of 4 in the conversations. This leaves the strategy of conventionalized formulas, which does not occur in the DCTs, but is more frequent in the conversations (9.4 % of all head acts) and more frequent still in the letters (21.3 %). What is more, whereas the speakers in the conversations produce non-genre specific formulations of the types mentioned by Blum-Kulka et al. (1989) (Examples 3 and 4), the letters typically use highly genre-specific strategies (Examples 5 and 6):

(3) <#123:1:C>Let 's have a good uh<#124:1:C>So **let 's play** Trivial Pursuit as well after or something (Con_148; s1A048)
(4) <#273:1:A>I 'm trying so hard to concentrate on this<#274:1:C>Well **why don't you give up for** five minutes <,,> (Con_083; s1a038)
(5) **I await your further thoughts** on the subject. (letters_117)
(6) Your assistance **is** greatly **appreciated**. (letters_176)

One could even argue that the latter strategies (i.e. genre-specific formulas) constitute a category of conventional indirectness in their own right. The fact that such strategies are not mentioned at all in Blum-Kulka et al. (1989) show, once again, that it is very dangerous indeed to develop a categorization based on just one type of data (DCTs in this case), especially since the data used by them are elicited under controlled conditions rather than spontaneously produced and are based on one genre exclusively.

What seems clear from the results discussed so far is that spontaneous data exhibit a wider range of directive strategies than DCTs, which makes categorization much more difficult.

There is little more that can be said here in general terms about indirect requests, or hints, due to the fact that the addressee needs to work out the directive force here based on the contextual conditions. As a result, what counts a as a directive in one context may (and usually would) fail to do so in another context. Since hints are context-specific, one would of course also expect them to be genre-specific: this is borne out by the vastly different formulations found in the conversations and the letters (Examples 7 and 8).

(7) Flake and refined would also be of interest. (letters_145)
(8) <#19:1:C>I knew I know the phone number of the chap uhm (…)<#21:1:B>Yeah<#22:1:B>**But what I need is a personal intro to him** (con_060; s1a027)

Since our database of letters is of a formal nature (business letters), and the conversations are mostly informal, the formality level probably plays a part in the realization of the directive, but much more research is needed here.

A number of conclusions can already be drawn from this brief investigation into directive head acts, however. First of all, directives elicited through DCTs are on the whole more indirect than spontaneous directives: both types of naturally occurring directives exhibit direct strategies to a far higher degree. Secondly, the elicited data overwhelmingly show conventionally indirect directives, specifically realized through preparatory strategies. As a result, most non-spontaneous directives show

far less variation as to the realization of the head act than the spontaneous ones (even within the category of conventional indirectness, but also within the category of direct directives). Thirdly, hints are by far the least preferred strategy overall. In short, all three data sets show a different picture for the realization of head acts, but the controlled-elicitation data differ most strikingly from the spontaneous data. We will return to this issue further on.

4.2 Comparison of Modification Strategies

The comparison of head acts in directives does not, of course, tell the whole story as to their level of directness. In order to get a more complete picture, we also need to look at how this head act is potentially modified on other levels of the utterance. A speaker/writer may, for instance, choose to downgrade the directness of a head act through certain lexical or syntactic devices, thereby making the directive less direct than it would have been. A typical example of lexical downgrading is the use of the politeness marker *please*. An example of syntactic downgrading would be the use of a modal past. As an example of yet another type of downgrading, the speaker/writer may provide, the reasons for producing the directive. This strategy, which Blum-Kulka at al. (1989) refer to as 'grounders' (a terminology that we will borrow here; see also Table 2 above), refers with high regularity in all our data sets.

Conversely, a speaker/writer could potentially also upgrade a head act, i.e. increase the level of face threat implied by the directive by making it seem urgent, for instance, as in (9) below, or by spelling out the negative consequences of not complying with the request, as in (10):

(9) Please call me as soon **as you can**. (letters_054)

(10) <#242:1:A>Keep talking with me (…)<#244:1:B>Don't tell me to keep talking **or I'm going to keep quiet** (con_098; s1a041)

While such strategies might seem unexpected, in that they raise rather than lower directness, they do occur more regularly than one might expect.

In what follows, therefore, we will look at various types of modification of the head act, both downgrading and upgrading; we will not make any formal distinction between lexical and syntactic modification. It needs to be pointed out here that we will NOT go beyond utterance boundaries here, and only consider modification strategies to the extent that they occur either inside the head act or immediately following or preceding it. We are very much aware that this does not show the whole picture as to the modification of directive speech acts, since downgrading and upgrading may also be done on a broader (con)textual level, i.e. in the wider discourse surrounding the head act (and not necessarily adjacent to it). Investigating such so-called "supportive moves" outside the utterance would, however, lead us too far here; for a full discussion of different types of such strategies occurring in the letters sub-corpus we refer the reader to Geluykens (2011: 66–88).

The first order of business, then, should be to examine which upgraders and downgraders occur in the data, and to investigate the extent to which the three data sets differ quantitatively in their actual usage. Another, related question we need to ask concerns the possible correlations between modification and head act. In other words, do certain modification strategies typically occur with certain head acts and, if so, what types of patterns can be identified? We will examine the latter in the following section, and concentrate here on downgrading and upgrading per se.

In the interest of succinctness, we will focus here exclusively on those modification strategies that occur with some regularity in at least ONE of the data sets. The following, therefore, is not an exhaustive overview of all modification but does cover all the most frequent types. Figure 3 shows the absolute frequencies of downgraders and upgraders in the three types of data. While the number of head acts is identical in each data set (n = 235), the number of modifiers is quite dissimilar.

It is clear from this figure that downgrading occurs most often in the DCTs (353 tokens), and least often in the conversational data (121 tokens), with letters occupying the middle ground (232 tokens); what is striking here, then, is that no two data sets show outspoken similarities (the differences between all data sets are statistically significant: Letters and DTCs: $\chi^2 = 34.22$, dF = 1, $p < 0.001$; conversations and DCTs: $\chi^2 = 7.49$, dF = 1, $p < 0.01$; conversations and letters: $\chi^2 = 4.33$, dF = 1, $p < 0{,}05$). Upgrading, as expected, is far less frequent than downgrading. However, it is far more frequent, relatively speaking, in the business letters (38 tokens) than in either the conversations or DCTs. In short, all three data sets exhibit different modification behavior, at least from a purely quantitative point of view. This is in itself already a remarkable finding, which once again not only shows the importance not only of analyzing natural data, but also of doing so across discourse genres.

Once one starts examining the type of modification strategy in more detail, according to a slightly modified version of the categorization scheme introduced in Sect. 3.2 above (cf. Table 4), more differences emerge.

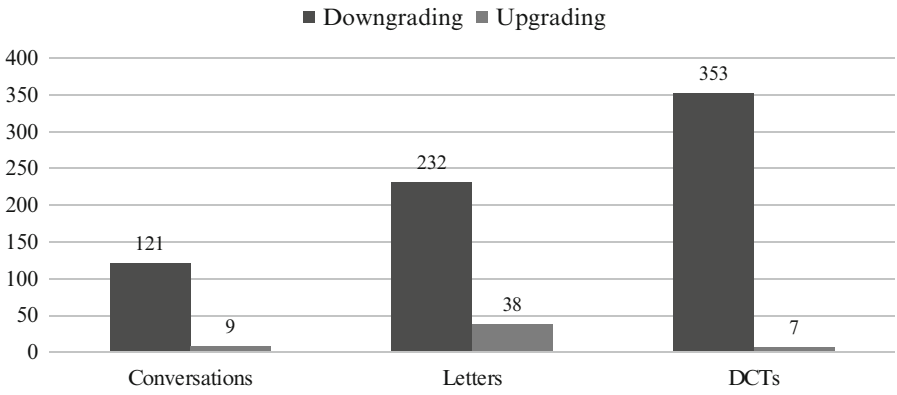


Fig. 3 Distribution of downgrading and upgrading modifiers

Table 4 Distribution of modification strategies (in absolute numbers)

Function	Modifier	Conversations (235 directives)	Letters (235 directives)	DCTs (235 directives)	
Downgrading	Modal past	22	64	126	212
	Politeness marker	12	83	49	144
	Downtoner	31	9	13	53
	Positive evaluation	5	4	40	49
	Post-grounder	32	25	23	80
	Pre-grounder	6	16	29	51
	Apologizing	4	0	9	13
	Condition	9	31	64	104
	Subtotal	*121*	*232*	*353*	*706*
Upgrading	Intensifier	6	7	6	19
	Time intensifier	1	29	0	30
	Consequences	2	2	1	5
	Subtotal	*9*	*38*	*7*	*54*
TOTAL		***130***	***270***	***360***	***760***

Within the categories of down- and upgrading, we can distinguish between functionally different subtypes. All of the downgrading strategies can be classified as politeness strategies within the classic politeness framework of Brown and Levinson (1987). They can be subclassified as to whether they appeal to the hearer's (or in fact the speaker's) positive or negative face. The most frequent type of downgrading appeals to the hearer's negative face wants in that the speaker shows awareness of the hearer's desire to be free in action and choses linguistic strategies that give the hearer more freedom of choice (at least on the surface). This is to be expected, since directive speech acts primarily threaten the hearer's negative face in that they limit his/her freedom in action. A very common use of this kind of downgrader, especially in the DCTs, is the use of a modal past (*could*/*would* instead of *can*/*will*), as in Example 11. By employing the modal past (which is an optional choice) the speaker makes the action desired appear as an irrealis state and therefore linguistically provides the hearer with an opt-out.

(11) Perhaps you **could** let us know what date and time would be most convenient to you? (letters_190).

Downtoners are yet another way of limiting the imposition posed on the hearer by the directive speech act. They predominantly occur in the form of sentential or propositional modifiers (as in Example 12) and are most frequent in the conversational data.

(12) <#182:1:B>**Maybe** we should head there and then **just** head for the taxi queue or **just** walk in from there <,> without taking our bikes in case we meet her <,> (con_091; s1a031)

Other downgrading strategies appealing to the hearer's negative face include the giving of conditions pertaining to the directive (Example 13). Here, the hearer is

only asked to comply with a directive if a certain condition is met. This leaves the hearer an explicit choice to opt out and therefore increases his/her freedom of action.

(13) We do ask that **where our photos are used for promotional purposes**, that we receive a copy of the literature for our own records. (letters_063)

This is quite frequent in the letters and DCTs (i.e. in the written data), but less so in the conversations.

In positive politeness strategies, the speaker appeals to the hearer's desire to be liked and respected. The prototypical linguistic realization is that of positively evaluating either the hearer or the hearer's (future) actions, as in Example 14:

(14) I know it's a hassle but would it be at all possible for you to water my plants as I am going away on holiday. **It would be a great help**. (DCT_139)

Here, the speaker positively evaluates the hearer's prospective action as helpful and thereby tries to make the hearer more inclined to comply with the directive produced. Positive evaluations are not very frequent in the spontaneous data sets overall but occur with some regularity in the DCTs (n = 40).

Another, much more frequent, positive politeness strategy is that of providing reasons why the speaker wants the hearer to do something (grounder). Grounders are common in all three data sets. They may occur prior to the request head act, as in (15), or following it, as in (16):

(15) **My memory of the exact structure of the secretarial course you run has become a little hazy**, so I'd be grateful if you could supply me with a brief syllabus (…). (letters_208)

(16) <#414:1:A>Well what about Monday <,> The M C L Christmas party <,> You ought to go Louisa. **People keep saying where 's Louisa** (con_096; s1a040)

As in most other studies on directive speech acts (cf. e.g. Faerch and Kasper 1989; Breuer and Geluykens 2007; Barron 2008), grounders are one of the most frequent types of modification in the present study. While in both field data sets, post-grounders (i.e. grounder occurring after the head act) are prevalent, the DCT data show a different pattern. Here, pre-grounders are more frequent than grounders after the head act. A more thorough look at the pre-grounders in the DCTs reveals that they seem to carry a slightly different function than post-grounders. While post-grounders only provide the addressee with the reasons why the speaker wants a certain action to be carried out, pre-grounders in DCTs also serve to establish a context and context in which the ensuing directive is embedded (cf. Example 17); the lack of context and context in the DCTs might therefore trigger the higher numbers of pre-grounders in this particular methodological condition:

(17) **I'm going on holiday for a few weeks** and I was wondering if you could possibly water my plants. (DCT_137)

One of the most frequent downgraders in all data sets is the politeness marker *please*. The function of *please* has been discussed widely (and somewhat controversially) in the pragmatics literature. It can be used as a lexical downgrader, but also

as an illocutionary force indicating device (IFID) serving the opposite function in that it makes the directive force of the speech act more transparent (cf. Sadock 1974; House 1989). In her analysis of *please* in the *London-Lund Corpus of Spoken English*, Aijmer (1996: 166) finds that *please* is predominantly found in situations in which formal politeness is required. This pattern is also confirmed by the distribution of the politeness marker in the present data sets. While the conversational data (as the most informal data) show the lowest occurrence of please (n = 12), the number of occurrence rises with the increasing formality of the data sets (49 in the DCTs and 83 in the business letters). An example:

(18) **Please** let us know on the attached form how many copies of NAME you require in YEAR and return the form by DATE. (letters_004)

Apologizing for making the request, the final type of downgrader, does not occur very frequently, and is used only in the conversational and DCT data:

(19) **Sorry to be a pain**, but would you please water my plants while I'm away. (DCT_118)

The relative frequencies of downgrading strategies discussed thus far do not, however, tell the whole story. If one examines in detail the amount of downgrading associated with each head act, in other words how many of the above-mentioned modifiers occur per head act (Fig. 4), a striking pattern emerges. In the conversational data, most head acts (58 %) are not modified at all. In both letters and DCTs on the other hand, the majority of head acts are accompanied by one or two modifiers (65 % and 67 %, respectively). Some DCT head acts even carry three or four downgraders.

We can conclude from all this, first of all, that both frequency and type of downgrading used for directives are genre-specific to some degree. Overall, downgrading occurs most often in DCTs. Since we already observed, in the previous section, that DCTs mostly use conventionally indirect strategies, this cannot but result in directives

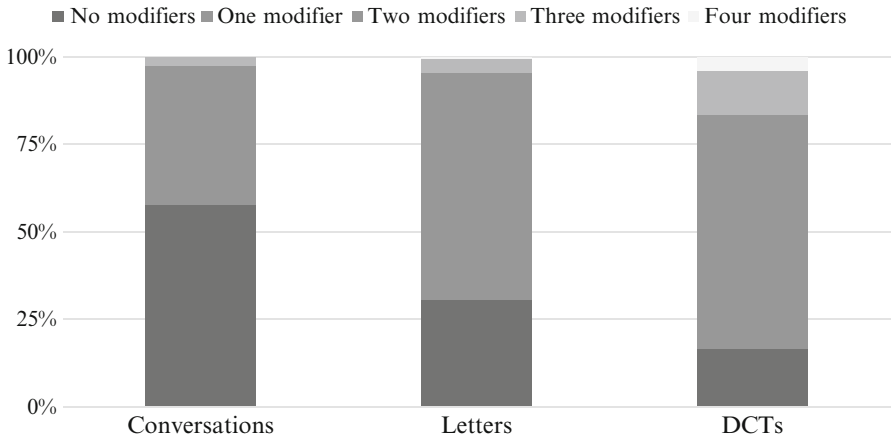


Fig. 4 Number of downgrading modifiers per head act

being far less direct on the whole than in the spontaneous data. Both types of spontaneous data also differ in that downgrading is significantly less frequent in conversations, making the directives in this genre potentially the most direct of all.

As shown in Table 4, upgrading is significantly more frequent in the letters than in the other data (38 tokens, or 16 per 100 directives, as opposed to a mere 9 and 7 in conversations and DCTs, respectively). Table 4 also shows that here, too, the types of upgrader used differ across the three data sets. In business letters, by far the most common type of upgrader is a time intensifier, as in:

(20) Please respond **as soon as possible** to COMP and let them work out any problems with COMP. (letters_222)

Given the fact that directives in a business context are often time-sensitive, this is not surprising: compliance in a given timeframe is often essential, making the directive pointless otherwise. One could thus argue that this type of upgrading is unavoidable rather than optional in this context.

The only other two types of upgraders that occur with any frequency in the data (about equally in the three data sets, incidentally) are other types of intensifiers, as in (21) and warnings about the consequences of non-compliance, as in (22):

(21) In your opinion, should any material weakness exist (…), please **be sure** to indicate such weakness in your response (…). (letters_139)
(22) <#244:1:B>Don't tell me to keep talking **or I 'm going to keep quiet** (con_098; s1a041)

The latter, however, only occurs five times overall, which is understandable, given the highly face-threatening nature of this particular upgrade.

4.3 Correlation of Head Acts and Modification

The final piece of empirical analysis that remains to be discussed here is the potential correlations between the types of head act employed and their modification devices. Within the scope of this paper, we can only provide a sample of analysis of one such correlation. As an illustration of why such an approach would be a useful enterprise, we will look at one particularly frequent downgrader, viz. the politeness marker *please*, and investigate the type(s) of head act it correlates with.

Before embarking on this, let us first have a look at how downgraders are distributed across head act categories. Figure 5 gives an overview of the relative number of downgraders across the two most frequent head act supercategories, direct and conventionally indirect directives. Due to their low overall numbers of occurrence, indirect head acts were not included in the analysis. We remind the reader (i) that downgraders are, on the whole, significantly less frequent in conversational discourse, and most frequent in DCTs, and (ii) that DCTs contain significantly fewer direct head acts than the spontaneous data. It is therefore important to look at relative rather than absolute frequencies here.

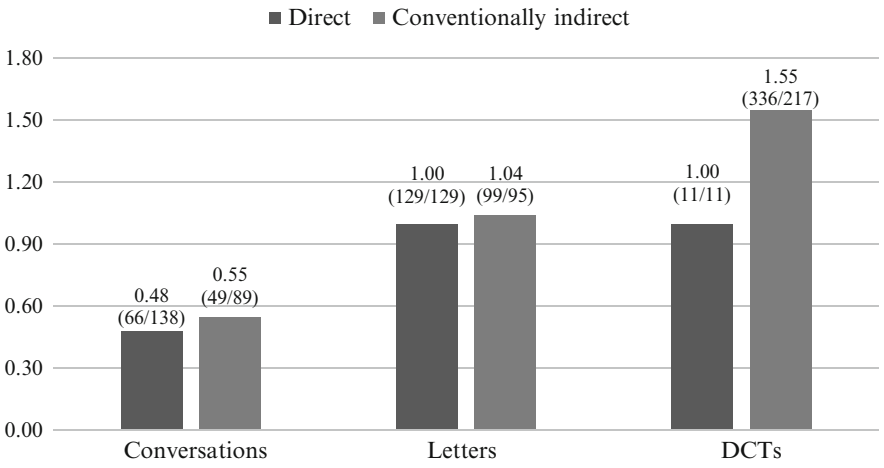


Fig. 5 Average number of downgraders per direct and conventionally indirect head act

The results are perhaps somewhat surprising: all things being equal, the ratio of downgraders per direct and conventionally indirect head act is very similar in the conversational and business letter data (0.48 and 0.55 downgraders per conversational head act and 1.00 and 1.04 downgraders per business letter head act). In other words, despite the difference in absolute numbers of the occurrence of downgraders in conversational and business letter directives, the relative frequencies among direct and conventionally indirect head acts does not differ significantly. However, this trend is different for the DCT directives. While every direct head act is modified by 1.00 downgrader, one conventionally indirect head act on average contains 1.55 downgraders. This is slightly counterintuitive, since one might expect more direct strategies to be in need of more downgrading. Clearly, some other factors are at work here, which clearly merit further investigation. It is clear once again, however, that whatever the reason for these tendencies, it would be dangerous to extrapolate results for one data-collection method or one discourse genre to other methods and/or genres.

In order to show variation in modification in more detail, let us now examine the distribution of *please*, in the first instance across head act types (Fig. 6).

This figure already shows that *please* is very unevenly distributed. Out of 83 occurrences of *please* in the letters, 62 (or 75.9 %) can be found accompanying direct strategies, whereas in the DCTs nearly all instances of *please* (n=47 or 95.9 %) co-occur with conventionally indirect ones. As pointed out in the previous section, *please* is rare in conversational discourse. Note also that it never co-occurs with hints (indirect head acts), which is to be expected, as this would rather defeat the purpose of using an off-record formulation in the first place, assuming that *please* is also a directive marker.

When looking at individual head act strategies, an even clearer usage pattern of *please* emerges (Table 5 provides an overview of the frequency of usage of *please* within all head act strategies). The politeness marker is predominantly used in imperatives and preparatory strategies, and only very infrequently used in locution derivable strategies, with merely one occurrence in a conventionalized formula.

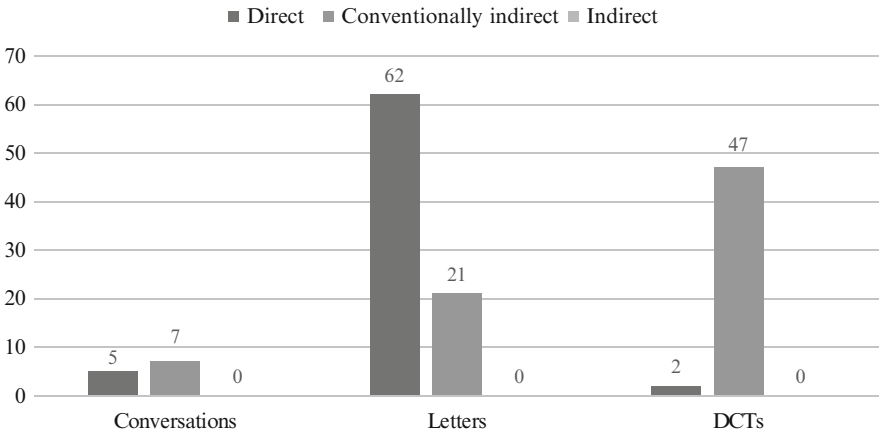


Fig. 6 Co-occurrence of the politeness marker *please* with head act superstrategies

Table 5 Co-occurrences of head act strategies and politeness marker *please*

	Head act strategy	Conversations	Letters	DCTs	*TOTAL*
Direct	Mood derivable	5	57	1	*63*
	Performative	0	0	0	*0*
	Locution derivable	0	5	1	*6*
	Subtotal	*5*	*62*	*2*	*69*
Convent. indirect	Conventional. formula	0	1	0	*1*
	Preparatory	7	20	47	*74*
	Subtotal	*7*	*21*	*47*	*75*
Indirect	Hint	0	0	0	*0*
	Subtotal	*0*	*0*	*0*	*0*
	TOTAL	***12***	***83***	***49***	***144***

These findings are in line with Aijmer's (1996) analysis of *please* in the *London-Lund Corpus of Spoken English*. The author (1996: 166) notes that

> *please* is especially frequent with imperatives. The large number of *please* after *could you* and after permission questions (*can I, may I, could I*) is also noteworthy. Since please is mainly used in situations in which formal politeness is needed (…).

While Aijmer's observation explains the absence of *please* in the conversational data set (the informality of the situation requires no formal politeness) and the high occurrence in the business letters (where formal politeness is indeed required), it cannot account for the differences found between the authentic data sets and the DCT directives. The politeness marker is used almost exclusively among preparatory head acts in situations where no formal politeness is required (everyday, informal scenarios between friends and acquaintances). It therefore stands to reason that *please* probably serves a different function in the DCTs than in the authentic data. One explanation might be that, since the politeness marker *please* seems to be one of the most salient politeness strategies used in directives, it therefore has a high

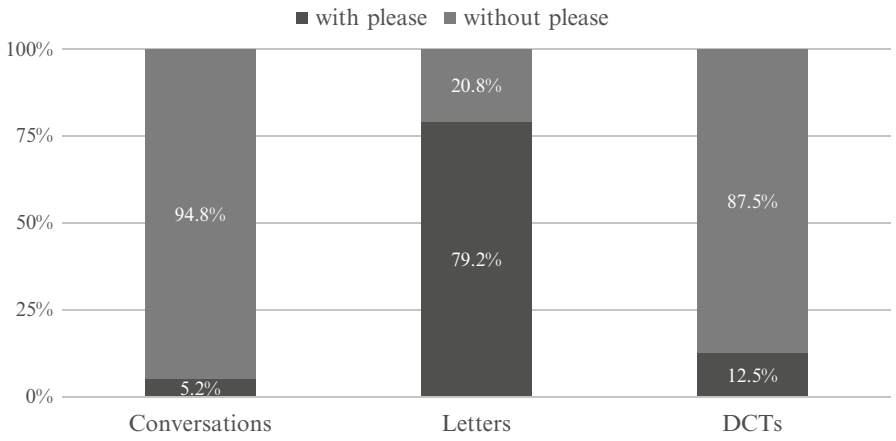


Fig. 7 Distribution of mood derivable strategies co-occurring with and without *please*

normative value (many children are told by their parents to use it in order to produce polite directives). It might be this salient role that causes the usage of *please* in situations where no formal politeness is required in DCTs. As Kasper (2000) points out, DCTs reveal speakers' sociopragmatic knowledge with regard to when which strategy is to be used in which context. It may well be the case, though, that informants rely on, and (over-) use, salient strategies in experimental conditions in which they are more aware of politeness requirements than in authentic situations.

Since frequencies of head acts differ in the three data sets, this does not tell the whole story. Figure 7 represents the relative frequency of the co-occurrence of *please* with so-called 'mood derivables' (mostly, and in letters exclusively, instances of imperatives).

Once again, we find highly significant variation. In conversations as well as in DCTs, imperatives occur overwhelmingly without *please*, whereas in the business letters they tend to co-occur (in 79.2 % of the cases, or 57 out of 73 tokens) with *please*. In other words, all others things being equal, imperatives in the letters corpus are mostly mitigated through a downgrading strategy, whereas in the other two data sets they are not. While imperatives are rare in DCTs anyway, the figures for conversational discourse show that this is a highly significant difference: 94.8 % of all imperatives in conversation are not modified by the politeness marker *please*, which makes them potentially bald-on-record.

4.4 Summary of Empirical Analysis

The results obtained in the present analysis show statistically significant differences for directives elicited in experimental conditions and naturally occurring ones. As we have argued in Sect. 3.1, our three data sets are situated in different locations

within the spoken – written and elicited – non-elicited dimensions. Whereas conversations and DCTs differ maximally with regard to these features (spoken and non-elicited vs. written and elicited), the business letters share one feature with both of the other data sets: they are written like the DCTs and non-elicited like the conversations. The results of our comparison of directive speech acts at least partially reflect this intermediate position of business letters. The similarities or differences between the data sets are displayed in Table 6 below.

First of all, on the head act level, the directives in business letters are maximally comparable to the conversations (row 1 in Table 6); however, they occupy an intermediate position between conversations and DCTs as to the number of downgraded head acts (row 4) and the ratio of downgraders per head act (row 5). For the head acts, we can therefore conclude that the written – spoken dimension did not influence speech act production quite as much as the elicited – non-elicited dimension did. The strong tendency for greater directness in the conversations and letters thus appears to be a typical feature of naturally occurring directives (at least in the particular discourse genres analyzed). The elicited directives, conversely, are typically realized with conventionally indirect head acts (row 2). Indirect head acts (i.e. hints) are rare in all three subcorpora (row 3). It should also be noted that the business letters exhibit some idiosyncratic features that cannot be explained by their intermediate position between conversations and DCTs, as is shown by the high frequency of co-occurrence of the politeness marker *please* and imperatives (row 6).

As to the distribution of modification strategies overall, the picture is less clear-cut; however, for most downgrading modifiers, business letters once again occupy a position in between conversations and DCTs (row 7). Whether this is caused by the influence of the spoken – written or the elicited – non-elicited dimensions or has different reasons altogether (e.g. genre-specific features) is impossible to tell from the data analyzed in the present study. A more detailed analysis of the correlation between specific types of head acts and individual modification strategies might shed more light on this matter, but falls outside the scope of the current paper. The fact that upgraders are significantly more frequent in the business letters than in either of the other two subcorpora (row 8), however, does appear to indicate that some idiosyncratic, genre-specific factors are at work here.

Table 6 Overview of similarities/differences in three data sets

	Linguistic variable	Convers.	Letters	DCTs
1	% of direct head acts	59	55	5
2	% of conventionally indirect head acts	38	40	92
3	% of indirect head acts	3	5	3
4	% of downgraded head acts	42	69	83
5	ratio of downgraders per head act	0.51	0.99	1.50
6	% of mood imperatives with *please*	5.2	79.2	12.5
7	Total # of downgrading modifiers	121	232	353
8	Total # of upgrading modifiers	9	38	7

What we can say, however, is that DCT directives exhibit the greatest degree of difference from conversational directives in almost all categories analyzed, which is remarkable given that the DCTs were constructed to mirror exactly that genre (i.e. spontaneous spoken interaction) maximally closely.

5 A Re-evaluation of Speech Act Research Across Data Collection Methods

As we have pointed out in Sect. 2.2, there are only few systematic comparisons of speech acts elicited under different conditions. Most of these studies involve a two-way comparison between elicited data (mostly DCTs, but occasionally other data such as roleplays) and non-elicited, spontaneous data. This raises the question as to the extent to which our results here are comparable to, or different from, earlier findings in the literature. What complicates matters even more is that the earlier studies referred to investigate a variety of different illocutions, and that the majority of them do not provide information about the frequency distributions of all the strategies studied.

The originality of the current study, in our opinion, resides in several factors. First of all, we have attempted a fairly thorough quantitative analysis of directives with regard to both head acts and modification strategies; furthermore, we have investigated at least a few instances of correlations between the two. Additionally, our analysis involves a three-way rather than two-way comparison between, on the one hand, the most widely analyzed type of elicited data (i.e. DCTs) and, on the other hand, not one but two types of spontaneous, non-elicited discourse (conversations and business letters, respectively). Our results clearly indicate that such a three-way comparison yields new insights. By contrasting three different subcorpora, we have shown not only that the data collection method has a major impact on the production of directive speech acts (our first hypothesis), but also that the discourse genre (and in particular the spoken/written dimension) impacts directive production in a substantial way. So how do our results compare to other contrastive empirical findings on speech act production?

Table 7 below offers an (non-exhaustive) overview of studies that compare speech acts elicited in DCTs to those found in naturally occurring discourse, ranging from informal face-to-face conversations (as in the present study and Pfingsthorn and Flöck 2014) to business telephone interactions (Economidou-Kogetsidis 2013) to academic advising sessions (Hartford and Bardovi-Harlig 1992).

As becomes clear from this overview, the results those studies yield are far from homogeneous, and the conclusions drawn about the representativeness of DCT data are even more diverse. Some authors argue that DCTs are "inappropriate for studying actual language use" (Golato 2003: 91) or that "an ethnographic approach is the only reliable method for collecting data about the way compliments, or indeed, any other speech act functions in everyday interactions" (Wolfson 1981: 115). The present

study certainly indicates that it would be dangerous to extrapolate from DCT findings, and draw conclusions about spontaneous interaction based on elicited data. Other studies, however, continue to claim that DCTs represent naturally occurring data closely enough to justify their further use as a data collection method (e.g. Schauer and Adolphs 2006; Economidou-Kogetsidis 2013; cf. also the discussion in Sect. 2.2).

At the very least, the divergence of results from these studies on the influence of data collection methods on speech act realization should make us skeptical about making claims about potential generalizations. Whilst the collection method clearly influences the linguistic variable under investigation, it does so in ways that need to be investigated further. Future research should, in our opinion, take at least three factors into account. First of all, Table 7 shows that the elicited/non-elicited dimension

Table 7 Comparison of contrastive meta-methodological studies

Study	Illocution	Methods compared	Relevant findings
Bodman and Eisenstein 1988	Expressions of gratitude	DCTs and field notes of conversations	DCTs elicit shorter and less complex strategies
Hartford and Bardovi-Harlig 1992	Rejections	DTCs and authentic spoken discourse (advising sessions)	DCT data contain fewer semantic formulae, status preserving and negotiation strategies
Beebe and Cummings 1996	Refusals	DCTs and authentic spoken data (telephone conversations)	DCT data contain same number of semantic formulae but are less complex and more direct
Yuan 2001	Compliment responses	DCTs and authentic spoken data (face-to-face conversations)	DCT data contain fewer markers of interaction and indirect strategies
Golato 2003	Compliment responses	DCTs and authentic spoken data (face-to-face conversations)	DCT data contain more turns, fewer markers of interaction and over-elicit routine formulae
Schauer and Adolphs 2006	Expressions of gratitude	DCTs and authentic spoken data (CANCODE: variety of spoken discourse)	DCT data are sequentially less complex but show similar head act strategy and modifier distribution
Economidou-Kogetsidis 2013	Requests (directives)	DCTs and authentic spoken data (business telephone encounters)	DCT data are more direct but show similar head act strategy and modifier distribution
Pfingsthorn and Flöck 2014	Directives	DCTs and authentic spoken data (face-to-face conversations)	DCT data are more indirect and contain more downgrading modifiers
Present study	Directives	DCTs and authentic spoken data (business letters and face-to-face-conversations)	DCT data are more indirect and contain more downgrading modifiers

indeed plays a substantial role. However, even within the category of elicited speech, the impact of the elicitation method has not been studied systematically (e.g. DCTs vs. roleplays). Secondly, the present study also shows that spontaneous, non-elicited discourse is not a homogeneous category, and that the contextual conditions of language production in various discourse genre (e.g. spoken/written, formal/informal, and the like) merit further examination. Finally, we should at least allow for the possibility that the type of illocution influences the production choices language users make: while directives exhibit a lot of variation as to realization patterns, other speech acts (e.g. thanking) might be much more routinized and stereotypical. What seems clear is that corpus pragmatics in the widest sense of the word has a major role to play in unraveling some of these complex issues.

6 Conclusion

In the present paper we set out to find answers to two basic hypotheses. First of all, we assumed a significant difference in directive production between elicited and spontaneous data. Secondly, we assumed that the realization of directives in non-elicited discourse is dependent on the contextual conditions of a particular genre (in this case the spoken/written dimension). Our three-way contrastive analysis of DCTs, conversations, and business letters essentially confirms both hypotheses.

First of all, our analysis shows that directive head acts in non-elicited conditions are significantly more direct than the ones found in the experimental condition. With regard to the usage of modifiers, conversations and DCTs can be found at the ends of the usage spectrum (lowest number in conversations, highest number in DCTs), with the business letters occupying an intermediate position. Secondly, apart from this general tendency, we find that the business letters display some specific usage patterns not present in either DCTs or conversations (e.g. the more frequent use of the politeness marker *please* in imperative structures). Generally speaking, the most notable and salient differences in usage patterns for directives were found between the DCTs (which are created to mimic spontaneous interaction) and naturally occurring conversations.

Since previous studies yield very heterogeneous and indeed contradictory results about methodology-induced differences in speech act production, it is difficult to draw any conclusions about the reliability of DCTs. In our data, however, we find clear evidence that DCTs do not reliably represent naturally occurring directives. The fact that language is used without any intrinsic communicative intent in DCTs does seem to be an influential factor here.

With regard to spontaneous discourse, the medium of representation (spoken vs. written) has been shown to play a role in our data sets, in head act realization as well as in the choice of modification strategies. It should not be forgotten, however, that more factors separate our conversations from our business letters, such as the respective formality levels. There are probably many (correlating) independent variables at work here, on top of the data collection method employed. Systematic

research into these variables is therefore necessary in order to find clear patterns and identify the interplay of different variables.

In the present study, we were able to compare three data sets from maximally different sources, but it is clear that data from more types of informant groups are sorely needed. The more data sets and subcorpora we can include, the easier it will be to arrive at an understanding of (the interplay of) the variables at work, and therefore to arrive at valid generalizations across discourse types. Especially with regard to naturally occurring speech, we need maximally comparable data sets that differ in only a limited number of variables. This has already been achieved partially by the advent of parallel corpora (e.g. the ICE series, set up mostly to study aspects of regional variation), but should be expanded even further to give researchers in corpus pragmatics the opportunity to make use of large data bases of naturally occurring discourse in various genres. Such corpora should thus allow us to use authentic as well as elicited materials for large-scale methodological comparisons.

References

Aijmer, K. (1996). *Conversational routines in English: Convention and creativity*. London: Longman.

Aijmer, K. (2002). *English discourse particles. Evidence from a corpus*. Amsterdam: Benjamins.

Barron, A. (2008). The structure of requests in Irish English and English English. In A. Barron & K. P. Schneider (Eds.), *Variational pragmatics. A focus on regional varieties of pluricentric languages* (pp. 35–67). Amsterdam/Philadelphia: Benjamins.

Beebe, L. M., & Cummings, M. C. (1996). Natural speech act data versus written questionnaire data: How data collection method affects speech act performance. In J. Neu & S. M. Gass (Eds.), *Speech acts across cultures: Challenges to communication in a second language* (pp. 65–86). Berlin/New York: Mouton de Gruyter.

Blum-Kulka, S. (1989). Playing it safe: The role of conventionality in indirectness. In S. Blum-Kulka, J. House, & G. Kasper (Eds.), *Cross-cultural pragmatics: Requests and apologies* (pp. 37–70). Norwood: Ablex.

Blum-Kulka, S., House, J., & Kasper, G. (Eds.). (1989). *Cross-cultural pragmatics. Requests and apologies*. Norwood: Ablex.

Bodman, J., & Eisenstein, M. (1988). 'May God increase your bounty'. The expression of gratitude in English by native and non-native speakers. *Cross Currents, 15*, 1–21.

Breuer, A., & Geluykens, R. (2007). Variation in British and American English requests. A contrastive analysis. In B. Kraft & R. Geluykens (Eds.), *Cross-cultural pragmatics and interlanguage English* (pp. 107–126). Munich: Lincom Europa.

Brown, P., & Levinson, S. C. (1987). *Politeness. Some universals in language usage*. Cambridge: Cambridge University Press.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide; Spoken and written English grammar and usage*. Cambridge: Cambridge University Press.

Clark, H. H., & Bangerter, A. (2004). Changing ideas about reference. In I. A. Noveck & D. Sperber (Eds.), *Experimental pragmatics* (pp. 25–49). Houndmills: Palgrave Macmillan.

Economidou-Kogetsidis, M. (2013). Strategies, modification and perspective in native speakers' requests. A comparison of WDCT and naturally occurring requests. *Journal of Pragmatics, 53*, 21–38.

Faerch, C., & Kasper, G. (1989). Internal and external modification in interlanguage request realization. In S. Blum-Kulka, J. House, & G. Kasper (Eds.), *Cross-cultural pragmatics. Requests and apologies* (pp. 221–247). Norwood: Ablex.

Farr, F., & O'Keeffe, A. (2002). Would as a hedging device in an Irish context: An intra-varietal comparison of institutionalised spoken interaction. In R. Reppen, S. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 25–48). Amsterdam: Benjamins.

Geluykens, R. (2011). *Politeness in institutional discourse. Face-threatening acts in native and nonnative English business letters*. Munich: Lincom Europa.

Geluykens, R., & Van Rillaer, G. (1995). Introducing ACID: The Antwerp corpus of institutional discourse. *Interface, 10*, 83–101.

Golato, A. (2003). Studying compliment responses: A comparison of DCTs and recordings of naturally occurring talk. *Applied Linguistics, 24*, 90–121.

Hartford, B. S., & Bardovi-Harlig, K. (1992). Experimental and observational data in the study of interlanguage pragmatics. In L. F. Bouton & Y. Kachru (Eds.), *Pragmatics and language learning* (pp. 33–52). Urbana: University of Illinois at Urbana-Champaign, Division of English as an International Language.

House, J. (1989). Politeness in English and German: The functions of 'please' and 'bitte'. In S. Blum-Kulka, J. House, & G. Kasper (Eds.), *Cross-cultural pragmatics. Requests and apologies* (pp. 96–119). Norwood: Ablex.

House, J., & Kasper, G. (1981). Politeness markers in English and German. In F. Coulmas (Ed.), *Conversational routine. Explorations in standardized communication situations and prepatterned speech* (pp. 157–185). The Hague: Mouton.

Jautz, S. (2008). Gratitude in British and New Zealand radio programmes. Nothing but gushing? In K. P. Schneider & A. Barron (Eds.), *Variational pragmatics. A focus on regional varieties of pluricentric languages* (pp. 141–178). Amsterdam/Philadelphia: Benjamins.

Jucker, A. H. (2009). Speech act research between armchair, field and laboratory: The case of compliments. *Journal of Pragmatics, 41*, 1611–1635.

Jucker, A. H., Schneider, G., Taavitsainen, I., & Breustedt, B. (2008). Fishing for compliments: Precision and recall in corpus-linguistic compliment research. In A. H. Jucker & T. Irma (Eds.), *Speech acts in the history of English* (pp. 273–294). Amsterdam/Philadelphia: Benjamins.

Kasper, G. (2000). Data collection in pragmatics research. In H. Spencer-Oatey (Ed.), *Culturally speaking. Managing rapport through talk across cultures* (pp. 316–341). London/New York: Continuum.

Kohnen, T. (2008). Tracing directives through text and time: Towards a methodology of a corpus-based diachronic speech-act analysis. In A. Jucker & I. Taavitsainen (Eds.), *Speech acts in the history of English* (pp. 295–310). Amsterdam/Philadelphia: Benjamins.

Labov, W. (1972). *Language in the inner city. Studies in the black English vernacular*. Philadelphia: University of Pennsylvania Press.

Manes, J., & Wolfson, N. (1981). The compliment formula. In F. Coulmas (Ed.), *Conversational routine* (pp. 115–132). The Hague: Mouton.

McCarthy, M. (2003). Talking back: "Small" interactional response tokens in everyday conversation. *Research on Language & Social Interaction, 36*, 33–63.

Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring natural language. Working with the British component of the international corpus of English*. Amsterdam/Philadelphia: Benjamins.

O'Keeffe, A., & Adolphs, S. (2008). Response tokens in British and Irish discourse: Corpus, context and variational pragmatics. In K. P. Schneider & A. Barron (Eds.), *Variational pragmatics. A focus on regional varieties in pluricentric languages* (pp. 69–98). Amsterdam/Philadelphia: Benjamins.

Pfingsthorn, J., & Flöck, I. (2014). Investigating and teaching pragmatics: A corpus-based approach. In W. Gehring & M. Merkl (Eds.), *Englisch lehren, lernen, erforschen* (pp. 155–174). Oldenburg: BIS-Verlag.

Sadock, J. M. (1974). *Toward a linguistic theory of speech acts*. New York: Academic.

Schauer, G. A., & Adolphs, S. (2006). Expressions of gratitude in corpus and DCT data: Vocabulary, formulaic sequences, and pedagogy. *System, 34*, 119–134.
Searle, J. R. (1975). Indirect speech acts. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (Speech acts, Vol. 3, pp. 59–82). New York: Academic.
Searle, J. R. (1976). A classification of illocutionary acts. *Language in Society, 5*, 1–23.
Trosborg, A. (1994). *Interlanguage pragmatics. Requests, complaints and apologies*. Berlin/New York: Mouton de Gruyter.
Vine, B. (2009). Directives at work: Exploring the contextual complexity of workplace requests. *Journal of Pragmatics, 41*, 1395–1405.
Wolfson, N. (1981). Compliments in cross-cultural perspective: A reader. *TESOL Quarterly, 15*, 117–124.
Yuan, Y. (2001). An inquiry into empirical pragmatics data-gathering methods. Written DCTs, oral DCTs, field notes, and natural conversations. *Journal of Pragmatics, 33*, 271–292.

Black and White Metaphors and Metonymies in English and Spanish: A Cross-Cultural and Corpus Comparison

Silvia Molina Plaza

Abstract Following a semiotically based concept of culture, this article presents an overview of the metaphoric conceptualizations of the colours *black* and *white* based on rich cross-linguistic empirical data from the BNC and *Corpus de Referencia del Español Actual* (CREA). There is a mixture of literal and figurative meanings in these two comparable corpora but the focus of attention is just on figurative meanings, relating multiword units such as collocations, idioms or proverbs to their different cultural contexts.

The general research question is: how and to what extent does culture actually show up in metaphors and metonymies related to *black* and *white* and their Spanish counterparts *negro* and *blanco?* To answer this question, Piraiinen's taxonomy (Piirainen E, Phrasemes from a cultural semiotic perspective. In: Burger H et al (eds), pp 209–219, 2007) will be used, in order to analyze different kinds of cultural phenomena from a qualitative point of view.

The results show that cultural metaphors appear to require an understanding of the input domains and their properties or connections with the output domains. The comparative outline of phrasemes containing *black/negro* and *white/blanco* clearly indicates the cultural foundation of phraseology (Wierzbicka, Semantics-primes and universals. Oxford University Press, Oxford, 1996). The uses of *black/negro* as 'bad, unhappy' and *white/blanco* as 'good, innocent' represent cultural facts and if taken as physical entities (colour terms), they symbolise these properties. English and Spanish 'black' and 'white' collocations, idioms and proverbs are powerful symbols in culture. The amount of knowledge that language users have on the relationship between the symbols BLACK and WHITE in language and culture allows that the 'right' reading can be activated in different contexts.

Keywords Contrastive phraseology • Colour terms • English • Spanish

S. Molina Plaza (✉)
Applied Linguistics Department UPM, Technical University of Madrid, Madrid, Spain
e-mail: silvia.molina@upm.es

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_3

1 Introduction

Soviet research in phraseology connected phrasemes with cultural knowledge (Dobrovol'skij 1998: 55ff.). A number of research traditions (Makkai 1978: 403ff.; Černyševa 1980: 11ff.) point out that idioms cannot be explained by means of just linguistic methods. Culture-based knowledge must be addressed to understand the differences between figurative and non-figurative multiword units. Culture is understood in the sense in which it is used by cultural anthropologists, according to whom culture is something that everybody has, in contrast with the 'culture' only found in cultivated circles, such as universities and the like. The term 'culture' is used differently by different researchers, but always it refers to some 'property' of a community that might distinguish it from other communities (Berry et al. 2002). Culture will also be used in this paper as socially acquired *knowledge*, including both the ideas of 'know-how' and 'know-that'. However, the knowledge included in a culture need not necessarily be factually or objectively correct in order to count. Therefore, both lay people's knowledge or common sense-knowledge and the specialist knowledge of scientists or scholars will be considered as forming part of both cultures.

More specifically, this study compares how Spanish and English figurative colour proverbs, idioms and collocations, and single lexical items to a lesser extent, preserve cultural knowledge, thus showing the points of contact and departure to conceptualize several phenomena using two central colours, *black* and *white*. Due to space constraints, this study only examines the types of metaphors and metonymies used in English and Spanish. Therefore, non-figurative uses of colours are outside the scope of this paper. Relevant approaches to the study of colour are Berlin and Kay (1991), Sherman and Clore (2009), Niemeier (2007), Vlajkovic and Stamenkovic (2013), Mey (2014) and Wierzbicka (2006), and they provide useful frameworks for the comparison of L1/L2 idioms and collocations.

A major contribution to our understanding on how metaphors are realised and perceived by speakers has come from the area of *Cognitive Linguistics* (Lakoff and Johnson 1980; Fauconnier 1997; Steen 1999). These authors put forward a constructivist approach that basically holds that metaphors are a phenomenon of thought and that metaphor creation forms part of the on-going process of communication.

In this paper, Lakoff and Johnson's (1980) definition of metaphors is adopted, conceived of as the result of the transfer of properties of the metaphorically used word or phrase from one cognitive domain to another unrelated domain. Metonymy relies instead on the juxtaposition of adjacent cognitive domains without the transfer of properties from one to the other but both may work together to capture meaning (Geeraerts 2002).

2 Method and Material

The present approach is data-driven and colour-based metaphors and metonymies are identified manually. The theoretical conclusions are worked out in a strict bottom-up way; a usage check is also carried out to see whether the metaphor is

frequently used or not in English or Spanish. The contrastive corpus consists of all colour expressions containing black and white in the BNC and the CREA[1] in order to check out the similarities and divergences in use in both Spanish and English phraseology. All the Spanish varieties have been included in the searches (Cuban, Mexican, Peninsular Spanish, etc.). The aim is to provide empirical evidence of how metaphorical meanings are expressed in lexical patterns.

This cross-linguistic comparison is a multidisciplinary enterprise as it has links with the Conceptual Theory of Metaphor, Corpus Linguistics, Contrastive Lexicology, Pragmatics, Semantics and Translation theory as the analysis of data may consequently benefit from meeting points between these various linguistic schools. The approaches mentioned above seek to combine two objectives: first, to describe the connection between figurative phraseology and culture as it becomes manifest in the phraseological data of the two languages related to the colours *black* and *white* and second, to outline trends in research on cultural features of these phrasemes.

Salient metaphorical patterns both that could be related to cultural differences are hand-searched in collocations, idioms[2] and proverbs in these two comparable corpora. The figures presented in the examples below are the raw data from the BNC and the CREA and are a mixture of literal and figurative meanings in adjectives and to a lesser extent, nouns. Though much of the conceptual colour system is metaphorical in English and Spanish, a significant part is non-metaphorical. Examples reveal that metaphorical understanding of colour expressions is grounded in non-metaphorical understanding.

The figures from the two corpora are indicative of the frequency of use of these colour phrasemes. To reduce this initial pool of materials to a more manageable size (33,232 citations for *negro* in Spanish versus 23,864 in English for *black* and 24,609 for *blanco* and 23,427 for *white*) the figurative uses of these expressions in the examples that follow will be the focus of this study.

However, prior to the study of our corpus data, I would like to illustrate a general point: fixed and semi fixed colour expressions encode cultural information motivated and grounded in a so-called 'common European Heritage' (Piirainen 2008). Colour phrasemes are made up by idioms originating from identifiable textual sources. Hence, it is not surprising that both cultures share common values in the use of black and white metaphors (good- white; bad-black) as the two countries belong to the same European Christian heritage. Traditionally, **black** perceptual input has been linked with negative affect (Mey 2005) as some examples from *Wordnet Search 3.1* reveal: death (*Black Death*), evil (i.e. black hens were used for

[1] CREA (*El Corpus de Referencia del Español Actual*) is a corpus of 160 million words from different genres which represents current Spanish usage. CREA's 90 % of texts are written and 10 % oral.

[2] According to Taylor (2002: 540), idioms are extremely important to master any language since everything turns out to be idiomatic to a greater or lesser extent. Idioms are also often used as evidence that conceptual mappings exist that are independent of language and govern the matching linguistic structures in their semantic and pragmatic behaviour (Dobrovol'skij and Piirainen 2005: 212).

Satanic rituals), famine, fear and the unknown (*black holes*), anger (*black looks, black words*) or sin. It is less commented on the literature that black is also related to some positive states like power, elegance, formality, or grief and mystery in both cultures. Black also denotes a positive semantic prosody: strength and authority; it is a very formal, elegant, and prestigious colour (*black tie,*[3] *black Mercedes*). But these examples do not rehabilitate black from a generally dysphemistic set of connotations. That is the reason why prominent professionals (lawyers, judges and priests) traditionally dressed in black and were reputed to bring about bad luck under some circumstances.

White symbolizes light, goodness, purity, righteousness, joy and virginity in the two cultures. This is strongly motivated by its contrast with black: *white magic* is good whereas *black magic* is bad. Babies are dressed in white at christenings and brides usually wear white dresses at weddings as a symbol of chastity and purity even if they are no longer virgin. It is the synthesis of all colours and symbol of the absolute, innocence and peace and represents positive values. The Bible as a target domain provides a wealth of examples: white is the colour of manna (Exodus 16:31); the beloved one (Song of Solomon 5:10; the shining garments of angels (Revelation 15:6) and of the transfigured Christ (Matthew 17:2); hair (Matthew 5:36) and the great throne of judgment (Revelation 20:11). White means free from moral blemish or impurity; unsullied "in shining white armour" in Western cultures. However, the same colour takes on a strongly negative semantic prosody in other cultural contexts such as Chinese theatre performances where it symbolizes slyness.

In advertising, white is associated with coolness and cleanliness because it is the colour of snow. Hence, the expression 'whiter than white' in its literal sense in washing powder ads: *One touch of a button and the family wash comes out "whiter than white"* (BNC, 16, A73). You can use white as a marker to suggest simplicity in high-tech products. White is an appropriate colour for charitable organizations; angels are usually imagined wearing white clothes. White is linked to hospitals, doctors, and sterility, so you can use white to suggest safety and giving the physiological impression of something cold and clean (Graumann 2007: 132) when promoting medical products as in the collocation 'white bandage' (in bold type) in this internet example: *A big* ***white bandage*** *on her head-the result of tonight's earlier craniotomy-overpowered the small, fragile face*.

3 Black and White in the BNC and CREA: Results and Discussion

This study takes the view that there are semantic correspondences and differences between individual L1/L2 idioms and collocations as the analysis of semantic networks reveals (Coseriu and Geckeler 1981) along with the way they are actually

[3] Example from Vogue 2002: 261 f as quoted in Wyler (2007: 125).

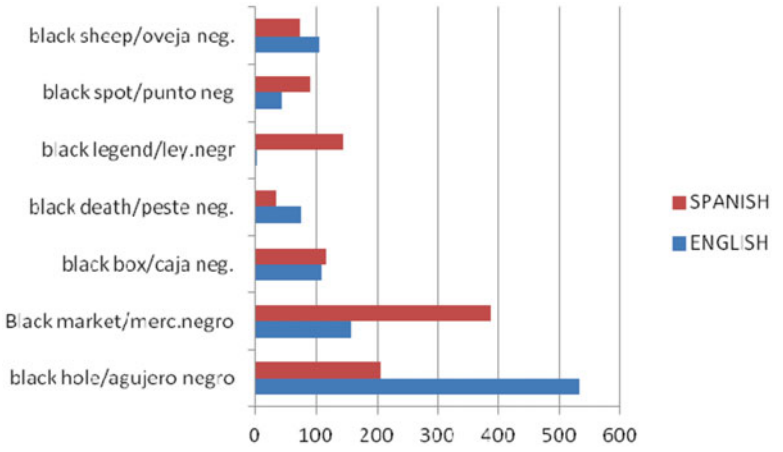


Fig. 1 Black metaphorical idioms and collocations in the BNC and CREA

used in corpora (Sinclair 1987). The quantitative analysis of the most frequent metaphors and metonymies sharing a common background is shown in Fig. 1. There are certain significant differences from a quantitative point of view in just two cases, notably in the use of "mercado negro" and "black hole". Both clearly outnumber their counterparts and denote a cultural focus.

The qualitative analysis of metaphorical and metonymic[4] multiword examples from the BNC and CREA (*Corpus de Referencia del Español Actual*) provides the starting point for uncovering some relevant aspects of the English and Spanish world view. Piirainen's model is briefly explained first in order to understand the different kinds of cultural phenomena which underlie phrasemes:

(a) **Intertextual phenomena** based on textual dependence. This group is made of phrasemes whose image components can be traced back to identifiable textual sources: direct references to particular texts, original quotations which gradually became idioms or proverbs as well as allusions to an entire text, summarizing the gist.
(b) **Fictive conceptual domains** are based conceptually on pre-scientific views of the world such as folk belief, superstitions or old folk medicine. An example of the latter is the non-metaphoric *ungüento blanco* [white ointment] used in the proverb *Como el ungüento blanco, que para todo sirve y para nada aprovecha*

[4] Metonymies are a prime factor in the generation and coining of colour word expressions, but it is misleading to suggest that colour-word expressions are metonymies proper; and it should also be stressed that metonymical motivation does not in any way preclude opacity of meaning. When the figurative meaning of a colour-word expression is not immediate, its etymology can be looked at in order to understand the cultural and connotative values that brought it into being. In doing so, however, the analyst has to be very careful to avoid over-interpretation, bearing in mind that the average language user's awareness of meaning is limited to the pragmatic function of the expression in a discourse situation, and does not generally extend into the diachronic dimension.

quoted in Don Quixote [literally, 'As the white ointment, useful for everything but good for nothing'] referring to an immemorial medical practice: the use of a white ointment for lesions and sores, practically useless for healing.

(c) **Cultural symbols** usually manifest themselves in one single key constituent that contains the relevant cultural knowledge (as opposed to the phraseme as a whole). The motivational link between the literal and figurative readings of these constituents is made by semiotic knowledge about the symbol in question, about its meaning in culturally sign systems other than language (e.g. religion, popular customs, etc.) The symbol is a sign whose primary content is used as a sign for denoting another concept. For example, the primary meaning of 'white' in the idiom *blanco como una patena* [literally, 'as white as a paten'] has shifted metonymically to meanings such as 'clean' or 'spick-and span'.

(d) **Aspects of material culture** are idiosyncratic elements of culture-specific artefacts but seem to be rare. The concept BULLFIGHTING is a source frame that is possibly unique to Spanish Phraseology. Thus, *sacar el pañuelo blanco* [literally, 'to take the white handkerchief'] is used to award the ears or the tail to a successful bullfighter and by extension to give an award to any successful person.

(e) **Culture-based interactions** This is an umbrella term to group phrasemes whose underlying cultural knowledge is related to social experiences and behaviours revealing cultural models, i.e. gestures (*to take one's hat off*), gender specifics (*to wear the trousers*) or recommendations such as the following: *vino tinto con la vaca, y blanco con espinaca*, suggesting literally 'drinking red wine with beef and white wine with spinach'. This proverb reveals a socially approved idea of former times that it was good to drink wine at meals. Wine has always been an idiosyncratic aspect of traditional Spanish culture and it is not surprising that it is a relevant source frame in more than 300 proverbs, giving information about which values and rules are upheld in Spanish social behaviour.

However, it is not always easy to draw sharp lines between these five cultural groups as they are often interconnected. Piirainen's model challenges the postulates of the Cognitive Theory of Metaphor (CTM) as proposed by Lakoff and Johnson in their ground-breaking book of 1980. Piirainen argues that although some conceptual metaphors are universal (e.g. LOVE IS FIRE) many others are based on historical knowledge, despite the fact that they are no longer current for many speakers, still contributing to the interpretation. She claims that CTM can only partially explain the wealth of phraseological data across languages and lacks relevant cultural inputs which profile the conceptual systems of different languages. Thus, in order to determine the metaphorical meaning of multiword units, the analyst needs to rely on several factors such as cognitive mechanisms (conceptual mapping between and within domains, conceptual metaphor and metonymy), knowledge of the language (semantics, syntax, contextual clues) and also knowledge of the world (cultural and historical background, symbols, etc.). In what follows, *black* is studied in both languages following Piirainen's taxonomy (2007) in point Sect. 3.1 and *white/*

blanco examples pertaining to aspects of material culture will be the focus of Sect. 3.2. Lastly, point Sect. 3.3 will deal with *white/blanco* examples belonging to the last category proposed by Piirainen: culture-based interactions.

3.1 Aspects of Black/Negro According to Piirainen's Model

3.1.1 Intertextual Phenomena

Burger used this term to refer to the relationship between phrasemes and identifiable textual resources of various types: written texts, quotations from poetry, folk tales, songs, etc. Many are related to the Bible or fables, such as *the black sheep*.

Black is used in the Bible to describe the colour of the middle of the night (Proverbs 7:9); diseased skin (Job 30:30); healthy hair (Song of Solomon 5:11; Matthew 5:36); the sky (Jeremiah 4:28); the darkening of the sun and the moon (Joel 2:10); horses (Zechariah 6:2; Revelation 6:5); and marble (Esther 1:6).

3.1.2 Fictive Conceptual Domains

These refer to old folk theories and pre-scientific or fictive conceptions of the world – including religion, superstition, common belief, etc. Black has been associated to sheer bad luck in some Mediterranean countries. If a black cat gets in your way, you are supposed to get into deep trouble as black cats were connected with witchcraft and bad luck by the Seventeenth Century in some European countries, including Spain. However, in some places which saw few witch hunts, black cats were ascribed to good luck. Many years ago, fishermen's wives kept black cats in their homes while their husbands went away to sea in their fishing boats to prevent any danger in the U.K and an unfamiliar black cat on the porch of a house was regarded as a sign of upcoming prosperity for its inhabitants in Scotland.

Despite these points in common, there are clear differences in idioms and collocations in fictive conceptual domains. Three relevant examples are shown below.

The first one is the Spanish metaphorical idiom, '*tener la negra*' [literally, 'have the black'] whereby black in the feminine stands for bad luck. Therefore, it is translated into English as "HAVING BAD LUCK OR BEING VERY UNLUCKY". The origin of the Spanish expression dates back to the Eighteenth Century when the town council members drew lots by choosing a black bean from a sack full of white beans. Therefore, *la negra* stood for the black bean that meant taking on a responsibility and there has been a 'domain shift' nowadays in Spanish (Charteris Black 2004) as it is no longer used in this original sense in examples 1a and 1b below. Today's meaning is *misfortune.* According to Dobrovol'skij and Piirainen (2005: 97), the black bean has become an *inactive symbol* that is no longer comprehensible for Spanish speakers today:

(1a) *Lupe vuelve a salir de la choza para gritar su desgracia a los cuatros vientos (…) la pobre "tenía la negra*".[Lupe comes out again of her hut to shut her disgrace to all and sundry (…) the poor woman had a run of bad luck].

La negra refers vaguely to a disgrace. Lupe, the speaker, makes the colour predicate precise enough for the purpose at hand, describing a run of bad luck. Metaphor is an alternative route to achieve optimal relevance. Whether this utterance is literally, loosely or metaphorically understood will depend on the mutual adjustment of content, context and cognitive effects in order to satisfy the overall expectation of relevance.

(1b) Como toda España, Barcelona ha sufrido la recesión de 1993. Si los efectos de la crisis se retrasaron por el boom de los Juegos Olímpicos, sus consecuencias han sido más profundas.`(…). Barcelona "tenía la negra" [Barcelona has undergone the 1993 recession. If the crisis effects were held up by the Olympic Games boom, its consequences have been deeper. Barcelona had a run of bad luck]

Tener la negra is used as a conventional idiomatic phrase with a fairly stable set of co-textual features (disgrace, economic crisis and other misfortunes are the principal collocating words). Once the language user becomes aware of the origin and metonymical motivation for the phrase, his/her perception of the meaning may change slightly, as the new image contributes an additional layer of literal, compositional meaning to the otherwise non-decomposable string.

The second example is the collocation *trabajo en negro* [literally, 'illegal work'] with 19 occurrences in CREA, closely related to the collocation *dinero negro* [illegal money] with 193 occurrences as in Example 2. It was a racist common belief that the money owned by black people was phony. Therefore, 'black money' meant "a black person's money". On the contrary, money owned by whites was real, trustworthy, hence 'white money'. *Dinero negro (black money)* in example two has a figurative meaning available context-free: illegal money. Example three below is an on-line construction of meaning using *blending* (Fauconnier and Turner 2002) whereby two input mental spaces (reliable money and Real Madrid's football club money, whose players are usually clad in white) create a blended mental space, framed by the importance of worthy money and its use to hire in new players by Real Madrid's CEO, Vicente Calderón. In the selective projection, the race characteristics of white money are lost and a metonymic blend is instantiated (Coulson and Oakley 2003):

(2) El juez inicia la intervención de todo el "dinero negro" procedente del fraude del IVA. [The judge seized all the black money coming from the VAT fraud]
(3) Vicente Calderón no negociará directamente con Ramón Mendoza el traspaso de Francisco Llorente. (…) La única posibilidad de culminar el fichaje es que la empresa Dorna se lo compre al Atlético, incluso con "dinero blanco", para que el extremo vuelva a casa. [Vicente Calderón will not negotiate Francisco Llorente's transfer with Ramón Mendoza directly. The only possibility to succeed is that the Dorna firm buys it to the Athletic football team, even with "white money" to ensure the footballer comes home]

The fourth example is an idiosyncratic metaphor to refer to depression. Depression is a *black dog*[5] (example 4)*:*

(4) The fullest and most fascinating case study is that of Churchill, whose famous 'Black Dog'; depressions are shown to have sprung ineluctably from childhood traumas (A1F 247)

This animal metaphor which means *in a black mood* popularly attributed to Churchill implies both familiarity and an attempt at mastery, because while that dog may sink his fangs into one's person every now and then, it is still, after all, only a dog, and it can be cajoled sometimes and locked up other times. The animal metaphor is powerful as depression, like a dog crouches in the corner of the room, waits for the person to make a move or lies at the foot of the bed, like a shadow, until s/he tries to get up.

The metaphor dates back to Samuel Johnson, James Boswell and Stevenson. They used the phrase to refer to a similar state in their prolific correspondence and writing. It is interesting to note that the very word 'melancholia' stems in part from the Greek word *melaina*, meaning 'black'. Of particular relevance to the present investigation is also the claim by the eleventh-century Hebrew astrologer Ibn Ezra that the *canes nigri* – or black dog – is the beast of Saturn, the melancholy god. It is widely used nowadays in music bands such as Led Zeppelin (*Black Dog*), The Manic Street Preachers (*Black Dog*) or recent fiction and memoir by writers such as Ian McEwan, (*Black Dogs* [*New York: Doubleday, 1992*]). Spanish uses the metaphorical collocations "caer en un agujero negro" [to fall into a black hole] which

[5] See Steinvall (2002) for an in-depth study of English colours in context.

stands for FALLING IS GETTING DEPRESSED and "verlo todo negro" [to see it all black], where depression and sadness is represented by dark colours. This expression prompts a dynamic cognitive process, which includes conceptual connections (happy is up whereas DEPRESSION IS DOWNWARD MOVEMENT) and a mapping (DEPRESSION IS A HOLE).

3.1.3 Cultural Symbols

There are phrasemes with colours containing cultural symbols (Dobrovol'skij 1998; Dobrovol'skij and Piirainen 2005) where the motivational link between the literal and figurative readings is established not only by semiotic knowledge about the symbol in question, but about its meaning in cultural sign systems other than language (e.g. in mythology, religions, fine arts, popular customs, etc.). The theoretical framework semiotics of culture, created by Lotman (1990) and others, allows relating different occurrences of symbols to each other. Dobrovol'skij and Piirainen (2005: 253ff.) describe their main features:

A Symbol Has Undergone a Metonymical Shift

It is a sign whose primary content is used as a sign for denoting a different content, usually of higher value than the primary content. Cf. the idiom *whiter than white.* The primary meaning of white has shifted metonymically to abstract meanings, such as "honest, true, pure" as in example (5) through the connotative values of white in religion:

(5) would you not say that this has shifted the onus of responsibility very much on to the financial institutions, the banks and others who had dealings with Mr Maxwell, and isn't the judge in effect saying in your interpretation er of that, that these institutions er really have got to show they were "whiter than white" in their dealings and actually went in and investigated him thoroughly? (K73 309).

Another variation to this set phrase (*white as snow,* AFF 715) relies on the comparison of metaphorical whiteness with that of the physical manifestation of the colour. This dead metaphor requires no further elaboration on the part of the speaker as the intended meaning is both truthful and immediately accessible.

Finally, there are also other English expressions with a metonymic basis: the similes *white as chalk/white as a sheet* indicate pallor from fear or shock and *white as driven snow* is used to refer to moral purity and it is of Biblical origin.

Symbols Tend to Occur in Groups or Symbolic Domains

Black and white are contrasted in the following old-fashioned Spanish proverbs as elements of a code of folk culture which relate animals and racist attitudes. The figurative examples below compare *white sheep* to parents and *black sheep* to a son or daughter who is considered undesirable or disreputable (first proverb).[6] The second proverb uses a concrete image from culture-specific realia, a black donkey, which activates an abstract situation (do not expect the impossible: finding white hair in a black animal):

(6) De "ovejas blancas" nacen corderos negros [Black sheep are born from white sheep]
A burro negro no le busques el pelo blanco [Do not look for white hair in a black donkey]

Finally, black and white are related in the Spanish collocation *negro sobre blanco,* [literally, 'black over white'] (26 occurrences in CREA) which means that something is official, in writing or print so that everyone can see exactly what it is and leaves no room for discussion (exactly the same as the English idiom 'in black and white'[7]). By extension, it also alludes to a serious or relevant affair as in the following example where a Spanish socialist politician, Bono, was posing a question about the best candidate to win at the polls. This utterance can be interpreted in Gricean terms as a literal assertion (Bono wrote it down on a piece of paper). A speaker should try this literal interpretation first and consider the figurative (a serious affair) only if the literal interpretation blatantly violates the maxim of truthfulness. Yet, there are several studies (Glucksberg 2001; Gibbs 1994) with empirical evidence suggesting that literal interpretations are not necessarily tested and rejected by speakers before figurative interpretations are considered; indeed, in interpreting example 7, it would not even occur to the Spanish reader to wonder whether Bono literally wrote it in black:

[6] There is also a fairly common expression in current Peninsular Spanish to replace black sheep: a black chick pea (*garbanzo negro*) which represents the unruly son or daughter and by extension to any troublesome person in a group.

[7] This expression related to the Latin maxim *verba volant, scripta manem [actions speak louder than words]* also has another meaning: if you see, judge or describe a situation in black and white, you think about it in a way that seems too simple, as if everything or everyone in it was completely good or completely bad (*Longman Idioms Dictionary*, page 28).

(7) Seis días después de la presentación de la eurodiputada, Bono puso "negro sobre blanco" la clave con la que esperaba conquistar el apoyo de los delegados, durante una comida en el Club Internacional de Prensa, ante cuyos miembros -y todos los periodistas especializados en información del PSOE- esbozó su proyecto: "¿Cuál de los candidatos ganaría mejor a Aznar?". [Six days after the European member of parliament was introduced, Bono put in black and white the key he meant to use in order to obtain the delegates' support during a meal at the International Press Club where he explained his project to its members, all experts in the Spanish Socialist Party (PSOE): Which candidate would beat Aznar best?]

3.1.4 Aspects of Material Culture

They are embedded in everyday life of the past and present. Examples first will show metaphorical and metonymic meanings in single lexical units (section "Black/Negro as a Single Lexical Unit"), and also idioms (section "Idioms with Black/*Negro*") and collocations (section "Collocations with Black/*Negro*") related to *black/negro.*

Black/Negro as a Single Lexical Unit

The total number of citations in CREA for *negro* makes up a total of 33,232 citations in Spanish. Again, there is a mixture of literal (i.e. '*el techo está negro*' [the ceiling is black]) and metaphorical meaning like *estar/ponerse negro*, a metaphor corresponding to the metaphor formula BLACK IS EXTREMELY ANGRY or UPSET in English illustrated in example 8:

(8) *González* "está negro" *porque sabe que en realidad, lo que ahora proponga es provisional y que su sucesor se llama Aznar.* [González is extremely angry as he knows that anything he puts forward is provisional in fact and his successor is Aznar].

Although colour predicates are paradigmatic cases of vague predicates, the verb "to be" in this context expresses a Situational Process of Being so it cannot have the referent quality of "blackness" coming from the domain of colour as the subject refers to former Socialist Spanish President Felipe González, who did not fancy the idea at all his successor was going to be the Conservative Party Leader, José María Aznar. This idiomatic collocation "estar negro" is mostly figurative when applied to persons and is rarely used when somebody is stained with dirt, soiled. Finally, it is

not to be confused with "ser negro"[be black], referring to black people[8] or "un/el negro", an idiomatic and metaphoric expression, translated by a non-colour metaphor in English, *ghostwriter* (example 9):

(9) *Su nombre comenzó a sonar como" el negro" de Ana Rosa Quintana.* [His name started to sound as Ana Rosa Quintana's ghostwriter]

Idioms with Black/*Negro*

There is a metonymic idiom in English, *in the black*, (10) meaning 'solvent, in profit'. This reading is activated by world knowledge, and not by the symbolic function of BLACK. It refers back to former customs of banks to record amounts on the credit side in black type.

Black and white alludes metaphorically to SIMPLICITY and CLARITY (examples 11 and 12):

(10) *PFS expects to be back* "in the black" *for the full year.*
(11) *The* "black and white" *decisions are easy.*
(12) *Positions, of course, are never as* "black and white" *as that – consolidators have radical streaks; sensible reformers know when to consolidate.*

Both convey an array of weak implicatures. Thus, "The black and white decisions are easy" weakly implicates that clear-cut decisions are somewhat easy to make.

Black looks (13) and *in a black mood* (14) metaphorically relate darkness with the negative state of anger and depression. According to Barcelona (2000: 39–40), the metaphor NEGATIVE IS DARKNESS in both examples, develops out of the generalization of the metonymy DARKNESS FOR NEGATIVE STATES CAUSED BY DARK. A black look appears on a face clouded with anger, threatening, frowning and capable of other baneful matters. This collocation is translated non-metaphorically as "mirada sombría" [a dark look].

(13) *'Good' dealers will derive motivation from the frowns of the establishment, the* "black looks" *of the wary and the total disinterest of 90 % of the population.*
(14) *Donald and Jean had disappeared and Mary was* "in a black mood"*, striding along and making old Donald gasp.*

[8] Black is a calque of Spanish or Portuguese *negro* (Allan 2009: 628).

Black sheep, a reduced version of "the black sheep in the family", is a fairly common idiom in English (103 tokens in the BNC) and other European languages, such as German, Dutch or Finnish (Dobrovol'skij and Piirainen 2005: 173) meaning UNDESIRABLE. It has an equivalent translation in Spanish with a similar metaphorical use and frequency as the semantic reinterpretation of the phenomenon is the same. Its original figurative use according to the Online Etymology dictionary is supposedly because a real black sheep had wool that could not be dyed and was thus worthless. In this case, the change of context and domain places this idiom at the metaphorical end as it refers to a member of a family or group who is unsatisfactory in some way as examples 15 and 16 in both languages reveal. Example 15 highlights the idea that the referent is less successful than the rest (speakers arrive at this interpretation using *symbolic knowledge* whereby expressions with BLACK have a secondary reading, 'bad') whereas the Spanish example is vaguer than the English. However, the co-text in the CREA example blocks vagueness as it refers to Colombia's hindering other Latin American countries initiative at a Conference. The idea is to highlight that Colombia is untypical of the group of Latin American countries and this property of being different is judged to be bad.

(15) *But he has to be careful not to be* "the black sheep" *of the family* (KGP 425).
(16) *Colombia (…) se convirtió en* "la oveja negra" *del continente* [Colombia became the black sheep of the continent].

There are several metaphorical idioms in Spanish with the word *negro,* which structure a complex and abstract target domain in terms of a more concrete and familiar domain of experience. One of the most popular ones is *bestia negra* [black beast] with 88 occurrences in 82 documents in CREA which refers to a person or thing strongly detested or avoided:

(17) *Ha sido la* "bestia negra" *de la oposición catalana.* [He has been the black beast of the Catalan opposition]

'Black beast' refers to someone or something unwanted or even hated, a pet peeve or strong annoyance in both cultures. Therefore, in this case *black* stands figuratively for individuals who act suspiciously or maliciously acquiring darkness as an attribute. The etymological origin of the phraseme comes from French *bête noir,* used to describe a person or thing strongly avoided or detested since 1828, according to the Merriam Webster Dictionary online.

Estar/ponerse negro un asunto [literally, a situation is/becomes black] means that a situation has gone awry (example 18). Once more, EXTREMELY NEGATIVE IS BLACK (Barcelona 2003: 40):

(18) *Esto* "se pone negro" *-dije. -No seas miedoso -dijo Teresa*. ["This is black" I said- "Don't get scared" Teresa said]

Collocations with Black/*Negro*

Collocations are understood as the habitual co-occurrence of words in general, also called *restricted collocations* by some scholars (Burger et al. 2007). There are many examples of collocations in the BNC whose origin is quite transparent and close to metonymy based on encyclopaedic knowledge, but whose usage nowadays points to metaphor, mainly due to their change of context. Cases in point are *black hole* and *black spot*. Both have got literal translations into Spanish and they also share the same metaphorical values. A *black hole* is an area in outer space into which everything near it, including light itself, is pulled and it becomes a place which seems to pull something, especially money, into it (examples 19 and 20 which describe financial problems, that is, failed states in the economy). There is also a third meaning in both languages whereby a BLACK HOLE IS A DEPRESSION which just swallows you up.

A *black spot* is either a part of the road where many accidents have happened in Spanish and especially British English or any place or area of serious trouble or difficulties (example 21). The interpretation of this idiom relies on knowledge of the world, as speakers know that certain areas of highways and roads concentrate more accidents than others.

(19) *Mr. Barron described it as 'an inevitable financial* black hole*', but reiterated his party's plans to raise investment to ensure that all services were free*.
(20) *Todos los indicadores señalan que la economía colombiana finalmente comenzó a salir del* "agujero negro" [All indicators point at Colombia's economy coming out of the black hole].
(21) *Within a month of nationalisation, the Authority's commercial manager had identified the more serious* "black spots" *in which tariffs for additional domestic kWh were below ¾d. -*.

Black Monday is associated to FATEFULNESS alluding to the most notorious day in financial history: Oct 19, 1987 when The DJIA fell 508 points, that is, almost 22 %. Since Black Monday, there have been multiple mechanisms built into the market to prevent panic selling, such as trading curbs and circuit breakers (examples 22 or 23):

(22) *The rise was the second largest on record, after a 142 point bounce in October 1987 when the market recovered from the* "Black Monday" *crash.*
(23) *es decir muy por debajo de los 11.144,34 puntos del índice Hang Seng al cierre del* "lunes negro".[That is to say, far lower than 11.144, 34 points of Hang Seng's index at the Black Monday closure]

Black is also linked to NEGATIVE VALUES as in *black legend* (example 24) with a hyperbolic value with the same translation in Spanish (example 25):

(24) *The* "black legend" *of a closed society, proud in its resistance to modern ideas, is transposed from the spheres of intellectual intractability into the lower regions of economic necessity; (…).*
(25) *El grave accidente ha resucitado la* "leyenda negra" *de* Superman [The serious accident has revived Superman's black legend]

Black market is a metaphor (example 26) where black is related to ILLEGALITY, its origin dates from 1930 to 1935, according to the Random House Unabridged Dictionary, and was expanded to other European languages, such as Spanish (example 27):

(26) *Even for some government cars, diesel is only available on the* "black market".
(27) *El valor de la droga encontrada hasta el momento, ascendería a unos tres millones de euros en el* "mercado negro".[The value of the drugs found until now would amount to three million euros in the black market]

Cuban Spanish has an alternative collocation for black market, *bolsa negra* [*black bag*]. It reveals the difficult situation 90 % of Cubans have to face every day to obtain basic products for survival, according to the Spanish newspaper *La Vanguardia*. Therefore, it is not unexpected to find this geographical variation of *mercado negro* [black market] in CREA in 12 examples, such as 28:

(28) *Bárbara Castillo justificó las reglas como un mecanismo para evitar la* "bolsa negra" *y proteger al consumidor.*[Barbara Castillo justified the norms as a mechanism to avoid "the black bag"]

Although the English *black cloud* and its Spanish rendering *nube negra* have correspondence in translation, these may have different meanings. A BLACK CLOUD IS A DEPRESSION in English (29) whereas A BLACK CLOUD IS A THREAT in Spanish (example 30):

(29) *Usually she managed to keep the looming black cloud of misery at bay, but there were times when her thoughts would drift away (…).*
(30) *La amenaza de las interferencias ajenas se empezaba a configurar como una* "nube negra" *sobre aquella ilusión alimentada a solas.* [The threat of external meddling was starting to look like a black cloud over that hope cherished alone].

Black comedy is closer to metonymy because black is related to black humour. It has its origins in comedy, where black slaves played an important role for parody and satire. Nowadays it has extended its meaning to refer to humour dealing with the unpleasant side of human life (examples 31 and 32):

(31) *His TV play Rotten Apples, to be shown over Christmas, is* "a black comedy" *set within a Northern Irish police station* (A9T 297).
(32) *Recibida con frialdad por la crítica, Lolita era no obstante una buena* "comedia negra"*, tejida como sátira del American way of life.*[Although received coolly by critics, Lolita was nevertheless a good "black comedy" woven as a satire of the American way of life].

Black death (75 tokens) and *black box* (108 tokens) are originally based on a metonymy. *Black Death* refers to outbreaks of bubonic and pneumonic plague that ravaged many European countries in the Fourteenth Century causing the death of millions of people. A *black box* refers to the crash-resistant steel container ('black box') that holds instruments that record performance data in airplanes which nowadays is orange (example 34):

(33) *Dale's the farm just over the hill. And they all had their flails with them. So they went to the Kildingy Well which was s supposed to have some kind of magical properties you see and er I don't ken if it was a a holy well or exactly but it certainly was reputed to have some kind of properties that could cure supposed to cure any disease save* "the black death".
(34) *This £175* "black box"*, produced by Leeds University, records measurements from pieces of laboratory equipment like digital thermometers, oscilloscopes, resistors, timers and so on.*

However, *black box* also takes on a more general metaphorical meaning in other contexts, data gathering, more frequent in English than in Spanish. See example 35 alluding to the collecting data used to analyse the causes of ethnographic behaviour or new information about theatre (example 36):

(35) *But the critics of collectivism face a formidable task because collectivism does seem to fit so well with commonsense ideas about Japan, and the identification of unique cultural values offers a convenient residual* "black box" *which can be used to explain away those aspects of Japanese experience which don't quite fit with social science models.*
(36) *Un estudio (250 asientos), que es una verdadera* "caja negra"*, especial para teatro experimental.* (…) [A study of 250 seats which is truly a black box, specially meant for experimental theatre]

Black death is a fairly curious case as there are not metaphorical examples in English, but there is a non-literal equivalent in Spanish meaning ACCUSED/REJECTED. This linguistic expression can be looked upon as a tradition of conceptualization which forms part of the Spanish culture and its legacy (example 37):

(37) *Llegan muchas personas a la estación para despedirnos y hasta aparecen algunos de mis compañeros de escuela para quienes yo, evidentemente, ya había dejado de ser* "la peste negra"*, y con quienes, sin mucho entusiasmo de mi parte, prometí mantener correspondencia*.
[Many people came to see us off at the station, even some of my school mates turn up and I obviously had stopped to be the Black Death for them and I promised, without much enthusiasm, to keep up correspondence]

*3.2 Aspects of White/*Blanco *According to Piirainen's Model: Material Culture*

The aim of this part is to describe white/blanco examples in both languages divided into two subcategories: single lexical units (Sect. 3.2.1) and collocations and idioms (Sect. 3.2.2).

3.2.1 Single Lexical Units

No examples of "white" used metaphorically as an adjective have been found in CREA. However, there is a metonymic use of *blanco* as a noun, origin of certain idiomatic phrases. In this sense, *blanco* means a target shot (38) as the dartboard

bulls-eye was traditionally white in colour but also, in a more metaphorical sense, refers to the goal which desire or actions are directed to (39). In the latter example, the success of scoring a pretty woman is metaphorically equated to hitting the target in a dartboard:

(38) *las pruebas realizadas en laboratorio han demostrado que un proyectil de teflón que no acierte en el* "blanco" *sí es capaz de penetrar el fuselaje interno del aeroplano.* [Tests in laboratory have shown that a Teflon projectile which does not hit the target is able though to penetrate the inner fuselage airplane]

(39) *Una mujer que es el* "blanco" *de los retorcidos deseos de unos jóvenes dispuestos a la trasgresión.* [A woman who is the target of the twisted desires of youngsters ready to break the norms]

3.2.2 Idioms and Collocations

The use in both languages of *The White House* coupled with the verb "to say" is a well-known example of metonymy with the collocation "White House" actually referring to the authorities who work there. Our encyclopaedic knowledge tells us about the additional meaning of this compound. Barkema (1996: 139) calls this type of collocation "pseudo-compositional". The white flag is also a symbol of surrender in both countries since at least the Roman times.

Metaphorical idioms containing *blanco* as target are *dar en el blanco* [literally, 'to hit the white'] coming from archery whose meaning is 'to reach the goal' where OBJECTIVES ARE TARGETS and FINDING A SOLUTION IS 'dar en el blanco[9]' or the idiom *ser el blanco de todas las miradas* [literally, 'to be the white of all looks'], where *blanco* stands here for the target as the centre or focal point for stares, criticism, etc. These idioms may not be semantically transparent to speakers, and even if they are transparent, this does not mean that speakers subscribe to the views or beliefs on which a particular idiom is based:

(40) *Creo sinceramente que Disney ha* "dado en el blanco*"del preciso gusto de su época.* [I sincerely believe that Disney hit the target of the very taste of his epoch].

(41) *Ser el líder sólo tiene una desventaja: estás sometido a la crítica, eres* "el blanco de todas las miradas" *y se te juzga con mayor rigor.* [Being the leader only has only one disadvantage: You are subjected to criticism, you are the target of all looks and you are judged more strictly]

[9] This idiom is also expressed with an alternative lexicalization: *dar en el clavo* [hitting the nail].

Other metaphorical idioms in Spanish with white are a) 'quedarse en blanco/con la mente en blanco' or the more formal variant 'quedarse in albis' [somebody's mind goes blank], a metaphor related to the conventional metaphor THE MIND IS LIKE A BLANK BOARD (TO WRITE ON) because what it is written may be erased; and b) 'no distinguir lo blanco de lo negro' [literally, 'not to distinguish white from black', meaning "to be ignorant"] in examples 42 and 43 respectively:

(42) *Rodrigo Egea volvió a* "quedarse en blanco". [Rodrigo Egea went blank again].
(43) *Cuando Tarsis recibió en su barracón del campo de trabajo su primer paquete no estaba para resolver enigmas: tanto es así que se quedó alelado, sin reacción, (…), ni* "distinguía lo blanco de lo negro"... [When Tarsis got his first packet at the work field he was in no mood for solving enigmas: he was in a daze, could not react, and could not distinguish white from black…].

Another idiom is *estar sin blanca* meaning 'to to be broke'. Nowadays it may be understood metaphorically because the object *blanca* which refers to is lost. Speakers do not know the image or are even aware that there is an image involved. It was originally metonymic as it was a whitish coin made of silver and copper called "Agnus Dei Blanca" [Agnus Dei White] minted in 1386. This coin lost its value with the passing of time and was finally minted in copper and became practically worthless. Hence, a person who did not have "a blanca" was bankrupt.

(44) *La Caixa de Tarragona sabía que GT* "estaba sin blanca". *Y pidió más garantías.* [Caixa Tarragona knew that GT was penniless and asked for more guarantees]

If the relative frequencies of the various categories of sets phrases are considered (cf. Moon 1998) the incidence of this metaphorical idiom in the CREA is striking: twenty-six occurrences whereas the idiomatic *no parecerse en el blanco de los ojos* [literally, 'two persons resemble just in the white of the eye', whose meaning is they are not in the least alike] does not even appear once despite the fact that it is present in several Spanish dictionaries like the prescriptive *Diccionario de la Real Academia Española* and has 282 occurrences in Google. One possible explanation to account for this difference is that the more idiomatic set phrases tend to be rather infrequent in written texts, which make up 90 % of the CREA.

In English, there are metaphorical idiomatic expressions, such as *white heat, white hot* and *white elephant*. All of them have a metonymic origin but their current use is metaphorical. Admittedly, these idioms cannot be translated into

Spanish using a colour expression except *white heat*, translated into Spanish by the terminological locution 'al rojo blanco' just in electrical contexts.

(45) *The Conservative hegemony of the interwar years still awaits an adequate explanation, and the Conservative dominance between 1951 and 1964 – Harold Wilson's 'thirteen years of Tory misrule'- has escaped the* "white heat" *of historical investigation.*
(46) *'Dipping', with its dark waves of guitar and whirlpool melody, is like a* "white hot" *punk version of the opening scenes from 9½Weeks.*

The last collocation under scrutiny in this section is *a white elephant*, a costly possession. It has a historical background but nowadays may be regarded as metaphorical by most speakers because they are not aware of its origin. White elephants heralded the birth of Buddha and were regarded as holy in ancient times in Thailand and other Asian countries. To keep a white elephant was a very expensive business, since you had to provide the elephant with special food and that is the reason why it was frequently given away as a present. The gift would, in most cases, ruin the recipient (example 47):

(47) *The achievements are well recorded: the Brabazon airliner, the abortive TSR2 fighter plane and Concorde – the fastest* "white elephant" *of modern times.*

3.3 White/Blanco*: Culture-Based Social Interaction*

Social interaction and patterns of behaviour play a large part in the cultural semiotic foundation of phraseology. Consequently, there is a certain shared knowledge of culture-specific social phenomena among members when they process phrasemes. Piirainen distinguishes among four subgroups related to each other: (a) semiotised gestures: *to take off one's hat,* lexicalized exactly the same in Spanish; (b) gender specifics: *to wear the trousers/los pantalones,* pointing out that a person dominates in a household; (c) cultural models; (d) and bans and taboos. The two latter subgroups have been found in our colour phrasemes and are discussed in Sects. 3.3.1 and 3.3.2.

3.3.1 Cultural Models

Many proverbs give information about which values are upheld in a given culture and express norms which govern social behaviour. Four proverbs have been found in Spanish but attention will be paid just to the most relevant metonymic example,

estar blanco como la cal [Somebody is as white as whitewash], still visible in many Spanish villages in the South and Castille because there are still many dwellings painted every year with whitewash to stand off the summer heat. From this literal usage, its meaning was extended metonymically to outward signs of fear, such as losing colour in your face and is still common in everyday speech in Peninsular Spanish to highlight that a person is pale with horror (example 48). This emotion concept WHITEWASH IS FEAR is thus culturally-bound (Glässer 1999: 156) and it is not clearly universal, as Wierzbicka (1999) points out. The same concept would be expressed in English with an idiom based on a different image: "as white as a sheet". When applied figuratively in English, *whitewash* is to create a cover-up in order to make the bad seem good.

(48) Su rostro cetrino se había vuelto "blanco como la cal". [His dark face had turned as white as a sheet].

3.3.2 Bans and Taboos

There are also phrasemes which include euphemisms and allusions used to avoid talking about something directly. Many speakers avoid saying a word openly in order not to offend decency. Such seems to be the case with the Spanish proverb *Eres un viejo cebolla, la cabeza blanca y el rabo verde* [literally, 'you are like an old spring onion: your head is white and your tail green'], whose meaning indicates that you act by instinct. *White head* is a metonymy referring to old-age but it does not have a metaphorical correspondence in English, 'dirty old man'.

To conclude this cultural models section, an example from English will be commented upon. The collocation *white collar crime* is used metonymically to refer to a crime committed by a person who works in an office or professional job and used to dress with white collar shirts in the past considered as socially superior to 'blue-collar' workers.

(49) *Can the Crown avoid the excesses and mistakes of which* "white collar crime" *fighters in the US are accused?*

4 Concluding Remarks

Cultural metaphors appear to require an understanding of the input domains and their properties or connections with the output domains. The comparative outline of phrasemes containing *black/negro* and *white/blanco* clearly indicates the cultural foundation of phraseology. *Black/negro* as 'bad, unhappy' and white as 'good, innocent' are cultural facts and, taken as physical entities (colour terms),

symbolise these properties. *Black* and *white* in collocations, idioms and proverbs are powerful symbols in English and Spanish culture. The knowledge about the link between the symbols BLACK and WHITE in language and culture triggers the 'right' reading to be activated.

Piirainen's taxonomy has proved useful for this corpora contrastive analysis as proverbs, idioms and collocations absorb and accumulate cultural elements from intertextual phenomena. These fictive conceptual domains, symbols, and mainly aspects of material culture become mostly visible on the level of rich images of the source domains; food, dwelling style or elements of modern society like sports (*white money*), traffic (*black spot*) or banking (*black money*) can play a part in the literal, metaphorical and metonymic reading of multiword units.

Idioms and collocations sometimes converge in both languages as they share a fairly similar material culture and belong to the common European cultural background as shown in Fig. 1. However, some subtle differences have emerged in the cross-linguistic comparison of the conceptions that underlie the material culture. Specific cultural background knowledge and a diachronic perspective is a must in order to be able to explain the metaphors and metonymies, whose origin is frequently unknown by native speakers.

Our data of linguistic-cultural analysis projects a preliminary empirical insight for verifying the linguistic relativity hypothesis, sustaining the view that some metaphorical and metonymic colour phrasemes are cultural relevant signs *per se*, reflecting a specific national mentality which has been passed on from generation to generation as shown in the Spanish proverb example of the spring onion.

This paper demonstrates some of these idiosyncratic phraseological features which set English and Spanish colour phrasemes apart often forming a part of a specific situational context. Decoding these *black* and *white* multiword units helps to understand world knowledge and cultural socialization, frequently metaphorically and metonymically expressed.

References

Allan, K. (2009). The connotations of English colour terms: Colour-based X-phemisms. *Journal of Pragmatics, 41*(2009), 626–637.

Barcelona, A. (Ed.). (2000). *Metaphor and metonymy at the crossroads. Cognitive Perspective*. Berlin/New York: Mouton de Gruyter.

Barcelona, A. (Ed.). (2003). *Metaphor and metonymy at the crossroads: A cognitive perspective*. Berlin: Mouton de Gruyter.

Barkema, H. (1996). Idiomaticity and terminology: A multi-dimensional descriptive model. *Studia Linguistica, 50*(2), 125–160.

Berlin, B., & Kay, P. (1991). *Basic colour terms- their universality and evolution* (2nd ed.). Berkeley: University of California Press.

Berry, J. W., et al. (2002). *Cross-cultural psychology: Research and applications*. Cambridge: Cambridge University Press.

Burger, H., Dobrovol'skij, D., Kühn, P., & Norrick, N. R. (2007). Phraseology: Subject area, terminology and research topics. In H. Burger, D. Dobrovol'skij, P. Kühn, & N. R. Norrick (Eds.),

Phraseology. An international handbook of contemporary research (Vol. 1, pp. 10–18). Berlin: Mouton de Gruyter.
Burguer, H. (1991). Phraseologie und Intertextualität. In Christine Palm (Ed.), *"Europhras 90". Akten der internationalen Tagung zur germanistischen Phraseologieforschung*, 13–27.
Černyševa, I. I. (1980). *Feste Wortkomplexe des Deutschen in Sprache und Rede*. Moskva: Verl. Vyssšaja škola.
Coseriu, E., & Geckeler, H. (1981). *Trends in structural semantics*. Tübingen: Narr.
Coulson, S., & Oakley, T. (2003). Metonymy and conceptual blending. In K.-U. Panther & L. L. Thornburg (Eds.), *Metonymy and pragmatic inferencing* (pp. 51–79). Amsterdam: John L. Benjamins.
Charteris-Black, J. (2004). *Corpus approaches to critical metaphor analysis*. Basingstoke/New York: Palgrave Macmillan.
Dobrovol'skij, D. (1998). On culture component in the semantic structure of idioms. In Ďurčo, P. (Ed.), *Europhras 97: International Symposium*. September 2–5, 1997. Bratislava: Liptovský Jan. Phraseology and paremiology.
Dobrovol'skij, D., & Piirainen, E. (2005). *Figurative language: Cross-cultural and cross-linguistic perspectives*. Amsterdam: Elsevier.
Fauconnier, G. (1997). *Mappings in thought and language*. Cambridge: Cambridge University Press.
Fauconnier, G., & Turner, M. (2002). *The way we think*. New York: Basic Books.
Geeraerts, D. (2002). The interaction of metaphor and metonymy in composite expressions. In R. Dirven & R. Pörings (Eds.), *Metaphor and metonymy in comparison and contrast* (pp. 463–465). Berlin: Mouton de Gruyter.
Gibbs, R. (1994). *The poetics of mind: Figurative thought, language and understanding*. Cambridge: Cambridge University Press.
Glässer, R. (1999). Indigenous idioms and phrases in Australian and New Zealand English. In U. Carls & P. Lucko (Eds.), *Form, function and variation in English. Studies in honour of Klaus Hansen* (pp. 155–168). Frankfurt: Peter Lang.
Glucksberg, S. (2001). *Understanding figurative language*. Oxford: Oxford University Press.
Graumann, A. (2007). Colour names and dynamic imagery. In M. Plümacher & P. Holz (Eds.), *Speaking of colours and odors* (pp. 129–140). Amsterdam/Philadelphia: John Benjamins.
Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
Lotman, J. M. (1990). *Universe of the mind: A semiotic theory of culture* (A. Shukman, Trans.). Bloomington: Indiana University Press.
Makkai, A. (1978). Idiomaticity as a language universal. In J. Greenberg (Ed.), *Universals of human language* (Word structure, Vol. 3, pp. 401–448). Stanford: Stanford University Press.
Mey, J. L. (2005). Horace and colors: A world in black and white. In H. Dag & W. Eirik (Eds.), *Haptačahaptaitiš: Festschrift for Fridrik Thordarson on the Occasion of his 77th Birthday* (pp. 163–176). Oslo: Novus Forlag and Instituttet for sammerlignende kulturforskning.
Mey, J. (2014). Horace, colors and pragmatics. In J. Romero-Trillo (Ed.), *Yearbook of corpus linguistics and pragmatics 2014: New empirical and theoretical paradigms* (pp. 99–122). Cham: Springer International Publishing.
Moon, R. (1998). *Fixed expressions and idioms in English: A corpus -based approach*. Oxford: Clarendon Press.
Niemeier, S. (2007). From psychophysics to semiophysics: Categories as acts of meaning. In M. Plümacher & P. Holz (Eds.), *Speaking of colours and odors* (pp. 141–154). Amsterdam/ Philadelphia: John Benjamins.
Piirainen, E. (2007). Phrasemes from a cultural semiotic perspective. In H. Burger, D. Dobrovol'skij, P. Kühn, & N. R. Norrick (Eds.), *Phraseology. An international handbook of contemporary research* (Vol. 1, pp. 209–219). Berlin: Mouton de Gruyter.
Piirainen, E. (2008). Phraseology in a European framework: A cross-linguistic and cross-cultural research project on widespread idioms. In S. Granger & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective* (pp. 243–258). Amsterdam/Philadelphia: John Benjamins.

Sherman, G. D., & Clore, G. L. (2009). The color of sin: White and black are perpetual symbols of moral purity and pollution. *Psychological Science, 20*(8), 1019–1025.
Sinclair, J. (1987). Collocation: A progress report. In R. Steele & T. Threadgold (Eds.), *Language topics* (pp. 319–331). Amsterdam/Philadelphia: John Benjamins.
Steen, G. J. (1999). From linguistic to conceptual metaphor. In R. W. Gibbs Jr. & G. J. Steen (Eds.), *Metaphor in cognitive linguistics* (pp. 57–78). Amsterdam/Philadelphia: John Benjamins.
Steinvall, A. (2002). *English colour terms in context*. Umeå: Umeå University.
Taylor, J. R. (2002). *Cognitive grammar*. Oxford: Oxford University Press.
Vlajkovic, I., & Stamenkovic, D. (2013). Metaphorical extensions of the colour terms 'black' and 'white' in English and Serbian. In S. Zivancevic (Ed.), *Sesti medunarodni interdisciplinarni simpozijum Susret kultura-Zbornik radova* (pp. 547–558). Novi Sad: Filozofski fakltet.
Wierzbicka, A. (1996). *Semantics-primes and universals*. Oxford: Oxford University Press.
Wierzbicka, A. (1999). *Emotions across languages and cultures: Diversity and universals*. Cambridge: Cambridge University Press.
Wierzbicka, A. (2006). The semantics of colour- a new paradigm. In C. P. Binggam & C. J. Kay (Eds.), *Progress in colour studies* (Language and culture, Vol. 1, pp. 1–24). Amsterdam/ Philadelphia: John Benjamins.
Wyler, S. (2007). Colour terms between elegance and beauty: The verbalization of colour with textiles and cosmetics. In M. Plümacher & P. Holz (Eds.), *Speaking of colours and odors* (pp. 113–128). Amsterdam/Philadelphia: John Benjamins.

Electronic Resources

REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) Corpus de referencia del español actual. Retrieved 22 October, 2014, http://www.rae.es
Retrieved 22 July, 2014, en.wikipedia.org/wiki/Beijing_opera
The BNC corpus. Retrieved 22 October, 2013, http://www.natcorp.ox.ac.uk/
Online Etymology dictionary. Retrieved 22 July, 2014, www.etymonline.com

Making Informed Healthy Lifestyle Choices: Analysing Aspects of Patient-Centred and Doctor-Centred Healthcare in Self-Help Books on Cardiovascular Diseases

Georg Marko

Abstract Using a corpus-based Critical Discourse Analytical approach, this paper examines the relation between doctor-centred and patient-centred elements in self-help books on cardiovascular diseases, represented by a 3.4-million-word corpus of self-help books. The analysis of two structures, viz. acronyms and imperatives realizing the speech act of advice, suggests that contrary to its own claims, self-help health promotion represents a doctor-centred approach rather than focusing on people's lifeworlds.

Keywords Critical discourse analysis • Health promotion • Acronym • Advice • Doctor-centred discourse • Corpus linguistics

1 Introduction

There is something contradictory about the phrase *informed healthy lifestyle choices* in the title of this article. The noun *choice* suggests that individuals are given the freedom to decide for themselves what they want to do, based on what they assume is good for them. The three premodifying elements *healthy*, *informed*, and *lifestyle*, on the other hand, tell a different story, one in which this choice is constrained: *healthy* indicates that the participants' wish to retain or regain health is presupposed. *Informed* says that their decision should be based on information that they have obtained or that they will obtain. And finally, the noun *lifestyle* (in the compound *lifestyle choices*) also narrows the scope of participants' choice to practices subsumable under this concept. The phrase thus highlights the tension between an individual person's (or patient's) freedom and confidence in matters of health and illness and doctors' (or other healthcare professionals') authority trying to control

G. Marko (✉)
Department of English Studies, Karl-Franzens-University Graz,
Heinrichstr. 36, A-8010 Graz, Austria
e-mail: georg.marko@uni-graz.at

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_4

this freedom, an opposition captured in the expressions *patient-centred healthcare* and *doctor-centred healthcare*.

Assuming that linguistic choices such as those just mentioned contribute to our conception of this difference, I will examine some of these in a corpus of health promotion texts, a discourse considered highly influential in the social domain of health today. The purpose of this analysis is to find out whether the frequencies and distributions of such linguistic elements and structures can reveal which of the aforementioned trends are more relevant in the discourse under scrutiny.

The research presented here combines a Critical Discourse Analytical approach with corpus linguistic tools, examining two structures considered to play a role in promoting a doctor-centred perspective, viz. acronyms and imperatives realizing the speech act of advice, in a corpus of self-help books on cardiovascular diseases, partly in comparison with medical textbooks on the same topic. Following the tradition of Critical Discourse Analysis, I will start with contextualizing the social issue to be examined. I will then introduce my approach, my methods, and my data. Finally, I will present the analyses proper, describing and interpreting the results.

2 Doctor-Centred vs. Patient-Centred Healthcare

Doctor-centred healthcare is characterised by an imbalance between an active and benevolent doctor and a passive and compliant patient, a relationship that is based on a – real or perceived – competence gap between formal education and professional experience resulting in recognized expertise, on the one hand, and subjective semi-ignorance, on the other.

This model has come under pressure in the last decades. With the rise of chronic diseases, many of which can only be managed but not cured, doctors have lost some of their 'magic' curative powers. At the same time, patients have become experts on their own conditions, obtaining and exchanging information about their diseases through and on the World Wide Web (cf. Nettleton 2006: 10, 137–139). As a consequence, a patient-centred model of healthcare has developed, in which doctors talk to patients in order to find out what the disease and what particular therapy options mean to them, to their immediate social environment (family, friends, workplace), and the wider social context (community, culture). Based on medical information and an awareness of subjective interpretations, doctors and patients come to a shared decision of how to proceed (cf. Nettleton 2006: 149; Clarke 2010: 317).

The relationship between the two models can be best represented by positioning them in a two-dimensional conceptual matrix (a modified version of the scheme presented in Beattie 1991 for describing health-promotional strategies, cf. Clarke 2010: 359f.). The two dimensions are:

- **Mode of intervention:** Who decides on the definition of the situation and on the course of action to be taken? This is a scale with two poles:
  - **Authoritative:** Does one person decide on the two aspects mentioned on the basis of her or his access to resources – mostly information and knowledge –

and her or his higher position in a social hierarchy? Does the person highlight the competence gap and the status difference and present them as given and/or as legitimate?
  - **Participative:** Are decisions discussed and negotiated so that those involved can come to an agreement? Do they share and value each other's resources and downplay the competence gap and the status difference?

- **Domain:** In terms of which concepts is the situation and the course of action defined? This is a scale with two poles:
  - **Expert systems:** Are such definitions based on conceptualizations of the world that are abstract (from time, place, persons), objective (focusing on the empirical), and fragmented (individual aspects are observed independently from their interrelations with other aspects)?
  - **Lifeworld:** Are definitions based on conceptualizations that are concrete (concerned with specific situations and specific persons), subjective (focusing on the social and the emotional and considering the interpretations by those involved and affected), and holistic and ecological (individual aspects are observed in their interactions with and dependencies on other aspects).

Graphically, the matrix looks as follows (Fig. 1).

Doctor-centred approaches to healthcare can be located in the lower left box, patient-centred approaches in the upper right box.

The opposition between doctor- and patient-centredness is primarily relevant to traditional healthcare, which focuses on the treatment of individuals with concrete conditions. However, what is interesting is whether the opposition is also useful for understanding self-help health promotion (henceforth just *self-help*), a non-individual approach that has become an additional defining force in the social domain of health in the past decades. Self-help encompasses public and private programmes and initiatives that try to motivate the population in general or people from predefined risk groups (high blood pressure, smokers, etc.) to take better care of their health, especially through changes in nutrition, exercise, alcohol and tobacco consumption, and stress management. As self-help thus assumes that health starts in everyday life and in lifestyles and that people themselves can instigate beneficial changes, it could theoretically be located towards the bottom right corner of the healthcare matrix above, making it a patient-centred rather than a doctor-centred approach. Whether this is really the case remains to be seen as I will analyse the discourse of self-help health promotion in the practical part of this article to see whether the way language is used underlines or relativizes authority and whether it highlights lifeworld aspects of health or defines health in terms of expert systems.

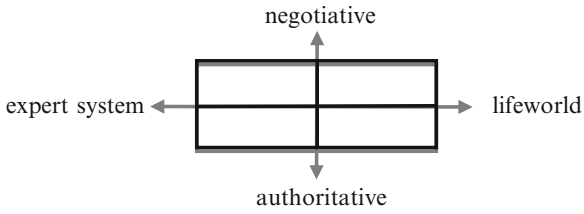


Fig. 1 The healthcare matrix

3 Approach, Method and Data

This chapter will be concerned with how the tension between doctor- and patient-centredness can be turned into a linguistically researchable question, what this question would be, and how it can be answered. Reversing the order of these aspects, I will start with the How, describing my approach (Critical Discourse Analysis) and my method (corpus analysis), to finally formulate a more concrete research question.

3.1 The Approach: Critical Discourse Analysis

Critical Discourse Analysis (= CDA) is an approach to the study of discourse that examines the role that language use plays for the construction of socially problematic and/or contested beliefs about, and attitudes towards, the world, social identities, and social relationships. CDA assumes that language normally conceals the relevance of these beliefs and attitudes, representing them as natural and as taken for granted even though they are always biased and work in favour of some groups. Only an analysis of the linguistic details of texts and their contexts can reveal the ideological operation of language. This form of linguistic critique is CDA's main objective. (For my version of CDA, cf. Marko 2008, for approaches that have been influential for my approach, cf. Fairclough 1992; Mautner 2012; van Dijk 1998; Reisigl and Wodak 2009).

3.2 The Method: Corpus Analysis

Tracing patterns in language use by means of the analysis of large electronic corpora is the method applied in my approach to CDA, here and elsewhere. The software used is Wordsmith Tools 6, created by Mike Scott (available at http://www.lexically.net/wordsmith/version6).

There have been people before me using corpus linguistic tools in CDA, most notably Gerlinde Mautner, who established the field almost 20 years ago (1995 (Hardt-Mautner), 2012), and Paul Baker (2006; Baker et al. 2008), to name just two researchers most prominent in defining the discipline. What I share with them and others in the field is the conviction that working with special corpora and quantitative methods adds a strongly empirical and thus objective element to an otherwise hermeneutic approach prone to subjectivity. My research, like theirs, combines quantification with qualitative interpretation, and particularly at the informal stages of a study alternates between concordancing and individual text analysis in order to find the patterns relevant for the research questions at hand. There are, however, three major differences between other researchers' approaches (pretending that they form

a homogeneous group) and mine regarding (a) quantification, (b) inductive vs. deductive procedures, and (c) the importance of grammar. I will discuss these three points below. Despite a bias in my own favour, this discussion is not intended as critique, just as a highlighting of differences.

The first difference is that I go a step further in quantification. It is common practice in corpus-based discourse analysis to establish quantitative relations first, e.g. by looking for a specific element or structure and determining the number of occurrences, and then to take a closer look at the occurrences in order to make qualitative statements about the co-text, illustrated with examples, following Stubbs' (1994: 212) suggestion that there is "the need to combine the analysis of large-scale patterns across long texts with the detailed study of concordance lines." Aspects found in concordance lines can, however, also be quantified, which I normally do. This can, for instance, be seen in the way I break down the co-text of imperative *eat* into countable elements (see chapter 4 "Women and Men Talking About Men and Women in Greek").

Secondly, the approaches mentioned above mostly apply inductive procedures. They prefer to examine general corpus data such as lexeme frequencies and comparative keywords and interpret these or derive information on which individual expressions deserve more profound analysis from such data. By contrast, I proceed deductively, starting with assumptions about specific beliefs and attitudes. I then select linguistic elements and structures that – more or less immediately – may contribute to constructing these beliefs and attitudes and examine them in the corpus. In the present article, I, for instance, argue that acronyms and advice realized by imperatives play a role in creating doctor-centredness and undermine patient-centredness in healthcare.

Thirdly, there is a strong lexical bias in corpus-based research in CDA, with grammar not figuring prominently on any level. In the analysis of collocations introduced in Baker et al. (2008: 286–289), for instance, the collocates of nouns are not differentiated with respect to their grammatical relationships, i.e. whether they are premodifying adjectives, verbs which take the noun as a subject or as an object, etc. This allows for a wider perspective, as elements from various dimensions can be subsumed under the same semantic or functional heading – after all, a verb and an adjective might add similar aspects of meaning to the noun. However, a grammatical differentiation might also help in finding finer semantic nuances. As demonstrated in my research in this article, I am more interested in grammatical structures – both acronyms and imperatives (as the main tool for realizing advice) are grammatical categories in the first place[1] – and their lexical realizations than in lexemes in their own right, and I do not analyse collocates without grammatically distinguishing different classes of these.

[1] As acronyms are concerned with lexemes, it could be objected that they are a lexical rather than a grammatical category. The category of acronyms, however, refers to a specific, formally-definable way of creating lexemes rather than with concrete lexemes and this is why I subsume the said category under grammar rather under lexicon.

While my research might deviate from 'standard' corpus-based CDA, it must appear much more marginal if compared to the core of corpus pragmatics. I will briefly explain why there might be a mismatch between – especially my version of – corpus-based CDA and corpus pragmatics. I would still maintain that differences are superficial rather than substantial.

At the foundation of the said mismatch is the fact that prototypical corpus pragmatic studies appear more coherent, conclusive and comprehensive than research in CDA using corpora. This, I would argue, is a consequence of their different research goals. Corpus pragmatic studies – or pragmatic studies in general, for that matter – seek to gain an understanding of what the general functions and effects of particular linguistic elements or structures are or, conversely, how specific functions are realized linguistically. Typically, "the starting point is either a discourse particle with a fixed form that can easily be retrieved from a large corpus, or a speech function that is generally realized in a small number of variant patterns" (Jucker et al. 2009: 4). These are then analysed with respect to their immediate modal and interactive meanings (judging from collections such as Partington et al. 2004, the 36/2 and the 36/9 issues of the *Journal of Pragmatics* (2004); Romero-Trillo 2008, and 2013[2]). CDA's general objective, on the other hand, is for the analysis of language use to find beliefs and attitudes constructed in and through texts. As an exhaustive analysis of all potentially relevant linguistic elements and structures is not possible, CDA is selective and confines itself to the study of a few of these. This is all the more relevant in work with corpora, which requires researchers to define search criteria precisely and which leaves less space for *ad hoc* interpretations than the qualitative analysis of individual texts. But analysing only a small proportion of elements and structures potentially interesting and relevant – for instance acronyms and the speech act of advice, as in the current article – means that the approach cannot do full justice to the objective and research will consequently always remain – or appear to remain – sketchy and incomplete.

The impression of incompleteness may also be the result of the different roles of theorizing. For CDA, the amount of theory concerning a linguistic element or structure that needs to be considered or developed is only a preparatory step for the analysis of the social questions about beliefs, identities, and relationships. In comparison to prototypical pragmatic research, this will appear insufficient or even inadequate. Theorizing imperatives and advice in this article will thus almost necessarily fall short of the standards required of purely pragmatic work as it will fail to consider and examine all aspects relevant to advice giving and imperatives.

[2]The other topic of corpus pragmatics consistently to be included is what Romero-Trillo (2008: 7) describes as "[c]orpus linguistics and Intercultural Pragmatics from a theoretical and language learning stance."

3.3 The Data: The Self-Help Corpus and the Textbook Corpus

This article will focus on medical self-help books on cardiovascular diseases as a prominent genre of self-help health promotion. Written by experts, these books combine medical information about specific conditions with advice on how to behave in order to prevent the onset or the progression of these conditions. I have decided to concentrate on the topic of cardiovascular diseases (= CVDs) because of their epidemiological prominence – approximately a third of all deaths are attributed to CVDs, at least in the EU as a whole, in the United Kingdom, and in the United States (cf. OECD 2012: 22; Murphy et al. 2013: 3–5, 7) – and because of their close association with lifestyles, the starting point for any health promotion activities.

The corpus consists of 50 self-help books. I have included books that specifically focus on the most important conditions subsumed under CVDs, with the number of books in the respective categories dependent on the relative weight I attribute to the conditions. The corpus includes 20 books on heart disease, 10 book on cerebrovascular (especially stroke) and peripheral vascular diseases (especially deep vein thrombosis), and 5 each on hypertension, cholesterol and diabetes. To stress the practical side of the issue, 5 books on dietetics and food in relation to health have been added.

The books were obtained from Amazon.co.uk, found with simple search phrases, e.g. *heart disease*, *stroke*, or *high blood pressure*. Books were not randomly selected because 'Popularity' rank, a category based on sales and availability, was taken as a criterion. The process was undertaken in early 2010.

I have also compiled a comparative corpus in order to relativize quantitative results. I have decided to use introductory medical textbooks on cardiovascular diseases. These are similar to the self-help books in that they, too, introduce a topic even though the relationship between writer and reader is more expert-to-expert. However, as the intended readership consists of experts in other fields, e.g. psychiatrists, general practitioners, or nurses, or students, these textbooks also involve a certain hierarchical relationship based on a competence gap.

The comparative corpus is smaller, consisting of only 20 books, again selected with reference to the six topics mentioned above and according to the same quantitative relations (8 heart diseases, 4 cerebrovascular and peripheral vascular diseases, 2 each for high cholesterol, high blood pressure, diabetes and dietetics and nutrition). The reason for this difference in size was the difficulty to find textbooks on all topics covered. High cholesterol, for instance, seems to be a topic for popular discussion but in science it is not normally treated independently. It would therefore have been impossible to include 5 textbooks on this issue.

The books were scanned and transformed into text files using OCR software.

The size of the self-help corpus is 3,435,354 word tokens, that of the medical textbook corpus 1,437,265 word tokens.

The corpora have been structurally annotated, especially for marking textual elements that do not belong to the normal, linear running text or that do not

represent the author's words. These elements include tables, diagrams and pictures, references, quotes and recipes. Assuming that such textual elements are not read in the same way as the main text, I exclude them from all the analyses below. Imperatives – one of the structures to be analysed – therefore do not include the instructions used in recipes. The corpora have also been tagged for parts of speech using the automatic tagging programme CLAWS via Wmatrix (cf. Rayson 2009). For illustrative purposes, I have added a passage from the textbook corpus, complete with structural tags (in bold, with explanations between brackets) and word-class tags.

<gr> [= opening tag for graphic elements in general, excluded from analysis] *<diagram>* [= tag for the category of the graphic element; non-binary, as only the category is relevant, everything else is covered by the binary tags <gr></gr>] *Figure < NN1 > 3.2 < MC > The < AT > SCORE < NN1 > relative < NN1 > risk < NN1 > chart < NN1 >* **<p>** [= paragraph]

Note < VV0 > :<:> Risk < NN1 > is < VBZ > expressed < VVN > as < II > a < AT1 > multiple < NN1 > of < IO > the < AT > lowest < JJT > risk < NN1 > (<(> one < MC1 >)<) > and < CC > not < XX > as < II > a < AT1 > percentage < NN1 >. <.> **<p>** [...]

<ref> [= opening tag for reference; excluded from analysis] *Eur < NN1 > Heart < NN1 > J < ZZ1 >,<,> 28(19) < FO > :<:> 2375-2414 < MCMC >. <.>* **</ref>** [= closing tag for references] **</gr>** [= closing tag for graphic element]

3.4 Research Question

The general research question is: How and to what extent does the discourse of self-help health promotion construct the information conveyed as part of expert systems and the relationship between writers (= doctors) and readers (= patients) as authoritative? How and to what extent does it include aspects that modify, enhance or mitigate these dimensions?

My hypothesis: Assuming that self-help at least partly takes a patient-centred perspective, I expect the self-help books on CVDs to feature linguistic elements and structures that contribute to a participative and lifeworld-oriented form of healthcare, especially in comparison to the medical textbooks examined.

This hypothesis is the reference point for the analyses below. Given the critique of health promotion especially in health sociology concerned about its exclusive focus on individuals' lifestyles and its medical orientation (cf. Nettleton and Bunton 1995: 44), it is, however, not likely to be supported throughout.

I will take a language-centred approach, focusing on two structures that on closer inspection appear relevant to the questions posed above, viz. acronyms, which present knowledge in a condensed and only selectively accessible form, and imperatives, which are used in advice and therefore add to the social gap between writer and reader.

4 Acronyms

Accessibility and transparency of information is closely associated with the opposition between authoritative vs. participative. The basic questions in this context are: Do writers present information in an inaccessible and technical way to highlight the gap in knowledge between them and their readers – equivalent to doctors and patients – or do they avoid this in order to reduce the vertical distance between them?

Technical vocabulary is of central importance for this question. It is defined as terms that within a particular social domain, e.g. within science, have specific, unambiguous meanings, free of evaluative connotations, with particular groups being granted 'primary access to them', i.e. members learn about the meanings in their education or training and only they are allowed to use them (cf. Marko 2012). In medicine and the natural sciences, technicality is mostly achieved through opacity. This involves using elements that are so far removed from everyday language that lay people are not familiar with their meanings and cannot easily deduce them from other evidence. Acronyms can be argued to achieve this effect.

Acronyms are abbreviations perceived as – usually pronounceable – words in their own right. The category covers alphabetisms, where individual letters are named (cf. Booij 2007: 20), e.g. *CVD*, and acronyms in the narrow sense, where letters form new word pronounced according to regular grapheme-phoneme-correspondence (cf. Plag 2003: 127), e.g. *PET* 'positron emission tomography'.

The major function of acronyms is to condense multi-word expressions into single lexical units that are easier to handle in practice. An intended or at least welcome side effect of acronyms is, however, also semantic opacity, i.e. the fact that it is not immediately obvious which elements, for instance, *PET* consists of. Acronyms can thus be argued to keep and declare expert information as inaccessible to lay people.[3] People more immediately affected may know some of these words because they feature in sources read. But overall, an 'overuse' of acronyms will probably have the effect described as even more knowledgeable people will still feel that the words are part of science rather than part of their lifeworlds.

Following my hypothesis, I predict acronyms to be more common in the textbook corpus, as it represents language used by experts for experts and the addressees can be expected to know or be about to learn the meanings of these words. Self-help books, on the other hand, will not foreground a competence gap through the use of words that will remain difficult to understand for the intended lay readership.

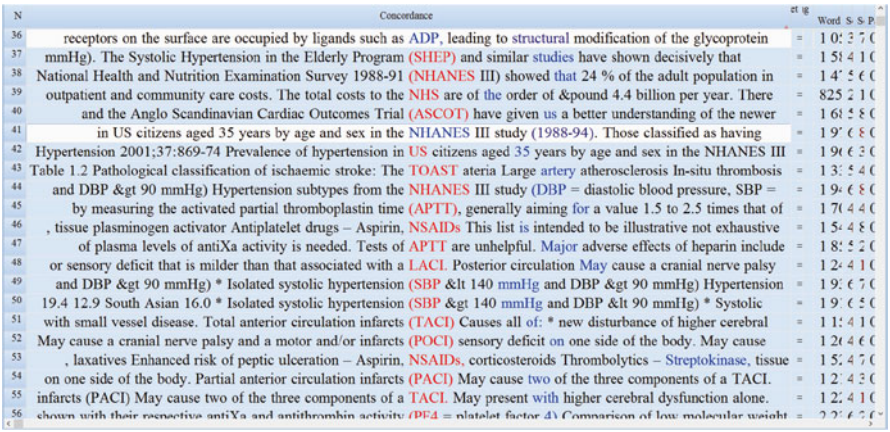
The corpora were searched for words spelt in capitals, as this is a major formal feature of acronyms.[4] To track these, I searched for words in which the *second* letter

[3] Informal and widely known acronyms popular in computer-mediated communication such as *lol* or *IMHO* do not occur in the corpora.

[4] Lower-case abbreviations – for example *e.g.*, *i.e.*, and units of measurement – are not considered acronyms here because they are always used with full stops between letters and are not even theoretically pronounceable.

was capitalized. The assumption was that while capitalization of the first letter is not distinctive, capitalization of the second letter almost always means that all letters are capitals (or that the word mixes upper- and lower-case letters, a feature also predominantly found in acronyms, e.g. *tPA* 'tissue plasminogen activator'). Using this search string produced concordances such as the following one (extract):

Concordance 1: Concordance of words featuring a second letter that is capitalized in the textbook corpus (extract)



After manually deleting irrelevant examples, I derived lists of the acronyms found from the remainders of the concordances.

Homonymous acronyms, where the same form is an abbreviation for different expressions, e.g. *ED* 'erectile dysfunction', 'eating disorder', and 'emergency department', have been disambiguated. *ED* is thus, for instance, counted extra for all three meanings, representing three different lexemes.

Table 1 presents token frequencies and the numbers of different lexemes (= types; they have been lemmatized, e.g. *CT* and *CTs* are part of the same entry).

The textbook corpus contains almost two thirds more different acronyms than the self-help corpus and these acronyms are used three times as often, relatively speaking. So the data supports the assumption that the discourse of self-help will avoid presenting itself as too technical and inaccessible so as not to widen the competence gap, at least in comparison to the discourse of medical textbooks.

This conclusion, however, has to be taken with a grain of salt since despite this relatively large gap, acronyms are far from rare in the self-help books considering that more than 1,000 different ones have been found and that we encounter a

Table 1 Type and token frequencies of acronyms in the self-help corpus and the textbook corpus

	Types	Tokens	Per 10,000 words
Self-help	1,019	21,611	62.9
Textbooks	1,680	26,787	186.4

member of the class every 166 words of running text. So while clearly not the most prominent feature of self-help discourse, there still is an element of inaccessibility present.

Even a cursory look at the data in the self-help corpus will reveal that the acronyms are very technical and their use will not make comprehension easy. As an example, all acronyms for diagnostic procedures and devices found in the corpus are listed below.

> *CAT* [Computed Axial Tomography/Computer Assisted Tomography]; *CMR* [Cardiac/Cardiovascular Magnetic Resonance (Imaging)]; *CT* [Computer Tomography]; *CTA* [Computed Tomography Angiography]; *CXR* [Chest Radiograph/Chest X-Ray]; *DEXA* [Dual Energy X-Ray]; *DSA* [Digital Substraction Angiography]; *EBCT* [Electron Beam Computerized Tomography; *EBT* [Electron Beam Computerized Tomography]; *ECG* [Electocardiogram]; *EEG* [Electro-encephalogram]; *EKG* [Electrocardiogram]; *EMG* [Electromyography]; *EP* [Electrophysiology]; *IVIDCT* [64-Channel Multi-Detector Computed Tomography]; *IVU* [Intravenous Urogram]; *IVUS* [Intravascular Ultrasound]; *MDCT* [Multidetector Computed Tomography]; *MPS* [Myocardial Perfusion Scintigraphy]; *MR* [Magnetic Resonance]; *MRA* [Magnetic Resonance Angiography]; *MRI* [Magnetic Resonance Imaging]; *MUGA* [Multigated Acquisition (scan)]; *NMR* [Nuclear Magnetic Resonance]; *PCT* [Perfusion Computerized Tomography]; *PET* [Positron Emission Tomography]; *RVG* [Radionuclide Ventriculography]; *SPECT* [Single Photon Emission Computed Tomography]; *TCD* [Transcranial Doppler]; *TEE* [Transesophageal Echocardiogram]; *VQ* [Ventilation Quotient/Perfusion (scan)]

While we may be familiar with *CT*, *ECG*, or *MRI* and know what to expect if we have to undergo one of these procedures, being told about the possibility or the necessity of an *IVU*, an *MPS*, or a *DEXA*, we may feel lost and exposed to a form of healthcare far removed from our everyday experience.

A follow-up question is whether there are any semantic domains where acronyms are more common than in others and thus whether their effects is confined to particular areas of knowledge. For this purpose, the expressions found were assigned to different semantic categories, drawing on a combination of universal and discourse-specific classes, all of which should be self-explanatory. The relative sizes of these categories were then calculated with respect to both lexical variability (how many different expressions in one category) and frequency of occurrence (how many tokens in one category).

As technical terms, acronyms will be used most extensively in those conceptual domains in which – especially medical – expertise plays a role, viz. anatomy & physiology, pathology, (bio)chemistry, healthcare and scientific practice, in both corpora. As health promotion should be interested in leaving people's lifeworlds intact, according to my hypothesis, I tentatively predict that the sizes of the 'non-expert' categories, e.g. lifestyle, society, and communication, will be smaller, relatively speaking, in the self-help corpus than in the textbook corpus, and that there will be a higher concentration of such words in the aforementioned 'expert' categories, i.e. the latter will be larger.

The results are contained in Table 2. Percentages represent the proportions of the number of terms assigned to one category in the overall number of terms, e.g. if 20 of 50 terms belong to one category, then the percentage of the latter is 40 %.

Table 2 Sizes of different semantic categories for acronyms in the self-help corpus and the textbook corpus

	Self-help				Textbooks			
	Types		Tokens		Types		Tokens	
		%		%		%		%
Anatomy & physiology	74	7.3 %	2,345	10.9 %	225	13.4 %	4,378	16.3 %
Pathology	162	15.9 %	2,950	13.7 %	335	20.0 %	8,101	30.2 %
(Bio-) Chemistry	175	17.2 %	7,464	34.5 %	242	14.4 %	4,289	16.0 %
Scientific practice	74	7.3 %	430	2.0 %	341	20.3 %	1,781	6.6 %
Healthcare	308	30.2 %	5,793	26.8 %	421	25.1 %	6,768	25.3 %
Communication	29	2.8 %	278	1.3 %	7	0.4 %	12	0.0 %
Lifestyle	27	2.6 %	409	1.9 %	32	1.9 %	295	1.1 %
Society	131	13.0 %	1,660	7.7 %	50	3.0 %	1,007	3.8 %
Technology	3	0.3 %	8	0.0 %	3	0.2 %	33	0.1 %
Time	4	0.4 %	113	0.5 %	2	0.1 %	13	0.0 %
Undefined	32	3.1 %	161	0.7 %	22	1.3 %	110	0.4 %
TOTALS	1,019		21,611		1,680		26,787	

Given the difference in the numbers of acronyms found in the two corpora, a comparison must be treated with caution since the same percentage still means that the respective category is more important in the textbook corpus because of the higher overall frequency of acronyms.

As expected, the categories immediately concerned with medical expertise together claim the vast majority of elements in both corpora. Not consistent with my predictions, some of the relative sizes of these categories are considerably smaller in the self-help corpus than in the textbook corpus. The only exception is (bio-) chemistry, which is significantly larger in the self-help corpus, especially with respect to tokens, mostly as a result of very high frequencies of terms concerned with cholesterol and blood glucose, e.g. *LDL* 'low density lipoprotein' (= bad cholesterol), *HDL* 'high density lipoprotein' (= good cholesterol), *BG* 'blood glucose'. As these are bodily parameters that can be directly linked to lifestyles – blood sugar and cholesterol are tightly associated with too many carbohydrates and too much animal fat in someone's diet – (bio-)chemistry becomes a lifeworld-related domain. This in turn means that, contrary to my hypothesis, the self-help books 'export' some technicality and inaccessibility and thus aspects of the expert systems to the lifeworld in the shape of acronyms. This assumption receives further support by the fact that the self-help corpus contains more acronyms in the lifeworld-oriented categories, especially those concerned with society.

In sum, the general data on acronyms examined in this chapter suggests that technicality and inaccessibility created by acronyms are a stronger element in medical textbooks than in self-help books. The hypothesis nevertheless is not fully supported by the results, firstly because the self-help corpus still contains a

large number of acronyms, and secondly, if analysed according to semantic categories, the data further shows that self-help books use technical acronyms in – and thus impose conceptualizations associated with these onto – areas normally attributed to people's lifeworlds.

5 Imperatives

This section will deal with giving advice. Assuming that advice is a speech act that contributes to vertical social distance between speaker and addressee, I will look at how extensively it is used in the self-help corpus (no comparison to the textbook corpus here). This chapter, however, will also be concerned with the problems involved in pursuing this issue beyond mere overall figures on imperative use and with how they could be resolved.

Self-help, as mentioned, primarily tries to make people, especially those defined as being at risk, change their lifestyles, adopt a healthy diet, engage in regular physical exercise, stop smoking, drink less alcohol, and manage their stress. The incentive for the intended change is information about its benefits and about the risks of continuing to live as before. A self-help text is thus like one large piece of advice, and advice generally can be assumed to be a dominant speech act on the micro-level, too.

Searle (1969: 67) defines advice as "telling you what is best for you." At least in the prototypical case, what is best is based on a rational conclusion from information and knowledge available to the addressee.

As the outcome of advice is in the addressee's best interest, the imperative, which directly and without redress commits the addressee to some form of behaviour (*Do X!*), should be the preferred linguistic tool for performing this speech act. This, however, only applies to an idealized scenario in which:

(a) the speaker has a higher social status than the addressee, at least in the social domain in which the situation is set,
(b) the status of the knowledge on which to base the advice is unproblematic,
(c) there is a certain urgency for the addressee to engage in the targeted behaviour,
(d) there are no other constraints on the targeted behaviour.

The conditions are not necessarily pre-given but may be implied by the formulations used. Imperatives thus may also serve to convey to the addressee that at least the speaker assumes and wants the addressee to assume, too, that conditions (a.)–(d.) are met. Regarding the distinction between doctor-centred and patient-centred healthcare, this could mean that *Do X!*, if interpreted as advice, will make the discourse more authoritative as the speaker puts her- or himself above the addressee, even if temporarily and restricted to the current situation, mostly as she or he is in possession of knowledge required to solve the addressee's problems. The status of the knowledge and of the authority based on it is – indirectly – presented as

unproblematic. While information is concerned with the addressee's lifeworld – after all, the intended effect is a change of lifestyles, which are by definition part of the lifeworld – the foundation is more likely to be some form of expert knowledge than personal experience. This knowledge may or may not be made explicit as part of the attempt to legitimate the advice.

Any deviation from (a.) to (d.), on the other hand, i.e. if speakers do not feel superior to the addressee, if they believe that they do not have enough knowledge or that the knowledge they have is questionable, if there is no immediate urgency, or if speakers think that performing the intended behaviour will not be easy for the addressee, they will resort to more indirect formulations, e.g. constructions involving modality, e.g. *you must/should do X*, or even more tentative ones. The latter may include using performative verbs, usually *advise* and the more intense *ask*, *urge*, or *implore*, in plain (*I advise you*) or mitigated (*I'd advise you*) form, e.g.:

If you already have established heart disease, I advise you to choose a large hospital with experience in cardiac care and a training program, if possible

If you have hypothyroidism, I urge you to go natural

I'd still advise you to save your money for real food.

The need to name the speech act explicitly in the performative verb may not be indirectness, technically speaking, but it suggests that the speaker does not feel absolutely confident with respect to her or his authority or expert status. Performative verbs are very rare in connection with advice in the corpus, though.

More common – and certainly the most common form of indirect advising – is the use of evaluative adjectives as descriptors for the act, e.g.:

It is best if you treat the stain while the blood is still wet […]

Since no nutrients work in isolation, it is a good idea to take a good high-strength multivitamin and mineral supplement.

[…] *it is critically important that you are aware of your risk of developing and dying of coronary heart disease.*

I will not include such realizations of advice in the analysis. Most importantly, indirect advice is more tentative, indicating problems with conditions (a.)–(d.) as defined above. Such constructions therefore cannot be assumed to have the same effect of highlighting authority. Besides, the constructions are not very common: *be a good idea to do*, for instance, occurs 68 times, *it is good/better/best to do* (with grammatical variations) 154 times, figures that appear insignificant if compared to almost 22,000 imperatives (see below).

Although modal and semi-modal verbs with second person *you* as the subject are the second most important tool for realizing advice, I will not include this structure in the analysis either, also because they do not express the illocution as directly as imperatives, without relativizing one of the four conditions mentioned. In comparison with indirect constructions, modals and semi-modals are used much more often. Table 3 presents the frequencies of the different forms.

These figures are 'raw' in that they do not differentiate between deontic and epistemic usages of the respective constructions (the latter seems to be very rare, though), and do not include advice with a subject other than *you*, e.g. *Blackouts must be treated seriously.*

Table 3 Frequencies of deontic constructions (with 2nd person subjects) in the self-help corpus

Imperative	21,968
you should	1,512
you need (to)	1,318
you have to	764
you must	358
you could	271
you ought to	9

It has been tacitly assumed so far that imperatives in self-help books have no other function than giving advice. Broadly speaking, this is probably the case. The other main functions of imperatives, viz. to realize orders and requests (e.g. *help me*) and offers (e.g. *help yourself*), are irrelevant in self-help books. However, there is also the related category of instruction. It differs from advice in that instructions normally involve a series of acts leading to a certain goal. Here is an example of instructions from the corpus, describing how to perform a certain exercise.

> *Then inhale deeply and push your abdomen out like a balloon, hold your breath for about five seconds while contracting your abdomen, and then let your abdomen completely relax as you exhale, through your mouth. Repeat the same sequence with your upper chest.*

In contrast to advice, the individual instruction does not have an immediate benefit for the addressee. *Stop smoking* thus qualifies as advice since discontinuing the habit is seen as having several direct health benefits. This does not apply to *Hold your breath for about five seconds*, the third act mentioned in the example, which will only have some positive effect in combination with the other acts.

Imperatives that have a formulaic status and whose illocutionary force is consequently relatively weak (e.g. *believe it or not*, *guess what*, etc.) may also fall outside the typical imperative-as-advice pattern.

Despite these differences, I have decided not to discard imperatives realizing instructions and formulaic imperatives for two reasons. Firstly, it is very difficult if not infeasible to differentiate between these two usages and imperatives realizing 'genuine' advice. And secondly, in all the cases discussed, the illocution – or whatever is left of the illocution in formulaic imperatives – is closer to the advice pattern than to the order pattern in that the outcome of the act is supposed to be beneficial for rather than 'costly' to the addressee. I will therefore not further distinguish between different imperative functions.

How frequently do imperatives now occur in my corpus of self-help books? According to my hypothesis about self-help being more patient-centred, I should expect advice not to be very commonly realized as imperatives, as the latter emphasize the competence gap between doctors and patients and thus could be argued to support a doctor-centred view of healthcare.

Technically speaking, imperatives can be traced in the corpus with the help of the word-class tags. The set of tags used by the automatic tagger CLAWS, however, provides only one tag covering the base form (the unmarked – i.e. not 3rd person

Table 4 General statistical data on imperatives in the self-help corpus

Number of imperatives:	21,968
Relative frequency:	0.6 % (6,395 per 1 million words)
Comparative values:	1,000 in news and academic writing
	2,000 in fiction
	10,000 in conversation
	(all per 1 million words) (cf. Biber et al. 1999: 221)
Number of (non-modal) verb forms:	569,599
Percentage of imperatives:	3.9 %

singular – present simple tense form) and the imperative. Base forms therefore had to be manually eliminated from the concordance list.

Table 4 contains all the general figures relevant for the use of imperatives in the self-help corpus. There will not be a comparison with the textbook corpus as scientific textbooks rarely directly address their readers, let alone give them advice in the form of imperatives.

As shown in Table 4, imperatives are very frequent in the self-help books, being three to six times as common as in other forms of written language according to Biber et al. (1999: 221). This suggests that, in contradiction to the hypothesis, direct advice is a characteristic feature of self-help books, which will emphasize rather than mitigate doctor-centredness.

In comparison to the figures cited in Biber et al. (1999: 221), the self-help books are closer to spoken than to written language with respect to their use of imperatives. This is partly due to the books' interactive and personal style. But as advice constitutes the main superordinate purpose of the books, we may wonder why the number of imperatives is not even higher, exceeding that for spoken language. There is probably a simple pragmatic reason: imperatives in conversations are not just used for advice, but also for orders, requests and offers, the other main functions of this structure. So advice realized by imperatives is certainly more common in self-help books. But as orders, requests and offers are not normally found in the latter, imperatives overall occur less often than in spoken interaction.

The next relevant step in a comprehensive analysis of imperatives is a closer look at the verbs actually used in this form. This allows insights into where doctor-centredness is more important. My hypothesis would suggest that this should not be domains, especially lifeworld domains, where medical expertise normally is irrelevant. As will become clear very soon, an exhaustive analysis of all verbs or even a selection of high-frequency verbs appears infeasible. To explain why this should be the case, I will present the twenty most frequent imperative verbs in the self-help corpus in Table 5.

Some of the verbs allow conclusions about the semantic domains that they foreground, e.g. *eat*, *add*, *use*, or *check*. Others, however, are structurally highly variable, i.e. they occur in different constructions. *Get* could, for instance, occur as *get on*, *get off*, *get out*, *get back*, *get along with*, *get something done*, *get somebody to do something*, *get something* ['obtain'], *get somewhere*, or *get* + adjective. All these constructions would have to be counted separately as they often are very distinct in

Table 5 The 20 most frequent verbs used in the imperative in the self-help corpus

see	2,502	*keep*	465
take	1,060	*be*	391
make	713	*check*	376
try	699	*think*	309
remember	608	*go*	307
avoid	603	*do*	276
ask	582	*let*	267
eat	536	*consider*	251
use	529	*look*	248
get	466	*add*	242

meaning with no real common semantic core. If there was a limited set of options – e.g. *see* being used as either 'visit', as in *see your doctor*, or 'check, compare', as in *see Chapter 4* – this would still be feasible. But with some verbs, there is no theoretical limit to the constructions and the concomitant meaning differences that must be considered.

Catenative verbs, i.e. verbs taking verbal complements, pose another problem to a more comprehensive analysis of imperatives. With words such as *try*, *remember*, *avoid*, *keep* ('continue'), *let*, or *consider*, the scope of the advice could be expanded to the complement. In *Try to eat more fruit*, for instance, the addressee is advised to eat more fruit, not just to try. We may even argue that the verb in the complement is more central to the intent of the imperative than the main verb. So should we count one of them or both? What complicates matters even further is that catenative verbs serve as mitigators, i.e. writers may downtone the obligation inherent in the advice by not telling someone *to act* but rather telling them *to try to act*, *begin to act*, *consider acting*, etc. It remains unclear whether we are then allowed to subsume these constructions under the same heading as straightforward imperatives at all.

As I have not worked out a solution to these problem, I have decided not to analyse the whole – or a select – set of verbs in the imperative. As an alternative, I will look at individual verbs and their environment, structurally and semantically, examining whether such verbs tend to go together with elements from particular semantic domains. If the domains are indeed restricted, then we speak of a semantic preference. Semantic preferences may be discourse-specific, revealing links relevant in the conceptualizations of the world created in the respective discourse.

Semantic preferences become manifest in the collocational patterns that linguistic units show, i.e. the group of words with which they co-occur in a corpus. The task of corpus analysis is to find such patterns, always with respect to particular grammatical and semantic functions, e.g. subjects, direct objects, adverbials, and analyse the semantic categories that they can be assigned to.

I will do an exemplary analysis of the verb *eat* as one of the most frequent verbs in the corpus (not just as an imperative). Analysing the linguistic environment means looking at the linguistic structures that realize the following semantic aspects:

(a) **Food:** What should the addressee eat?
(b) **Attributes:** Which attributes does the food have?

(c) **Quantity:** How much (of the food) should the addressee eat?
(d) **Frequency & duration:** How often or for how long should the addressee eat (the food)?
(e) **Other aspects:** Is the situation of eating further defined, e.g. by when, where, or how? Is the advice giving further defined, e.g. by why, for which purpose, despite what?

These elements tend to be expressed by certain structures in the imperative clause (but are not necessarily restricted to those) (Table 6).

I will not do an exhaustive analysis of all the aspects mentioned, but will confine myself to some salient points. The most important category in the understanding of *eat* in advice is, of course, the object of the activity, i.e. the food. Food can be conceptualized on four different levels, (i) as a meal (e.g. *lunch*), (ii) as a dish (e.g. *spaghetti carbonara*), (iii) as an aliment (the individual food item, e.g. *carrot*), or (iv) as a nutrient (e.g. *fructose*). As pointed out in Marko (2009), in meals, eating is conceptualized primarily as a social event, in dishes as a matter of taste, while in aliments and nutrients, food is invested with ecological and social values and a rational – normally health-related – functionality. We thus eat to practice a certain form of morality, and we eat because we rationally know that food fulfils a function in our bodies.

Can we derive any concrete expectations from my hypothesis concerning which of these categories will be prominent in the corpus? The fact that the construction

Table 6 Linguistic structures used to represent different semantic aspects in imperative clauses with *eat*

Semantic aspect	Linguistic structure	Example
Food:	Direct object	*Eat more walnuts*
Attributes:	Modifiers of the direct object (mostly adjectives and relative clauses)	*Eat fresh fruit*
		Eat vegetables that are rich in potassium
Quantity:	Quantifiers or quantifying expressions in the direct object	*Eat more green vegetables*
		Eat 1 clove of garlic every day
	Modifiers in the direct object	*Eat small meals*
	Adverbials	*Eat red meat in moderation*
		Eat lightly
Frequency & duration:	Adverbials	*Eat fish at least twice a week*
	Adverbial clauses	*Do not eat anything before the test has been done*
Other aspects:	Adverbials	*Eat slowly*
		Do not eat in front of the TV
		Eat little and often, with plenty of fruit as snacks in between
	Adverbial clauses	*Eat less fat to encourage weight loss*
		Eat a multi-coloured variety of foods, as each natural colour contains different health-related benefits

Table 7 Sizes of the semantic categories of the food occurring as direct objects of the verb *eat* in imperative clauses in the self-help corpus

	Types		Tokens	
		%		%
General	4	1.5 %	19	2.3 %
Meal	5	1.8 %	65	7.9 %
Dish	6	2.2 %	6	0.7 %
Aliments	221	80.7 %	639	77.5 %
Nutrient	25	9.1 %	47	5.7 %
Undefined	13	4.7 %	48	5.8 %
Totals	**274**		**824**	

analysed realizes advice on health-related behaviour constrains the options. Eating is thus likely to be presented as nutrition based on the rational regimen of food consumption. Eating as pleasure-oriented and/or social activity, aspects that also characterise people's everyday life experience of food, cannot play a significant role. The vast majority of expressions will therefore fall into the categories of nutrients and aliments. However, if the initial hypothesis is plausible, we should expect this majority to be at least partly balanced by a certain number of expressions denoting meals or dishes.

Type and token frequencies are contained in Table 7.

As predicted, the vast majority of terms – with respect to both different expressions and token frequencies – belongs to the semantic class of aliments. Meals may be mentioned – the numbers are only relevant in the token column – but in disagreement with the hypothesis and its interpretation, not very often and if so, less as social phenomena centring on food, but rather as entries in a food schedule (as in: *Eat breakfast every day* and *Eat dinner early*), which lends further weight to the assumption about a conception of food not really related to the experience of the lifeworld. The conclusion is also supported by the quantitative irrelevance of the level of the dish. With its association with the subjective dimensions of taste and pleasure, it could have created a stronger lifeworld connection in the discourse. However, even in the rare occurrences of dishes, there is an addition that rationalizes such a scenario, as in *Eat your hamburger without the bun*. The only relativizing aspect is the low frequency of expressions denoting nutrients, which could have made the notion of food presented even more rational and health-oriented. Their rarity in this position could therefore be taken to at least not further undermine a lifeworld-related conception of eating.

Many clauses with imperative *eat* contain whole catalogues of items, usually in the form of coordinated objects, or illustrating examples are added to objects, often introduced by *such as*, *for instance*, *including* or in parenthesis or between brackets. This sometimes takes rather extensive forms, e.g. in the following sentences, each containing ten or more food items.

> Eat one cup a day of bok choy, escarole, Swiss chard, collard greens, kale, watercress, spinach, or dandelion, mustard, or beet greens.

Eat five servings a day of dark green, leafy and root vegetables such as watercress, carrots, sweet potatoes, broccoli, Brussels sprouts, spinach, green beans or peppers, raw or lightly cooked.

Such additive constructions strengthen the impression of a tight rational regimen of food created in the discourse of self-help.

As aliments represent by far the largest semantic category in the analysis of food items occurring with imperative *eat*, it is also interesting to see which types of aliments are mentioned frequently. If health promotion wants to be patient-centred and orient towards the addressee's lifeworld, we should expect a semantic distribution that corresponds, roughly speaking, to our daily menus, with a strong emphasis on meat and fish, and some typical side dishes such as potato-based products, pasta, bread, rice and certain types of vegetables and fruit. As above, we, however, have to concede that advice on food will probably not primarily focus on existing patterns, but will rather seek to replace them. It is therefore reasonable to assume that there will also be expressions in categories that do not normally feature that prominently in our prototypical ideas of food, especially categories often mentioned in connection with health, e.g. legumes (beans, peas, and lentils), grains other than rice (from quinoa to millet), and oils.

I have assigned expressions for aliments to such categories, with the following results (see Table 8).

The figures contained in Table 8 contradict the hypothesis that the composition of food items should correspond to our lifeworld experience of food. The categories encompassing aliments most prominent in Western diets – meat & fish, pastry,

Table 8 Sizes of the semantic categories of aliments occurring as direct objects of the verb *eat* in imperative clauses in the self-help corpus

	Types		Tokens	
		%		%
General	6	2.7 %	77	12.1 %
Grains	29	13.1 %	64	10.0 %
Meat & fish	18	8.1 %	57	8.9 %
Pastry, bread & pasta	16	7.2 %	24	3.8 %
Dairy & eggs	17	7.7 %	29	4.5 %
Vegetables	56	25.3 %	142	22.2 %
Fruit	25	11.3 %	112	17.5 %
Nuts & seeds	16	7.2 %	50	7.8 %
Legumes	15	6.8 %	50	7.8 %
Oil	9	4.1 %	14	2.2 %
Spices & herbs	8	3.6 %	11	1.7 %
Other	4	1.8 %	4	0.6 %
Aliment parts	2	0.9 %	5	0.8 %
Totals	**221**		**639**	

bread & pasta, or dairy & eggs – are underrepresented in both the type and token columns. Fruit and vegetables, on the other hand, are by far the most common items mentioned, claiming more than a third of all types and tokens. Compared to real life constraints, this proportion appears exaggerated as even finding the different types of fruit and vegetables mentioned and preparing dishes from them will cause difficulties. The assumption that advice on aliments abstracts from normal diets is further supported by the fact that the relatively narrow – i.e. not covering a large number of possible members – categories grains, legumes, and nuts & seeds are comparable in size to the meat & co. categories mentioned above. Now these categories are bag and parcel of any health-related discourse on food as their members are renowned for their high contents of healthy fats, proteins, and complex carbohydrates. However, apart from some items, especially rice, these categories do not figure prominently in standard Western diets. Overrepresenting them as the desired objects of eating hence puts rational, health-related aspects of eating over lifeworld aspects, further undermining any attempt to be patient-centred.

Advice on food is not only qualitative, i.e. concerning the type of foods to be consumed, but also quantitative, i.e. concerning how much of these foods the addressee should consume, how often, for how long, etc. As a final aspect, I will examine to what extent the quantification of food (including number and size) plays a role in the imperative clauses with *eat*.

Quantification is not irrelevant in our lifeworld experience of food – especially in cooking, quantities are central. But we may assume that in our understanding of eating, quantification is not the most important factor, in contrast to medicine and other expert systems, where it is a pivotal principle. According to my hypothesis, I should therefore expect quantification to play a minor role in the semantic preferences of imperative *eat*.

The following table contains all quantifying expressions related to the food items mentioned in the *eat*-clauses (Table 9).

Overall, 108 different quantifying expressions which occur 240 times have been found. This means that 45 % of the 536 clauses with imperative *eat* and almost a third of the 824 direct objects denoting food in these clauses contain a quantifying element. Quantification thus seems to figure prominently in the advice on eating. This can certainly not be interpreted as support for my hypothesis. On the contrary, it rather suggests that in giving advice on food, self-help emphasizes a rational and regulating approach, which again stresses the scientific dimension of the discourse rather than an orientation towards people's lifeworlds.

All in all, the data on the use of advice realized by imperatives analysed in this chapter mostly contradicts my hypothesis. Imperatives, which are supposed to highlight the competence gap and hierarchical distance between the writer-expert and reader-lay person are used extensively, especially in comparison to forms that might mitigate this effect. The semantic preferences of the verb *eat* used in the imperative also point to a re-interpretation of food in terms of expert systems rather than to a conception close to our everyday experience of eating.

Table 9 Quantifying expressions used for food items in imperative clauses with *eat* in the self-help corpus

more (of) (26); *plenty of* (20); *small* (16); *less (much)* (14); *lots of* (9); *little (very)* (7); *some* (7); *in moderation (only)* (5); *only* (5); *one* (4); *five servings (at least)* (3); *lightly* (3); *six servings (at least)* (3); *smaller* (3); *three or more servings* (3); *three pieces (at least)* (3); *two servings (at least)* (3); *1 cup* (2); *1–2 tablespoons* (2); *15 g* (2); *2 tablespoons* (2); *20–25 g* (2); *3 g* (2); *a* (2); *a full portion* (2); *a handful* (2); *five portions (at least)* (2); *four or more servings* (2); *one cup (at least)* (2); *one heaped tablespoon* (2); *one tablespoon* (2); *three* (2); *½ cup*; *1 clove*; *1 ounce*; *1 teaspoon*; *1-½ ounces*; *11 servings (up to)*; *170 g (6 oz) of (no more than)*; *20%*; *2–3 g*; *2–3 tablespoons*; *3–10 g*; *4 to 5 servings*; *5 portions (at least)*; *60 g (only)*; *6–18 capsules*; *8 to 10 servings (at least)*; *80%*; *a clove*; *a few*; *a few grams*; *a piece*; *a serving*; *a spoonful*; *as much as possible*; *as much as you want*; *cautiously*; *dose*; *eight portions (at least)*; *eight to nine portions*; *enough*; *excessively*; *extra slice*; *fewer*; *generous amounts*; *half a cup*; *half or a quarter of your usual*; *half-portion*; *in large quantities*; *in moderate quantities*; *in small quantities*; *less than 6 g*; *light*; *medium*; *moderate amounts*; *moderately*; *more than once*; *nine servings (up to)*; *normally*; *not too much*; *one clove*; *one half cup*; *one or two servings*; *one or two tablespoons*; *one servings*; *one small serving*; *one to three servings (at least)*; *plentifully*; *primarily*; *same amount*; *seven servings (up to)*; *six*; *six pounds*; *small amounts*; *small portions*; *small serving*; *smaller portions*; *sparingly*; *sufficient*; *three or four servings*; *three to four (only, at the most)*; *too much*; *twenty-three*; *two*; *two to three pieces*; *two to three portions*; *whole*

6 Conclusion

The main question this paper set out to answer was whether health promotion, especially in the shape of self-help sources aiming at individual lifestyle modification as the main preventative measure, is a patient-centred approach in healthcare, mitigating the competence and status gap between doctors and patients and conceiving of health problems also in terms of people's subjective experience rather than just in terms of expert systems such as the medical sciences. For this purpose, a combination of Critical Discourse Analysis and corpus linguistics was used. Two linguistic items, viz. acronyms and advice realized as imperatives, were examined in a 3.4-million-word corpus of self-help books on cardiovascular diseases – partly in comparison to a 1.4-million-word corpus of medical textbooks on the same topic.

Some of the data analysed support the assumption that self-help health promotion is indeed patient-centred in the above sense, especially the fact that acronyms are much less common in self-help books than in medical textbooks. However, most of the results – whether the overuse of technical acronyms for everyday domains, the high frequency of hierarchy-emphasizing imperatives as the preferred tool for giving advice, and the linguistic elements that co-occur with imperative *eat*, all presenting process as a rational and health-related regimen rather than as a pleasurable social activity – do not support this conclusion, suggesting that self-help is another form of healthcare based on the doctor's expert status and the abstract and technical nature of her or his expertise that does not particularly focus on people's subjectivity.

References

Baker, P. (2006). *Using corpora in discourse analysis*. London/New York: Continuum.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society, 19*(3), 273–305.

Beattie, A. (1991). Knowledge and control in health promotion. A test case for social policy and social theory. In J. Gabe, M. Calnan, & M. Bury (Eds.), *The sociology of the health service* (pp. 166–202). London: Routledge.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Booij, G. (2007). *The grammar of words. An introduction to morphology* (2nd ed.). Oxford: Oxford University Press.

Clarke, A. (2010). *The sociology of healthcare* (2nd ed.). Harlow/New York: Longman.

Fairclough, N. (1992). *Discourse and social change*. Cambridge: Polity Press.

Hardt-Mautner, G. (1995). *'Only Connect.' Critical discourse analysis and corpus linguistics. UCREL* (University Centre for Computer Corpus Research on Language) *Technical Papers* 6: Lancaster University. http://ucrel.lancs.ac.uk/papers/techpaper/vol6.pdf. Accessed 9 Oct 2014.

Journal of Pragmatics (2004), *36*(2) and *36*(9) (Special issue: Corpus Linguistics Part III).

Jucker, A. H., Schreier, D., & Hundt, M. (2009). Introduction. Corpus linguistics, pragmatics and discourse. In H. Andreas, D. S. Jucker, & M. Hundt (Eds.), *Corpora: Pragmatics and discourse* (pp. 3–9). Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29). Ascona, Switzerland, 14–18 May 2008. Amsterdam/New York: Rodopi.

Marko, G. (2008). *Penetrating language. A critical discourse analysis of pornography*. Tübingen: Narr.

Marko, G. (2009). Eat the right thing. A critical analysis of orthorexic discourses. In E. Lavric & C. Konzett (Eds.), *Food and language. Sprache und Essen* (pp. 261–274). Frankfurt am Main: Peter Lang.

Marko, G. (2012). My painful self. Health identity construction in discussion forums on headaches and migraines. *AAA – Arbeiten aus Anglistik und Amerikanistik, 37*(2), 243–270.

Mautner, G. (2012). Die kritische Masse. Korpuslinguistik und kritische Diskursanalyse. In E. Felder, M. Müller, & F. Vogel (Eds.), *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen* (pp. 83–114). Berlin/Boston: De Gruyter.

Murphy, S. L., Xu, J., & Kochanek, K. D. (2013). Deaths. Final data for 2010. *National Vital Statistics Reports, 61*(4). http://www.cdc.gov/nchs/data/nvsr/nvsr61/nvsr61_04.pdf. Accessed 1 Jan 2014.

Nettleton, S. (2006). *The sociology of health and illness* (2nd ed.). Cambridge/Malden: Polity Press.

Nettleton, S., & Bunton, R. (1995). Sociological critiques of health promotion. In R. Bunton, S. Nettleton, & R. Burrows (Eds.), *The sociology of health promotion: Critical analyses of consumption, lifestyle, and risk* (pp. 41–58). London: Routledge.

OECD. (2012). *Health at a glance: Europe 2012*. No place given: OECD Publishing. http://dx.doi.org/10.1787/9789264183896-en. Accessed 1 Apr 2014.

Partington, A., Morley, J., & Haarman, L. (Eds.). (2004). *Corpora and discourse.* Proceedings of CamConf 2002. Bern, etc.: Peter Lang.

Plag, I. (2003). *Word-formation in English*. Cambridge: Cambridge University Press.

Rayson, P. (2009). *Wmatrix: A web-based corpus processing environment*. Computing Department: Lancaster University. [Online] http://ucrel.lancs.ac.uk/wmatrix. Accessed 14 Apr 2014.

Reisigl, M., & Wodak, R. (2009). The discourse-historical approach (DHA). In R. Wodak & M. Meyer (Eds.), *Methods of critical discourse analysis* (2nd ed., pp. 87–121). Los Angeles: Sage.

Romero-Trillo, J. (2008). Introduction. Pragmatics and corpus linguistics – A mutualistic entente. In J. Romero-Trillo (Ed.), *Pragmatics and corpus linguistics* (pp. 1–10). Berlin/New York: Mouton de Gruyter.

Romero-Trillo, J. (Ed.). (2013). *Yearbook of corpus linguistics and pragmatics 2013*. Dordrecht: Springer.

Searle, J. R. (1969). *Speech acts. An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Stubbs, M. (1994). Grammar, text, and ideology: Computer-assisted methods in the linguistics of representation. *Applied Linguistics, 15*(2), 201–223.

van Dijk, T. (1998). *Ideology. A multidisciplinary approach*. Thousand Oaks/London: Sage.

Women and Men Talking About Men and Women in Greek

Georgia Fragaki and Dionysis Goutsos

Abstract This study examines the frequency and meaning distinctions of gender-related nouns for man, woman, boy and girl in Greek in approx. 600,000 words of spontaneous interaction between, mainly, young women and men, drawn from the *Corpus of Greek Texts* (CGT). Data has been annotated for speaker gender and age so that speaker preferences can be closely studied. The comparison of our findings from conversational data with our previous research in newspapers and magazines suggests that the former are generally much less stereotypical in the construction of gender identity and much less biased against women. A statistical analysis indicates that groups of speakers distinguished in terms of age and gender tend to talk more about their own members. Furthermore, Greek women in our data, as opposed to men, tend to talk about both men and women in terms of specific persons rather than as examples of their gender, in agreement with what has been found in other languages. On the basis of our discussion it is argued that corpus linguistics, despite its shortcomings, can be fruitfully applied to the study of gender in conversation, preferably in conjunction with micro-analytic approaches.

Keywords Collocates • Conversation • Conversation analysis • Corpus linguistics • Gender-related nouns • Greek • Speaker preferences

1 Introduction: Corpus Approaches to Language and Gender

Corpus linguistic methods have increasingly been applied to the study of language and gender, as can be seen e.g. in the recent virtual special issue of *Gender and Language* on corpus approaches (Baker 2013). Corpus linguistics allows us to

G. Fragaki (✉) • D. Goutsos
Department of Linguistics, University of Athens, Zographou, 15784 Athens, Greece
e-mail: efraga@phil.uoa.gr; dgoutsos@phil.uoa.gr

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_5

answer questions about the multiple relations of language and gender by looking at large quantities of authentically occurring data in systematic ways that reveal typical patterns of use. Since a basic method of analyzing corpus data involves the search for specific forms in their immediate co-text through concordancing, gender-related lexical items and phrases can be thoroughly studied in large amounts of data. At the same time, the systematic combination of data with metadata concerning e.g. the gender of text producers allows researchers to investigate gender-related language usage.

Thus, corpora can be used to resolve the tension between earlier formal studies of gender in language and the more recent emphasis on gender as socially constructed in interaction. Both *representation-based* and *usage* corpus approaches, as Baker (2013) calls these two lines of research respectively, have already been followed in a considerable number of studies in the case of English. As regards the latter, Schmid (2003), for instance, has studied the markers and semantic fields preferred by men and women in specific corpora, while Charteris-Black and Seale (2009) and King (2011) have investigated the language usage of particular gender-related groups in large collections of spoken interview and online data. In parallel, there have been several papers on how males and females are represented in corpora through the use of specific lexicogrammatical choices (Leech and Fallon 2004 [1992]; Holmes 2000, 2001; Romaine 2001; Stubbs 2001: 161–164; Holmes and Sigley 2002; Sigley and Holmes 2002; Gesuato 2003; del-Teso-Craviotto 2006; Johnson and Ensslin 2007; Baker 2008, 2010; Pearce 2008; Caldas-Coulthard and Moon 2010; Macalister 2011), although Baker has recently remarked that corpus approaches to gender and language "seem to be in the minority" (2014: 6).

Representation-based corpus approaches to gender seem to share Holmes' assumption (2000: 141) that social identity is constructed through semantic distinctions codified in the vocabulary and the grammar of a language. Its emphasis has been on studying contrasting male vs. female vocabulary pairs such as *man-woman, boy-girl, his-her, Mr-Miss/Mrs/Ms* etc., by examining their frequency and use, mainly in newspapers and magazines or general reference corpora of English, and to a much less extent spoken data. Most studies converge on the finding that there are more occurrences of words for men than women, and usually more mentions of boys than girls, in English, something which has been taken to reflect the asymmetrical influence of the two genders in English-speaking societies (e.g. Leech and Fallon 2004 [1992]; Sigley and Holmes 2002). Furthermore, an asymmetry has been observed in the collocates for male and female terms, revealing stereotypical roles and associations for both males and females. For example, an asymmetry has been found in the age range of boys and girls, with the term *girl* being used to refer to a larger age span that reaches well into middle age (e.g. Bolinger 1980: 100; Sigley and Holmes 2002; Holmgreen 2009: 12; Taylor 2013).

Comparable research is not easy to find for other languages, so as to test the cultural-specificity of findings. For Greek, in particular, there have only been a few corpus-based gender studies. Gender-related words have been investigated by Goutsos and Fragaki (2009), focusing on nouns, and Fragaki and Goutsos (2005), focusing on adjectives in data from newspapers and magazines. Hatzidaki (2011) has also studied lexical clusters in a corpus of Greek men and women magazines from a critical discourse analysis perspective. This scarcity of corpus studies goes hand in hand in the case of Greek with an earlier emphasis of gender studies on language descriptions of the grammatical system and the vocabulary, using data mainly based on intuition or anecdotal evidence. More recent studies that draw on empirical evidence or spoken corpora (e.g. Pavlidou 2003a, 2006; Archakis and Papazachariou 2008) do not follow a corpus linguistic approach in their treatment of gender and language.[1]

The present study aims at looking into authentic conversational data by following a corpus linguistic approach. We investigate the frequency and meaning distinctions of several gender-related nouns in Greek, namely the basic pairs *άνδρας/άντρας*[2] /'anðras-'andras/(‘man’) vs. *γυναίκα* /ji'neka/(‘woman’) and *αγόρι* /a'ɣori/ (‘boy’) vs. *κορίτσι* /ko'ritsi/, along with the two frequent words *κοπέλα* /ko'pela/and *κοπελιά* /kope'ʎa/(‘young girl’),[3] in approximately 600,000 words from everyday, spontaneous discourse. Our corpus consists of conversational interactions between, mainly, young female and male university students, talking to each other or interacting with older people. Our study seeks to broaden the scope of representation-based corpus linguistic approaches to gender by looking at spoken, non-English data. At the same time, we attempt to investigate speaker preferences by identifying the users of these nouns in terms of gender (male–female) and age (younger-older). In Baker’s (2013) terms mentioned above, representation findings would thus be looked at from a usage-based perspective.

A further purpose of our study is to compare gender representation in conversation with our previous research on written genres (newspapers and magazines). Gesuato (2003) has already pointed out with regard to English the need for gender

[1] Pavlidou (2006) offers a comprehensive view of previous language and gender studies in Greek and brings together approaches based on empirical data. The issues concerning gender in Greek are concisely discussed in English in Pavlidou (2003b).

[2] The lemma for ‘man’ has two phonologically and orthographically distinct forms, the more formal /'anðras/ *άνδρας* and the informal /'andras/ *άντρας*. In all nouns, as is standard practice, we use the lemma form to subsume all different morphological variants for case and number, unless we explicitly comment on a particular form.

[3] Since there cannot be a one-to-one correspondence between Greek and English, we have opted to translate these two nouns as ‘young girl’, in order to indicate their potential overlap with *ko'ritsi*. Other possible translations include ‘young woman’, ‘young lady’ and ‘woman’, depending on the emphasis given in each context.

terms to be examined through corpora in a wide range of genres. Although there are evident shortcomings of corpus linguistics when dealing with spoken data that have already been pointed out in the literature (e.g. Adolphs 2008: 5ff.), here we aim at exploring the potential of corpus approaches in the study of interactional material. In addition, our comparison with written data is expected to reveal whether patterns of gender representation can be extrapolated to a language as a whole, as opposed to specific sub-sets of it. Finally, this comparison will allow us to identify genre particularities with regard to how gender-related nouns are used that would otherwise remain unearthed and thus to suggest differences in interaction through speech and writing.

The following section concisely presents our previous research on Greek newspaper and magazine corpora in order to facilitate comparison with spoken interactions. We then proceed to a discussion of our data and methodology, before presenting our findings from conversation. Finally, before concluding, we discuss the implications of our research for the contribution of corpus linguistics to the analysis of spoken data, pointing out the complementariness of quantitative and qualitative analysis.

2 Gendered Terms in Written Corpora of Greek

Goutsos and Fragaki (2009) and Fragaki and Goutsos (2005) have studied the use of gender-related nouns and adjectives, respectively, in 2.5 million words from newspapers and magazines, drawn from the Corpus of Greek Texts (CGT), a general reference corpus of Greek. CGT has collected 30 million words from a variety of spoken and written text types, ranging from academic texts, literature and non-fiction to e-mails, private letters, parliament talk etc., collected from two decades of Modern Greek, 1990–2010. (For details, see Goutsos 2010).[4] In these previous studies of gender in Greek we have focused on news and opinion articles from daily and Sunday newspapers published in Greece, on the one hand, and Greek magazines, on the other hand, including (a) general interest magazines, mainly with a social and political focus, (b) men's magazines and (c) women's magazines.[5] The number of words studied from each genre is shown in the following Table 1.

As regards the frequency of gender-related lexis, Greek written data seems to draw a different picture to that of English, since, as can be seen in Table 2, words for women and girls are much more frequent than male nouns.[6]

[4] The corpus is freely available at: www.sek.edu.gr.

[5] These categories refer to the intended audience of magazines, rather than their actual readership, as follows from their analysis.

[6] A similar difference was found for gender-related adjectives, with *ανδρικός*/*αντρικός* /anðri'kos-andri'kos/ 'male' occurring 54 times vs. *γυναικείος* /jine'cios/ 'female' 165, *αρσενικός* /arseni'kos/ 'masculine' 47 vs. *θηλυκός* /θili'kos/ 'feminine' 59 etc.

Table 1 Size of data (in tokens) investigated in studies of written Greek

Newspapers		Magazines		
News articles	Opinion articles	General	Men's	Women's
600,000	600,000	600,000	300,000	300,000

Table 2 Frequency of gender-related nouns in written genres of Greek

Lemma	Frequency	Lemma	Frequency
'andras ('man')	716	*ji'neka* ('woman')	1,468
a'γori ('boy')	141	*ko'ritsi* ('girl')	203
Total	857		1,671

It is worth noting here that this preponderance of female nouns also holds for each individual sub-corpus studied, and particularly so for women's magazines in which occurrences of *ji'neka* are almost three times as many as those of *'andras*.

Our account for this discrepancy with findings in other languages has been that writers may stress the gender of a referent in Greek data, mainly when it is female, e.g. in professional or, more generally, occupational names, in which the female member of the pair is associated with a marked, non-typical use (e.g. γυναίκα αστυνομικός 'woman policeperson'). In addition, gender terms seem to be particularly frequent in women's magazines, in which, as mentioned above, the difference between male and female-related nouns is especially pronounced, adding heavily to the overall discrepancy. A comparison with other genres would be useful in explaining this difference of Greek data, as women do not seem to be more visible or influential in Greek society than in others.

Furthermore, qualitative data analysis has pointed out a similar asymmetry as that found in research on other languages with regard to the representation of men and women. For instance, the meaning distinctions found for gendered terms by and large depict men in a positive and women in a negative light. Table 3 presents the basic meaning distinctions identified for 'man' and 'woman' with illustrative examples for each gender (cf. Goutsos and Fragaki 2009).

Meaning distinctions were identified on the basis of formal and semantic criteria in a data-driven fashion, as exemplified by Kilgariff (1997: 16, cf. the distinction between corpus-driven and corpus-based in Tognini-Bonelli 2001). In particular, concordance lines were thoroughly analyzed and meanings were clustered according to criteria arising from the corpus itself.[7] Thus, the label GENDER was assigned to references to males in juxtaposition to females and the opposite, found in patterns of singular or plural nouns with definite article in Greek ('the man' i.e. all men, 'the women' i.e. all women). The PERSON meaning refers to specific male or female

[7] However, it is interesting that most, though not all, of these meaning distinctions can also be found in the two major reference dictionaries of Modern Greek in the respective lemmas (Triandaphyllidis 1998; Babiniotis 2002).

Table 3 Meaning distinctions in written genres of Greek

Gender	(1) στον άντρα το κόμπλεξ μου τη σπάει 'a man with a complex gets on my nerves'	(2) το Κοράνι σέβεται τη γυναίκα 'the Koran respects women'
Person	(3) ένας ρακένδυτος άνδρας μαζεύει τα απομεινάρια 'a man in rags picks up leftovers'	(4) συνάντησα μια γυναίκα σε αξιοθρήνητη κατάσταση 'I met a woman in a lamentable state'
Spouse	(5) ο άνδρας της αγόρασε τη δεκατριάχρονη 'her husband [lit. man] bought the thirteen year-old'	(6) μπήκαμε σε μεγάλο δίλημμα με τη γυναίκα μου 'me and my wife [lit. woman] were in a big dilemma'
Occupational	(7) να δουλέψεις με πέντε άντρες ηλεκτρονικούς 'to work with five male [lit. men] electrical engineers'	(8) το Λουξεμβούργο δεν έχει καμία γυναίκα ευρωβουλευτή 'Luxemburg has no woman Euro-MP'
Adult	(9) τα «παιδιά» σας γίνανε άνδρες 'your "children" have become men'	(10) είναι γυναίκα, δεν είναι κοριτσάκι πλέον η Δήμητρα 'Dimitra is a woman, no longer a little girl'
Personnel	(11) οι άνδρες της προσωπικής του φρουράς 'the men of his guard'	
High status person	(12) οι δύο άνδρες συζήτησαν θέματα Μεσογειακής συνεργασίας 'the two men discussed issues of Mediterranean co-operation'	
Stereotypical role	(13) έχει να κάνει με άντρα και όχι με… ρόμπα 'he had to deal with a real man [lit. man] and not with a puppet'	
Family member	(14) οι άντρες της οικογένειας το έχτισαν με τα χέρια τους 'the men in the family built it with their own hands'	

(continued)

Table 3 (continued)

Illegal person		(15) η παράνομη διακίνηση γυναικών 'the illegal trafficking of women'
Housemaid		(16) …τις δουλειές του σπιτιού σε κάποιαν άλλη γυναίκα 'the house chores to some other woman'
Special constructions		(17) για να έχει την ησυχία της η γυναίκα 'so that she, the woman, be left in peace'

individuals, in patterns of the type: indefinite article/determiner + singular noun ('a man', 'this woman'). The words used for man and woman in Greek are also employed for the meaning of SPOUSE in constructions with possessive pronouns or genitive noun phrases ('my woman' meaning 'my wife', 'the man of X' meaning 'the husband of X'). Male or female nouns can also accompany OCCUPATIONAL nouns, as in (7) and (8) above.

The other meaning distinctions are mainly distinguished by semantic criteria, including collocations, although there usually is some correspondence with formal patterns; for instance, the meaning of STEREOTYPICAL ROLE ('a real/authentic man') is usually found with nouns in the predicative position, whereas phrases with nouns in the HIGH STATUS PERSON meaning usually refer backward to proper names of MPs, ministers etc. Finally, particular meanings are assigned to SPECIAL CONSTRUCTIONS, i.e. patterns that deviate from canonical word order in Greek. For example, (17) above shows a pattern of right-dislocation, that has been characterized as *tail* in Greek (Valiouli 1990) and involves an evaluative or epithet noun phrase at sentence final position. As Valiouli (1990: 173) notes, tail phrases in Greek function as "an evaluative comment on the preceding part of the sentence" and express the speaker's subjective evaluation. In (17) the noun phrase *η γυναίκα* /i ji'neka/('the woman') is not the subject (or a reduplication of it) but a "redundant" expression, added in order to express the writer's sympathy with the person referred to, similarly to appositions of the type "the poor thing" in English.[8]

As can be seen in Table 3, apart from five meaning distinctions which are common for the two members of the pair 'man' vs. 'woman', the distinctions that are exclusive to women are not prestigious, including ILLEGAL PERSON or HOUSEMAID,

[8] For ambiguous cases in the assignment into a meaning distinction the larger context beyond the concordance line was consulted. Cross-checking with several other native speakers of Greek was also employed in order to ensure that the categorisation was reliable and consistently done.

whereas those found with men are (e.g. HIGH STATUS PERSON, FAMILY MEMBER). In addition, different frequencies found for these meaning categories suggest an emphasis on stereotypical views of professions for men and women, e.g. by linking men with the armed forces (see PERSONNEL) or by drawing attention to the exceptional character of some occupations for women (e.g. *γυναίκα ευρωβουλευτής* 'woman Euro-MP').[9]

The same asymmetry was observed in the studies of Greek written genres for the collocates of male- and female-related nouns and adjectives: thus, to put it concisely, collocates that are common to both men and women refer to their appearance, sexuality, equality, relationships and power; collocates of men refer to public life and conflict, while collocates of women are associated with the obstacles they face, with abuse, prostitution, work/business, participation in society or are linked with other less privileged groups of people such as children, the elderly, black people etc.

Finally, as regards the age range for boys and girls, an asymmetry like that found for English data (see above) was also observed in the Greek written sub-corpora, in which mature women are referred to as 'girls' in contrast to men, who are much less often referred to as boys after their teens. The noun *ko'ritsi* 'girl' was thus found to function more as a characterization or label, rather than or in addition to being a marker of age.

3 Greek Conversational Data

The same methodology as that used in the studies on written data is employed in this paper for the analysis of gender-related nouns in spoken data from conversation interactions. The data is drawn from the spoken sub-corpus of CGT (see above), which includes face-to-face, spontaneous conversations between friends in informal settings. In particular, 185 texts (573,904 tokens) are used in this study.[10] Most speakers are university students talking to their family or friends; because of the preponderance of female students in the Department of Linguistics (National and Kapodistrian University of Athens), who participated in the conversations, most speakers are young women.

[9] Obviously, the absence of meanings like ILLEGAL PERSON for men or HIGH STATUS PERSON for women does not mean that there are no references to illegal men or prestigious women, but implies that the words *a'nδras* and *jine'ka* are not used for them respectively. In these cases more specific words, indicating the gender of the person involved, like *ο μετανάστης* (the-MASC immigrant) or *η υπουργός* (the-FEM minister) may be found instead.

[10] All data has been audio recorded by one of the participants in the conversation, who also provided comments on gestures, facial expressions etc., especially when relevant. There were no restrictions as to the time and place of the interactions, which were recorded throughout the year. Some of the transcription conventions used in the sub-corpus are given in the Appendix. The amount of data in the spoken corpus is roughly the same as in the written sub-corpora, something which allows for the comparison of raw data.

Table 4 Number of speakers in the conversational data of Greek studied

	Men	Women	Total
Young	112 (MY)	367 (FY)	479
Old	40 (MO)	79 (FO)	119
Total	152	446	598

Since detailed metadata are kept in CGT about speaker gender, age, profession, dialect used etc., it is possible to compare between gender and age groups. Texts were thus annotated for the type of gender interaction between participants (F: female only, M: male only, X: mixed)[11] and age of participants in interaction (Y: young only, O: old only, D: mixed). These labels can also be applied to speakers in four combinations, namely male young (MY), female young (FY), male old (MO) and female old (FO), allowing us to investigate speaker preferences. Since conversations come from young speakers of a university age, we have decided to label all speakers over 30 as 'old', so as to give emphasis on a fairly homogeneous group of speakers. In practice, then, our distinction corresponds to below and above 30 year old speakers of Greek.

Table 4 presents the number of speakers for each category in the 185 conversations studied.

As can be seen, our data is biased towards younger speakers (80 % of the whole), as noted above, and women speakers (74 % of the whole) in all types of interaction. The cross-category of this, i.e. young women speakers takes up 61 % of the total number of speakers. All our findings should be considered with this proviso in mind.

4 Findings

This section presents the findings of our study with regard to the frequency, meaning distinctions and collocates of gender-related nouns, as was done with written data. Apart from the two basic pairs, as mentioned above, we also investigate the words *κοπέλα* /ko'pela/and *κοπελιά* /kope'ʎa/('young girl'), which seem particularly prominent in conversation. In addition, gender nouns are related to speaker age and gender (who says what). The software programmes *Wordsmith Tools* and *Antconc* were used for the analysis of the corpus. Tokens were searched with the help of wildcards, while lemmas were manually identified.[12]

[11] The dimension of gender-only vs. mixed conversation is not discussed in this paper, although the corpus offers such a possibility, since 90 conversations were female-only, 5 male-only and 90 were mixed-gender. We thank an anonymous reviewer for pointing out this possibility, which, however, cannot be further explored here.

[12] Diminutives like αντράκι /a'ndraci/, γυναικούλα /jine'kula/ etc. were excluded.

4.1 Frequency of Gender-Related Nouns

Table 5 and Fig. 1 present the frequency of each gender-related noun in our conversation data.

Of the 660 gender-related nouns in the spoken data corpus, female-related nouns are much more frequent (477; 72 %) than male-related nouns (183; 28 %). This is true even if we compare only the four main items that were studied in written genres (i.e. *'andras-ji'neka*, *a'γori-ko'ritsi*), for which there are 183 (44 %) male-related as opposed to 231 (56 %) female-related nouns.

This preference for female-related nouns concurs with what has been observed in written genres of Greek (see Table 2), suggesting thus a tendency of the language as a whole in contrast to English. At the same time, spoken data seem to be much less biased in the representation of the basic noun pairs, approaching the numbers found in general magazines (cf. Goutsos and Fragaki 2009: 321), the most balanced sub-corpus in the written data studied. Figure 2 presents the relative contribution of each gender-related noun in the spoken data and the written genres studied before.

Apart from the almost balanced treatment of the two genders in our conversational data noted above, Fig. 2 shows that the largest contributions of *a'γori* ('boy') and *ko'ritsi* ('girl') in the total number of occurrences is also observed in conversation. Although this may be expected in a corpus in which young speakers, closer to

Table 5 Frequency of gender-related nouns in the spoken corpus

Lemma	Frequency	Lemma	Frequency
'andras ('man')	119	*ji'neka* ('woman')	153
a'γori ('boy')	64	*ko'ritsi* ('girl')	78
		ko'pela ('young girl' 1)	222
		kope'ʎa ('young girl' 2)	24
Total	183	Total	477

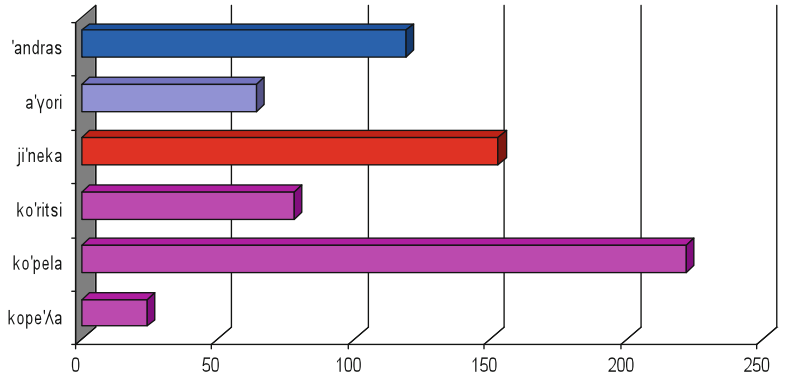


Fig. 1 Relative frequency of gender-related nouns in the spoken corpus

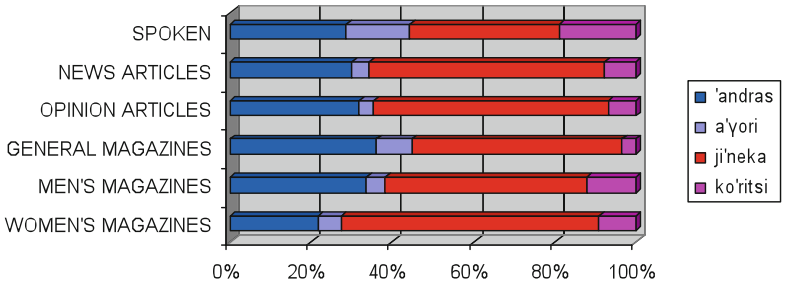


Fig. 2 Relative frequency of gender-related nouns in spoken and written genres

the age of 'boys' and 'girls', predominate, it points to a general lack of visibility of this age group in written data of the type studied before, that is newspaper and magazine articles. Conversely, it can be argued that the more topical in a genre gender issues appear to be, the more occurrences of female-related nouns are to be found.

Finally, it is notable that the most frequent gender-related noun in conversational data is *ko'pela* ('young girl'1) (see Fig. 1). Along with *kope'ʎa* ('young girl'2), they almost take up one in every five occurrences of gender-related nouns (37 %), while, along with *ko'ritsi* ('girl'), they almost take up half of all these occurrences (49 %). The predominance of words for 'girl' and 'young girl' (324 in total) certainly reflects the larger number of young women speakers in our data (see Table 4),[13] but can only be properly evaluated with reference to speaker preferences, i.e. with whether it is men or women, younger or older, who use these nouns (see Sect. 4.4).

In all, the frequencies of gender-related nouns in conversational data have revealed that references to the two genders occur almost equally, while the tendency of female-related nouns to exceed male-related nouns seems to remain the same across spoken and written genres in Greek. The composition of our spoken corpus can also account for an emphasis on nouns for younger men and women that was not found in other genres.

4.2 Meaning Distinctions

Based on the same criteria followed in the case of written data, fewer meaning distinctions were found in conversational data than in newspapers and magazines, although a couple of new meanings that are not found in the latter occur in the former. In particular, for the pair 'man' vs. 'woman' the following meanings were mainly identified:

[13] At the same time it is surprising that young women are referred to as 'girls', something which can be interpreted with reference to the larger age span for girls observed in the literature.

(a) GENDER: this is the same with the category found in written data (see Table 3), in which the noun refers to a male or female person in general with emphasis on their gender, as in:

(18) εγώ δε δικιολογώ τον άνδρα αλλά απλώς τώρα η διαφορά είναι ότι κάνουν αντίστοιχα και οι γυναίκες τα ίδια (MY)[14]

I'm not trying to excuse men [lit. the man] but it's just that now what's different is that women also do the same things

(19) Λ: πάντως (.) έχω παρατηρήσει Βάσω ότι εμείς οι γυναίκες είμαστε ΧΕΙΡΟΤΕΡΕΣ απ' τους άνδρες

A: NAI NAI NAI NAI
Λ: NAI NAI εγώ έτσι που βγαίνω στις καφετέριες (.) πιο πολλές ΓΥΝΑΙΚΕΣ καπνίζουν παρά ΑΝΤΡΕΣ (FY)

L: anyway I've noticed, Vasso, that we, women, are worse than men
A: YES YES YES YES
L: YES YES when I go out to cafés there's more WOMEN smoking than MEN

As can be seen in examples (18) and (19), the reference to one of the genders tends to trigger an upcoming reference to its opposite, so that nouns for both genders co-occur in the same or adjacent utterances.

(b) PERSON: this is also a category found in written data; in conversational data it is mainly found with 'woman', as in (20):

(20) είχα διαβάσει ένα βιβλίο κι ήτανε μια γυναίκα (.) η οποία το έκανε με τη θέλησή της (FY)

I had read a book and there was a woman who did it on her own will

(c) SPOUSE: this category, referring to a husband or wife, which was also found in written data, is quite common in conversation, as well:

(21) ανεβήκανε οι θείοι μου από κάτω η ξαδέρφη μου με τον άντρα της (FY)
my uncles came up from downstairs, my cousin with her husband [lit. the man hers]

(22) α είχε έρθει και η θεία σου η:: γυναίκα του Στέλιου; (MY)
oh, so your aunt also came, Stelios's wife? [lit. the woman of Stelios]

(d) RELATION: this meaning distinction is not found in written data and refers to lovers or partners, persons with whom one has an affair:

(23) E: και με ποιον άντρα θες να 'σαι αύριο ξέρω 'γώ δηλαδή (FY)

E: and with what man you'd like to be tomorrow with like

[14] The gender and age category of the main speaker is indicated in a parenthesis at the end of each example in an abbreviated form.

(e) SPECIAL CONSTRUCTION: uses of gender-related nouns in special constructions were only marginally found in written data, whereas they seem to be quite frequent in conversational data, mainly with women. Two patterns are mostly found:

(i) property (age or height) NP + gender noun in predicative uses:

(24) είχαμε φιλόλογο εξήντα χρονών άντρα έκανε μάθημα δεν βαριόσουνα (MY)
we had a literature teacher (who was a) 60 year old man, he used to teach well, he was not boring

(25) Β: κοίτα ντάξει είναι και δυο μέτρα άντρας δεν είπες;
Γ: πού να το χορτάσεις; (FO)
V: look, OK, he's two metres high [lit. (he) is two metres man] too, didn't you say?
G: you can't feed it [meaning 'him']

(ii) tail phrases, of the type found in (17) above (see Sect. 2):

(26) Μ: μετά στεκότανε πάνω απ' το κεφάλι μας σαν το Χάρο για να σηκωθούμε να φύγουμε με την καλή έννοια αλλά τέ[λος πάντων
Κ: [ε να μαζέψει η γυναίκα να πάει σπίτι της (FY)
M: and then she would stand over our shoulder like Hades so that we would get up and go, so to say, but anyway
K: so that she'd pick things up, the woman, and go home

(27) έμεινε ξαφνικά μόνη ε πώς να μη της κακοφανεί τώρα της γυναίκας; (FO)
she was suddenly left alone, well, how could she not feel bad, the woman?

It should be noted that tail phrases in examples like (26) and (27) are overwhelmingly found with women and, in contrast to written data, show a sympathetic or affiliative speaker stance rather than an ironical treatment, as in the case of newspapers (see Goutsos and Fragaki 2009: 323).

Finally, the meaning categories ADULT and STEREOTYPICAL ROLE are also found in conversations. An interesting example of the latter is found in (28):

(28) Μ: και το φοράει άντρας; και είναι άντρας; (FY)
M: and is this worn by a man? and is he (a) man?

Here, the first mention of *'andras* ('man') refers to the gender of the person referred to, whereas the second mention questions the gender-related behaviour of this person, according to its stereotypical role, since, according to the speaker, fur-coated boots cannot be worn by a man.

In general, it appears that many of the meaning distinctions found in written data do not occur in conversational data, including the meanings PERSONNEL, HIGH STATUS PERSON, HOUSEMAID, ILLEGAL PERSON etc., whereas the meaning RELATION is only found

Table 6 Frequency of meaning distinctions for the pair 'man' vs. 'woman' in the spoken corpus

	'anδras ('man')	*ji'neka* ('woman')
Gender	71 (60 %)	74 (48 %)
Person	6 (5 %)	26 (17 %)
Spouse	23 (19 %)	27 (18 %)
Special construction	3 (3 %)	21 (14 %)
Relation	9 (8 %)	2 (1 %)
Stereotypical role	5 (4 %)	1 (1 %)
Adult	2 (1 %)	2 (1 %)

in conversational data. Although the composition of the corpora may account for these differences, it still is important that particularly ingratiating meanings for men or demeaning meanings for women are avoided in conversational interaction.

Furthermore, a quite interesting picture arises when preferences for meaning distinctions are examined in conversational data, as can be seen in the following Table 6.

Although for both male and female terms most occurrences refer to GENDER, suggesting that the question of gender relations is foregrounded in conversation, there is a clear asymmetry with regard to the use of 'man' and 'woman' for reference to specific persons. Thus, the noun for 'woman' is used much more in order to refer to a specific individual than the noun for 'man'.

The other major difference between *'anδras* ('man') and *ji'neka* ('woman') concerns their use in special construction patterns. The three occurrences of 'man' in a special construction concern the pattern: property (age or height) NP + 'man' (see examples 24 and 25 above), used to express the speaker's admiration for a man. Special constructions are much more frequent with women and mainly concern tail phrases, as in examples 26 and 27 above, which express the speaker's affiliation with a woman.

These preferences for meaning distinctions may relate to the fact that both the PERSON and the SPECIAL CONSTRUCTION uses are taken over in the case of 'man' by the pseudo-generic *άνθρωπος*/*'anθropos* 'human being', as has already been suggested in the literature (Makri-Tsilipakou 1989). Further research on this lexical item is needed in order to confirm this.

With regard to the pair 'boy' vs. 'girl', including the items for 'young girl', several meaning distinctions are found in conversational data:

(a) GENDER: it is interesting that this meaning distinction also applies to this noun pair, as in the following example:

(29) και πόσο μάλλον για μας που είμαστε και κοπέλες και::
κάποια στιγμή θα 'ρθει και η εγκυμοσύνη:: (FY)
and especially for us who are girls ('young girl'1) too and at
some point pregnancy will come

(b) PERSON: this is a common meaning for both 'boy' and 'girl', as can be seen in (30) and (31) respectively:

(30) α καλά έχουν πάρει από το μπαμπά μου @@@ σ' αυτό το σπίτι η συζήτηση μονίμως για τη μπάλα και τα τρία τα αγόρια κολλημένα με την μπάλα (FY)
oh well, they've taken after my dad, in this house talk is always about football, all three boys are crazy with football

(31) τα ξανάφτιαξε πάλι με την κοπελιά τον έβαλε η κοπελιά και πήγε και της αγόρασε καναπέ (FO)
he made up again with this girl ('young girl'2), this girl ('young girl'2) made him go and buy her a sofa

(c) RELATION: this meaning corresponds to the sense of 'boyfriend' or 'girlfriend':

(32) X: δεν είχε αρραβωνιαστεί αυτός που λέω εγώ
A: μήπως είχε κοπέλα;
X: τσ ((κίνηση άρνησης)) (MY)
X: the one I'm talking about was not engaged
A: did he have a girl ('young girl'1)?
X: ((makes a sound for 'no'; shakes head))

(d) ADDRESS: this is a meaning distinction that has not been found in the written data, while it occurs in conversation in greetings, summons etc. with all nouns meaning 'girl' or 'boy':

(33) στην υγειά σου κορίτσι μου πολύχρονη (FO)
to your health, my girl ('girl'), many happy returns

(34) Π: εγώ έχω ψιλά περίμενε
ΠΑ: κοπελιά να σε πληρώσουμε; (MY)
P: I've got change, hang on
PA: girl ('young girl'2), can we pay?

(35) ναι κοπελιά άσε μας τώρα με τις ηλικίες (.) έχεις τρελαθεί κι εσύ (FY)
OK, girl ('young girl'2), let us be with age, you've gone crazy

(36) αγόρι μου εδώ πέρα έχουν βρει λύσεις ενεργειακές εδώ και δεκαετίες να μην πω εκατονταετίες (MY)
my boy, they've already found energy solutions decades ago, not to say centuries

(e) SPECIAL CONSTRUCTION: this involves tail phrases, also found with both 'boy' and 'girl':

(37) η Μαρία η κοπέλα του είχε γίνει:: εντάξει καθαρίστρια η κοπέλα είχε γίνει είχε χάσει φίλους είχε χάσει τα πάντα (FY)
Maria, his girlfriend, had become, well, a cleaner the girl ('young girl'1) had become, she had lost friends, she had lost everything

Table 7 Frequency of meaning distinctions for the pair 'boy' vs. 'girl' in the spoken corpus

	ko'ritsi ('girl')	*ko'pela* ('young girl 1')	*kope'ʎa* ('young girl 2')	Total 'girl'	*a'γori* ('boy')
Gender	13	31	0	44 (14 %)	17 (27 %)
Person	37	148	8	193 (60 %)	13 (20 %)
Relation	3	22	3	28 (8 %)	17 (27 %)
Address	21	9	13	43 (13 %)	15 (23 %)
Special construction	4	12	0	16 (5 %)	2 (3 %)

Table 7 sums up the frequencies for each meaning distinction of 'girl' and 'boy', indicating their relative importance for each noun.

As can be seen in Table 7, nouns for 'girl' are mainly used for referring to specific persons and, secondarily, for referring to the female gender or as a means of address, whereas meaning distinctions for 'boy' are more equally distributed. In particular, *ko'pela* ('young girl 1') is a basic noun, used to refer to a specific person in 46 % of all occurrences of 'girl' and 'young girl' (148 out of 324), whereas more than half of the occurrences of *kope'ʎa* ('young girl 2') are used as a means of address. This finding would suggest that the two lexical items for 'young girl' are relatively specialized in their function.

In comparison to the respective findings in written data for the 'boy' vs. 'girl' pair (cf. Goutsos and Fragaki 2009), the meaning CHILD is not found in conversational data, while the meanings GENDER and ADDRESS are only found in conversation. Moreover, quite a few occurrences of SPECIAL CONSTRUCTION have been found for both boys and girls. As found for 'woman' above, these uses also express the speakers' affiliation with the person referred to, rather than an ironical distancing as is the case with newspaper uses of SPECIAL CONSTRUCTION, exclusively with 'girl' (cf. Goutsos and Fragaki 2009: 328).

In sum, the analysis of the meanings found with gender-related nouns in conversational data, as well as their frequencies, suggests that conversation gives prominence to interactional and evaluative meanings that accord with its special functions. By contrast, written genres are more preoccupied with stereotypical characterizations of the two genders, for which a wider range of meanings is employed. Furthermore, while the discussion of gender is prominent in both spoken and written data, conversation also seems to give emphasis on the discussion of individual persons, and especially women and young girls.

4.3 Collocates

Only some preliminary indications can be given about the collocates of gender-related nouns in conversational data, since not many significant and meaningful collocates appear more than once with any gendered noun, in contrast to the written

Table 8 Collocates found with gendered nouns in the spoken corpus

'anδras (‘man’)	*ji'neka* (‘woman’)
γυναίκες ‘women’, κορίτσια ‘girls’, δουλειά ‘job’, δουλεύουνε ‘work’, τελευταίος ‘last’, δεν είναι ωραίος ‘not handsome’, αλκοολικοί ‘alcoholics’	άσχημη ‘ugly’, τρελή ‘crazy’, θεά ‘goddess’, καλοκάγαθη ‘naïve’, γλωσσού ‘chatterbox’, μορφωμένη ‘educated’, ώριμη ‘mature’, ανασφάλεια ‘insecurity’, μεγαλύτερη ‘older’, ηλικιωμένη ‘aged’
a'γori (‘boy’)	*ko'ritsi* (‘girl’)
συμπαθέστατο ‘very nice’, κολλητός ‘mate’	καλό ‘nice’, ωραίο ‘pretty’
	ko'pela (‘young girl 1’)
	καλή ‘nice’, σωστή ‘proper’, ωραία ‘pretty’, φλογερή ‘hot’, νέες ‘young’, εξωθούνται ‘are promoted’, χαζή ‘stupid’, σχέση ‘relation’, παντρεμένη ‘married’, μικρή ‘young’
	kope'ʎa (‘young girl 2’)
	φίλες ‘women friends’

data, where specific preferences were found.[15] Table 8 presents a selection of the collocates occurring with gendered nouns in our data.

Although there cannot be but tentative conclusions here, we can see that collocates of male nouns concern the other sex, work etc., whereas collocates of female nouns mainly have to do with age, relations, behaviour, appearance etc. Stereotypical characterizations of women (e.g. ‘chatterbox’, ‘proper’) are not avoided in conversational data.

At the same time and in contrast to data from newspapers and magazines, mainly positive collocates occur for women, while a few negative collocates are found for men. This may suggest a male-dominated view of written media, as opposed to casual conversation, in which women may directly influence the way gender is viewed. However, it is obvious that larger corpora of spoken data are necessary before we are able to draw any solid conclusions.

4.4 Speaker Preferences

In written data of the type studied (i.e. newspapers and magazines) the writer is quite difficult to identify; for this reason audience design was deemed to be more significant, referring to male- or female-oriented publications. By contrast, our

[15] Both t-score and MI were used to find the statistical significance of collocates in a span of 4 items left and 4 right. By “meaningful” we basically mean non-grammatical items that are closely related to the meaning of the noun in question.

Table 9 Speaker preferences for gender-related nouns in the spoken corpus

	Women	WY	WO	*Men*	MY	MO	Total
'andras (‘man’)	*78*	61	17	*41*	32	9	119
ji'neka (‘woman’)	*115*	90	25	*38*	26	12	153
a'ɣori (‘boy’)	*51*	41	10	*13*	10	3	64
ko'ritsi (‘girl’)	*68*	55	13	*10*	5	5	78
ko'pela (‘young girl 1’)	*183*	167	16	*39*	36	3	222
kope'ʎa (‘young girl 2’)	*19*	14	5	*5*	5	0	24
Total	*514*	428	86	*146*	114	32	660

spoken data is annotated for speaker sex and thus it is easier to identify the gender of speakers who make particular choices out of all gendered nouns. Numerical data regarding these preferences can be seen in Table 9.

A chi-square test on these data indicates a significant relationship between gender and the use of gender-related nouns at the .05 level ($\chi^2(5) = 17.89$, $p = .003$). In particular, the most significant differences are found in the use of words *'andras*, *ko'ritsi*, *ko'pela*. Men and women show a preference for the nouns related to their gender; *'andras* is the most frequently used gender-related noun by men (28.1 % of all nouns used by men), while *ko'pela* is the most frequently used noun by women (35.6 % of all nouns used by women). The percentage of these nouns is significantly lower in the opposite gender: *'andras* takes up 15.2 % of all nouns used by women, while *ko'pela* 26.7 % of all nouns used by men. Finally, this divergence is also significant in the case of *ko'ritsi,* in which the percentage of use is 13.2 % for women vs. 6.8 % for men.

With regard to the parameter of age statistical significance is even more pronounced ($\chi^2(15) = 41.85$, $p = .000$). Significant differences are observed not only between opposite genders of the same age, but also between the same gender of different age. In particular, young men (MY) show a significant preference for the noun *'andras* (28.1 % of all nouns used by young men), whereas for young women (WY) the percentage is much lower (14.3 %). It is also interesting that young women (WY) talk more about young women, especially using the noun *ko'pela* (39.0 % of all nouns used by young women), something which is in contrast with the much lower percentage of *ko'pela* for old women (18.6 % of all nouns used by old women). The same trend with old women is found in the case of old men, in which the percentage is very low (9.4 % of all nouns used by old men).

In all, groups of speakers distinguished in terms of age and gender tend to talk more about their own group, as suggested by statistical analysis. It would be even more revealing to study speaker preferences for specific meaning distinctions. Space does not allow for detailed analysis but we can summarize here a quite complex picture by pointing out that men use ‘man’ to mainly speak about GENDER, whereas women use the same word mainly with the meanings SPOUSE and RELATION. Women, on the other hand, mostly use ‘woman’ with the meaning PERSON, as well as in SPECIAL CONSTRUCTION phrases. Furthermore, men use ‘boy’ as a means of ADDRESS e.g. in the affectionate but challenging phrase ‘my boy’ when talking to their male friends (see example 36), while women mostly use it again with the meaning of PERSON.

Speaker preferences are also helpful in distinguishing the three synonym terms for 'girl'. It thus appears that women use *ko'ritsi* ('girl') with the meanings of PERSON and in phrases of ADDRESS as e.g. in the affectionate phrase 'my girl' when they talk to their female friends. Men use *ko'pela* ('young girl 1') to speak about GENDER, while women use it for PERSON and in SPECIAL CONSTRUCTION phrases. Lastly, *kope'ʎa* ('young girl 2') is used by both men and women in phrases of ADDRESS, sometimes in distancing or challenging uses (see 35 above).

The general picture that emerges about speaker preferences for meaning distinctions is that women speakers mostly talk about specific persons in our conversational data, whereas men tend to talk about gender in general. This typically happens by engaging in comparisons of men and women and their typical behaviour, as e.g. in example 18. Although this trend should be interpreted with care because of the special composition of our corpus, it accords with the research literature (e.g. Coates 2012) and suggests that such differences are part of Greek conversational practice, as well as of English-speaking communities. This data also suggests that conversation is predominantly oriented towards what happens to individual people (the female perspective) or how the two genders typically behave (the male perspective), whereas what seems to be important in written texts is their catering for different audiences (gender-specific or mixed).

5 Using Conversational Corpus Data in Gender Research

The contrastive analysis of conversational and written data has allowed us to draw some general conclusions about the treatment of gender in Greek discourse, which are summarized in the concluding section that follows. There are, however, several methodological and theoretical caveats that need to be taken into account before attempting to generalize on the basis of our findings.

First, as was repeatedly pointed out, our analytic findings depend crucially on the composition of the corpus, which reflects post-teenager talk. The lack of larger and general population Greek spoken corpora seriously restricts the scope of our findings and the questions that can be raised. This limitation reflects the nature of corpus studies (cf. McEnery and Hardie 2012: 2); however, the well-known difficulty of developing spoken corpora that are representative of the general population in ways similar to those in written corpora should not deter us from exploiting specialized spoken corpora of any kind.

Secondly, spoken corpora like the one examined in this paper seem not to be very revealing as regards the collocations of gender-related nouns. This may be related to the way spoken discourse functions in Greek or in most languages, namely that it does not allow for extensive patterning of this kind. However, spoken data seem to be useful in the study of special constructions, such as the tail construction with 'woman', found here with a different meaning and more frequently than in written data. They can also offer the opportunity to observe the use of a wide range of semi-synonymous linguistic items in text (see e.g. 39 below), which can then be studied in a larger corpus.

A more serious objection to the use of corpus linguistic approaches relates to the absence of close analysis of co-text. This is precisely one of the reasons why approaches like Conversation Analysis are inimical to quantitative approaches to data (see e.g. Schegloff 1988, 1993). Without doubt, corpus analysis tends towards generalizations that may conceal what actually happens in interactions, involving subtle differences that could be important for interpretation. For instance, in our corpus, because of psychological priming, speakers may continuously use the gendered term initially selected. For example, in extract (38) below all young male speakers adopt the noun *ko'pela* ('young girl 1') with the meaning PERSON. This noun is first introduced by speaker G in his characterization of Katerina for several turns, before another speaker (P) switches to the different meaning of GENDER, by making a generalization about all 'girls': [16]

(38) Γ: […] σοβαρά μιλάω αν δεν ήμουν με την Κατερίνα με τη συγκεκριμένη όμως κοπέλα η οποία με ανέχεται ανέχεται τις ζήλειες μου
Α: ναι
Γ: […] πολύ δύσκολα θα ήμουν με άλλη κοπέλα
((3 lines omitted))
Π: και καλά αυτή η κοπέλα πληροί τις προϋποθέσεις που έχεις βάλει εσύ
((2 lines omitted))
Α: είναι θέμα ότι (.) είναι μια κοπέλα (.) είναι μια κοπέλα η οποία ε:: κάνει αυτό το πράγμα […]
((11 lines omitted))

Π: όχι εγώ το λέω υπό την έννοια ότι:: η Κατερίνα είναι απ' τις κοπέλες οι οποίες θα:: την έχεις εμπιστοσύνη ρε παιδί μου ε:: […] γιατί σπανίζει ας πούμε να (.) οι κοπέλες να έχουν μια σχέση και είναι απόλυτα προσηλωμένες σ' αυτή
G: […] I'm serious, if I were not with Katerina, the specific girl who can stand me and my jealousy
A: yes
G: […] I would hardly be with another girl
((3 lines omitted))
P: that this girl meets the standards you've set
((2 lines omitted))
A: the point is (.) she's a girl (.) she's a girl who ehm does this thing […]
((11 lines omitted))
P: no, I mean that Katerina is one of the girls who:: you can trust, my friend, ehm […] because it's rare, say, to (.) for girls to have a relationship and stay absolutely faithful to it

[16] This example is thus characteristic of a gendered activity emerging from a previous noticing of gender, as Hopper and LeBaron (1998) show for English conversations.

This repeated use of the same noun increases the local frequency of the item and thus skews its even distribution throughout the corpus.

The opposite problem appears with subsequent references to the same person that are not realized by the same gendered term. There are instances where a re-negotiation of the initial term used by one of the speakers takes place, either by the same speaker or by others, as in extract (39):

(39) Γ: εδώ σου 'δωσαν ολόκληρο δίμετρο παλικάρι
((3 lines omitted))
Ε: Γιώργο πάντα:: πώς το λένε; (.) συνυπήρχα με πολύ ωραίους άνδρες
Γ: πιο ωραίος από μένα δεν υπάρχει
((story omitted))
Ε: […] θέλω ο άντρας που 'ναι δίπλα μου να με υπερασπίζεται
((4 lines omitted))
Ε: ρε άκου τι είπε ο άνθρωπος (.) κάνε πέρα
Β: καλά μωρέ κάνε πέρα είναι έκφραση
Ε: ρε Βάσω είσαι δίπλα με τον άνθρωπο που είσαι μαζί
Β: ναι
Ε: και λέει συγγνώμη ν' ανοίξω; έτσι θα 'λεγα; θα έλεγα τουλάχιστον συγγνώμη κάνε πέρα::
Β: καλά τώρα με τον άνθρωπο που είσαι μαζί δεν είναι απαραίτητο να πεις και το συγγνώμη το συγγνώμη το λες και με τον ξένο
((6 lines omitted))
Β: κοίτα ντάξει είναι και δυο μέτρα άντρας δεν είπες;
G: well, you were given a big, 2 metre high lad
((3 lines omitted))
E: George, I always, what's it called, (.) co-existed with very handsome men
G: there's no-one more handsome than me
((story omitted))
E: […] I want the man who's next to me to defend me
((4 lines omitted))
E: hey, listen to what the person [lit. human being] said (.) "move over"
V: that's OK, "move over" is just an expression
E: Vasso, you're next to the person [lit. human being] you're with
V: yes
E: and he says "sorry shall I open?", is this what I'd say?, I would at least have said "I'm sorry, move over"
V: OK now, to the person [lit. human being] you're with you don't need to say "sorry", "sorry" you say to a stranger
((6 lines omitted))
V: look, OK, he's a two metre high man, didn't you say?

Here the first speaker uses the noun *παλικάρι*/pali'kari 'lad' as a self-characterization in a SPECIAL CONTRUCTION phrase that is similar to the one found with the noun 'man' above (see e (i) in Sect. 4.2). After three turns his wife-to-be seems to accept his self-presentation by including him in the category of her boyfriends, who were generally very handsome, through the use of the word *άνδρες*/'anδres

'men' with the meaning of RELATION. She then tells a story presenting him as not having manners when talking to her and concludes with a generalization that she expects her boyfriend to defend her in a challenging situation. While in previous utterances the speaker continuously uses the gender-related noun 'man' with the meaning of RELATION, after four turns she uses the pseudo-generic *'anθropos* ('human being') for him in an SPECIAL CONSTRUCTION, like those found with 'woman'. She then repeats this word, when addressing a friend, probably because it is primed by its use in the previous phrase, but this time with the meaning of RELATION. Her friend takes up this word used with the same meaning and then at her last turn she goes back to the SPECIAL CONSTRUCTION use found at the very beginning of this long exchange with the word *'andras* ('man').

This extract suggests that speakers are influenced in their choice of gender-related nouns both by predominant patterns in the use of these nouns, as those that we have analysed in this paper, and by their need to cohere with a previous speaker. As a result, nouns like 'man' can shift from one meaning to another within a short span, while other nouns like *pali'kari* ('lad') or *'anθropos* ('human being') can be used with similar meanings.[17]

These particularities of spoken interaction, which are produced by its sequential and cumulative nature, are not easy to accurately capture through an exclusively corpus linguistic analysis. It is for this reason that researchers like Walsh (2013) point out the need for combining corpus linguistic approaches with other methodologies for the analysis of spoken discourse. As he suggests, a Corpus Linguistic Conversation Analysis methodology (CLCA, in his term) "goes some way at least in compensating for the deficiencies of each method when used alone": corpus linguistics can provide a means of generalization beyond the small sample of data that Conversation Analysis typically uses, while Conversation Analysis can offer the kind of interactional detail, usually left out in corpus linguistic approaches (Walsh 2013: 47).

We fully subscribe to this view, since our research suggests that corpus-based methods should be combined with an awareness of the use of gender-related terms in situated interaction. It is precisely this combination of macro- and micro-analysis that can open new perspectives to gender research.[18] In this combined effort, corpora offer a valuable asset, since they can reveal the resources drawn upon in recurrent discourses on gender, which are responsible for cumulative ideological effects on the representation of gender in micro-communities, such as the one studied in our data.

[17] In relation to this it is worth noting that a variety of nouns can be used in special construction phrases for men with an evaluative function, of which 'man' and 'lad' are used for positive evaluation, while *'anθropos* seems to be reserved for neutral or negative evaluation.

[18] Similar attempts to integrate quantitative (e.g. variationist) with qualitative (e.g. constructionist) approaches to gender become increasingly popular in the relevant literature (see e.g. Holmes 1997). The same call is increasingly heard within the frame of corpus linguistics (e.g. Mautner 2009).

6 Conclusions and Further Research

Our focus in this paper has been on the treatment of gender in Greek authentic spoken data. Our purpose has been to investigate the potential of corpus linguistic methods in the analysis of gender, and in particular to bring together our previous findings from the study of written Greek genres with the study of conversation. As is evident, our conclusions should be interpreted in the light of the restrictions that arise from the composition of our corpus. As was suggested above, a fuller analysis could be achieved by a combination of qualitative and quantitative methods in the analysis of spoken discourse, which would overcome the shortcomings of corpus linguistic approaches to conversation.

As a first general conclusion, we have established that women-related nouns tend to predominate in both spoken and written text types in Greek, in contrast to what literature has suggested about English data. Although gender-related nouns occur almost equally for each gender in conversational data, there is a clear tendency for female-related nouns to exceed male-related nouns. The comparison of different genres is thus suggestive of a general trend in the language as a whole.

In addition, Greek conversational data were found to be generally much less male-dominated and stereotypical in the construction of gender identity than newspapers and magazines. This is achieved, among else, by avoiding derogatory meaning distinctions for women or especially flattering meanings for men, by employing special constructions in an affiliative rather than ironical treatment of women etc. Although this could be expected of the female university students who predominantly interact in our corpus, it also suggests that spoken interaction may be subject to different processes of negotiation that are unavailable to readers of newspapers and magazines, who are exposed to portrayals of gender constituted of (and constituting) dominant discourses. Our suggestion is, at least with respect to our data, that access to discourse in everyday interactions between intimates is not controlled by specific people, as is the case with mass media, and this allows for more flexible gender representations in spoken discourse.

At the same time, the comparison of spoken with written corpus data has allowed us to identify the particularities of each mode of discourse. Thus, conversation is predominantly preoccupied with interactional and evaluative meanings in contrast to written genres, in which more specialized meanings are developed, especially with regard to the audience to which they are addressed. What seems to transpire through our conversational data is the prominence of social topics in them, and especially social experiences of people, either as examples of their gender or as individual cases (what we dubbed above the male and female perspective, respectively). This seems to underline the importance of conversational interaction in the life of language users and the role of gender in this interaction.

Finally, in quantitative terms, our statistical analysis has indicated that groups of speakers, distinguished in terms of age and gender, tend to talk more about their own group, while in qualitative terms women in our corpus were found to talk about both men and women in terms of specific persons rather than as representatives

of their gender, as opposed to male speakers, who tend to favour comparisons between men and women in terms of their "typical" behaviour. This certainly reflects the tendency of young men in our data to be preoccupied with questions of gender, but also concurs with the focus of women speakers on the specific and their avoidance of gender-based generalizations, which has already been found in the literature.

Further research on more varied groups is necessary in order to reach any definite conclusions on the basis of these findings. As more and more varied spoken corpora become available for Greek, it will be possible to test the extent to which our findings are representative of what happens in interactions between young women or are typical of larger discourse strategies. Moreover, the analysis of individual nouns should be complemented by a study of other gender-related words such as adjectives, diminutives or supposedly neutral words for referring to people such as *άνθρωπος* /'anθropos 'human being' or *παιδί* /pe'δi 'kid, guy' etc. Finally, more data will allow us to answer specific questions such as the relative age of boys and girls who appear in conversational data and will make possible a more extensive statistically sensitive treatment of our findings.

Concluding, it seems that a corpus linguistic approach to the study of gender-related terms in spoken discourse can be useful in drawing a general picture of qualitative (meanings of terms and meanings of collocations) and quantitative (frequency of use, frequency of meanings and collocations, statistical correlations between parameters) characteristics of the items under research. Apart from these representation findings (in Baker's terms mentioned in the Introduction), the metadata on each speaker's gender, which are available for spoken texts, offer the opportunity of a usage-based perspective into data by exploring speaker preferences. Thus, corpus linguistic analysis offers us the broad picture, by generalizing on the basis of quantitative and qualitative data, which can then be refined by resorting to close conversation analysis, in order to obtain more sensitive findings on the use and the meaning negotiation of gender-related terms.

Appendix: Transcription Conventions

[	overlapping talk
[	
(.)	small pause
@@	inaudible speech
::	sound or syllable lengthening
CAPS	louder voice
(( ))	transcriber comments
[…]	omitted segment
underline	point of discussion

References

Adolphs, S. (2008). *Corpus and context: Investigating pragmatic functions in spoken discourse*. Amsterdam/Philadelphia: John Benjamins.

Archakis, A., & Papazachariou, D. (2008). Prosodic cues of identity construction: Intensity in Greek young women's conversational narratives. *Journal of Sociolinguistics, 12*(5), 627–647.

Babiniotis, G. (2002). Μπαμπινιώτης Γ. 2002. *Λεξικό της Νέας Ελληνικής Γλώσσας* [= Dictionary of the Modern Greek Language] (2nd ed.). Athens: Lexicology Centre.

Baker, P. (2008). 'Eligible' bachelors and 'frustrated' spinsters: Corpus linguistics, gender and language. In K. Harrington, L. Litosseliti, H. Sauntson, & J. Sunderland (Eds.), *Gender and language research methodologies* (pp. 73–84). Basingstoke: Palgrave Macmillan.

Baker, P. (2010). Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender and Language, 4*(1), 125–149.

Baker, P. (2013). *Introduction: Virtual special issue of gender and language on corpus approaches*. https://www.equinoxpub.com/journals/index.php/GL/article/view/17185/13506. Accessed 15 May 2013.

Baker, P. (2014). *Using corpora to analyze gender*. London: Bloomsbury.

Bolinger, D. (1980). *Language, the loaded weapon*. London: Longman.

Caldas-Coulthard, C., & Moon, R. (2010). 'Curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse and Society, 21*(2), 99–133.

Charteris-Black, J., & Seale, C. (2009). Men and emotion talk: Evidence from the experience of illness. *Gender and Language, 3*(1), 81–113.

Coates, J. (2012). Gender and discourse analysis. In J. P. Gee & M. Handford (Eds.), *The Routledge handbook of discourse analysis* (pp. 90–103). London: Routledge.

del-Teso-Craviotto, M. (2006). Words that matter: Lexical choice and gender ideologies in women's magazines. *Journal of Pragmatics, 38*, 2003–2021.

Fragaki, G., & Goutsos, D. (2005). Gender adjectives and identity construction in Greek corpora. *Proceedings of the 7th international conference on Greek Linguistics, University of York, 8–10 September 2005*. http://icgl7.projects.uoi.gr/Fragaki-et-al.pdf. Accessed 15 May 2013.

Gesuato, S. (2003). The company women and men keep: What collocations can reveal about culture. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the corpus linguistics 2003 conference* (pp. 253–262). Lancaster: UCREL, Lancaster University.

Goutsos, D. (2010). The corpus of Greek texts: A reference corpus for Modern Greek. *Corpora, 5*(1), 29–44.

Goutsos, D., & Fragaki, G. (2009). Lexical choices of gender identity in Greek genres: The view from corpora. *Pragmatics, 19*(3), 317–340.

Hatzidaki, O. (2011). Greek men's and women's magazines as codes of gender conduct: The appropriation and hybridisation of deontic discourses. In D. Majstorovic & I. Lassen (Eds.), *Living with patriarchy: Discursive constructions of gendered subjects across cultures* (pp. 113–144). Amsterdam/Philadelphia: John Benjamins.

Holmes, J. (1997). Women, language, and identity. *Journal of Sociolinguistics, 1*(2), 195–223.

Holmes, J. (2000). *Ladies* and *gentlemen*: Corpus analysis and linguistic sexism. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory. Papers from the 20th international conference on English Language Research on Computerized Corpora (ICAME 20)* (pp. 141–155). Amsterdam: Rodopi.

Holmes, J. (2001). A corpus-based view of gender in New Zealand English. In M. Hellinger & H. Bussman (Eds.), *Gender across languages. The linguistic representation of women and men* (Vol. 1, pp. 115–136). Amsterdam/Philadelphia: John Benjamins.

Holmes, J., & Sigley, R. (2002). What's a word like *girl* doing in a place like this? Occupational labels, sexist usages and corpus research. In P. Peters, P. Collins, & A. Smith (Eds.), *New frontiers of corpus linguistics. Papers from the 21st international conference on English language research on computerized corpora. Sydney 2000* (pp. 247–263). Amsterdam: Rodopi.

Holmgreen, L.-L. (2009). Metaphorically speaking: Constructions of gender and career in the Danish financial sector. *Gender and Language, 3*(1), 1–32.

Hopper, R., & LeBaron, C. (1998). How gender creeps into talk. *Research on Language and Social Interaction, 31*(1), 59–74.

Johnson, S., & Ensslin, A. (2007). 'But her language skills shifted the family dynamics dramatically'. Language, gender and the construction of publics in two British newspapers. *Gender and Language, 1*(2), 229–254.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities, 31*(2), 91–113.

King, B. W. (2011). Language, sexuality and place: The view from cyberspace. *Gender and Language, 5*(1), 1–30.

Leech, G., & Fallon, R. (2004 [1992]). Computer corpora – What do they tell us about culture? In G. Sampson, & D. McCarthy (Eds.), *Corpus linguistics. Readings in a widening discipline* (pp. 160–171). London: Continuum.

Macalister, J. (2011). Flower-girl and bugler-boy no more: Changing gender representation in writing for children. *Corpora, 6*(1), 25–44.

Makri-Tsilipakou, M. (1989). The gender of άνθρωπος: An exercise in false generics. *Proceedings of the third annual symposium: Description and/or comparison of English and Greek* (pp. 61–83). Department of Theoretical and Applied Linguistics, School of English, Aristotle University of Thessaloniki, Thessaloniki.

Mautner, G. (2009). Checks and balances: How corpus linguistics can contribute to CDA. In R. Wodak & M. Meyer (Eds.), *Methods for critical discourse analysis* (pp. 123–143). London: Sage.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Pavlidou, T.-S. (2003a). Patterns of participation in classroom interaction: Girls' and boys' noncompliance in a Greek high school. *Linguistics and Education, 14*(1), 123–141.

Pavlidou, T.-S. (2003b). Women, gender and modern Greek. In M. Hellinger & H. Bussman (Eds.), *Gender across languages: The linguistic representation of women and men* (Vol. 3, pp. 175–199). Amsterdam/Philadelphia: John Benjamins.

Pavlidou, Th.-S. (2006). Γλώσσα-γένος-φύλο: Προβλήματα, αναζητήσεις και ελληνική γλώσσα [= Language-gender-sex. Problems, inquiries and the Greek language]. In Th.-S. Pavlidou (Ed.), *Γλώσσα-Γένος-Φύλο* [= Language-Gender-Sex] (pp. 15–64, 2nd ed.). Thessaloniki: Institute of Modern Greek Studies.

Pearce, M. (2008). Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine. *Corpora, 3*(1), 1–29.

Romaine, S. (2001). A corpus-based view of gender in British and American English. In M. Hellinger & H. Bussman (Eds.), *Gender across languages. The linguistic representation of women and men* (Vol. 1, pp. 153–175). Amsterdam/Philadelphia: John Benjamins.

Schegloff, E. A. (1988). Discourse as an interactional achievement II: An exercise in conversation analysis. In D. Tannen (Ed.), *Linguistics in context: Connecting observation and understanding* (pp. 135–158). Norwood: Ablex.

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction, 26*, 99–128.

Schmid, H.-J. (2003). Do women and men really live in different cultures? Evidence from the BNC. In A. Wilson, P. Rayson, & T. McEnery (Eds.), *Corpora by the Lune: A Festschrift for Geoffrey Leech* (pp. 185–221). Frankfurt: Peter Lang.

Sigley, R., & Holmes, J. (2002). Looking at girls in corpora of English. *Journal of English Linguistics, 30*(2), 138–157.

Stubbs, M. (2001). *Words and phrases. Corpus studies of lexical semantics*. Oxford: Blackwell.

Taylor, C. (2013). Searching for similarity using corpus-assisted discourse studies. *Corpora, 8*(1), 81–113.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.

Triandaphyllidis. (1998). Ίδρυμα Μανόλη Τριανταφυλλίδη. 1998. *Λεξικό της κοινής νεοελληνικής* [= Dictionary of Modern Greek]. Thessaloniki: Institute of Modern Greek Studies, Manolis Triandaphyllidis Foundation.

Valiouli, M. (1990). *Anaphora, agreement and the pragmatics of "right dislocations" in Greek*. Unpublished Phd dissertation, Aristotle University of Thessaloniki.

Walsh, S. (2013). Corpus linguistics and conversation analysis at the interface: Theoretical perspectives, practical outcomes. In J. Romero-Trillo (Ed.), *Yearbook of corpus linguistics and pragmatics 2013* (pp. 37–51). Dordrecht: Springer.

Does Speaker Role Affect the Choice of Epistemic Adverbials in L2 Speech? Evidence from the Trinity Lancaster Corpus

Dana Gablasova and Vaclav Brezina

Abstract This study investigates stance-taking strategies in a context of an examination of spoken English. The focus of the research is on the interaction between the candidates (advanced L2 speakers) and the examiners (L1 speakers of English). In particular, the study explores the use of epistemic adverbial markers such as 'maybe', 'certainly' and 'surely'. These markers are used not only to express speakers' position (certainty or uncertainty) towards a statement, but also to express speakers' position towards other interlocutors (e.g. to manage interpersonal relationships or to downplay strong assertions). The study is based on the advanced subsection of the Trinity Lancaster Corpus of spoken L2 production which currently contains approximately 0.45M words based on four speaking tasks: one mostly monologic task and three highly interactive tasks. The study compares the expression of epistemic stance by both the candidates and examiners and explains the differences between speakers' performance in terms of different speaker roles assumed by the candidates and examiners in three dialogic tasks. The study stresses the importance of looking at the contextual factors of speakers' pragmatic choices and demonstrates that when studying L2 spoken production it is important to go beyond characterising the speakers as 'native' or 'non-native' speakers of a language. Whereas the fact of being a 'native user' or a 'non-native user' can indeed be part of the speaker role and speaker identity, other equally important factors arising from the context of the exchange may play a role in speakers' stance-taking choices.

Keywords Epistemic stance • L2 pragmatics • L2 spoken production • Speaker roles • Learner corpus

D. Gablasova (✉) • V. Brezina
ESRC Centre for Corpus Approaches to Social Science, Lancaster University, Lancaster, UK
e-mail: d.gablasova@lancaster.ac.uk; v.brezina@lancaster.ac.uk

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_6

1 Introduction

Epistemic stance-taking is an important aspect of communicative skills, whether in one's first or additional language. It plays an essential role in conveying the epistemic perspective of the speaker (i.e. his or her certainty-related evaluation of what is said) as well as in managing and negotiating interpersonal relationships between speakers (Kärkkäinen 2003, 2006; Hunston and Thompson 2000). However, despite the significance of stance-taking in everyday discourse (Biber et al. 1999), so far there has been only a limited number of studies that address this issue in second language spoken production (e.g. Aijmer 2004; Fung and Carter 2007; Mortensen 2012). This study therefore aims to contribute to our understanding of this area by exploring how epistemic stance is expressed in the context of a spoken English exam by two groups of speakers – the (exam) candidates (advanced L2 speakers of English) and examiners (L1 speakers of English). In particular, this study focuses on how the speakers use epistemic adverbials to position themselves in three speaking tasks which differ in terms of speaker roles and aims of the communication.

The topic is addressed with the help of a new, growing corpus of L2 spoken production – the Trinity Lancaster Corpus (TLC). The corpus represents semi-formal institutional speech, and thus complements other corpora of L2 spoken language that may elicit more informal spoken production (e.g. LINDSEI). The corpus is described in greater detail in Sect. 2.1. The TLC allows us to study each L2 speaker in four different speaking tasks (three of which are highly interactive) while also taking into consideration individual differences between speakers, the importance of which has been repeatedly stressed in recent corpus-based studies of L2 language (Mukherjee 2009; Callies 2013). Overall, the study contributes to our understanding of pragmatic ability of advanced L2 speakers of English.

1.1 The Pragmatics of Epistemic Stance

Epistemic stance – the expression of different degrees of certainty and uncertainty in discourse – is a complex notion. Indeed, epistemic markers (i.e. linguistic signals of epistemic stance) represent a heterogeneous group of linguistic items both from the formal and the functional perspective (Coates 1987, 1990; Holmes 1990; Aijmer 2002; Simon-Vandenbergen and Aijmer 2007). Epistemic markers include a variety of linguistic forms (adverbs, adjectives, nouns, lexical verbs and modals) that in addition to the epistemic (i.e. certainty-oriented) function have also a number of social and discourse-oriented functions (e.g. managing interpersonal relationship and politeness strategies).

The focus of this paper is on epistemic adverbs which, arguably, are relatively stable from the functional perspective and therefore lend themselves more easily to quantitative corpus-based investigation. For instance, *certainly* expresses a high degree of certainty across different contexts. However, regardless of this semantic stability, we have to be mindful of the fact that adverbial epistemic markers (AEMs)

have also a social function and play an important part in the inter-speaker interaction, which, as Kärkkäinen (2006) argues, is crucial for the full appreciation of their meaning scope. This feature can be best demonstrated with the following example (1) taken from the Trinity Lancaster Corpus of L2 spoken production (for more information about the corpus, see Sects. 2.1 and 2.2) in which speakers discuss social change and its evaluation.

(1) [S1 first speaks about the fact that things were better in the past.]
S2: I agree with this point but don't you think **maybe** the ti=
fact that times are changing is a good thing?

In this example, the first speaker (S1) argues that things in the past were better than they are now. The second speaker (S2) disagrees with this opinion and in his reply employs the epistemic adverb 'maybe' signalling low certainty. However, it could be argued that in this reply the epistemic marker does not only function to express the exact degree of speaker's certainty (subjective function) but also plays an important role in managing the intersubjective relationship – i.e. downplaying the possibly face-threatening nature of the disagreement.

Simon-Vandenbergen and Aijmer (2007) discuss this issue with reference to the distinction between *conceptual* and *procedural* meaning. The term *conceptual meaning* is used to refer to the semantics proper of an epistemic form, while the notion of *procedural meaning* is reserved for the pragmatic meaning, i.e. speaker's "attitudes to discourse and the participants in it" (Traugott and Dasher 2002: 10). Simon-Vandenbergen and Aijmer (2007) therefore suggest that the function of epistemic forms such as the adverb *certainly* can be understood in terms of both of these meanings:

> *[C]ertainly* is contentful in that it means epistemic certainty and procedural when looked upon from the perspective of indexing the speaker's or writer's stance to the text or one of the participants. (p. 54)

Similarly, Coates (1987: 130) points out that epistemic markers operate in a rich pragmatic environment that enables them to take up different layers of meaning:

> I am not sure that it is possible to say exactly what any one modal form 'means' on any particular occasion. In informal conversation, where participants are trying to achieve, simultaneously, the goals of (a) saying something on the topic under discussion; (b) being sensitive to the face-needs of the various addressees; (c) qualifying assertions to avoid total commitment to a point of view which they may want to withdraw from; (d) qualifying assertions to encourage the flow of discussion; (e) creating cohesive text, then it does not seem feasible to conclude 'this form expresses x and that form expresses y'. Speakers exploit the polypragmatic nature of the epistemic modals to say many things at once.

Coates's comment applies not only to informal conversation, but can also be extended to other spoken and written contexts, which provide particular frameworks of speaker/writer roles and register expectations. It is precisely these normative frameworks that are the focus of this study. In particular, we explore the effect of speaker role on the choice and quantity of AEMs used by speakers in three different speaking tasks.

1.2 Epistemic Stance in L2 Spoken Production

With the rise of corpora capturing spoken L2 production, more attention has been given to the use of pragmatic markers by L2 speakers. Previous studies examined a variety of markers with different pragmatic functions (e.g. Aijmer 2004, 2011; Müller 2005; Fung and Carter 2007; Buysse 2012). Most of these studies described the use of pragmatic markers by L2 users in relatively broad terms and compared this to native speaker production.

There are also several studies on second language pragmatic ability that specifically focused on epistemic stance-taking, most of them addressing epistemicity in writing (e.g. Hyland and Milton 1997; McEnery and Kifle 2002; Fordyce 2014) with only a few investigating spoken production (e.g. Mortensen 2012; Baumgarten and House 2007, 2010). However, stance-taking patterns in writing are significantly different from expressions of stance in spoken (and often highly interactive) communication. The difference stems from the presence of at least one other active interlocutor in spoken exchange which creates demands on managing intersubjective relationships between two or more interlocutors.

Studies that focused on epistemic expressions used by L2 users in spoken communication observed that L2 speakers used epistemic markers for both subjective and intersubjective functions (Aijmer 2004; Mortensen 2012). Yet some researchers pointed out that L2 speakers' use of specific epistemic markers differed from that of native speakers with regard to the range of pragmatic functions (Baumgarten and House 2007; Aijmer 2004). Other researchers noted that non-native speakers used a more restricted range of epistemic forms compared to L1 users (Bardovi-Harlig and Salsbury 2004; Fordyce 2009). However, the focus on the cross-linguistic comparison (i.e. native versus non-native speakers of the language) did not allow for a more detailed investigation of how contextual factors such as speaker roles, the communicative goals and the topic of the exchange contribute to the linguistic choices of either group of speakers.

1.3 Research Questions

The study was motivated by one overarching question: what is the difference between epistemic stance as expressed through adverbials by candidates and examiners in spoken English? This general question was then divided into the following three specific research questions:

RQ 1: Is there a difference in the frequency of AEMs used by the candidates and the examiners?

RQ 2: Is there a difference between the frequency of certainty and uncertainty AEMs used by the two groups of speakers?

RQ 3: How does a particular interactional setting affect epistemic stance expressed by the candidates and the examiners?

A note on terminology: Although previous studies often used labels such as 'L2/non-native speakers' or 'learners' and contrasted this group with 'L1/native speakers', this dichotomy is not adopted in this study. While we are interested in how L2 speakers in the corpus express epistemic stance, their language use should not be automatically conceptualised through their status as 'non-native speakers' and all variation in the data should not be ascribed to the difference between native and non-native use. Instead, we want to focus on the roles that speakers perform in different speaking tasks. For this reason, the two groups of speakers in this study will be referred to as '(exam) candidates' and 'examiners' because these are their roles in the framework of the data that comes from an examination of spoken English.

2 Method

2.1 Corpus Description

The corpus used in this study is the advanced subsection of the Trinity Lancaster Corpus of spoken L2 production (TLC). The corpus is based on examinations of spoken English conducted by the Trinity College London, a major international examination board, and contains interactions between exam candidates (L2 speakers of English) and examiners (L1 speakers of English). At present, this sub-section of the corpus contains approximately 0.45 million words, with almost 300,000 tokens produced by the candidates and about 150,000 tokens produced by the examiners. These data came from 132 candidates and 66 examiners (some examiners participated in more examinations). The recordings come from six countries – 31 recordings were made in Italy, 31 in Mexico, 30 in Spain, 23 in China, 13 in Sri Lanka and 4 in India. The candidates included in the corpus were examined at Grades 10, 11 and 12 of the Graded Examinations in Spoken English (GESE) which correspond to C1 and C2 levels of the Common European Framework of Reference for Languages (CEFR). To ensure a minimal level of L2 proficiency, only candidates who fulfilled the requirements of the Grade and were awarded a Pass were included in the corpus.

2.2 Composition of the Corpus

In the advanced sub-corpus of the TLC, speech from each candidate was elicited in four speaking tasks – one monologic and three dialogic tasks. The four tasks are: a *presentation* (PRES), in which the candidates talk on a pre-prepared topic of their own choice. This task has the format of a formal presentation and apart from occasional back-channelling signals or comments from the examiner is largely monologic. *Presentation* is followed by the second task, *a discussion* (DISC), in which the examiner and the candidate further discuss some of the ideas and topics

Table 1 Overview of the sub-corpus (advanced L2 users)

Task	Candidates	Examiners	Total
Presentation	86,549	11,693	98,242
Discussion	61,913	43,440	105,353
Interactive task	50,093	41,314	91,407
Conversation	90,382	56,720	147,102
Dialogic tasks combined	202,388	141,474	343,862
Corpus total	288,937	153,167	442,104

introduced in the presentation. The third task, the *interactive task* (INT), is based on the prompt delivered by the examiner. The prompt is a statement that involves an observation or an issue (usually presented as a personal issue or belief of the examiner) about which the candidate should provide comments and suggestions. The candidate also needs to ask questions to find out more about the situation or issue presented. Finally, in the last task, *the conversation* (CONV), the candidate talks with the examiner on two topics of general interest. In Grades 10 and 11 these are chosen by the examiner from one of the two lists both of which the candidate is familiar with. For Grade 12, the most advanced level, any topic can be selected by the examiner. All three dialogic tasks described here are semi-formal in nature and highly interactive. Each sub-component of the exam lasts for about 5 min and altogether the corpus contains about 20 min of speech from each candidate at the C1/C2 level. Since the exam allows the candidates to bring in their own topics for the *presentation* and some aspects of this topic are also discussed in the *discussion,* the corpus contains spoken L2 production on a great variety of topics. A more detailed description of the exam and each speaking task can be found in the Exam Syllabus by Trinity College London (2009). Table 1 summarises the number of tokens produced in each task in the advanced subsection of the corpus.

As can be seen from the table, there is a lot of variation in terms of the amount of speech produced by the candidates and examiners in individual tasks. This variation is largely attributable to the nature of the task and the roles of the speakers. For example, in the *conversation,* candidates are given more space to develop their answers than in the *interactive task* in which one of their main requirements is to ask the questions with the examiner providing many of the answers.

2.3 Identifying Adverbial Epistemic Markers

When studying any aspects of pragmatics, which is notoriously sensitive to and dependent on context variables, we need to consider carefully the suitability of the tools and the procedure, especially if it involves automatic corpus analysis (Hunston 2007; Rühlemann 2011). The approach we have chosen was to combine automatic corpus searches with manual analysis to ensure high quality of the results. First, a list of candidate adverbial epistemic markers (AEMs) was compiled based on previous studies that focused on epistemicity, i.e. Holmes (1988), Biber et al. (1999) and

Table 2 Adverbial epistemic markers signalling certainty and uncertainty

Epistemic function	Adverbial markers
Expression of uncertainty	Maybe, perhaps, possibly, probably
Expression of certainty	Certainly, clearly, definitely, for sure, inevitably, no doubt, obviously, of course, surely, undoubtedly

Brezina (2012). The following AEMs were included in this list: *actually, apparently, beyond doubt, certainly, clearly, definitely, doubtless, evidently, for sure, indeed, indubitably, inevitably, unmistakably, kind of, maybe, necessarily, no doubt, obviously, of course, perhaps, plainly, possibly, predictably, probably , really, roughly, sort of , surely, undeniably, undoubtedly, without * doubt.*

On this list, there were several forms that are used also for other than epistemic functions. All of the expressions from this list were searched in the corpus and decisions to exclude some of the words from the list were made on the basis of their primarily non-epistemic functions. The expressions excluded from the search were: *really*, *indeed* and *roughly*.

In order to answer the second research question, the AEMs signalling certainty and uncertainty were selected. These can be seen in Table 2. Only the AEMs that unambiguously signal either certainty or uncertainty were included in this analysis. Thus, for example, 'necessarily' was left out as it appeared often in combinations such as 'not necessarily' in the corpus.

3 Results

3.1 RQ1: Difference Between the Number of AEMs Used by Candidates and Examiners

In this section, we examine the number of AEMs used by the candidates and examiners in their interactions. First, we compare the overall number of AEMs used by the two groups of speakers in each task and then we look more closely at individual AEMs used by the speakers. With respect to the overall number of AEMs, Table 3 shows the descriptive statistics for each group of users as well as the results of the Mann–Whitney U test conducted to compare the examiners and the candidates (the frequencies are normalised to 1,000 words). This analysis showed that the difference was statistically significant in one out of the three compared tasks, namely in the *interactive task.*

With respect to the average number of AEMs produced, the differences among the three dialogic tasks were relatively small: The candidates produced the largest number of AEMs in the *interactive task*, followed by the *discussion* and *conversation*; to some extent opposite trend was observed for the examiners with the highest number of AEMs recorded in the *discussion*, followed by the *conversation* and the *interactive task.*

Table 3 Number of AEMs produced per task by candidates and examiners

	CAND		EX		Comparison (N = 132[a])	
	Mean	SD	Mean	SD	Mann–Whitney U	Sig.
DISC	**8.68**	8.32	**7.41**	6.99	7888.0	.183
INT	**9.43**	8.04	**6.78**	6.76	6865.0	.003**
CONV	**8.00**	6.60	**6.87**	6.44	7620.5	.078

**p < .01

Notes

[a]Please note that as explained in Sect. 2.1, data from 132 L2 users and 66 examiners were included in this analysis since some examiners took part in more than one examination

Table 4 Individual AEMs produced by candidates and examiners: results for three dialogic tasks

Adverbial epistemic markers	Dialogic tasks				LL-score (significant results highlighted)
	CAND		EX		
	Freq.	RF	Freq.	RF	
1. Actually	263	**129.95**	140	**98.96**	6.95
2. Apparently	1	**0.49**	6	**4.24**	5.98
3. Certainly	6	**2.96**	50	**35.34**	57.04
4. Clearly	3	**1.48**	5	**3.53**	1.48
5. Definitely	36	**17.79**	13	**9.19**	4.56
6. For sure	8	**3.95**	0	**0.00**	8.48
7. Inevitably	0	**0.00**	2	**1.41**	3.55
8. Kind of	289	**142.80**	139	**98.25**	13.64
9. Maybe	752	**371.56**	210	**148.44**	160.61
10. Necessarily	1	**0.49**	28	**19.79**	42.1
11. No doubt	0	**0.00**	1	**0.71**	1.78
12. Obviously	34	**16.80**	43	**30.39**	6.73
13. Of course	182	**89.93**	52	**36.76**	37.4
14. Perhaps	43	**21.25**	112	**79.17**	61.47
15. Possibly	6	**2.96**	29	**20.50**	25.8
16. Probably	93	**45.95**	65	**45.94**	0.0
17. Sort of	38	**18.78**	97	**68.56**	51.11
18. Surely	10	**4.94**	52	**36.76**	48.18
19. Undoubtedly	0	**0.00**	1	**0.71**	1.78
Total types used	16		18		

Next, we looked at the individual AEMs used by the examiners and candidates. Table 4 shows the individual AEMs produced in the three dialogic tasks (i.e. *discussion, interactive task* and *conversation)*. The table presents both the absolute frequencies (Freq.) and the relative frequencies normalised to 100,000 words (RF). In addition, the results of a statistical test (Log likelihood) are reported comparing the use of the individual AEMs by the candidates and the examiners. Because 19 statistical

tests were performed on the dataset, a Bonferroni-adjusted significance level of 0.002632 was required to account for the increased possibility of type-I error. With this alpha level, the test statistic cut-off point is 9.05. In addition, further caution is necessary when interpreting the results of Log likelihood as this type of analysis may over-emphasise differences between compared groups (cf. Brezina and Meyerhoff 2014).

As can be seen from the table, as a group, examiners and candidates used a similar number of different AEMs (16 and 18) in their speech. Both groups of speakers used the following three AEMs very frequently – 'actually', 'kind of', 'maybe' while the AEMs such as 'apparently', 'clearly' and 'inevitably' were used only sporadically by speakers. The largest difference between the two groups of speakers was observed with respect to the following AEMs:

(a) AEMs used more often by examiners: 'certainly', 'necessarily', 'perhaps', 'sort of', 'surely', 'possibly'
(b) AEMs used more often by candidates: 'maybe', 'of course'

3.2 *RQ 2: Expression of Certainty and Uncertainty by Candidates and Examiners*

In this section we first investigate the number of certainty and uncertainty AEMs produced by the speakers in each of the examined tasks, before reporting on a case study of certainty markers used by the examiners and the candidates in two selected tasks, the *discussion* and the *interactive task.*

As the first step, the number of AEMs expressing uncertainty was compared for the two groups of speakers. The results of this analysis can be seen in Table 5, which shows both the descriptive statistics for the two groups of speakers as well as the results of the Mann–Whitney U test used to compare the two groups. The frequencies in the table are normalised to 1,000 words.

As can be seen from the table, most markers of uncertainty were used by both groups in the *interactive task*. This was followed by the *conversation* and the *discussion* for the candidates and by the *discussion* and *conversation* for the examiners. As

Table 5 Adverbial epistemic markers expressing uncertainty

	CAND		EX		Comparison (N = 132[a])	
	Mean	SD	Mean	SD	Mann–Whitney U	Sig.
DISC	**3.93**	6.32	**2.38**	3.66	7562.0	.048*
INT	**5.70**	6.49	**3.16**	4.39	6671.0	.001**
CONV	**4.26**	4.90	**2.73**	3.45	6919.0	.003**

**p < .01; *p < .05

Notes

[a]Please note that as explained in Sect. 2.1, data from 132 L2 users and 66 examiners were included in this analysis since some examiners took part in more than one examination

Table 6 Adverbial epistemic markers expressing certainty

	CAND		EX		Comparison (N = 132[a])	
	Mean	SD	Mean	SD	Mann–Whitney U	Sig.
DISC	**1.71**	3.16	**2.02**	2.93	8072.0	.248
INT	**.81**	1.74	**1.22**	2.25	7953.05	.115
CONV	**1.34**	2.58	**1.31**	2.14	8668.5	.936

Notes
[a]Please note that as explained in Sect. 2.1, data from 132 L2 users and 66 examiners were included in this analysis since some examiners took part in more than one examination

for the comparison of the two speaker groups, across all compared tasks the candidates used on average more markers of uncertainty than the examiners with the difference being statistically significant in all cases.

Next, the number of AEMs expressing certainty was compared for both groups of speakers. The descriptive statistics and the results of the Mann–Whitney U test used to compare the two groups can be seen in Table 6 (the frequencies are normalised to 1,000 words).

Although in the *discussion* and the *interactive task* the examiners employed on average more AEMs of certainty than the candidates and in the *conversation* the candidates used more of these markers than examiners, these differences were not statistically significant.

3.3 Case Study: Expressions of Certainty in the Discussion *and the* Interactive Task

The results in the previous sections showed that there was no statistically significant difference in the number of AEMs expressing certainty between the candidates and the examiners. However, to better understand the nature of the communicative strategies of these two groups of speakers, a close analysis examining the nature of interaction in which the certainty markers were used was conducted. In this study, two tasks, the *discussion* and the *interactive task*, were selected to contrast two different speaking scenarios: in the *discussion,* the speakers discuss a topic that the candidate is familiar with while in the *interactive task* it is the examiner who is the primary knower in the task, having all the relevant information which the candidate has to elicit from him or her.

Following grounded analysis of the data, three types of contexts in which the AEMs of certainty were employed were identified (cf. Kärkkäinen 2006; Simon-Vandenbergen and Aijmer 2007; Mortensen 2012).

1. Subjective use: In this case, the certainty markers indicate primarily the speaker's positioning towards his or her statement in terms of the degree of certainty. This use is demonstrated in Example 2 (E = Examiner, C = Candidate).

(2) E: and erm it reminded me of er erm an a male friend of mine who who once said to me well erm people only ask questions because they want to talk about themselves
C: mm
E: erm and I was absolutely gobsmacked by that comment because it's **certainly** not true for me

2. Intersubjective use: In this case, while also carrying subjective meaning and expressing a degree of certainty, the epistemic markers are explicitly used to negotiate the speaker's position with respect to the other interlocutor and to react to what he or she has said. Very often, the certainty markers are part of the agreement or disagreement speech acts. This use can be seen in Example 3.

(3) E: yeah okay but how far would you say? cos it it strikes me that the changing nature of the labour market you know with more women working everybody working longer hours the emphasis on money for consumer goods and so on is directly impacting on the role of the family
C: yeah
E: the traditional role
C: it **certainly** is you know I I come from a family in which my er it's a single parent my father […]

3. Other use: The markers in this category included AEMs whose function could not be clearly categorised as subjective or intersubjective. This was often due to the fact that speakers re-phrased their statement or abandoned the formulation of the statement. The example of this category can be seen in Example 4 below.

(4) C: I'm trying to tell you that Adidas has played a m-major role in our history erm
E: right
C: that's and er you erm and f= **of course** you I'm su=
E: okay
C: I'm sure that you have seen it in Olympics of <unclear> British

Following the identification of the typical contexts in which certainty AEMs occur, all of the AEMs of certainty were searched for in the two tasks separately for each group of speakers and then coded manually according to the three categories. The results of this analysis can be seen in Table 7 and Fig. 1 which show results for each of the two tasks and for each group of speakers.

As can be seen from Fig. 1, in both tasks the examiners' expressions of certainty consisted mostly of subjective statements – i.e. they used the markers of certainty to position themselves with respect to their statement. A considerably smaller propor-

Table 7 Different types of certainty expressed by examiners and candidates[a]

	CAND-INT		CAND-DISC		EX-INT		EX-DISC	
Type of certainty	Freq.	%	Freq.	%	Freq.	%	Freq.	%
Subjective	16	**39.0**	57	**54.8**	35	**68.6**	65	**71.4**
Intersubjective	18	**43.9**	33	**31.7**	10	**19.6**	9	**9.9**
Other	7	**17.1**	14	**13.5**	6	**11.8**	17	**18.7**
Total	41	**100**	104	**100**	51	**100**	91	**100**

[a]*CAND* … candidate, *EX* … examiner, *INT* … interactive task, *DISC* … discussion

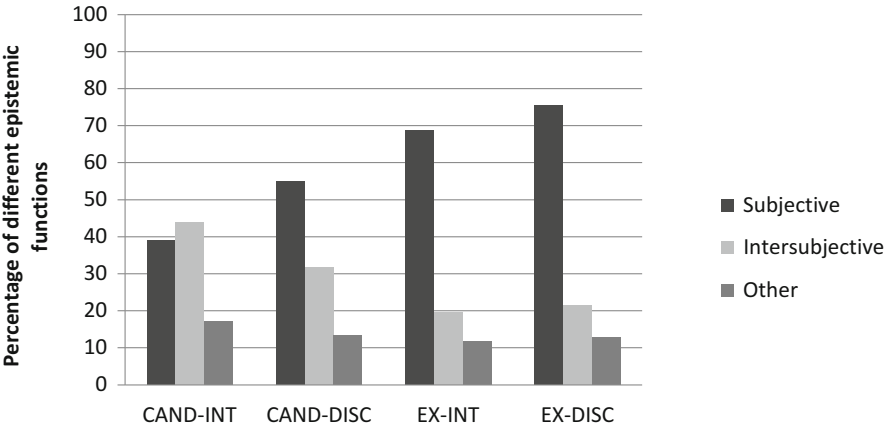


Fig. 1 Different types of certainty expressed by examiners and candidates

tion (about one fifth) of their certainty AEMs was part of the intersubjective positioning – in all instances this involved agreeing with the candidate. A similar trend was observed for the candidates in the *discussion*, in which the subjective positioning accounted for over a half of the uses of certainty AEMs and only about one third involved intersubjective positioning. The pattern, however, was very different for the candidates in the *interactive task* in which a similar proportion of AEMs of certainty was used for the subjective and intersubjective functions, with the latter accounting for over forty percent of expressions of certainty.

4 Discussion

4.1 Quantity and Range of the AEMs Used by Candidates and Examiners

In the first research question we asked whether the candidates and examiners differ in the number of AEMs produced in three different types of interaction. The results showed that whereas the candidates produced on average more AEMs in each of the

compared tasks than the examiners, the difference was statistically significant only in one of them: in the *interactive task.* This difference can be explained by the nature of the *interactive task* where the candidate takes the leading role in the interaction. This results in an increased need on the candidate's part for managing the epistemic position together with the interpersonal relationship with the other speaker (the examiner). This aspect is discussed in more detail in Sect. 4.3.

Next, the study examined the range of individual AEMs used by the two groups of speakers. When looking closely at the difference between the candidates and the examiners in the three dialogic tasks, there are both similarities and differences in the use of the individual markers. With respect to similar features, the same range of AEMs was found in the production of both groups with some AEMs appearing systematically with very high frequency (e.g. 'maybe', 'actually', 'kind of') and others occurring equally rarely in both groups (e.g. 'apparently', 'clearly', 'no doubt').

However, the results also highlighted some differences in the overall choice of AEMs, identifying several markers that were used considerably more often by only one group of speakers. The statistical test revealed that some of the individual AEMs occur significantly more often in the speech of examiners (*certainly, necessarily, perhaps, possibly, sort of* and *surely*) while others are used significantly more by the candidates (*kind of, maybe* and *of course*). A closer analysis of the frequencies of individual AEMs suggests that the examiners more systematically employed a greater range of AEMs, whereas the candidates used fewer markers with higher frequency instead. For example, the candidates used words such as 'definitely' and 'of course' to indicate strong certainty and 'maybe' to signal uncertainty. By contrast, the examiners employed a wider range of mid-range frequency markers to indicate similar degrees of certainty: These were expressions such as 'certainly', 'definitely', 'necessarily', 'obviously' and 'surely' for higher degree of certainty and 'perhaps', 'possibly' and 'maybe' for lower degree of certainty.

Previous studies attributed the limitation in the range of epistemic forms expressed by L2 users when compared to L1 speakers to lower proficiency of the L2 speakers in the target language (e.g. Bardovi-Harlig and Salsbury 2004; Fordyce 2009). However, while the L2 speakers in these studies were of relatively low L2 proficiency, the exam candidates in the present study are advanced users of English. Thus, while L2 proficiency could play a role in their choice of AEMs, other important factors should be also taken into consideration when explaining the difference between the two groups of speakers. First of all, it is important to realise that despite the fact that both the examiners and the candidates participated in the same interactions, their roles in these were different as stipulated in the task requirements (Sect. 2.2) and also confirmed in our analysis of their contributions (Sect. 3.3). It is therefore possible that the range of AEMs used by the candidates in the present study is not only related to their L2 proficiency but also to their role in the interaction which is more limited in scope than that of the examiners who employ a wider range of speech acts since in addition to acting as dialogue partners they also manage the structure of the examination (e.g. by opening and closing each speaking task) for which they are ultimately responsible.

Second, other characteristics such as age or level of education could contribute to the differences between the two groups of speakers with respect to their choice of words when expressing certainty or uncertainty. For example, while examiners are typically over 35 years old and with a university degree, one third of the candidates (45) are just 20 years old or younger. Many of the candidates are thus still completing either their upper secondary education or beginning their university degrees. As a result, their linguistic skills may still be developing, especially their skills appropriate for semi-formal or formal interaction. A higher reliance on a particular pragmatic marker is therefore not necessarily related only to L2 proficiency but may also be found in L1 speakers with similar characteristics. For example, Precht (2003) in her study observed the native speakers of British and American English using also a very limited range of epistemic markers.

4.2 Certainty and Uncertainty in the Speech of Examiners and Candidates

In order to better understand the use of the AEMs by the two groups of speakers, in the second research question we asked how the speakers expressed two broad categories of epistemic stance – certainty and uncertainty. The use of expressions signalling certainty and uncertainty was therefore examined quantitatively across the tasks. The results showed that both groups of speakers produced more AEMs indicating uncertainty than those of certainty.

When exploring the individual contexts in which uncertainty markers are used, we can see that the prevalence of these markers is largely due to politeness strategies employed by speakers where uncertainty markers serve as hedges to downplay the potentially face-threatening nature of speakers' utterances (Aijmer 2004; Precht 2003; Brown and Levinson 1987) in an interaction in which speakers exchange and discuss opinions and beliefs. This is especially true of the candidates who may perceive themselves as being in a less powerful position in the exchange than the examiner who is a native speaker of English and the person who evaluates their performance; as a result, the candidates hedge their assertions/statements more than the examiners. For example, in the *interactive task,* a task in which the candidates provided the highest number of uncertainty AEMs, many of these were found in the statements where the candidates expressed recommendations or suggestions to the examiner (according to the requirements of the task). This use of AEMs to make a tentative recommendation can be seen in Example 5. These occasions are similar to the category of 'hedged opinion' observed by Precht (2003) in her study of British and American English speakers and cannot be therefore considered as a special feature of L2 production.

(5) E: yeah but I know people bought the new Ipad and they had a lot of teething problems with it
C: yeah but **I think maybe you should actually** buy a laptop and just a basic phone basic things

4.3 The Effect of Speaker Role on the Expression of Certainty

Whereas the candidates used a higher number of uncertainty markers than the examiners and this difference was statistically significant, no such difference was found with respect to certainty markers in the dialogic tasks. In order to gain additional insight into why this is the case, the certainty markers used by the candidates and examiners in the *discussion* and *interactive task* were further analysed and compared. The two dialogic tasks were selected because they represent different interactional settings, with different expectations about the candidates' and examiners' speaking roles. The analysis identified two clear contexts for the use of certainty markers – that of positioning oneself primarily towards one's statement (subjective use) and that of positioning oneself towards the other speaker as well as towards the statement (intersubjective use) (Kärkkäinen 2006; Simon-Vandenbergen and Aijmer 2007; Mortensen 2012). The findings showed that while candidates and examiners demonstrated a similar pattern in the *discussion* with certainty AEMs appearing mostly with the subjective function, this trend was different for the candidates in the *interactive task.* In this task, the candidates produced a higher proportion of certainty markers with the intersubjective function, a pattern different from examiners in both of the tasks.

It seems that some of these stance-taking strategies can be attributed to the speaking roles of the candidates and examiners in these tasks. In the *discussion*, both the examiners and the candidates appear in relatively equal roles with the power imbalance equalised by the fact that the candidate has the knowledge of the topic – the candidate selected the topic for the *presentation* and the examiner listened to the presentation and then initiated the conversation. Thus, in the *discussion* the candidates were in the position of experts on the topic with their expertise established during the *presentation.* The following extract (Example 6) taken from the *discussion* demonstrates this role.

(6) C: er in Italy we don't have er a nuclea = we we < unclear > voted for erm for erm for for to not have nuclear power stations you in er in Great Britain what er
E: I was worried you were going to ask me that
C: <unclear>
E: <unclear> because I don't know I am sorry I ought to but I don't
C: if if I remember right you have three or four nuclear power stations in Great Britain
E: we don't have many yeah but we have some
C: yes
E: yeah
C: but you are still producing nuclear power so if in mm you e-even you cannot stop and just produ = because you don't know where to find the other energy
E: yeah yes

In this extract, the speakers discuss the topic of energy sources in the UK and Italy. We can see that a) the examiner openly declares his lack of knowledge on the subject ('because I don't know I am sorry I ought to but I don't') and b) the candidate actually demonstrates his expertise ('if I remember right you have three or four nuclear power stations in Great Britain'). On a general level, it appears that this knowledge distribution affects the nature of the interaction in the *discussion* with the examiners asking more open-ended questions (in many of which the candidates are acknowledged as the source of information) and the candidates more readily framing their statements without hedging. In these statements the candidates most often rely on two sources of knowledge: a) certainty based on personal conviction or preference and b) certainty stemming from candidates' experience in the area under discussion. The former type is demonstrated in Example 7 and the latter in Example 8 (in this example, the candidate is a teacher discussing the implementation of methods that support learner autonomy in the teaching institution where he works).

(7) E: you sh= you should go to to a big city like New York or London <unclear>
C: well I love New York I love New York but **definitely** I don't like big city except for New York New York is a very nice

(8) E: so towards autonomy and independence and I guess I want to ask you a very simple questions
C: yes
E: how effective has it been changing the paradigm?
C: well the situation is that there are there are a lot of aspects there there are a lot of things that we should consider because **obviously** we may find a lot of resistance to change
E: mm
C: we may see well not only in the students but as professors as teachers <unclear> **definitely** there's a great side that we might be very afraid of of changing our methods and and beliefs because in certain form beliefs the students' and teachers' beliefs <unclear>
E: so so are are you saying that you haven't actually put this paradigm in […]

While the subjective function was the dominant type of context for the use of certainty AEMs by the candidates in the *discussion,* in the *interactive task* the intersubjective function of certainty markers was equally frequent. This appears to be caused by the different role assumed by the candidate in this task, i.e. that of an information-seeker rather than an expert. As a result, in this task, the proportion of certainty markers accompanying subjective statements was much lower than in the *discussion* and nearly half of the expressions of certainty appeared in the context where candidates agreed with the examiner (the expert and primary knower) as demonstrated by Example 3 in Sect. 3.3. Thus, rather than expressing an independent opinion, in *the interactive task* the candidates stressed shared experience or shared opinion.

It is interesting to note that while the number of certainty AEMs used by the candidates in the *interactive task* was the lowest from among the three tasks, the number of uncertainty markers was at the same time the highest in this task (see Sect. 3.2). This twofold evidence suggests a considerably higher degree of tentativeness expressed by the candidates in the *interactive task* than in the other two dialogic tasks. This distinctive variation occurred and was observed despite the fact that the interaction takes place between the same two speakers across all tasks. This provides a strong evidence of the ability of the candidates, L2 speakers of English, to adjust their speaking style with some flexibility according to the speaker role and identity they adopt in each task.

These findings show that advanced L2 speakers' (as well as L1 speakers') stance-taking patterns are strongly affected by the context in which they produce language. These results are in line with other studies that reported variation in L1 and L2 speakers' choice of pragmatic markers which was linked to the speaker role in the discourse and the purpose of the conversation (Fuller 2003; Liao 2009). Similar to Fuller (2003), this study also found that the analysed markers of certainty did not appear in different speaking contexts (i.e. in the *discussion* and *interactive* task) with noticeably different frequency; rather they were used in these contexts by two groups of speakers with very similar frequency but with different functions.

These findings provide an interesting contribution to the discussion about what motivates the linguistic choices of L2 speakers. Very often, the differences found between a group of L1 and L2 speakers are attributed to L2 speakers' proficiency in the target language without taking into consideration various other factors that may affect speakers' linguistic choices. For example, in the present study, the exam candidates used some of the markers (such as 'maybe' or 'of course') significantly more often than the examiners. While this could be taken as a sign of lexical limitations on the part of L2 speakers, it was argued that the interactional setting in some cases required them to express more uncertainty or to indicate their agreement with the examiners more often (e.g. as part of politeness strategy). Also, despite the fact that some words in L2 speakers' production were more dominant than others, we should not overlook the fact that the L2 speakers did not solely rely on a single AEM of uncertainty or certainty, but used a range of synonymous expressions to convey these meanings.

With respect to the functional range of the epistemic markers used by the candidates (L2 speakers), the case study suggested that the candidates were able to use certainty AEMs for different functions. This is somewhat in contrast with the findings of the studies on L2 pragmatic development and use (e.g. Baumgarten and House 2007, 2010; Aijmer 2004; Romero-Trillo 2002) that identified limitations in the functional range of the pragmatic markers employed by L2 speakers (and attributed this to L2 proficiency). The present study thus pointed to the fact that the picture of epistemic marker use in (advanced) L2 is more complex than previously suggested. However, a much closer qualitative analysis of the AEMs used by both the candidates and examiners is needed to fully understand the nuances of the AEMs use by both groups of speakers.

5 Conclusion

This study sought to demonstrate the effect of different speaker roles and identity on the speakers' linguistic choices when expressing their position (stance) in interaction. We demonstrated that candidates (advanced L2 speakers of English) in an exam differed in their positioning according to the type of speaking task and their role in the interaction which was affected by factors such as familiarity or expertise with the topic discussed and the type of interaction (e.g. discussion of a topic or providing advice to the other speaker). These findings show that when studying L2 spoken production it is important to go beyond characterising the interlocutors as 'native' or 'non-native' speakers of a language. Whereas the fact of being a 'native user' or a 'non-native user' can indeed be part of the speaker role and speaker identity, there are other equally important factors that arise from the context of the exchange. As demonstrated, it would be too simplistic to say that the differences between the candidates and examiners in these interactions merely reflect the difference between L1 and L2 use of English. This study thus pointed out the complexity of factors that affect linguistic choices of speakers (whether of L1 or L2), which include the characteristics of the task or interactional setting as well as the role-related expectations and communicative aims.

Transcription conventions

Symbol	**Meaning**
E:	examiner
C:	candidate
=	marks unfinished words (e.g. 'I'm su=')
-	marks repeated sound (e.g. 'e-even')
<unclear>	marks speech that was unclear and could not be transcribed

Acknowledgments The research presented in this chapter was supported by the ESRC Centre for Corpus Approaches to Social Science, ESRC grant reference ES/K002155/1.

References

Aijmer, K. (2002). *English discourse particles: Evidence from a corpus*. Amsterdam: John Benjamins.

Aijmer, K. (2004). Pragmatic markers in spoken interlanguage. *Nordic Journal of English Studies, 3*(1), 173–190.

Aijmer, K. (2011). Well I'm not sure I think… The use of well by non-native speakers. *International Journal of Corpus Linguistics, 16*(2), 231–254.

Bardovi-Harlig, K., & Salsbury, T. (2004). The organization of turns in the disagreements of L2 learners: A longitudinal perspective. In D. Boxer & A. D. Cohen (Eds.), *Studying speaking to inform second language learning* (pp. 199–227). Clevedon: Multilingual Matters.

Baumgarten, N., & House, J. (2007). Speaker stances in native and non-native English conversation. In J. D. ten Thije & L. Zeevaert (Eds.), *Receptive multilingualism* (pp. 195–216). Amsterdam: John Benjamins.

Baumgarten, N., & House, J. (2010). 'I think' and 'I don't know' in English as lingua franca and native English discourse. *Journal of Pragmatics, 42*(5), 1184–1200.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English*. London/New York: Longman.

Brezina, V. (2012). *Epistemic markers in university advisory sessions: Towards a local grammar of epistemicity*. Unpublished PhD dissertation, University of Auckland, Auckland, New Zealand.

Brezina, V., & Meyerhoff, M. (2014). Significant or random?: A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics, 19*(1), 1–28.

Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language*. Cambridge: Cambridge University Press.

Buysse, L. (2012). 'So' as a multifunctional discourse marker in native and learner speech. *Journal of Pragmatics, 44*(13), 1764–1782.

Callies, M. (2013). Advancing the research agenda of interlanguage pragmatics: The role of learner corpora. In J. Romero-Trillo (Ed.), *Yearbook of corpus linguistics and pragmatics 2013* (pp. 9–36). Dordrecht: Springer.

Coates, J. (1987). Epistemic modality and spoken discourse. *Transactions of the Philological Society, 85*(1), 110–131.

Coates, J. (1990). Modal meaning: The semantic–pragmatic interface. *Journal of Semantics, 7*(1), 53–63.

Fordyce, K. (2009). A comparative study of learner corpora of spoken and written discursive language: Focusing on the use of epistemic forms by Japanese EFL learners. *Hiroshima Studies in Language and Language Education, 12*, 135–150.

Fordyce, K. (2014). The differential effects of explicit and implicit instruction on EFL learners' use of epistemic stance. *Applied Linguistics, 35*(1), 6–28.

Fuller, J. M. (2003). The influence of speaker roles on discourse marker use. *Journal of Pragmatics, 35*(1), 23–45.

Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics, 28*(3), 410–439.

Holmes, J. (1988). Doubt and certainty in ESL textbooks. *Applied Linguistics, 9*(1), 21–44.

Holmes, J. (1990). Hedges and boosters in women's and men's speech. *Language & Communication, 10*(3), 185–205.

Hunston, S. (2007). Using a corpus to investigate stance quantitatively and qualitatively. In R. Englebretson (Ed.), *Stancetaking in discourse: Subjectivity, evaluation, interaction* (pp. 27–48). Amsterdam: John Benjamins.

Hunston, S., & Thompson, G. (2000). *Evaluation in text: Authorial stance and the construction of discourse*. Oxford: Oxford University Press.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing, 6*(2), 183–205.

Kärkkäinen, E. (2003). *Epistemic stance in English conversation: A description of its interactional functions, with a focus on I think*. Amsterdam: John Benjamins.

Kärkkäinen, E. (2006). Stance taking in conversation: From subjectivity to intersubjectivity. *Text & Talk, 26*(6), 699–731.

Liao, S. (2009). Variation in the use of discourse markers by Chinese teaching assistants in the US. *Journal of Pragmatics, 41*(7), 1313–1328.

McEnery, T., & Kifle, N. A. (2002). Epistemic modality in argumentative essays of second-language writers. In J. Flowerdew (Ed.), *Academic discourse* (pp. 182–195). Harlow: Longman.

Mortensen, J. (2012). Subjectivity and intersubjectivity as aspects of epistemic stance marking. In N. Baumgarten, I. D. Bois, & J. House (Eds.), *Subjectivity in language and in discourse* (pp. 229–246). Bingley: Emerald.

Mukherjee, J. (2009). The grammar of conversation in advanced spoken learner English. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 203–230). Amsterdam/Philadelphia: John Benjamins.

Müller, S. (2005). *Discourse markers in native and non-native English discourse*. Amsterdam: John Benjamins.

Precht, K. (2003). Stance moods in spoken English: Evidentiality and affect in British and American conversation. *TEXT, 23*(2), 239–258.

Romero-Trillo, J. (2002). The pragmatic fossilization of discourse markers in non-native speakers of English. *Journal of Pragmatics, 34*(6), 769–784.

Rühlemann, C. (2011). Corpus-based pragmatics II: Quantitative studies. In W. Bublitz & N. R. Norrick (Eds.), *Foundations of pragmatics* (pp. 629–656). Berlin: Mouton DeGruyter.

Simon-Vandenbergen, A.-M., & Aijmer, K. (2007). *The semantic field of modal certainty: A corpus-based study of English adverbs*. Berlin: Walter de Gruyter.

Traugott, E. C., & Dasher, R. B. (2002). *Regularity in semantic change*. Cambridge: Cambridge University Press.

Trinity College London. (2009). Exam information: Graded examinations in spoken English (GESE).

Part II
Current Approaches to Translation Studies

Source Language Interference in English-to-Chinese Translation

Richard Xiao

Abstract Translational language as a "third code" has been found to differ from both source and target languages. Recent corpus-based studies have proposed a number of translation universal (TU) hypotheses including, for example, simplification, explicitation and normalisation. This article investigates the "source language shining through" hypothesis put forward by Teich (2003: 207) by exploring source language interference in translated texts, at both lexical and grammatical levels, in English-to-Chinese translation on the basis of comparable corpora and parallel corpora of the two languages. The evidence from the two genetically distant languages is of critical importance in generalising the source language interference as a potential translation universal.

Keywords Translation universal • Source language interference • Translational Chinese • English-to-Chinese translation

1 Introduction

Pragmatics, in its broad sense, is related to and depends on choices of linguistic features in a text to achieve meaning in discourse. Translational language has been shown to exhibit a variety of linguistic properties which indicate that it is a "third code" different from both source and target languages (Frawley 1984). As Hansen and Teich (2001: 44) observe, "It is commonly assumed in translation studies that translations are specific kinds of texts that are not only different from their original source language (SL) texts, but also from comparable original texts in the same language as the target language (TL)". Recent studies of linguistic features at lexical, syntactic and discourse levels, which are mainly based on translated English, have motivated the formulation of TU hypotheses such as simplification, explicitation, normalisation, sanitisation, under-representation, levelling out/convergence, and source language shining through.

R. Xiao (✉)
Department of Linguistics and English Language, Lancaster University, Lancaster, UK
e-mail: r.xiao@lancaster.ac.uk

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_7

Simplification refers to the "tendency to simplify the language used in translation" (Baker 1996: 181–182), and as a result translated language is expected to be simpler than non-translated target language lexically, syntactically and/or stylistically (*cf.* also Blum-Kulka and Levenston 1983; Laviosa-Braithwaite 1997; Laviosa 1998). Explicitation is manifested by the tendency in translations to "spell things out rather than leave them implicit" (Baker 1996: 180), for example, through more frequent use of connectives and increased cohesion (*cf.* also Pym 2005; Chen 2006; He 2003; Dai and Xiao 2010; Xiao 2010). Normalisation means that translational language displays a "tendency to exaggerate features of the target language and to conform to its typical patterns" so that translated texts tend to avoid creative language use and thus appear more "normal" than non-translated texts (Baker 1996: 183).

The TU hypothesis of sanitisation suggests that translated texts, with lost or reduced connotational and hidden meaning, are "somewhat 'sanitised' versions of the original" (Kenny 1998: 515). Under-representation, which is also known as the "unique items hypothesis", means that the linguistic features that are unique in the target language but do not exist or are rarely used in the source language may be under-represented in translations in comparison with comparable non-translated texts in the target language (Mauranen 2007: 41–42; Xiao 2012). Levelling out refers to "the tendency of translated text to gravitate towards the centre of a continuum" (Baker 1996: 184), which Laviosa (2002: 72) calls "convergence", i.e. the "relatively higher level of homogeneity of translated texts with regard to their own scores on given measures of universal features."

Another common feature of translations, which is to be focused upon in the present study, is the "source language shining through" hypothesis put forward by Teich (2003: 207), which states that "[in] a translation into a given target language (TL), the translation may be oriented more towards the source language (SL), *i.e.* the SL shines through." For example, Teich (2003: 207) finds that in both English-to-German and German-to-English translations, both target languages exhibit "a mixture of TL normalisation and SL shining through".

Hopkinson (2007: 13) also notes, in translation from Czech (L1) into English (L2), that "[the] product of L1 – L2 translation will thus usually contain examples of what is colloquially termed 'translationese', *i.e.* a non-standard version of the target language that is to a greater or lesser extent affected by the source language." His analysis focuses on three key factors in interference: poor reference materials, translators' generalisations of false hypotheses, and systemic-structural differences between Czech and English. The examples analysed cover interference in lexis, word-formation, grammar and syntax. All of his analysis is within the framework of the interlanguage model, but does not pay attention to the interference from the source to target language in translation.

Indeed, source language interference is prevalent in translation. As Toury (1979: 226) notes, "virtually no translation is completely devoid of formal equivalents, *i.e.*, of manifestations of interlanguage." According to Toury's (1995: 275–276) "law of interference":

> In translation, phenomena pertaining to the make-up of the source text tend to be transferred to the target text. […] The more the make-up of a text is taken as a factor in the formulation of its translation, the more the target text can be expected to show traces of interference.

Toury's law gives a vivid description of the feature of translations and casts new light on translation studies. However, Toury does not explicitly deal with his law of interference (*cf.* Teich 2003). Teich suggests that one of the factors that makes translations different from comparable native texts in the target language is that the source language—to a greater or lesser extent—"shines through" in translation. In other words, the language used in translation is not as idiomatic and prototypical as it is in texts originally composed in the same language, for the translated language contains deviations from the general TL patterns, with SL being their source of such deviations.

Nevertheless, the TU hypothesis of source language interference has not attracted much attention in translation studies, possibly because TU research has until recently focused on, or indeed confined to translation involving closely related European languages, which may display less marked contrasts than typologically different languages. As English, German and Czech all belong to the Indo-European language family and are thus related languages, the studies reviewed above arguably provide only limited evidence for generalising source language shining through as a "universal" feature of translation.

This article seeks to approach the phenomenon of source language interference on the basis of evidence from two genetically distant languages, namely English and Chinese, with the aim of answering the following two questions:

1. Is the phenomenon of source language interference also observable in translational Chinese?
2. If so, to what extent does source language interference occur in English-to-Chinese translation?

In addressing these research questions, the present study investigates source language interference in translated texts, at both lexical and grammatical levels, in English-to-Chinese translation on the basis of comparable corpora and parallel corpora of the two languages. The evidence from the two genetically distant languages is of critical importance in generalising source language interference as a potential translation universal.

Following this introduction, the chapter first introduces the research method and corpora used in this study (Sect. 2). Sections 3 and 4 are respectively concerned with contrastive analyses of a range of linguistic features at lexical and grammatical levels, in translational and native Chinese, which demonstrate evidence of source language interference in English-to-Chinese translation. A case study of passive constructions is also undertaken on the basis of parallel corpus data in an attempt to quantify the extent of source language interference. Section 5 concludes the study by summarising the major research findings.

2 Research Method and Data

In order to address the research questions set in Sect. 1 above, the present study will take a composite approach that integrates monolingual comparable corpus analysis and parallel corpus analysis as advocated in McEnery and Xiao (2002). The monolingual comparable corpus approach compares matching corpora of translated Chinese and native Chinese in an attempt to uncover salient features of translations, while the parallel corpus approach compares source and target languages on the basis of English-to-Chinese parallel corpora to establish the extent of source language interference, *i.e.* the extent to which the features of translated texts are transferred from the source language. Four corpora are used in this study, including two comparable corpora and two parallel corpora, which are introduced as follows.

The monolingual comparable corpora are the Lancaster Corpus of Mandarin Chinese (LCMC) and the ZJU Corpus of Translational Chinese (ZCTC), which represent native and translational Chinese respectively. LCMC is designed as a Chinese match for the FLOB corpus of British English (Hundt et al. 1998) for use in cross-linguistic contrast of English and Chinese (McEnery and Xiao 2004), while ZCTC is created as a translational counterpart of LCMC with the explicit aim of studying features of translational Chinese (Xiao et al. 2010).

These two Chinese corpora are each composed of ca. one million words in five hundred approximately 2,000-word text samples which are taken proportionally from 15 text categories published in China in the 1990s as shown in Table 1. As can be seen, the two corpora are comparable in terms of both overall size and proportions for different genres. English is the source language of about 99 % of text samples

Table 1 LCMC and ZCTC corpus design

Type	Register	Code	Genre	Samples	Proportion (%)
Non-literary	Press	A	Press reportage	44	8.8
		B	Press editorials	27	5.4
		C	Press reviews	17	3.4
	General prose	D	Religious writing	17	3.4
		E	Instructional writing	38	7.6
		F	Popular lore	44	8.8
		G	Biographies and essays	77	15.4
		H	Reports & official documents	30	6
	Academic prose	J	Academic writing	80	16
Literary	Fiction	K	General fiction	29	5.8
		L	Mystery & detective fiction	24	4.8
		M	Science fiction	6	1.2
		N	Adventure fiction	29	5.8
		P	Romantic fiction	29	5.8
		R	Humour	9	1.8
Total				**500**	**100**

included in the ZCTC corpus, which also includes a small number of texts translated from other languages to mirror the reality of the world of translations in China.

Of the 15 genres covered in the corpora, text categories A–C are press material; D–H represent general prose; H is academic writing while K–R represent various types of fiction. These registers can be further merged into two broad text categories, namely, non-literary (A–J) versus literary (K–R). The contrastive analyses to be presented in the following sections will be based on more fine-grained genres or broader text categories as appropriate.

In addition to these comparable corpora of Chinese, two English-Chinese parallel corpora are also used in a case study that attempts to determine the extent of source language transfer of passive constructions in English-to-Chinese translation. They are the Babel English-Chinese Parallel Corpus (Babel) and the General Chinese-English Parallel Corpus (GCEPC), which are both annotated with part-of-speech information for English and Chinese texts and aligned at the sentence level.

The Babel corpus consists of 327 English articles and their translations in Mandarin Chinese. Of these 115 texts were collected from the bilingual magazine *World of English* between October 2000 and February 2001 while the remaining 212 texts were collected from the *Time* magazine from September 2000 to January 2001. The corpus contains a total of 253,633 English words in the source texts and 287,462 Chinese words in the translations (see Xiao 2005). As this corpus comprises mixed genres which are not encoded in the corpus, it can only be used to investigate translation patterns in English-to-Chinese translation but cannot be used to explore genre variation.

The GCEPC corpus created by Beijing Foreign Studies University allows for such variation study. It is the largest existing parallel corpus of English and Chinese, containing approximately 20 million English words and Chinese characters. This is a bidirectional parallel corpus which comprises four subcorpora, namely Chinese-to-English Literature, Chinese-to-English Non-literature, English-to-Chinese Literature, and English-to-Chinese Non-literature (Wang 2004; Wang and Qin 2010). As we are interested in how Chinese translations are affected by English source texts, only the two English-to-Chinese subcorpora will be used, amounting to 12 million words/characters, 60 % of which are for English-Chinese Literature, and 40 % for English-Chinese Non-literature (*cf.* Wang 2004: 40).

Having introduced the research method and corpora used, we will move on to explore, in the sections that follow, the features of translational Chinese that are of relevance to the investigation of source language interference. We will first consider linguistic features at lexical level.

3 Lexical Features

This section compares four lexical features of translational and native Chinese as represented in the ZCTC and LCMC corpora, namely mean word length, prefixes and suffixes, pronouns, and word clusters. Mean word length is considered because it is often a lexical indicator of text readability and thus it is related to the

simplification hypothesis. Affixes are included because Chinese is a non-inflectional language and is therefore expected to be less productive in its use of prefixes and suffixes than English, the source language of the translated texts in the translational Chinese corpus ZCTC. As pronouns are a linguistic device used for achieving cohesion, their frequency of use reflects the extent of explicitation. Finally a frequent use of word clusters is sometimes associated with the translational tendency to strive for fluency at discourse level (e.g. Baker 2004).

3.1 Mean Word Length

Mean word length is a basic statistic in text analysis which is readily available in a wordlist generated using WordSmith. It has also been used in translation studies to compare native and translated Chinese texts. For example, Wang and Qin (2010: 168) observe that the mean word length is marginally greater in translated Chinese than in native Chinese, with a higher proportion of monosyllabic words in native Chinese and a higher proportion of disyllabic words in translated texts. This observation is supported by our data.

As illustrated in Fig. 1, the mean word length in translated Chinese is slightly greater than in native Chinese (1.59 vs. 1.57, a statistically insignificant difference), which is true in both non-literary (1.63 vs. 1.61) and literary (1.47 vs. 1.42) texts, with an even more marked contrast between native and translated Chinese in literary texts possibly because these genres contain more proper nouns such as transliterated personal names and place names, which are longer than similar words in native Chinese.

Table 2 shows the distribution of words of various lengths in LCMC and ZCTC across two broad categories, namely literary and non-literary texts, and their mean

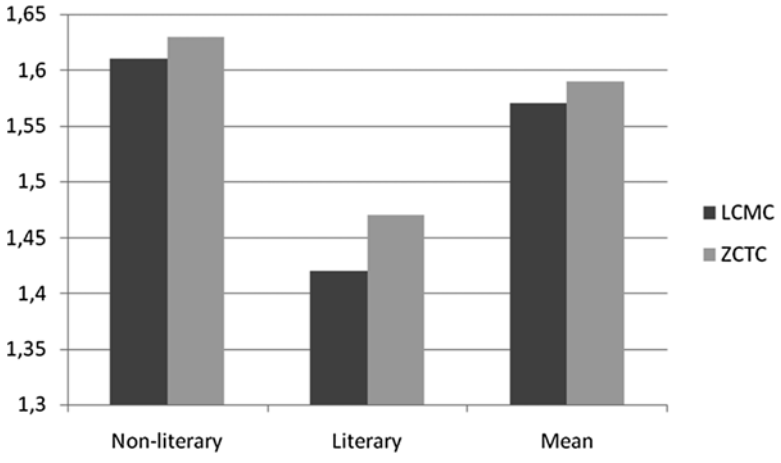


Fig. 1 Mean word length in LCMC and ZCTC

Table 2 Proportions of words of various lengths in LCMC and ZCTC

	Non-literary texts		Literary texts		Mean score	
Length	LCMC	ZCTC	LCMC	ZCTC	LCMC	ZCTC
1 syllable	46.76	46.06	62.45	58.61	50.68	49.14
2 syllables	47.65	48.04	34.02	37.5	44.25	45.45
3 syllables	3.60	3.90	2.54	2.73	3.34	3.62
4 syllables	1.59	1.43	0.91	1.05	1.42	1.34
5 syllables	0.31	0.37	0.06	0.09	0.25	0.30
6+ syllables	0.09	0.19	0.01	0.02	0.07	0.15

scores. As the subcorpora are of different sizes, relative frequencies in the form of percentages will be compared. As can be seen, no matter whether native and translational corpora are taken as a whole, or the two broad text categories are considered separately, monosyllabic and quadrisyllabic words are generally more common (except for quadrisyllabic words in literary texts) in native Chinese. A key word analysis suggests that monosyllabic words are more common in LCMC because native Chinese texts make more frequent use of Chinese surnames, which are typically monosyllabic, as well as high frequency monosyllabic words such as 元 *yuan* 'Chinese currency unit' and 党 *dang* '(Communist) Party', though many monosyllabic function words are more frequently used in the translational corpus, *e.g.* the structural auxiliary 的 *de* and personal pronouns 你 *ni* 'you', 我 *wo* 'I, me' and 她 *ta* 'she, her', which are all negative keywords in LCMC in relation to ZCTC. Quadrisyllabic words are more common in LCMC because non-literary texts in native Chinese tend to make significantly more frequent use of idioms (see Xiao and Dai 2010), which are typically four-character words. In contrast, disyllabic and trisyllabic words are more frequent in translated texts. The translational tendency for long words is particularly marked in words containing five or more syllables, though these words *per se* are infrequent in both native and translated Chinese.

A key part-of-speech analysis shows that transliterated foreign personal names and place names, which are typically much longer than Chinese names, are on the key part-of-speech list of ZCTC in relation to LCMC. While the higher proportion of monosyllabic words in native Chinese makes the contrast in the mean word lengths in native and translational Chinese less marked, the inevitably more frequent but less varied use of transliterated foreign names in the translation process still results in a marginally greater mean word length in translated Chinese texts. In this sense, the general tendency in translational Chinese to use slightly longer words can be taken as evidence of source language interference.

3.2 Prefixes and Suffixes

A key part-of-speech analysis of the translational Chinese corpus in relation to the native Chinese corpus suggests that the suffix tag (K) is a key part-of-speech in the translational corpus. As can be seen in Fig. 2, which compares the frequencies of

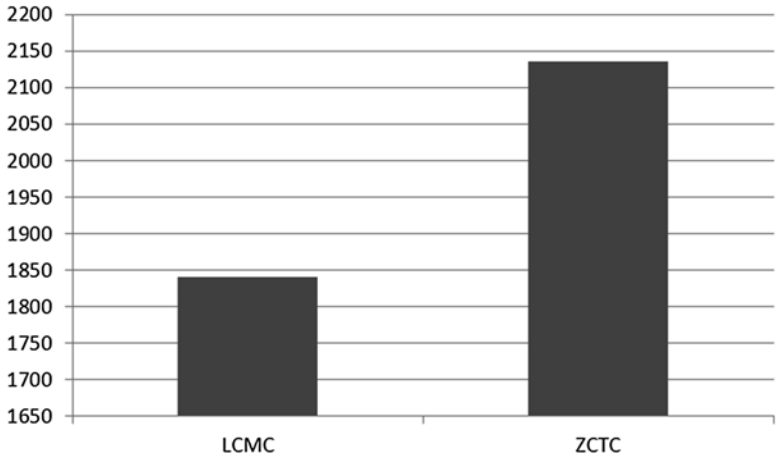


Fig. 2 Prefixes and suffixes in LCMC and ZCTC

prefixes and suffixes in LCMC and ZCTC, there is a marked contrast between the two corpora in their frequency of affixes. The log-likelihood test (LL) indicates that the difference is highly significant (LL = 23.01, $p < 0.001$).

Because Chinese is not a morphologically inflectional language, the more frequent use of prefixes and suffixes in translated Chinese texts is arguably a result of source language interference. This finding is in accord with Wang and Qin's (2010: 175) observation that some morphemes in translated texts, *e.g.* suffixes such as *-xing* (a nominal suffix indicating property, similar to *-ness* / *-ity* in English), are so productive in Chinese translations because of the influence of English source texts that there is a tendency for them to replace the original expressions in native Chinese.

3.3 Pronouns

Among all parts-of-speech, pronouns are the one that displays the most marked contrast between LCMC and ZCTC, which contain 49,582 and 70,401 instances in the two corpora (LL = 3,707.69, $p < 0.001$). As pronouns have the function of making discourse more cohesive (*cf.* Xiao 2012), translational language is hypothesised to make more frequent use of pronouns. This section seeks to test this hypothesis on the basis of LCMC and ZCTC by exploring the overall distribution of pronouns in the two corpora.

Figure 3 shows the overall distribution of pronouns in the two corpora. As can be seen, pronouns are distributed in native and translated Chinese in a similar pattern, with the most frequent use in fiction, followed by general prose and news, and the least frequent use in academic prose. On the other hand, translated Chinese makes more frequent use of pronouns no matter whether the two corpora are taken as a whole

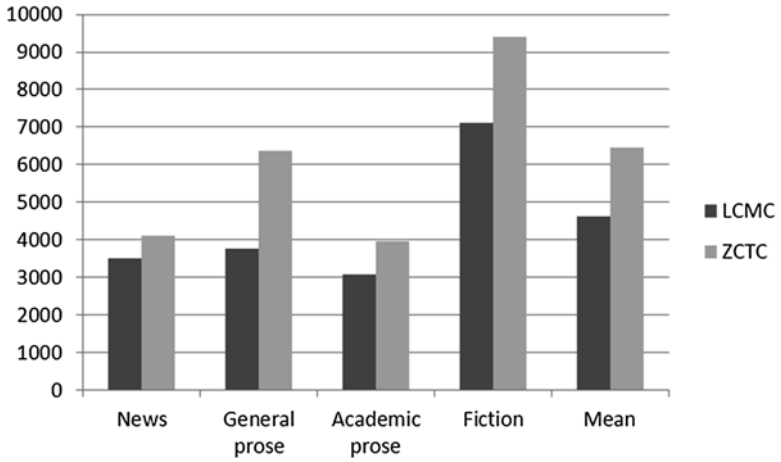


Fig. 3 Pronouns in LCMC and ZCTC

or individual registers are considered separately. All differences are statistically significant (at $p<0.001$) according to the results of log-likelihood tests.

The significantly more frequent use of pronouns (especially personal and demonstrative pronouns) can be taken as an indicator of translational explicitation (*cf.* Xiao 2012). The relatively low frequency of pronouns in LCMC in comparison with ZCTC can also be regarded as a reflection of source language interference. This is because in native Chinese, unlike in English, the grammatical subject can be dropped because of its connective discourse function, whereas the subject in the English source text is likely to be transferred to the translated text. This point is well illustrated in example (1a), which is excerpted from *A Madman's Diary* by the renowned Chinese writer Lu Xun.

(1a) **我**看不见他 ， 已经三十多年了 ； 今天**[我]**见了 ， 精神分外爽快。**[我]**才知道以前的三十多年 , **[我]**全是发昏 , 然而**[我]**须十分小心。
Wo kanbujian ta, yijing sanshi duo nian le; jintian **[wo]** jian le, jingshen fenwai shuangkuai. **[Wo]** cai zhidao yiqian de sanshi duo nian, **[wo]** quan shi fahun, ran'er **[wo]** xu shifen xiaoxin.

(1b) **I** have not seen it for over thirty years, so today when **I** saw it **I** felt in unusually high spirits. **I** begin to realize that during the past thirty odd years **I** have been in the dark; but now **I** must be extremely careful.

In example (1a), which is originally written in Chinese, the subject pronoun 我 *wo* 'I' is dropped after its occurrence in the first sentence. Although the passage comprises more than one sentence, the subject pronoun in the first sentence functions to glue the ensuing discourse in the excerpt together. Because of the cohesive function of pronouns in Chinese, a competent Chinese speaker would hardly have any difficulty in understanding the passage. However, if the same message is translated into Chinese from English (1b), the translator is very likely, under the

influence of the English source text, to include all of the dropped subjects as highlighted and included in the brackets in (1a). This is because English and Chinese have different conventions of using pronouns: English tends to repeat personal pronouns, which is dispreferred in Chinese so that where a pronoun is repeated in a text in English, Chinese either drops the pronoun or repeats a noun instead (*cf.* Liu 1991: 371).

3.4 Word Clusters

Word clusters are fixed and semi-fixed formulaic expressions based on collocations, which are also known as 'lexical bundles', 'multiword units', 'prefabs', and '*n*-grams' and so on. Scott (2009: 286) observes that "all words have a tendency to cluster together with some others". Word clusters are purely structurally defined on the basis of co-occurrences with no regard to their semantic contents. They can be computed automatically using corpus exploration tools such as WordSmith (Scott 2009). Generally speaking, the frequency of word clusters tends to drop sharply as their length grows. For example, the frequency of 4-word clusters is significantly lower than that of 3-word clusters, which are in turn substantially less frequent than 2-word clusters. The statistical significance of word clusters is usually measured by their recurring rate, *e.g.* 5 or 10 occurrences in a million words. Another useful parameter in computing word clusters is their coverage rate, which measures how widespread a word cluster occurs in a given corpus. It is expressed as a percentage of the number of text samples containing a particular word cluster in the total number of text samples in that corpus.

In translation studies, Baker (2004) and Nevalainen (2005, cited in Mauranen 2007) both find that recurring word clusters are more commonly used in translations in comparison with non-translated texts. This finding echoes Baroni and Bernardini's (2003: 379) observations based on their investigation of collocations in translated and native texts, which even differentiate between two types of repetition patterns:

> […] translated language is repetitive, possibly more repetitive than original language. Yet the two differ in what they tend to repeat: translations show a tendency to repeat structural patterns and strongly topic-dependent sequences, whereas originals show a higher incidence of topic-independent sequences, *i.e.* the more usual lexicalised collocations in the language.

Xiao (2011) notes that this finding is also applicable in translational Chinese, which demonstrates that word clusters composed of 2-to-6 words are significantly more frequent in translational Chinese than in native Chinese. The higher use of word clusters in the translational corpus is also evidenced by a keyword cluster analysis. The more frequent use of word clusters in translational Chinese is probably a result of the translation process in which "translators are likely to opt for safe, typical patterns of the target language and shy away from creative or playful uses", and consequently, translators tend to be make heavy use of "pre-packed, recurring stretches of language" (Baker 2007: 14). However, an equally plausible

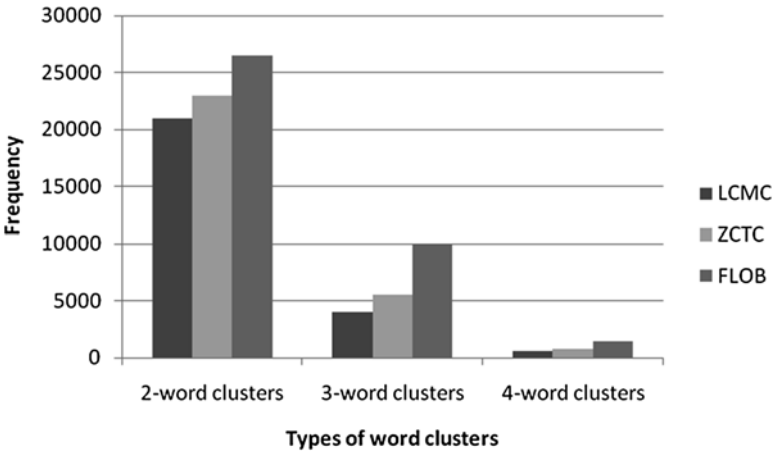


Fig. 4 Word clusters in Chinese and English

alternative explanation for the more frequent use of word clusters in translational Chinese, in our view, is that translations are under the influence of the English source language.

As can be seen in Fig. 4, which compares the use of 2–4-word clusters (clusters of more than four words are infrequent in the million-word corpora and thus excluded in the graph) in the three comparable corpora, all of the three types of word clusters are most frequent in FLOB and least frequent in LCMC, with ZCTC between the two.

In addition to their significantly higher frequencies in translational Chinese, word clusters demonstrate two other interesting characteristics. On the one hand, high-frequency word clusters (defined here as those accounting for at least 0.01 % of the respective corpus) are more common in Chinese translations. As can be seen in Fig. 5, the number of high-frequency word clusters in ZCTC (a total of 413, including 403 2-word clusters and ten 3-word clusters) is greater than that in LCMC (a total of 291, including 287 2-word clusters and four 3-word clusters), which is a statistically significant difference ($LL = 21.96$, $p < 0.001$). Given that translated Chinese tends to use high-frequency words (Xiao 2010), it is hardly surprising to find a more common use of high-frequency word clusters in ZCTC. While word clusters in different languages cannot be directly compared against each other, it is also of interest to note in Fig. 5 that the use of high-frequency word clusters in the ZCTC corpus of translational Chinese is more similar to the English corpus FLOB, which yields 522 instances of high frequency clusters (498 2-word clusters and 24 3-word clusters).

On the other hand, word clusters have a much wider coverage in translated Chinese in comparison with native Chinese, which is possibly a result of the influence of English (see Figs. 6 and 7). As can be seen in the figures, because of the low overall frequencies of 2-word clusters with a minimum coverage rate of 50 % (18,

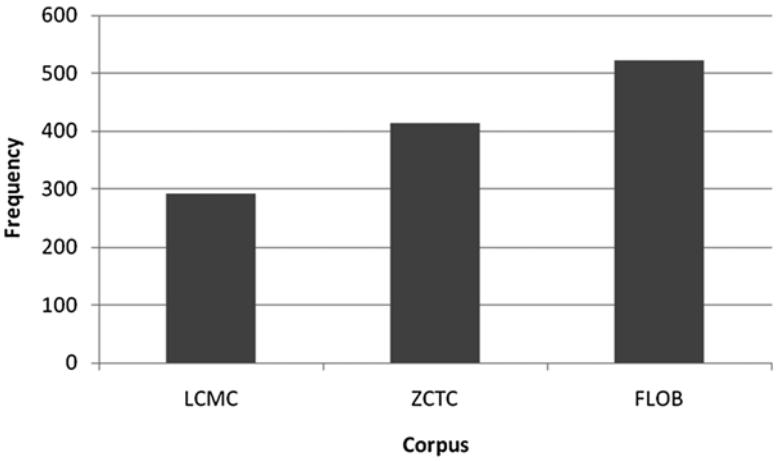


Fig. 5 High-frequency word clusters in English and Chinese

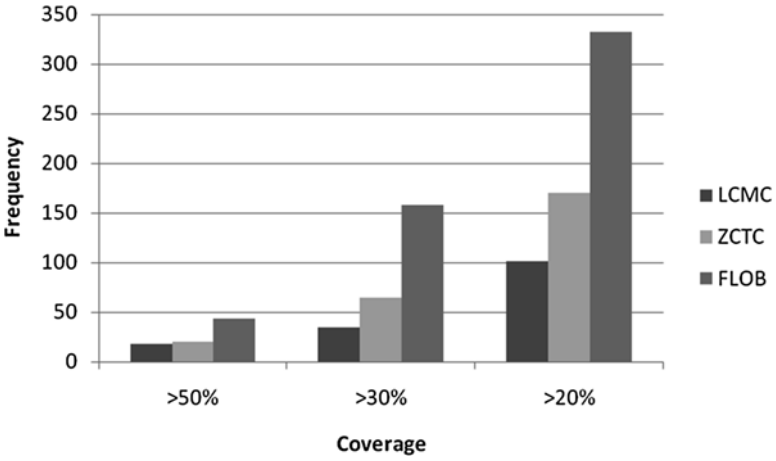


Fig. 6 Coverage of 2-word clusters in English and Chinese

20 and 44 instances in LCMC, ZCTC and FLOB respectively) and 3-word clusters with a minimum coverage rate of 20 % (zero, four and 13 instances respectively), their frequencies are quite similar in the three corpora. However, there is a marked contrast in the frequencies of 2-word clusters with a minimum coverage rate of 30 % (35, 65 and 158 instances respectively) and 3-word clusters with a minimum coverage rate of 10 % (eight, 23 and 90 instances respectively) in the three corpora. This contrast displays an accelerating tendency as the coverage rate drops: there are

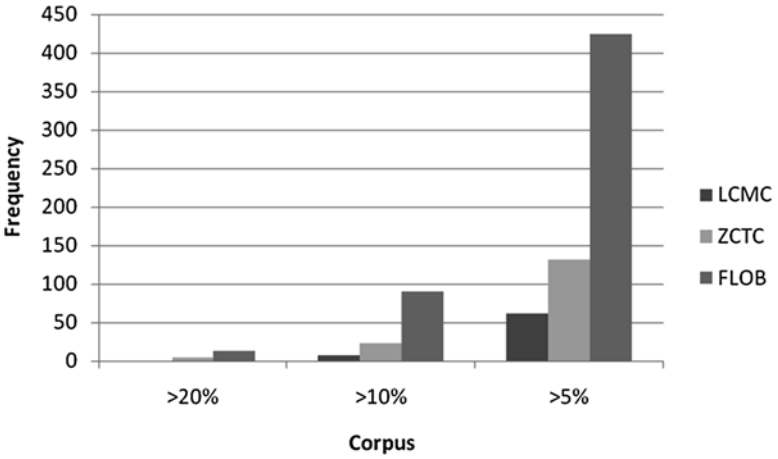


Fig. 7 Coverage of 3-word clusters in English and Chinese

101, 170 and 332 occurrences of 2-word clusters with a minimum coverage rate of 20 %, and 61, 132 and 425 instances of 3-word clusters with a minimum coverage rate of 5 %, in the native and translated Chinese corpora and the comparable English corpus respectively.

The higher frequency and wider coverage of word clusters in translational Chinese suggest that translators demonstrate a higher propensity for striving for fluency than writers of native Chinese texts. Translators are also likely to be under the influence of English, the principal source language of the ZCTC corpus.

This section has explored four lexical features, namely mean word length, affixes, pronouns and high-frequency high-coverage word clusters in the native and translational Chinese corpora. The results show that these features are all significantly more frequent in translated texts. While alternative accounts are plausible (*e.g.* translational explicitation for the overuse of pronouns in translated Chinese), the significantly more frequent use of all of these lexical features provide evidence in support of the TU hypothesis of source language interference at lexical level. The next section will explore three linguistic features at grammatical level.

4 Grammatical Features

This section examines three grammatical features, namely mean sentence segment length, the predicative 是 *shi* ("be") structure, and the 被 *bei* passive construction.

4.1 Mean Sentence Segment Length

Mean sentence length has often been used as a parameter in research of translational language. However, different results have been reported in different studies. For example, Malmkjær (1997) observes that using stronger punctuation in translation entails shorter sentences in translational language, while Laviosa (1998) notes that the mean sentence length is lower in translated newspaper articles than comparable original texts but higher in translated literature than original narrative texts. According to Xiao (2010), while the mean sentence length is slightly greater in LCMC than in ZCTC, the difference has no statistical significance ($t = -1.41$ for 28 d.f., $p = 0.17$).

In native Chinese texts complete sentences do not always end with full stops, because commas are often used to replace full stops. In translated Chinese texts, by contrast, full stops in English source texts tend to be transferred into the translations, which explains why full stops are significantly less frequent (LL = 202.29, $p < 0.001$) but commas are substantially more common (LL = 2,555.28, $p < 0.001$) in LCMC, as shown in Fig. 8.

Chen (1994) finds that three quarters of sentences ending with a full stop and semi-colon contain two or more structurally complete sentence segments. For instance, in example (2),[1] while the Chinese version only contains one sentence, it actually expresses three relatively complete meanings, and as such, three sentences are used in the English version.

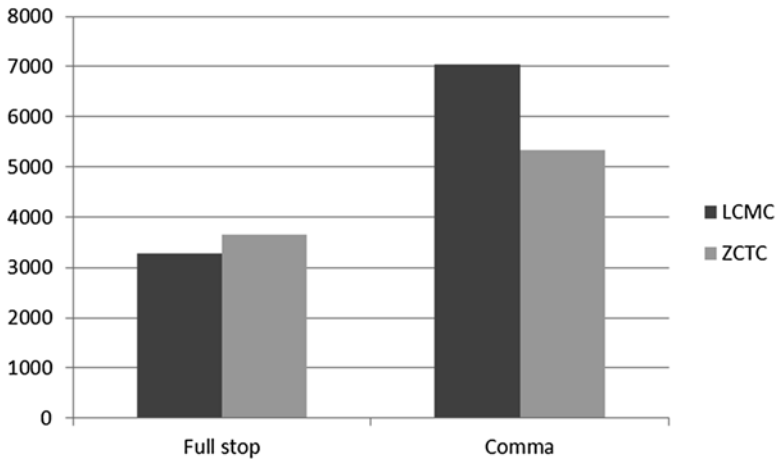


Fig. 8 Full stops and commas in LCMC and ZCTC

[1] This example is taken from a bilingual magazine (www.taiwan-panorama.com).

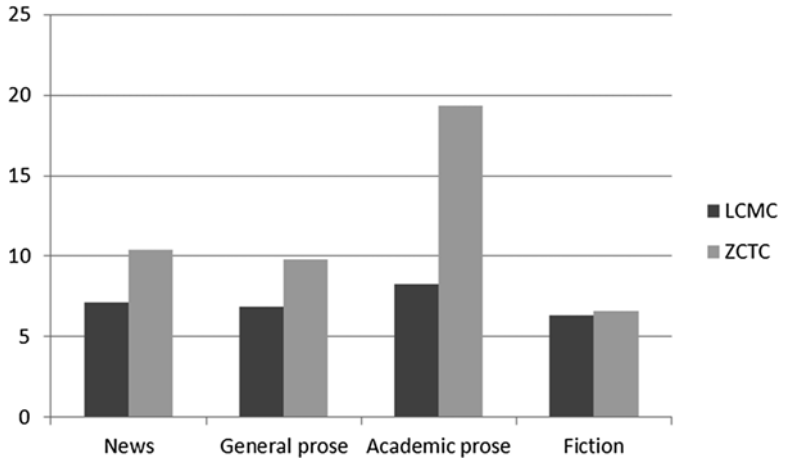


Fig. 9 Mean sentence segment lengths in LCMC and ZCTC

(2a) 人们大多通过电影认识叶锦添 ,(2)尤其在2001年凭着《卧虎藏龙》中典雅清幽的东方意象 , 夺下华人世界第一座奥斯卡最佳美术指导奖后 , (3)各地蜂拥而至的邀约 , 更快速将他推向全球舞台。
Renmen daduo tongguo dianying renshi Ye Jintian, (2) youqi zai 2001 nian ping-zhe "Cang Long Wo Hu" zhong dianya qingyou de dongfang yixiang, duoxia huaren shijie diyi zuo Aosika zui jia meishu zhidao jiang hou, (3) gedi fengyong'erzhi de yaoyue, geng kuaisu jiang ta tuixiang quanqiu wutai.

(2b) Most people know Tim Yip through films. (2) In particular, in 2001 he became the first ever person from the Chinese world to win the US Academy Award for Best Art Direction, received for the elegant Oriental imagery he brought to Crouching Tiger, Hidden Dragon. (3) Since then, demand for his services has gone into hyperdrive, accelerating the spread of his fame and appeal worldwide.

Hence, Wang and Qin (2010: 169) argue that for languages that are characterised by parataxis such as Chinese (Liu 1991), sentence segment length is more meaningful than sentence length. This section will compare the mean sentence segment lengths in native and translated Chinese. In this study, the number of sentence segments is equivalent to the sum of the sentence number and the number of commas. Figure 9 compares the mean sentence segment lengths of native and translated texts. Clearly, the mean sentence segment length is greater in translated Chinese than in native Chinese, in all of the four registers, particularly in academic prose. This finding is in line with Wang and Qin's (2010: 169) observations of literary and non-literary translations. One possible explanation is source language interference, because the mean sentence segment length in English is greater than in Chinese (the

mean sentence segment length is 25.59 words in FLOB but only 13 words in LCMC) while corresponding registers in English (*e.g.* academic writing) customarily make use of long sentences.

4.2 *The Predicative* shi *Structure*

The predicative 是 *shi* ("be") is the most frequently used verb and also the second most frequent word in the Chinese language (Xiao et al. 2009). The predicative structure is a sentence with the predicative *shi* as the main predicate. This section compares the distribution of the predicative structure in native and translational Chinese.

Figure 10 shows the normalised frequencies (per 100,000 words) of the predicate structure in LCMC and ZCTC. As can be seen, when the two corpora are taken as a whole, the *shi* structure is significantly more frequent in translated texts (LL = 16.96, $p < 0.001$). The structure is also more frequent in translations in both literary and non-literary subcorpora, though the contrast in literary texts is not as marked as in non-literary texts, possibly because the predicative structure is very frequently used in both original and translated literary Chinese texts.

The more frequent use of the predicative *shi* structure in translated Chinese is a result of transfer of the copular verb *be* in English source texts. Like the predicative *shi* in Chinese, the copular *be* is also a high-frequency verb in English, which is used in a broader range of contexts than the predicative *shi* in Chinese. For example, the verb *be* can be used as a main verb or an auxiliary whereas the predicative *shi* is not used as an auxiliary. Consequently, native English learners of Chinese tend to overuse the predicative *shi* structure, *e.g.* *我是饿了 *Wo shi e le* 'I am hungry'; *我今年是 20 岁 *Wo jinnian shi ershi sui* 'I am 20 years old this year'. In examples like these, native Chinese speakers would not use the predicative *shi*, but rather say

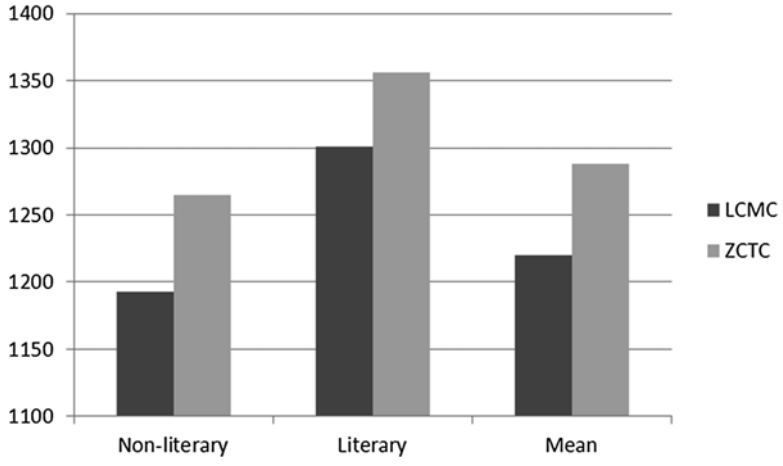


Fig. 10 The predicative *shi* structure in LCMC and ZCTC

我饿了 *Wo e le*; 我今年20岁*Wo jinnian ershi sui*, unless they want to use the predicative structure for emphasis. This is because the correspondence between word classes and syntactic functions is not as rigid in Chinese as in English so that adjectives and even nouns can be used directly as predicates without using the predicative verb *shi* whereas in English the copular verb is mandatory in such cases. Like English learners' Chinese interlanguage, Chinese texts translated from English are also characterised with the overuse of the predicative *shi* structure, as illustrated in the following (a) examples cited from the ZCTC corpus.

(3a) …从来 都 不 是 容易 的。
conglai dou bu shi rongyi de
ever all not SHI easy DE
'…has never been easy.'

(3b) …从来 都 不 容易。
conglai dou bu rongyi
'…has never been easy.'

(4a) 这 种 心悦神怡 的 感觉 是 非常 美妙 的。
zhe zhong xinyueshenyi de ganjue shi feichang meimiao de
this kind joyful DE feeling SHI very beautiful DE
'This kind of joyful feeling is very beautiful.'

(4b) 这 种 心悦神怡 的 感觉 非常 美妙。
zhe zhong xinyueshenyi de ganjue feichang meimiao
this kind joyful DE feeling very beautiful

(5a) 其 效果 也 是 不错 的。
qi xiaoguo ye shi bu-cuo de
its effect also SHI not-bad DE
'Its effect is also not bad.'

(5b) 其 效果 也 不错。
qi xiaoguo ye bucuo
its effect also not-bad

In the above examples, although it cannot be said that it is grammatically incorrect to use the predicative *shi*, (3a–5a) are certainly not as natural and idiomatic as (3b–5b) when there is no need for emphasis or contextual contrast; these sentences simply read like translations.

The LCMC and ZCTC corpora respectively contain 456 and 578 instances of the pattern "*shi* + (adverb) + adjective + DE + punctuation mark", a typical predicative structure in Chinese. The quantitative difference between the two corpora is statistically significant (LL = 12.19, $p < 0.001$), suggesting that the predicative structure is much more frequently used in Chinese translations. Then to what extent is the structure transferred from the English source texts? The Babel parallel corpus shows 568 occurrences of the structure "*be* + adjective", of which 197 are translated into Chinese as the predicative *shi* structure, accounting for more than one third of the total instances. Examples similar to (3a–5a) are abundant in the Babel corpus. These are clearly a result of source language interference.

4.3 *The* bei *Passive Construction*

This section considers Chinese passives marked with *bei*. Passives that profile the agent are conventionally called long passives while those that do not are known as short passives (Xiao et al. 2006). *Bei* passives can take either long or short form.

Figure 11 shows the proportions of short and long passives in native and translated Chinese, and for comparative purposes, the corresponding figures in the native English corpus FLOB are also included. As can be seen, although short passives are more frequent than long passives in both native and translated Chinese, the proportion of short passives in ZCTC is significantly greater than in LCMC (LL = 63.1, $p < 0.001$). The higher proportion of short passives in translated Chinese is clearly a result of source language interference, because the short passive is the statistical norm of passive use in English (see Xiao et al. 2006), which accounts for over 90 % of the total, as shown in Fig. 11. The passive in English is a strategy for expression in that it is used when the agent is unknown or there is no need to mention the agent. In Chinese, in contrast, three out of the five syntactic passive markers (*wei…suo*, *jiao*, *rang*) can only occur in long passives, while the proportions of short passives for the other two (60.6 % and 57.5 % for *bei* and *gei* respectively) are considerably lower than that of English passives (Xiao et al. 2006). As earlier Chinese grammarians Lü and Zhu (1979) and Wang (1985) noted, the agent must be included in the Chinese passive, though this constraint has become more relaxed. When it is hard to identify the agent, vague expressions such as *ren* 'person, someone' or *renmen* 'people' is specified as the agent, which seldom occur in English passive use. In cases where English uses the passive but does not profile the agent, Chinese tends to avoid the passive.

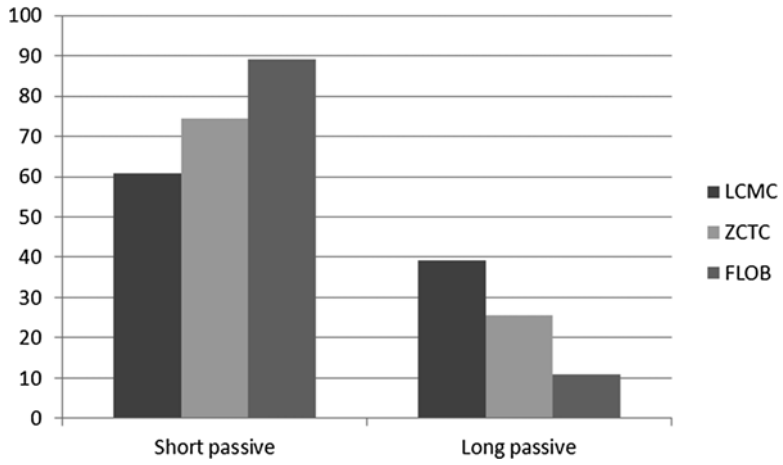


Fig. 11 Short and long forms of *bei* passives

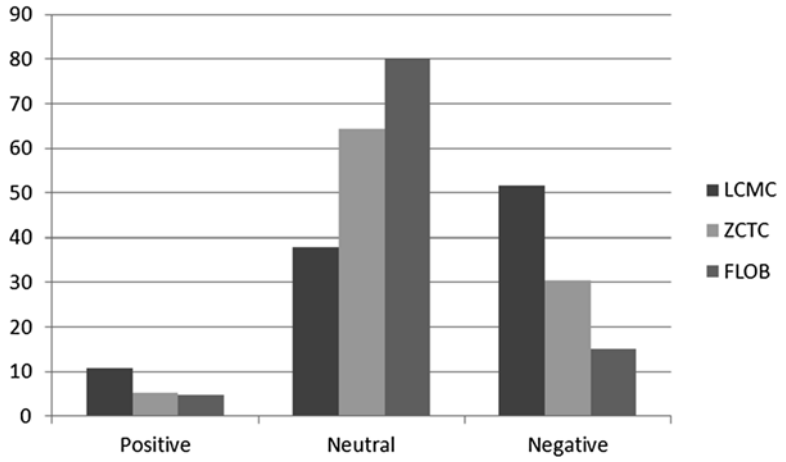


Fig. 12 Pragmatic meanings expressed by *bei* passives

Figure 12 compares the pragmatic meanings expressed by *bei* passives in LCMC and ZCTC and by English *be* passives in FLOB. As can be seen, there are significant differences in the proportions of different pragmatic meaning categories between the three corpora (LL = 212.28 for 2 d.f., $p < 0.001$), with the translated Chinese corpus positioned between the native Chinese and native English corpora, and particularly marked contrasts in neutral and negative meaning categories. Passives in English and Chinese have different functions. English passives primarily function to mark a formal, objective and impersonal style, and are thus pragmatically neutral whereas Chinese passives are an "inflictive voice" that tends to express a negative pragmatic meaning, evaluating the event being described as undesirable, unfavourable or adversative (Xiao et al. 2006). This is because the prototypical passive marker *bei* is derived from a verb in ancient Chinese which meant 'suffer'. Consequently, many disyllabic words with *bei* in modern Chinese refer to something undesirable, *e.g. beibu* 'be arrested', *beifu* 'be captured', *beigao* 'the accused', *beihai* 'be victimised', and *beipo* 'be forced', though the semantic constraint on passive use in modern Chinese is no longer as rigid as before (Xiao et al. 2006).

Native and translated Chinese texts also differ in the frequency of passives and in the distribution of passives across genres. Figure 13 shows the normalised frequencies of passives in different genres in the two Chinese corpora. It is clear that the overall mean frequency of passives is significantly greater in translated Chinese than in native Chinese (LL = 69.59, $p < 0.001$). Given that passives are over ten times as frequent in English as in Chinese (Xiao et al. 2006: 141–142), it is hardly surprising that translated Chinese texts in ZCTC (99 % translated from English) make more frequent use of passives than original Chinese writings. It can also be seen that the most marked contrasts between native and translated Chinese in the distribution of passives are displayed in reports and official documents (H), news reviews (C) and academic prose (J), where passives are significantly more common in translated

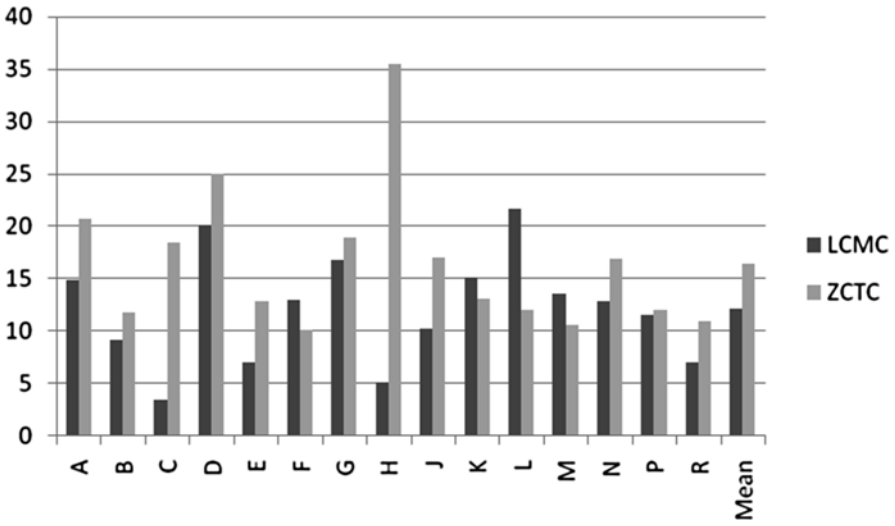


Fig. 13 Distribution of *bei* passives in LCMC and ZCTC

Chinese, and in detective stories (L), where passives are substantially more frequent in native Chinese.

Such distribution patterns of passives in native and translational Chinese are closely related to the different functions of passives in Chinese and English, the overwhelmingly dominant source language in our translational corpus. Since mystery and detective fiction (L) is largely concerned with victims who suffer from various kinds of inflictive events that are usually described using passives in Chinese, it is hardly surprising to find that the inflictive voice is more common in this genre in native Chinese. On the other hand, expository genres like reports and official documents (H), press reviews (C), and academic prose (J), where the most marked contrast is found between translational and native Chinese, are all genres of formal writings that make greater use of passives in English. When texts of such genres are translated into Chinese, passives tend to be carried over and overused in translations because of source language interference or shining through. In such cases, a native speaker of Chinese would not normally use the passive when they express similar meanings. For example, the translated example 该 证书 就 必须 被 颁发 (this certificate then must PASSIVE issue) (ZCTC_H) is clearly a direct translation of the English passive "Then the certificate must be issued". To express this meaning, a native Chinese is very likely to avoid using the passive: 该 证书 就 必须 颁发 (this certificate then must issue) (Xiao and Dai 2010; Xiao 2010: 28).

The differences between native and translated Chinese in their use of *bei* passives as discussed above can reasonably be regarded as the result of source language interference arising from cross-linguistic differences between English and Chinese (Dai and Xiao 2011). Then to what extent does source language interference occur in English-to-Chinese translation, i.e. the extent to which *bei* passives in Chinese translations are transferred from English source texts? We will seek to

answer this question on the basis of English-Chinese parallel corpora in the remainder of this section.

A search for the Chinese passive marker *bei* in the Babel parallel corpus returned 526 instances in Chinese translations, which can be divided into two categories according to whether a passive form is used in the English source text. A total of 446 instances of passives are transferred from the English source texts (including the structure of *be* or other copular verbs such as *get*, *become*, *feel*, *look*, *remain* and *seem* followed by a past participle). For the remaining 80 instances of passives in Chinese translations, a passive form is not used in the English source texts (*cf.* Dai and Xiao 2011). It can be seen that the majority of the passives (about 85 %) in Chinese translations are transferred from English source texts, a finding which is in line with Teich (2003: 196). Furthermore, even those instances of passives in Chinese translations which are not directly carried over from English passives can be traced back to the influence of English source texts (*e.g.* the past participial constructions).

As noted earlier, there are considerable variations in the distribution of passives across genres. In genres of expository writing passives are significantly more frequent in translational Chinese while the contrast is less marked in genres of imaginative writing. This suggests that literary and non-literary texts behave differently in terms of their use of passives in English-to-Chinese translation. As Babel is a corpus of mixed genres, it cannot be used to investigate how source language interference varies in literary versus non-literary texts. In order to explore source language interference in literary and non-literary texts, we will compare the distribution of passives in the English-to-Chinese Literature and English-to-Chinese Non-literature components of the GCEPC parallel corpus.

Figure 14 compares the frequencies of transfer and non-transfer of passives from the source texts in literary and non-literary subcorpora. Of the 553 instances of passives in the literary component, 405 instances are derived from English passives, accounting for 73 % of the total; and of the 768 occurrences in the non-literary component, 712 instances are transferred from English source texts (93 %). This means that as far as English-to-Chinese translation is concerned, source language transfer of passive constructions is more likely to occur in the non-literary than literary translation. This is because a large part of non-literary work relates to genres in English that tend to overuse passives including, for example, official documents and scientific writing.

5 Conclusions

This article has investigated the phenomenon of source language interference in English-to-Chinese translation by undertaking a contrastive study of a range of lexical and grammatical features in translational Chinese in relation to comparable native Chinese. The lexical features investigated include mean word length, affixes, pronouns and word clusters while at grammatical level mean sentence segment length, the predicative *shi* structure and passive constructions are considered. The

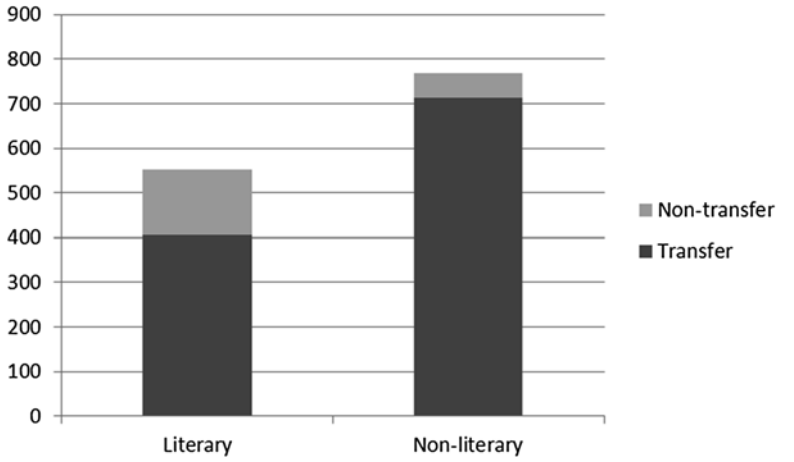


Fig. 14 Source language interference in literary and non-literary texts

results demonstrate that the source-induced difference between translational and native Chinese at both lexical and grammatical levels indicates that the phenomenon of source language interference is observable in English-to-Chinese translation. This study has thus uncovered a fresh body of evidence from translation involving two genetically distant languages, English and Chinese, which supports the hypothesis of source language interference or shining through that has previously been studied only in closely related languages including English, German and Czech. Our case study of the translation of passive constructions in English-to-Chinese parallel corpora suggests that source language interference or shining through typically occurs in 85 % cases in data of mixed genres, with a higher transfer rate of 93 % for non-literary translation in comparison with 73 % for literary translation.

Given the typological distance between the two languages involved in the translation under consideration, the evidence revealed in this study is of critical importance if source language interference or shining through is to be generalised as a universal feature of translation. Future research in this area will benefit from the investigation of a wider range of linguistic features of translational Chinese, and from more language pairs given the availability of appropriate corpus resources. As regards the parallel corpus approach, this study has only considered the direction of English-to-Chinese translation. As source language interference is asymmetrical in different directions of translation because of cross-linguistic differences, it will also be worth investigating the direction of Chinese-to-English translation.

References

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager* (pp. 175–186). Amsterdam: John Benjamins.

Baker, M. (2004). A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics, 9*(2), 167–193.

Baker, M. (2007). Patterns of idiomaticity in translated vs. non-translated text. *Belgian Journal of Linguistics, 21*, 11–21.

Baroni, M., & Bernardini, S. (2003). A preliminary analysis of collocational differences in monolingual comparable corpora. In D. Archer, P. Rayson, A. Wilson, & A. McEnery (Eds.), *Proceedings of the corpus linguistics 2003 conference* (pp. 82–91). Lancaster: UCREL, Lancaster University.

Blum-Kulka, S., & Levenston, E. (1983). Universals of lexical simplification. In C. Færch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 119–139). London: Longman.

Chen, H. (1994). The contextual analysis of Chinese sentences with punctuation marks. *Literary and Linguistic Computing, 9*(4), 281–289.

Chen, W. (2006). *Explication through the use of connectives in translated Chinese: A corpus-based study.* PhD thesis, University of Manchester.

Dai, G., & Xiao, R. (2010). A corpus-based investigation of explicitation in translation. *Chinese Translators Journal, 2010*(1), 76–80.

Dai, G., & Xiao, R. (2011). SL shining through in translational language: A corpus-based study of Chinese translation of English passives. *Translation Quarterly, 62*, 85–108.

Frawley, W. (1984). Prolegomenon to a theory of translation. In W. Frawley (Ed.), *Translation: Literary, linguistic and philosophical perspectives* (pp. 159–175). London: Associated University Press.

Hansen, S., & Teich, E. (2001). Multi-layer analysis of translation corpora: Methodological issues and practical implications. In D. Cristea, N. Ide, D. Marcu, & M. Poesio (Eds.), *Proceedings of EUROLAN 2001 workshop on multi-layer corpus-based analysis* (pp. 44–55). Iasi.

He, X. (2003). Explicitation in English-Chinese translation. *Journal of PLA University of Foreign Languages, 2003*(4), 63–66.

Hopkinson, C. (2007). Factors in linguistic interference: A case of study in translation. *SKASE Journal of Translation and Interpretation, 1*, 13–23.

Hundt, M., Sand, A., & Siemund, R. (1998). *Manual of information to accompany the Freiburg-LOB corpus of British English*. Freiburg: University of Freiburg.

Kenny, D. (1998). Creatures of habit? What translators usually do with words. *Meta, 43*(4), 515–523.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta, 43*(4), 557–570.

Laviosa, S. (2002). *Corpus-based translation studies: Theory, findings, applications*. Amsterdam: Rodopi.

Laviosa-Braithwaithe, S. (1997). Investigating simplification in an English comparable corpus of newspaper articles. In K. Klaudy & J. Kohn (Eds.), *Transferre necesse est. Proceedings of the second international conference on current trends in studies of translation and interpreting* (pp. 531–540). Budapest: Scholastica.

Liu, M. (1991). *Han Ying Duibi Yanjiu yu Fanyi* (Contrastive of Chinese and English and Translation). Nanchang: Jiangxi Educational Press.

Lü, S., & Zhu, D. (1979). *Yufa Xiuci Jianghua* (Talks on grammar and rhetoric). Beijing: Chinese Youth Press.

Malmkjær, K. (1997). Punctuation in Hans Christian Andersen's stories and in their translations into English. In F. Poyatos (Ed.), *Nonverbal communication and translation: New perspectives and challenges in literature, interpretation and the media* (pp. 151–162). Amsterdam: John Benjamins.

Mauranen, A. (2007). Universal tendencies in translation. In M. Rogers & G. Anderman (Eds.), *Incorporating corpora. The linguist and the translator* (pp. 32–48). Clevedon: Multilingual Matters.

McEnery, T., & Xiao, R. (2002). Domains, text types, aspect marking and English-Chinese translation. *Languages in Contrast, 2*(2), 211–229.

McEnery, T., & Xiao, R. (2004). The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. In M. Lino, M. Xavier, F. Ferreire, R. Costa, & R. Silva

(Eds.), *Proceedings of the fourth international conference on language resources and evaluation (LREC)* (pp. 1175–1178). Lisbon: The Cultural Centre of Belém.

Nevalainen, S. (2005). Köyhtyykö kieli käännettäessä? Mitätaajuuslistat kertovat suomennosten sanastosta. In A. Mauranen & J. Jantunen (Eds.), *Käännössuomeksi* (pp. 141–162). Tampere: Tampere University Press.

Pym, A. (2005). Explaining explicitation. In K. Károly & Á. Fóris (Eds.), *New trends in translation studies* (pp. 29–43). Budapest: Akadémiai Kiadó.

Scott, M. (2009). *The Wordsmith tools* (version 5.0). Oxford: Oxford University Press.

Teich, E. (2003). *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*. Berlin: Mouton de Gruyter.

Toury, G. (1979). Interlanguage and its manifestations in translation. *Meta, 24*(2), 223–231.

Toury, G. (1995). *Descriptive translation studies and beyond*. Amsterdam: John Benjamins.

Wang, L. (1985). *Zhongguo Xiandai Yufa* (Modern Chinese grammar). Beijing: Commercial Press.

Wang, K. (Ed.). (2004). *Shuangyu Duiying Yuliaoku Yanzhi yu Yingyong* (The development of the compilation and application of parallel corpora). Beijing: Foreign Language Education and Research Press.

Wang, K., & Qin, H. (2010). A parallel corpus-based study of translational Chinese. In R. Xiao (Ed.), *Using corpora in contrastive and translation studies* (pp. 164–181). Newcastle: Cambridge Scholars Publishing.

Xiao, R. (2005). *The Babel-English-Chinese parallel corpus*. Lancaster: UCREL, Lancaster University.

Xiao, R. (2010). How different is translated Chinese from native Chinese. *International Journal of Corpus Linguistics, 15*(1), 5–35.

Xiao, R. (2011). Word clusters and reformulation markers in Chinese and English: Implications for translation universal hypotheses. *Languages in Contrast, 11*(1), 145–171.

Xiao, R. (2012). *Ying Han Fanyi zhong de Hanyu Yiwen Yuliaoku Yanjiu* (Corpus-based studies of translational Chinese in English-Chinese translation). Shanghai: Shanghai Jiao Tong University Press.

Xiao, R., & Dai, G. (2010). Idioms and word clusters in translational Chinese: A corpus-based study. *Foreign Language Research, 2010*(3), 79–86.

Xiao, R., McEnery, T., & Qian, Y. (2006). Passive constructions in English and Chinese. *Languages in Contrast, 6*(1), 109–149.

Xiao, R., Rayson, P., & McEnery, T. (2009). *A frequency dictionary of Mandarin Chinese: Core vocabulary for learners*. London: Routledge.

Xiao, R., He, L., & Yue, M. (2010). In pursuit of the "third code": Using the ZJU Corpus of translational Chinese in translation studies. In R. Xiao (Ed.), *Using corpora in contrastive and translation studies* (pp. 182–214). Newcastle: Cambridge Scholars Publishing.

From the Other Side of the Looking Glass: A Cognitive-Pragmatic Account of Translating Lewis Carroll

Francisco Javier Díaz-Pérez

Abstract The present paper offers a cognitive-pragmatic account of the translation of *Alice in Wonderland* and *Through the Looking Glass*. More specifically, its main objective is to analyse the translation of puns in a corpus consisting of one Galician and six Spanish versions of the mentioned novels from a relevance-theoretic perspective. The analysis is based on 959 textual fragments which correspond to the 137 ST extracts which contained wordplay. The results show that translation technique selection is determined, among other factors, by the principle of relevance. In those cases in which there is a coincidence in the relation between the levels of signifier and signified across source and target language, translators normally opt to translate literally and reproduce a congenial pun. In the rest of the cases, translators still strive to produce a pun which, even if it is not able to reproduce the meanings of the ST pun, at least gives rise to some of the cognitive effects intended by the original author, particularly those associated with the processing of wordplay. Other solutions adopted by translators include the sacrifice of secondary information, a non-selective translation containing the different meanings of the ST pun in a non-punning context, the resort to diffuse paraphrase, punoid, editorial means, or transference. Variables such as the specific version considered or the type of pun have been found to have an effect on the choice of translation technique. Moreover, it has also been proved that choice of translation technique and use of editorial means were interconnected.

Keywords Relevance Theory • Wordplay • Lewis Carroll • Spanish • Galician

F.J. Díaz-Pérez (✉)
Departamento de Filología Inglesa, Facultad de Humanidades y Ciencias de la Educación, Universidad de Jaén, Jaén, Spain
e-mail: fjdiaz@ujaen.es

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_8

1 Introduction

One of the most defining characteristics of the language of Lewis Carroll's *Alice's Adventures in Wonderland* and *Through the Looking Glass and What Alice Found There* is the overwhelming abundance of wordplay. In consonance with their ludic tone, language is one of the elements which is used to play with in those two novels.[1] In fact, in both books 137 puns have been identified. Partly because of that, the translation of those two literary works into any language represents a real challenge. The main objective of the present study involves analysing the techniques used by the translators of six different Spanish versions and one Galician version of the two mentioned novels to render wordplay in their target texts (TTs)[2] from a cognitive-pragmatic standpoint. More specifically, the framework used in this study is Sperber and Wilson's Relevance Theory.[3] All in all, then, our corpus consists of 959 textual fragments corresponding to 137 source text (ST) extracts containing wordplay.

The second section of this chapter will be devoted to a brief explanation of the relevance-theoretic account of translation and specifically of the translation of wordplay, Sect. 3 will focus on the different techniques used by the translators to face wordplay, and Sect. 4 presents and discusses the results of the study. Finally, the chapter is closed with the conclusions section, followed by the bibliographical references.

2 Relevance Theory and the Translation of Wordplay

Relevance Theory, which originated in the late 1980s, is one of the most influential theoretical frameworks within the field of pragmatics. It departs from the assumption that human beings are programmed to address their attention to that which is relevant to them, or in other words, to that which may produce changes in their cognitive environment. Those changes are technically called cognitive effects. From a relevance-theoretic standpoint, the more cognitive effects a given stimulus gives

[1] As highlighted by Weissbrod (1996: 222–223), the tendency to use wordplay in children's literature is both a long-lived literary convention and an answer to children's linguistic development. Moreover, the use of wordplay in the Carrollian texts which are the concern of this study also accounts for their appeal to an adult audience, since they were conceived as ambivalent texts, functioning simultaneously in the children's and adults' literary systems.

[2] The acronyms used in this paper are: ST – which stands for source text, or original text –, TT – which stands for target text, or translated version –, SL – source language, or original language –, and TL – target language or language into which the ST is translated –.

[3] See, for instance, Sperber & Wilson (1986, 1995) and Wilson & Sperber (2004).

rise to, the more relevant it will be. However, those cognitive effects must be put in relation to the effort needed to derive them, since an increase in the effort needed to process a given utterance will go to the detriment of its relevance. Thus,

(a) Other things being equal, the greater the positive cognitive effects achieved by processing an input, the greater the relevance of the input to the individual at that time.
(b) Other things being equal, the greater the processing effort expended, the lower the relevance of the input to the individual at that time (Wilson & Sperber 2004: 609).

In this sense, one of the main principles of Relevance Theory is the *principle of relevance*, according to which, "[h]uman cognition tends to be geared to the maximization of relevance" (Wilson & Sperber 2004: 610). In other words, an addressee will make the effort to process a given statement if s/he considers that the statement will be relevant, or in relevance-theoretic terms, will be able to modify his/her cognitive environment. As will be seen below, many of the decisions taken by a translator can be explained by resorting to the principle of relevance.

Another assumption which is particularly important in the case of translation is the difference between the descriptive use and the interpretive use of language. In this sense, language is said to be used descriptively when a given utterance is intended to be taken as true of a state of affairs in some possible world. On the contrary, language is said to be used interpretively when an utterance represents what someone else has said or thought. Thus, in (1a) below Alice uses the utterance "I'm not a Monster" to claim that the state of affairs that the utterance describes is true. In (1b), however, Alice does not necessarily claim that the state of affairs described by the same utterance is true. That is to say, whereas in (1a) the utterance is being used descriptively, in (1b) it is being used interpretively. As regards the interpretive use of language, there is a relation of interpretive resemblance between the original utterance and that other utterance used to represent it. The degree of interpretive resemblance will depend on the amount of implicatures and explicatures shared between the two utterances.[4]

(1) a. Alice: "I'm not a Monster."
    b. Alice: The Unicorn said, "I'm not a Monster."

It has been pointed out more than once (Gutt 1990, 1991, 1998, 2000, 2004, 2005; Rosales Sequeiros 2002, 2005; Alves & Gonçalves 2003, 2007, 2010; Zhonggang 2006; Jing 2010; Martínez-Sierra 2010; Yus 2012; Díaz-Pérez 2013, 2014) that

[4]The content explicitly communicated by means of an utterance is an explicature, whereas the content which is derivable from the proposition expressed by the utterance together with the context is called an implicature.

Relevance Theory can be applied to translation. From a relevance-theoretic perspective, translation involves interpretive use across languages. In this connection, Relevance Theory allows the study of intra- and inter-lingual verbal communication as manifestations of the same underlying concepts, and in this sense, offers a unified theory of verbal communication.

Amongst the different applications of Relevance Theory to translation, Gutt's has been the most influential one. According to him,

> From the relevance-theory point of view, translation falls naturally under the interpretive use of language: the translation is intended to restate in one language what someone else said or wrote in another language. In principle it is, therefore, comparable to quoting or speech-reporting in intra-linguistic use. One of its primary distinctions setting it off from intra-lingual quoting or reporting is that original text and translation belong to different languages (Gutt 1998: 46).

From the perspective of this relevance-theoretic view of translation, the relation between a translation and its ST is considered to be based on interpretive resemblance. After analysing the original author's assumed intentions and assessing the cognitive environment shared by ST addresser and TT addressee, the translator may adopt different techniques to try to recreate the cognitive effects intended by the original writer with the lowest possible processing effort by the TT receptor. In a subsequent expansion of his application of Relevance Theory to translation, Gutt (2004) claims that when translation brings into contact a communicator and an audience with different cognitive environments, additional sophistication is required, namely the human beings' capacity of metarepresentation.[5] Metarepresentation involves the ability to represent in one's mind how other human beings represent states of affairs in the world in their minds. The translator needs to metarepresent not only the shared cognitive environment between the original communicator and his/her audience, but also the target receptors' cognitive environment. In Gutt (2005), translation is defined as a higher-order act of communication (HOAC), "an act of communication that is about another (lower-order) act of communication" (Gutt 2005: 25). Since the lower-order act of communication consists of a stimulus and an intended interpretation, according to Gutt (2005: 34) two modes of higher-order communication can be distinguished, namely the stimulus-oriented mode (or s-mode) and the interpretation-oriented mode (or i-mode).

As regards the particular case of the translation of puns, the difficulty it entails is something obvious, which has been highlighted on several occasions. According to Delabastita, the reason for this difficulty is that

> the semantic and pragmatic effects of source text wordplay find their origin in particular structural characteristics of the source language for which the target language more often than not fails to produce a counterpart, such as the existence of certain homophones, near-homophones, polysemic clusters, idioms or grammatical rules. (Delabastita 1994: 223)

[5] Wilson (2012) has defined metarepresentation as "a representation of a representation: a higher-order representation with a lower-order representation embedded within it" (Wilson 2012: 230).

Particularly those cases in which there is lack of symmetry between the levels of form and meaning across languages are the most challenging ones for the translator, since they demand a higher degree of creativity, as emphasized more than once (Levý 1969; Gutt 2000; Sanderson 2009; Marco 2010). The translator will have to decide whether it is more important to be faithful to content or to the effect produced by wordplay. It has been argued (Asimakoulas 2004; Díaz-Cintas & Remael 2014; Yus 2012) that the preferable solution in these cases is that which involves the creation of a new pun, even if (part of) the content had to be sacrificed. From the perspective of Relevance Theory, this solution is said to recreate the cognitive effects produced by the processing of wordplay.

Yus (2012) presents several examples of jokes based on puns. Adopting a relevance-theoretic standpoint, he defends that a translator's most important task is to preserve those inferential strategies which made the derivation of humorous effects possible in the source language (SL). That task very often demands that the semantic content should be changed completely. The pragmatic scenario predicted by the SL communicator would then be preserved in the target language (TL), not only in the quantity and quality of inferential strategies, but also in the balance of cognitive effects and mental effort (Yus 2012: 144).[6] One of the examples used by Yus (2012) is the following joke from the film *The Naked Gun*, which plays on different senses of the subsentential utterance *Goodyear?*, which can encode both the explicature "Was the typical blimp with the Goodyear logo on it?" and the explicature "Was it a good year?" For the cognitive effects associated with the processing of wordplay triggering humour to be reproduced in the TT, the translator had to change the cultural and semantic scenarios. Cultural references such as *Orange Bowl* and *Goodyear*, which are unlikely to be part of a Spanish speaker's cognitive environment,[7] have been changed into *Conserva del Norte* ("fish cans from the North of Spain") and *Bonito* (signifier meaning both "variety of tuna fish from the North of Spain" and "nice"). This change of scenario has allowed the translator to include a pun on the subsentential utterance *¿Bonito?*, which could encode the explicatures "were the cans of bonito fish?" and "was it nice?".

[6] With respect to the translation of humour, Yus (2012) devises the existence of three parameters, which he calls scenarios, namely the cultural, pragmatic, and semantic scenarios. The same parameters could also be applied to the translation of puns, as discussed in Sect. 3.

[7] A cultural reference may be defined, following González Davies & Scott-Tennent (2005), as

> Any kind of expression (textual, verbal, non-verbal or audiovisual) denoting any material, ecological, social, religious, linguistic, or emotional manifestation that can be attributed to a particular community (geographic, socio-economic, professional, linguistic, religious, bilingual, etc.) and would be admitted as a trait of that community by those who consider themselves to be members of it. Such an expression may, on occasions, create a comprehension or a translation problem. (González Davies & Scott-Tennent 2005: 166)

(2) a. SL JOKE.
DREBIN: lt's the same old story: boy finds girl, girl finds boy, boy loses girl, girl finds boy, boy forgets girl, boy remembers girl, girl dies in a tragic blimp accident over the Orange Bowl on New Year's Day.
JANE: Goodyear?
DREBIN: No, the worst. (*The Naked Gun*)

b. TL TRANS.
DREBIN: La historia de siempre. Chico conoce chica, chico pierde chica, chica conoce chico, chico olvida chica, chico recuerda chica, chica muere en trágico accidente en globo anunciando pescado en Conserva del Norte.
JANE: ¿Bonito?
DREBIN: No, fue horrible.

c. BACK TRANS.
DREBIN: The same old story: boy meets girl, boy loses girl, girl meets boy, boy forgets girl, boy remembers girl, girl dies in a tragical blimp accident while making publicity for canned fish from the North (of Spain).
JANE: Tuna fish? (Or: Was it nice?)
DREBIN: No, it was horrible . (Yus 2012: 140–141)

3 Techniques for the Translation of Puns

Following Hurtado Albir (2001: 268) and Marco (2010: 265), the term *technique* has been employed in this paper instead of other labels which have also been used to refer to the same notion, such as strategy, method, solution-type, or procedure. Taking Hurtado Albir's (2001: 268) definition as a basis, a translation technique can be described as a procedure, normally at the verbal level, which has a functional character and which refers to a textual fragment. According to Hurtado Albir (1996, 1999, 2001) and Molina & Hurtado Albir (2002), a translation strategy, in turn, is a conscious or unconscious, verbal or non-verbal procedure used during the translation process with an objective in mind. Translation strategies include strategies for comprehension and strategies for reformulation. As argued more than once (Zabalbeascoa 2004; Marco 2004, 2007, 2010), typologies of translation techniques for specific translation problems are better suited to explaining the particularities of each problem than general classifications, considered valid for any textual fragment.[8] Ten different techniques have been identified to translate puns in the corpus used in this study, which in turn have been grouped into six categories, as shown in Table 1.[9] These techniques are explained in the sub-sections below from the point of view of Relevance Theory.

[8] The translation techniques proposed in Molina & Hurtado Albir (2002: 509–511) are adaptation, amplification, borrowing, calque, compensation, description, discursive creation, established equivalent, generalization, linguistic amplification, linguistic compression, literal translation, modulation, particularization, reduction, substitution, transposition, and variation.

[9] Compensation – dealt with in Sect. 3.7 – is not included here, since strictly speaking, it is not a technique used to translate puns, as it is not applied to punning textual fragments.

Table 1 Techniques for the translation of puns in the corpus

Translation technique	Category
Punning correspondence	Preservation of the pragmatic and semantic scenarios
Change of Pun	Preservation of the pragmatic scenario
Punoid	
Sacrifice of secondary information	Preservation of the semantic scenario
Non-selective translation	
Transference	Preservation of the cultural scenario
Direct copy	
Omission	None of the scenarios Preserved
Diffuse paraphrase	
Editorial means	Amplification (used in combination with any of the above)

3.1 Preservation of the Pragmatic and Semantic Scenarios: Punning Correspondence

Although often considered to be a very difficult task, a ST pun more often than expected finds a punning counterpart in the TT. In this way, in relevance-theoretic terms, all those ST-intended cognitive effects associated with the processing of wordplay will be accessible to the TT receptor as well. Following Yus (2012), it could be said that when this is the case, the pragmatic scenario is preserved, even though very often the semantic scenario may have to be sacrificed, as will be discussed in Sect. 3.2.1. On some other occasions, however, as presented in this sub-section, a coincidence in the relation between the levels of signifier and signified across languages will allow the translator to adhere to both the ST semantic and pragmatic scenarios.

In those cases in which there is a lucky coincidence in the relation between form and content across SL and TL, the translator frequently takes the opportunity and reproduces a congenial pun in the TT. The term *congenial pun* has been used by Delabastita (1993: 196) to refer to a TT pun which reflects the same semantic ambiguity as its ST counterpart and which is based on the same linguistic phenomenon. Previous to the application of this technique, the translator will have to correctly metarepresent the cognitive environments of ST communicator and TT receptor.

The following excerpt contains a polysemic and horizontal pun on the verb *find*,[10] whose meaning in its first and third occurrences is "to consider or regard as" (s1), whereas in the other occurrences the content is "to come upon by chance or in the course of events" (s2).[11] By translating the punning fragment literally into Spanish (3b) and Galician (3c), a congenial pun has been reproduced in the TT, since the

[10] A horizontal pun, according to Delabastita (1993: 79, 1996: 128), is that in which the relationship between the components is of a syntagmatic type, that is to say, the components are one after the other lineally in the sequence in which the pun is inscribed. When the two components are co-present in the same portion of text, however, the pun is said to be vertical (Delabastita 1996: 128).

[11] http://www.oed.com/view/Entry/70348?rskey=Y8cJjW&result=2&isAdvanced=false#eid

verbs *encontrar* and *atopar* are also polysemic respectively in Spanish and Galician, and they transmit the same senses as their English counterpart.

(3) a. 'Edwin and Morcar, the earls of Mercia and Northumbria, declared for him; and even Stigand, the patriotic archbishop of Canterbury, **found** it advisable—'"[12]
"**Found** what?" said the Duck.
"**Found** it," the Mouse replied rather crossly: "of course you know what 'it' means."
"I know what 'it' means well enough, when I **find** a thing," said the Duck: "it's generally a frog, or a worm. The question is, what did the archbishop **find**?" (Carroll 2000/1865: 32)[13]

b. Edwin y Morcar, condes que eran a la sazón de los condados de Mercia y Northumbria, se pusieron de su parte. Incluso Stigand, honra y prez de patriotas, arzobispo que era de la sede episcopal de Canterbury, lo **encontró** oportuno en aquellas circunstancias…
Pero ¿se puede saber *qué* es lo que **encontró**? –preguntó el Pato.
Encontró *«lo»* –respondió irritado el Ratón–, ¿o es que acaso no sabe usted lo que significa «lo»?
¡Pues claro que sé lo que significa «lo»! -contestó el Pato-. ¡Pero he de ser *yo* el que «lo» **encuentre**! Y «lo» que yo encuentro suele ser una rana o algún gusano. Pero aquí se trata de averiguar «lo» que **encontró** ese arzobispo… (Buckley 2005: 130–131)[14]

c. … Edwin e Morcar, Condes de Mercia e Northumbria, apoiárono, e mesmo Stigand, o patriota arcebispo de Canterbury, foi con Edgardo Atheling ó encontro de Guillermo para ofrccerlle a coroa, **atopándo**o ben aconsellable …
¿**Atopando** o que? – dixo o Parrulo.
Atopándoo – contestou o Rato enfurruñado-; vostede sabe perfectamente o que significa *o* nestes casos.
Ben sei o que significa *o* cando son eu o que atopo algo, que é case sempre un sapo ou un verme. Pero o que digo eu é, ¿que foi o que **atopou** o arcebispo? (Barro & P. Barreiro 2002: 46 and 48)

In (4) there is a horizontal pun based on the homophony between *tea* and the name of the letter *T*. Both in Spanish and Galician *té* and *T* are also homophones and, leaving aside certain different connotations, the semantic content is basically the same in the occurrences of the two lexical items both in ST and TT, so that a

[12] The fragments involving wordplay in the ST and TT in all the examples appear in bold. Emphasis is mine.

[13] In the examples, the ST excerpts are identified as Carroll 2000/1865, which stands for *Alice's Adventures in Wonderland* in the edition used in this study, by Gardner, and as Carroll 2000/1871, which corresponds to *Through the Looking Glass and What Alice Found There* in the same edition.

[14] In the examples the TT excerpts are identified by the name of the translators, except in the case of the versions published by El Cid Editor, which are referred to by the name of the publishing house, since

word for word translation has reproduced a congenial pun in the TT. Even though the presence of a pun demands a higher processing effort, this is compensated by the yielding of additional cognitive effects, as signalled more than once (Tanaka 1992, 1994; Yus 2003, 2008; Van Mulken et al. 2005; Higashimori 2011; Solska 2012). The additional cognitive effects are not only derived from the existence of at least two meanings, but also from the presence of a pun and from its processing. As Solska (2012: 180) puts it, "cognitive effects are not limited to the additional propositional content, but include such benefits as the appreciation of wittiness or the enjoyment of humour".[15]

(4) a. "I'm a poor man, your Majesty," the Hatter began in a trembling voice, "and I hadn't but just begun my **tea**—not above a week or so—and what with the bread-and-butter getting so thin—and the twinkling of the **tea**—"
"The twinkling of *what ?"* said the King.
"It began with the **tea**," the Hatter replied.
"Of course twinkling begins with a **T**!" said the King sharply. "Do you take me for a dunce? Go on!" (Carroll 2000/1865: 79)

b. –Soy un pobre hombre, Su Majestad –empezó con voz temblorosa el Sombrerero–, y aún no había empezado el **té**… hará cosa de una semana… y con las pocas tostadas… y con el titilar del **té**…
–¿El titilar de qué? –preguntó el Rey.
–La cosa empezó con **té**, y… –replicó el Sombrerero.
–¿Titilar? ¡Claro que empieza con **T**! –le cortó el Rey–. ¿Me tomas porzopenco? ¡Sigue! (Maristany 1986: 120)

c. -Eu valer non vallo cousa. Maxestade –empezou o Sombreireiro, con voz tremente– e aínda non empezara a merendar… non haberá máis dunha semana ou así… e co pan con manteiga máis fino de cada vez, e o tintilar do **té** …
-¿O tintilar do que? -dixo o Rei.
-Empezou co **té** -replicou o Sombreireiro.
-¡Ben sei que tintilar empeza cun **T** -dixo o Rei asperamente-. ¿Coidas que son un simplorio? ¡Continúa! (Barro & P. Barreiro 2002: 148–149)

Whereas in (4) the ST contained a phonetic pun based on homophony, the ST pun in (5) is a syntactic one. Thus, the phrase *a minute* can be analysed as a time adjunct – which is the most likely interpretation – or as a direct object of the verb *stop*. This second analysis gives rise to a much more unlikely but also possible

the name(s) of the translator(s) is not provided for. This latter case represents an extreme case of what Venuti (1995) called the *translator's invisibility*, or a "weird self-annihilation" (Venutti 1995: 8). In the bibliographical references section, however, all the versions from which the excerpts have been extracted appear under the name of the ST author: Carroll.

[15] In this sense, as argued by Kosińska (2005: 77), Dynel (2010: 106), and Seewoster (2011: 71), the relevance of puns also lies in humour and wit, in such a way that the addressee may choose to devote more effort in order to obtain, for instance, humorous effects.

interpretation, and it is in fact the king's interpretation of Alice's utterance, which produces humorous effects. The literal translation of that sequence into Galician has produced a congenial pun which will allow the TT addressee to retrieve the ST-author-intended cognitive effects without investing unnecessary processing effort.

(5) "Would you—be good enough—" Alice panted out, after running a little further, "to **stop a minute**— just to get—one's breath again?"
"I'm good enough," the King said, "only I'm not strong enough. You see, a minute goes by so fearfully quick. You might as well try to stop a Bandersnatch!" (Carroll 2000/1871: 144)
–¿Tería a bondade… de… **parar un minuto**… xusto… para coller folgos? –arquexou Alicia, despois de correr un pedazo máis.
–Bondade teño –dixo o Rei–·o que non teño é a forza. ¡Un minuto pasa tan axiña! Iso é coma querer parar a un Bandarpillán! (Barro & P. Barreiro 1985: 124–125)

In the three previous examples, the translators have decided to keep the original puns by translating the sequences in which the puns are inscribed word for word, so that the target addressees could recover from their cognitive environments the encoded meanings of the lexical items *encontrar*, *atopar*, and *té* and of the phrase *parar un minuto*, Thus, the target audience would be able to recognize the existence of a pun and to recover the cognitive effects intended in the ST. The degree of interpretive resemblance corresponding to this translation technique is a very high one, because of the high amount of implicatures and explicatures shared by ST and TT. As mentioned above, apart from the lucky coincidence that the correspondence between form and content is identical or almost identical in SL and TL, the translator decides to apply this translation technique after metarepresenting the cognitive environments of source writer and target reader. With regard to (5), the analysis could be summarized in the following way:

Cognitive environment and Effects (source culture)

Existing Assumptions (EA)

1. In the English sequence *to stop a minute* the phrase *a minute* is an adverbial which refers to duration.
2. Although much more unlikely, in the sequence *to stop a minute* the NP *a minute* can also be interpreted as the direct object of the verb *stop*.
3. The two previous interpretations can be combined in a pun.

Contextual Assumptions (CA)

1. Both *Alice's Adventures in Wonderland* and *Through the Looking Glass and What Alice Found There* abound in puns.
2. Much of the humour in the two novels is based on puns.
3. Many of the characters in both novels interpret linguistic utterances in unusual ways, sometimes nonsensical or literal.
4. The King's answer indicates that he has misunderstood Alice's request, interpreting the sequence *stop a minute* in an unlikely but possible way.

Cognitive Effects (CE)

1. CA1 reinforces mainly EA3, but also EA2 and EA1.
2. CA2 reinforces EA2 and EA3.
3. CA3 reinforces EA2.
4. CA4 reinforces EA2.
5. Contextual implication: the combination of CA3 and CA4 with EA2 and EA3 produces a surprising and amusing effect, because what might seem an unlikely interpretation – that in which *a minute* is the direct object of *stop* – is relevant in this context, and this produces humorous effects.

The five cognitive effects derived would also be accessible to the target reader without gratuitous processing effort, as s/he would depart from the same assumptions. As a result of the technique adopted, then, the target addressee can have access to roughly the same cognitive effects intended by the source communicator.[16] Had the translators opted to reflect the meaning in the previous fragments without reproducing wordplay in the TT, the target addressee would have had to invest less processing effort, but conversely the ST-intended cognitive effects would have been sacrificed. The target receptor, then, would have been deprived of the processing of wordplay and, consequently, of the cognitive effects – humorous or of any other type – associated with that processing.

3.2 Preservation of the Pragmatic Scenario

3.2.1 Change of Pun

Despite the fact that, as seen in the previous sub-section, a ST pun occasionally may have a congenial TT counterpart, in the majority of cases the relation between form and content across SL and TL is an asymmetrical one. It is in these cases that the translator will have to decide whether content or the cognitive effects associated with the processing of wordplay should prevail. If the translator decides to preserve those effects associated with the processing of wordplay, a new pun will have to be created, at the expense of a larger or smaller sacrifice of the semantic content.

In (6) there is a phonologic pun on *Tortoise* and *taught us*. As the literal translation into either Spanish or Galician would not reproduce the pun in the TT and the pun is highly relevant in this case due to the humorous cognitive effects it gives rise to, the translators of 6 out of the 7 versions studied decided to create a new pun in the TT. Thus, in (6b) there is a morphological pun on *galápago*, "fresh water turtle", which is interpreted as though it were composed of the morphemes *gala* and *pago*,[17]

[16] This situation represents, in Gutt's (2004: 83) opinion, the translator's ideal, since, given that original communicator, translator, and receptors share a mutual cognitive environment, there is no need to overcome differences in cognitive environments.

[17] In addition, in this excerpt *gala* appears in the set phrase *tener a gala*, "to be very proud of", and *pago* is part of the phrase *escuela de pago*, "private school".

in (6c) the pun is idiomatic, playing on the literal and idiomatic senses of the set phrase *tener más conchas que un galápago*, respectively "to have more shells than a turtle" and "to be a sly one", and in (7d) the Galician TT pun is based on the paronymy between *Sapoconcho*, "turtle", and *sabio chocho*, "doddering wise man".

(6) a. "The master was an old Turtle—we used to call him **Tortoise**—"
"Why did you call him **Tortoise**, if he wasn't one?" Alice asked.
"We called him **Tortoise** because he **taught us**," said the Mock Turtle angrily. (Carroll 2000/1865: 70)

b. El maestro era una vieja tortuga a la que llamábamos **Galápago**.
– ¿Por qué lo llamaban **Galápago**, si no era un **galápago**? – preguntó Alicia.
– Lo llamábamos **Galápago** porque siempre estaba diciendo que tenía a «**gala**» enseñar en una escuela de «**pago**» –explicó la Falsa Tortuga de mal humor – (El Cid 2009: 131–132).

c. «El maestro era una vieja tortuga al que llamábamos Galápago».
«Y ¿por qué lo llamaban '**Galápago**' si no lo era?», preguntó Alicia.
«Lo llamábamos '**Galápago**'», replicó muy molesta la Tortuga Artificial, «por las muchas **conchas que tenía**, ¡naturalmente! ¡Vaya pregunta! (Ojeda 1976: 152)

d. O mestre era un **Sapoconcho** xa vello (que nós chamabásmolle **Sabiochocho**) …
– ¿E logo por que lle chamaban así? –preguntou Alicia.
– Chamabámoslle **Chocho** porque ás veces, cando se ía da clase, estaba ido, e **sabio**, porque cada un sabe de si –dixo a Tartaruga de Imitación, toda enfadada–. (Barro & P. Barreiro 2002: 127)

Whereas in (6) at least one of the senses in the TT puns coincided with one of the senses realized in the ST pun, on some other occasions the TT pun is completely unrelated to its ST counterpart from a semantic point of view. In other words, neither of its meanings coincides with the meanings realized by the original pun. That is the case of (7), which contains two instances of wordplay. The first ST pun is a horizontal and phonologic one, based on the homophony between *flour* and *flower*. As a literal translation into Spanish would not reproduce any pun, the translator of (7b) decided to create a new pun on *harina*, the Spanish word for *flour*, and a new element, *arena* ("sand"). It is in the translation of the second ST pun – that between the past participle of the verb *to grind* and the noun *ground* ("soil") – that its TT counterpart introduces completely new senses, as it plays on two senses of the polysemic word *grano* in Spanish: "grain (of wheat)" and "spot, pimple". In the Galician TT (7c), there are also two puns, one on *fariña* ("flour") and *fouciña* ("sickle") and another one on *moer* ("to grind") and *mover* ("to move").

(7) a. "I know *that*!" Alice cried eagerly. "You take some **flour**—"
"Where do you pick the **flower**?" the White Queen asked. "In a garden or in the hedges?"
"Well, it isn't picked at all," Alice explained: "it's **ground**—"
"How many acres of **ground**?" said the White Queen. (Carroll 2000/1871: 160)

b. – ¡*Esto* sí que lo sé! – se apresuró a decir Alicia–. Se pone harina…
– ¿***Arena***, dices? – Preguntó la Reina Blanca – .
– ¿Dónde la pones? ¿En el jardín o en la playa?
– No dije **arena**, sino **harina** – corrigió Alicia – y, propiamente, primero se muele el **grano**…
– ¡Moler el **grano**! – exclamó horrorizada la Reina Blanca–. ¿De la cara? ¡Qué método más salvaje! (Maristany 1986: 259–260)

c. – ¡Iso seino! –exclamou Alicia moi animada–. Móese un pouco de **fariña**…
– ¿E como moves a **fouciña**? – preguntou a Raíña Branca–. ¿De esquerda a direita ou de direita a esquerda?
– Non a ***moves***; a ***moes*** no moiño. (Barro & P. Barreiro 1985: 160)

In (6) and (7), and whenever this technique is used, the translator decides to sacrifice (part of) the semantic scenario and to preserve the pragmatic one, in such a way that the cognitive effects derived from the processing of wordplay may be accessible to the TT audience without gratuitous processing effort. Judging from the translation technique chosen by the translators in these cases, they must have decided that, rather than the specific meanings communicated by the ST puns, what was really relevant was the presence of wordplay in the ST. The degree of interpretive resemblance in this case was lower than in the case of the previous technique.

3.2.2 Punoid

Occasionally, the translator decides to tackle the translation problem which constitutes the object of this study by means of the resort to some type of rhetorical device, such as rhyme, alliteration, repetition, etc. Delabastita (1993, 1994) brings together all those devices under the term *Punoid*. In (8) the ST pun is a phonologic one based on the homonymy between *well* as a noun, meaning "a deep hole that is dug in the ground to provide a supply of water",[18] and as an intensifying adverb. The ST puns in (9) and (10) are both phonologic puns based on homonymy, respectively between the noun *miss* ("[a] form of address to a (usually young) woman"[19]) and the verb *miss* ("[n]ot to be in time for"[20]) and between the noun *mine* – "[a]n excavation in

[18] http://www.macmillandictionary.com/dictionary/british/well_61

[19] http://www.oed.com/view/Entry/119940?rskey=JoNqf3&result=2&isAdvanced=false#eid

[20] http://www.oed.com/view/Entry/119943

the earth for extracting coal or other minerals"[21] – and the pronoun *mine* -"[u]sed to refer to a thing or things belonging to or associated with the speaker"[22] –. The rhetorical devices used by translators have been rhyme in (8) and (9) and alliteration – specifically of nasal sounds – in (10).

(8) "But they were in the well," Alice said to the Dormouse, not choosing to notice this last remark.
"Of course they were," said the Dormouse: "**well in**."
This answer so confused poor Alice, that she let the Dormouse go on for some time without interrupting it. (Carroll 2000/1865: 59)
«Pero ¡es que estaban dentro del pozo!», insistió Alicia dirigiéndose al Lirón y no queriendo darse por enterada del calificativo que le acababa de propinar el Sombrerero.
«Pues claro que estaban **dentro, ¡y bien en el centro**!», declaró el Lirón. Esta contestación dejó a Alicia tan aturdida que no volvió a interrumpir al Lirón durante algún rato. (Ojeda 1976: 126)

(9) "That would never do, I'm sure," said Alice: "the governess would never think of excusing me lessons for that. If she couldn't remember my name, she'd call me '**Miss**,' as the servants do."
"Well, if she said '**Miss**,' and didn't say anything more," the Gnat remarked, "of course you'd **miss** your lessons. That's a joke. I wish *you* had made it." (Carroll 2000/1871: 114)
No está tan claro –repuso Alicia–. La institutriz encontraría la manera de salvar esa dificultad..Se inventaría algún nombre para llamarme… Diría, por ejemplo, ¡Venga aquí… señor**ita**!
– Pues entonces tú le contestas: ¿Dice usted que hay… vis**ita**?. ¡Pues entonces no hay clase! –exclamó el Mosquito–. Bueno… ¿Qué te ha parecido el chiste? ¡Se te podía haber ocurrido a ti! (Buckley 2005: 274)

(10) "there's a large mustard-**mine** near here. And the moral of that is—'The more there is of **mine**, the less there is of yours.'" (Carroll 2000/1865: 68)
aquí pretiño hai u**nh**a gra**n** mi**n**a de **m**ostarda. E a lección **m**oral diso é … "Ca**n**ta **m**áis hai **n**a **m**iña mi**n**a **m**e**n**os haberá **n**a túa." (Barro & P. Barreiro 2002: 123)

As highlighted by Marco (2010), this technique – which he calls *pun → related rhetorical device* – "implies using some kind of rhetorical compensation for the loss of the pun proper — even though the borderline between the pun proper and such devices as rhyme or alliteration is far from clear-cut" (Marco 2010: 280). From a relevance-theoretic perspective, it could be said that by means of the resort to both rhyme and alliteration some of the cognitive effects derivable from the processing of puns have been reproduced in the TT, particularly those related to using language in a playful way. In this sense, the translator – in this technique as well as in the previous one – has given prevalence to the pragmatic scenario over the semantic one in accordance to what a translator should do in Yus's (2012: 130) opinion.

[21] http://www.oxforddictionaries.com/definition/english/mine#mine-2

[22] http://www.oxforddictionaries.com/definition/english/mine#mine

3.3 Preservation of the Semantic Scenario

On some other occasions, no pun appears in the TT and consequently the pragmatic scenario is sacrificed. The semantic scenario, however, is often (partially) preserved, as shown in 3.3.1 and 3.3.2.

3.3.1 Sacrifice of Secondary Information

In (11) there is a vertical polysemic pun on *head*, which simultaneously means both "[t]he uppermost part of the body of a human" (s1) and "*Brit. colloq.* A postage stamp" (s2).[23] According to different studies in the field of lexical pragmatics, the meanings of words are very often pragmatically adjusted and fine-tuned in context. As stated in Wilson & Carston (2007: 238), a theory of lexical pragmatics within the framework of Relevance Theory can account for pun-like cases, such as this one, by saying that the interpretation of the noun *head* in this context involves the construction of an ad hoc concept HEAD* whose denotation includes both s1 and s2.[24] Of those two senses contained in the ST pun, however, only the first one is retained in the Galician TT.

(11) "She must go by post, as she's got a **head** on her–" (Carroll 2000/1871: 219)
Ten que ir por correo, que leva unha cabeza… (Barro & P. Barreiro 1985: 59)

Likewise, in (12), the Spanish TT keeps only one of the two senses reflected in the paronymic pun on *Laughing and Grief* and *Latin and Greek*. This pun corresponds to the third type devised by Yus (2003: 1323), in which "[n]o interpretation consistent with the principle of relevance is reached (initially) due to absurd and/or nonsensical punning associations. Only the reliance on a mutually manifest joking intention keeps the hearer searching for a relevant interpretation." Nevertheless, the translator decided in this case to keep the initial interpretation in a non-punning textual fragment, which gives rise to a textual fragment which does not make much sense.[25]

(12) "He taught **Laughing and Grief**, they used to say." (Carroll 2000/1865: 71)
Creo que enseñaba la Risa y la Pena. (Alba 1982: 52)

The translator must have considered that the presence of wordplay is not relevant enough to demand an extra processing effort from the TT receiver. The cognitive

[23] http://www.oed.com/view/Entry/84896?rskey=WlOGX6&result=1&isAdvanced=false#eid

[24] For further discussion of ad hoc concept construction within the relevance-theoretic view of utterance understanding, see for instance Carston (2002a, b), Wilson & Carston (2006), or Wilson and Sperber (2004). The standard practice represents ad hoc concepts as starred concepts, e.g. FIND*.

[25] Although the approach adopted in this study is an unevaluative and descriptive one, for translation assessment or evaluation, the reader is referred to Hrala (1994), Cámara Aguilera (1999), or Vázquez et al. (2011).

effects derived from the processing of wordplay in this case would not offset the extra processing effort the TT recipient would have to invest, the translator must have decided. Therefore, the application of this technique could also be explained by the principle of relevance. Other intervening factors, however, should not be disregarded, such as the translator's unawareness of the existence of wordplay, his/her inability to find a punning solution, or his/her personal attitude towards punning in general, among others.

3.3.2 Non-selective Translation

Although both techniques coincide in the preservation of the semantic scenario to the detriment of the pragmatic one, in the non-selective translation, contrary to the technique presented in 3.3.1, both of the semantic layers of the ST pun are kept in the TT extract corresponding to it. Unlike the previous technique, the non-selective translation seems to indicate that the translator must have thought that both meanings are equally relevant, and therefore, has decided to convey the two of them to the TT. This implies an increase in the interpretive resemblance between ST and TT. All the originally intended cognitive effects derivable from the processing of a pun, however, will not be accessible to the TT audience in this case either.

In (13a) the signifier *draw* contains two different meanings, namely s1: "[t]o cause (anything) to move toward oneself by the application of force; to pull" and s2: "[t]o make (a picture or representation of an object) by drawing lines; to design, trace out, delineate".[26] Both meanings are present in the TTs in the sequences *dibujar, sacando* and *sacar debuxos*, as *dibujar* and *debuxar* are respectively the Spanish and Galician verbs for s1 and *sacar* is both Spanish and Galician for s2.

(13) a. "And so these three little sisters—they were learning to **draw**, you know—"
"What did they **draw**?" said Alice, quite forgetting her promise.
"Treacle," said the Dormouse, without considering at all, this time.
(Carroll 2000/1865: 58)

b. Así pues, nuestras tres hermanitas… estaban aprendiendo a dibujar, sacando…
– ¿Qué sacaban? – preguntó Alicia, que ya había olvidado su promesa.
– Melaza -contestó el Lirón, sin tomarse esta vez tiempo para reflexionar.
(El Cid 2009: 104)

c. – Pois logo estas tres irmás… que estaban aprendendo a sacar debuxos, sacaron…
– ¿O que sacaron?– dixo Alicia, que esquecera xa que dera palabra de estar calada.
– Melaza– dixo o Leirón, desta vez sen pararse a pensalo. (Barro & Barreiro 2002: 102–103)

[26] http://www.oed.com/view/Entry/57534?rskey=5GDRcp&result=2&isAdvanced=false#eid

3.4 Preservation of the Cultural Scenario

3.4.1 Transference

By means of the *Transference* technique, a TT word or sequence acquires the meaning associated with its counterpart in the ST, even if it is not its normal meaning.[27] That is the case of *catching a crab* and its calqued translation *apañar un caranguexo* in (14). The ST sequence, via lexical broadening, also refers, in a figurative or loose sense, to "making a faulty stroke in rowing whereby the oar becomes jammed". Both the literal and the figurative interpretations are relevant, and in fact Alice's literal interpretation of the phrase gives rise to humour. Therefore, the resource to an ad hoc concept can explain the interpretation of the pun, as the verb phrase *catch a crab* – understood as CATCH A CRAB* – may be interpreted literally and figuratively at the same time. The sequence has been calqued or translated word for word into Galician as *apañar un caranguexo* (14b) and into Spanish as *coger un cangrejo* (14c), and these TT phrases have adopted the two meanings of the original pun explained above. The figurative meaning is explained in a footnote in the Galician version and in the Spanish version used to illustrate this technique.[28] By resorting to this editorial means, the translator makes sure that the target addressee interprets the phrase in the intended way.

(14) a. "You'll be **catching a crab** directly."
"A dear little crab!" thought Alice. "I should like that." (Carroll 2000/1871: 130)

b. "¡Que vas apañar un caranguexo!"
"¡Un caranguexiño pequeniño!," pensou Alicia. "¡Gustábame ben coller un!" (Barro & P. Barreiro 1985: 98)

c. O no tardarás en coger un cangrejo.
«¡Un cangrejito encantador!», pensó Alicia. «Me encantaría». (Torres Oliver 1984: 240)

Similarly, the English proverb *a cat may look at a king*, which is used to indicate that "there are certain things which an inferior may do in presence of a superior",[29] is source of wordplay, as the sequence is also interpreted in a literal sense. The translators of the seven versions analysed have applied *Transference* to deal with this pun, as the English proverb has been calqued in all of them, even though the same proverb does not exist in Spanish or Galician. In four of the versions an editorial means is added to explain the meaning of the original proverb, whereas in the remaining three, which do not use editorial techniques, the translators may have decided that the meaning can be easily inferred from the context.

[27] This translation technique corresponds to literal translation in Hurtado (2001: 271) and in Molina & Hurtado (2002: 510). The example they provide to illustrate that technique is to translate *They are as like as two peas* as *Se parecen como dos guisantes*.

[28] The use of footnotes and other editorial means will be dealt with in Sect. 3.6.

[29] http://www.oed.com/view/Entry/28649?rskey=7svzq2&result=1#eid10062650

(15) a. "**A cat may look at a king**," said Alice. (Carroll 2000/1865: 64)
b. – Un gato puede mirar a un rey – dijo Alicia –. (Torres Oliver 1984: 109)
c. – "Un gato pode ollar para un rei —dixo Alicia—" (Barro & P. Barreiro 2002: 115)

3.4.2 Direct Copy

Direct copy involves, as its name indicates, a reproduction of the ST pun in the TT in its original form in the SL. It is a technique normally used when at least one of the ST pun semantic layers coincides with a cultural reference. In different typologies of techniques for the translation of cultural references, other names used instead of *Direct copy* are *exoticism* (Haywood et al. 2009), *loan* (Díaz-Cintas and Remael 2014), *repetition* (Franco Aixelà 1996), or *retention* (Pedersen 2011). The technique has been rarely used in our corpus, mainly when dealing with proper nouns which involve a pun,[30] as in (16) and (17). In the first case there is a pun on *hatter*, whereas in the second case the ST pun is on *L. C.*, the initials of Lorina Charlotte.

(16) "The other Messenger's called **Hatta**" (Carroll 2000/1871: 142)
El otro mensajero se llama Hatta. (Maristany 1986: 225)[31]
(17) "and their names were **Elsie**, Lacie, and Tillie;[11] and they lived at the bottom of a well—" (Carroll 2000/1865: 58)
– Había una vez tres hermanitas –empezó apresuradamente el Lirón–, que se llamaban Elsie, Lacie y Tillie, y vivían en el fondo de un pozo… (Torres Oliver 1984: 96)[32]

A more surprising case of the application of this technique can be found in (18), a TT fragment corresponding to the ST pun introduced in (9). Instead of devising a solution in the TL, in this case the translator has decided to reproduce the pun in the SL accompanied by a footnote.[33] The adoption of the *Direct copy* technique gives rise to an ungrammatical textual fragment in the TL in this case.[34] The adoption of this translation technique may produce a *barbarism* or "[a] translation error where the translator uses an inappropriate calque, borrowing, or literal translation that is perceived as foreign to the linguistic sensibilities of the target audience" (Delisle et al. 1999: 121), or an *Anglicism* – "[a] borrowing from English into another language" (Delisle et al. 1999: 118) –.[35] *Direct copy* involves an extreme case of foreignizing

[30] Proper nouns are considered cultural references (See in this respect, for instance, Franco Aixelà 1996: 59).

[31] The same strategy is used to deal with this name in the Spanish versions by Ojeda and El Cid Editor.

[32] The same technique is used to deal with this name in the Spanish version by Maristany.

[33] The footnote reads as follows: "juego de palabras con miss, señorita y to miss, perder o eludir la asistencia a las clases" (Alba 1982: 104) [pun on *miss*, form of address, and *to miss*, not to be in time for class; my translation].

[34] For different typologies of translation errors, see for instance Cámara Aguilera (1999: 99–145), Vázquez et al. (2011), or Diéguez (2001: 209).

[35] Vinay & Dalbernet (1958) include borrowing, calque and word for word translation as procedures of literal translation.

translation in Venuti's (1995: 20) terms, since the cultural values and, in this case, the language of the source or foreign culture are present in the TT. The linguistic and cultural difference of the ST is thus registered and the "cultural other is manifested" (Venuti 1995: 20), in such a way that the target reader is sent abroad. At the opposite end of the scale, *domestication* (Venuti 1995: 23) or *naturalization* (Jaskanen 1999: 44) involves replacing the ST cultural referent by a more local or accessible one.

(18) Nunca sucederá eso, estoy segura –dijo Alicia–. La institutriz nunca pensaría en no darme lecciones por eso. Si no pudiera recordar mi nombre, diría para llamar "¡Señorita!", como hacen los sirvientes.
Bueno, si dice *miss* y no dice nada más –observó el Mosquito– por supuesto tú *miss* tus clases. Esto es un chiste. Me gustaría que tú lo hubieras hecho. (Alba 1982: 104)

The fact that in this case the punning words have been left in the SL may also contribute to the creation of cognitive effects associated with different aspects of the source culture. Particularly in the case of *Direct copy*, but also in the case of *Transference*, the extent to which the target addressee derives cognitive effects intended by the source communicator will depend on the target addressee's knowledge of the English language and culture. In this sense, Martínez-Sierra (2010: 202–203) highlights the importance of shared knowledge of the world between the source and the target audience when translating humour. A higher quantity of existing assumptions shared by both audiences will increase the probability of obtaining an analogous degree of relevance to the target addressee. Zabalbeascoa (2005: 204), in this sense, mentions contrastive differences in the background knowledge of source and target audiences as one of the obstacles which will have to be overcome during the translation process.

The *Direct copy* technique represents a clear case of stimulus-oriented mode or s-mode, according to Gutt (2005), as the higher-order communicator – or translator in this case – reproduces another token of the original stimulus. In s-mode the target audience "is practically independent of the interpretive activities of the higher-order communicator" or translator (Gutt 2005: 38). The decisive factor which will determine how close the target receptor's interpretation gets to that of the source addressee is the extent to which s/he can have access to the originally intended context.

3.5 None of the Scenarios Preserved

3.5.1 Omission

In this case, the textual fragment which contains the original pun is simply omitted in the translation. This may imply a decision on the part of the translator that neither the pun nor the meanings realized by that pun are relevant enough to be rendered in the TT. The textual fragment which activates the pun between the noun *mine* and the possessive pronoun *mine* in the ST in (19) has disappeared from the TT. This

solution deprives the target audience of the possibility to access the cognitive effects related to the processing of wordplay and also those cognitive effects derivable from at least one of the meanings realized in the original pun.

The ST pun in (20) is a morphologic one, as it is based on a word-formation process, namely blending, by means of which two words are fused into an only word in such a way that their boundaries merge.[36] Thus, *galumphing* is a blend of *galloping* and *triumphing*. Carroll himself invented the term *portmanteau word* to refer to these new coinages.[37] This ST word, however, has no textual correspondence at all in the Galician TT.

(19) Hay una gran mina de mostaza cerca de aquí. Y la moraleja de esto es…
(El Cid 2009: 126)

(20) One, two! One, two! And through and through
The vorpal blade went snicker-snack!
He left it dead, and with its head
He went **galumphing** back. (Carroll 2000/1871: 102)
¡Zis, zas! ¡Zis, zas! A espada orzal
bateu de bris, cortou de bras,
deixouno morto no piñeiral,
levoulle a testa coas dúas mans. (Barro & P. Barreiro 1985: 35)

3.5.2 Diffuse Paraphrase

When a non-punning textual fragment corresponds to the ST pun, another possibility involves keeping neither of the meanings realized in the original pun. Following Delabastita (1993: 206), in these cases the TT is said to offer a *diffuse paraphrase* of the original. In (21) the ST offers an idiomatic pun on *much of a muchness*, but the counterpart of *muchness* in the Spanish TT excerpt is *maullido*, "miaow", which does not involve a pun and which reflects neither the idiomatic nor the literal meaning of the original sequence. In (22) the ST pun is based on the homonymy between the proper noun *Bill* and the common noun *bill* ("[a] note of charges for goods delivered or services rendered").[38] Neither of the semantic layers of this pun is reflected in the non-punning TT extract ("The Rabbit sends a Pancho down the chimney").

[36] This type of pun is very frequent in James Joyce's works, to the extent that it has been sometimes called Joycean pun, as explained by Gardner in one of his notes to his edition of *Through the Looking Glass and What Alice Found There*, when he says:

> *Portmanteau word* will be found in many modern dictionaries. It has become a common phrase for words that are packed, like a suitcase, with more than one meaning. In English literature, the great master of the portmanteau word is, of course, James Joyce. *Finnegans Wake* (like the *Alice* books, a dream) contains them by the tens of thousands (Carroll 2000/1871: 321; Editor's note).

[37] Humpty Dumpty says in *Through the Looking Glass and What Alice Found There*: "You see it's like a portmanteau— there are two meanings packed up into one word." (Carroll 2000/1871: 137)

[38] http://www.oed.com/view/Entry/18987?rskey=7ZE8F7&result=3&isAdvanced=false#eid

(21) "—that begins with an M, such as mousetraps, and the moon, and memory, and muchness—you know you say things are '**much of a muchness**' —did you ever see such a thing as a drawing of a **muchness**!" (Carroll 2000/1865: 59)
… todo lo que empieza con M, como: memoria, mostaza, minino, maullido… ¿Has visto alguna vez el dibujo de un maullido? (Alba 1982: 40)

(22) The Rabbit Sends in a Little **Bill** (Carroll 2000/1865: 36)
O Coello manda un Pancho pola cheminea abaixo[39] (Barro & P. Barreiro 2002: 55)

Neither the ST pun nor the senses it contains must have been considered relevant enough from the point of view of the translator to be reflected in the TT. Compared with the previous strategies, this one is that in which the degree of interpretive resemblance between ST and TT is lowest. None of the ST scenarios – pragmatic, semantic, or cultural – has been preserved in the TT.

3.6 Amplification: Editorial Means

Translators may also decide to make themselves visible by resorting to a technique known as amplification, which involves the inclusion of specifications which did not appear in the ST. As indicated by Hurtado Albir (2001: 269), editorial means, such as footnotes, can be considered as a special type of amplification. The term used by Franco Aixelà to refer to those cases which involve amplification and in which the explanation is not mixed with the text is *extratextual gloss* (Franco Aixelá 1996: 62). Under the general label of editorial means, several devices can be included, such as footnotes, endnotes, or commentaries about the translation by means of an introduction or epilogue. The editorial techniques used in the translations analysed in this study fulfil the functions of explaining or commenting on the ST pun, which the translator reproduces literally, paraphrases or explains. The footnote in (23), corresponding to the Galician version of *Alice's Adventures in Wonderland*, explains and paraphrases the ST pun.

(23) Melaza dise en inglés "treacle," que tamén significa antídoto (da mesma raíz grega, theriake, có galego *triaga*). As fontes medicinais de Oxfordshire chamábanse *treacle-wells*, ou «pozos de triaga». Vivien Greene, a muller de Graham Greene, que moraba en Oxford, foi a primeira en comunicarlle a Martin Gardner, o anotador de *Alicia*, que nos tempos de Carroll habia un deses pozos en Binwy, preto de Óxford. (Barro & P. Barreiro 2002: 102)
[*Melaza* is "treacle" in English, which also means antidote (from the same Greek root, theriake, as the Galician word *triaga*). The medicinal springs of Oxfordshire were called "treacle-wells." Vivien Greene, Graham Greene's wife, who lived in Oxford, was the first person to tell Martin Gardner, the annotator of *Alice*, that in Carroll's age there was one of those wells in Binwy, near Oxford.]

[39] *Pancho* is a hypocorism for *Francisco*.

In addition, the editorial means may explicitly reflect on the relationship between the ST and the TT, whether the latter contains a pun or not, as in (24):

(24) [N. del T.] En inglés *to draw* significa tanto "dibujar" como "sacar o extraer." La melaza la "sacaban" y a la vez la "dibujaban." Por más que se ha estrujado los sesos, el traductor no ha encontrado una palabra castellana que expresara el doble juego del inglés. (Buckley 2005: 177)
[In English "to draw" means "to sketch" as well as "to extract" or "to pull out." They both "pulled out" and "sketched" the treacle. However much he racked his brains, the translator has not found a Spanish word which reflected the English pun.]

With respect to the use of editorial techniques, Gutt (2000: 96) says that in those cases in which complete interpretive resemblance is not achieved, due for instance to linguistic differences between the two languages, strategies for preventing communicative failure may be resorted to. Thus, for instance, the translator may alert the audience to the problem and correct the difference by some appropriate means, such as footnotes, endnotes, comments on the text, and so on. The translator, of course, will have to consider in each case whether the correction will be adequately relevant to his or her audience. In other words, a decision will have to be taken as to whether the benefits derived from the correction or editorial technique will outweigh the processing effort required by it.

Jing (2010: 94), in turn, considers that the use of editorial means presents a number of disadvantages, since not only does it disrupt the smoothness of the TT, increasing the target reader's processing effort, but it also destroys the punning effect and fails to match the source writer's intention with the target reader's expectation. Therefore, in her opinion, this solution should be regarded as the last resort for the translation of puns. However, it should be remembered that editorial means are necessarily combined with other strategies, even with the creation of a pun in the TT. And in those cases in which the TT presents no pun, the editorial means may serve to explain the original pun or its lost sense, for the reader to become aware of the source writer's punning intention.

3.7 Compensation

Aware that on some occasions the cognitive effects derived from the processing of a pun in the ST will not be accessible to the target audience, the translator may also decide to offer a TT pun corresponding to a textual fragment which does not contain any pun or even with no textual counterpart at all in the ST.

The first case may be found in (25), where the Spanish word *pena* means both "penalty" and "pity."

(25) "She's under sentence of execution."
"What for?" said Alice.
"Did you say 'What a pity!'?" the Rabbit asked.
"No, I didn't," said Alice. "I don't think it's at all a pity. I said 'What for?'"
(Carroll 2000/1865: 63)
¡Está bajo **pena** de muerte!
– ¿Qué **pena**?– preguntó Alicia.
– ¿Has dicho "¡Qué **pena**!"?– le preguntó a su vez el Conejo.
– No, no he dicho eso– repuso Alicia–, porque a mí la Duquesa no me da ninguna **pena**… He querido decir ¿por qué le han dado esa **pena**? (Buckley 2005: 184)

The second possibility referred to above may be illustrated by means of (26). In this case, the TT contains a pun for which it is impossible to find any corresponding textual material in the ST, as happens with the inserted set phrase *a toda costa* in the following fragment, which plays on the meanings "along the whole coast" and "at any price":

(26) "The reason is," said the Gryphon, "that they *would* go with the lobsters to the dance." (Carroll 2000/1865: 73)
"La razón es," dijo el Grifo, "que querían bailar con las langostas ***a toda costa***…" (Ojeda 1976: 163)

Finally, extract (27) serves to illustrate both possibilities, since the ST contains a morphologic pun on *bough-wough* and *bough*, which has as a counterpart another morphologic pun on *fungar* ("to produce a continuous and dull sound, like the wind") and *fungueirazo* ("blow struck with a stick") in the Galician TT, but in addition the Galician extract contains two other puns. One of them is a polysemic one on the noun *paos* – with the semantic layers "sticks" and "blows" – and the other one, also polysemic, is triggered by the noun *leña*, meaning both "firewood" and "a beating". Of those two new TT puns, the first one corresponds to a non-punning ST fragment, whereas the second one does not correspond to any textual fragment at all.

(27) "It says 'Bough-wough!'" cried a Daisy. "That's why its branches are called boughs!" (Carroll 2000/1871: 104)
– E como funga co vento, pode dar fungueirazos –berrou unha Margarida– e máis dá **paos** tamén, e **leña**. (Barro & P. Barreiro 1985: 42)

The point of the resort to this strategy is to make accessible to the target audience those cognitive effects which are derivable from the processing of wordplay and which in many other cases have been lost in the TT. Even if the use of this strategy has as a consequence an increase in the processing effort demanded from the target reader, that additional effort will be compensated for by the creation of new cognitive effects. This is particularly relevant if we take into consideration that the target audience has been deprived of the possibility to access cognitive effects of the same type in many other fragments of the TT.

4 Results and Discussion

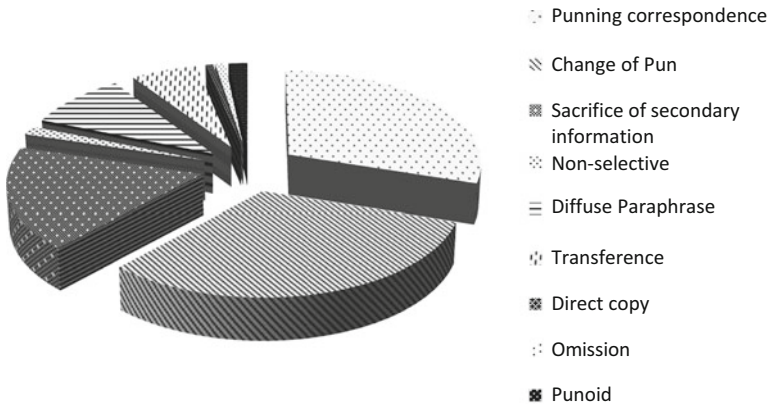
4.1 General Overview

As the results of this study reflect, the most widely used technique to translate puns in the corpus analysed is *Change of pun* – with 31.9 % –, followed closely by *Punning correspondence* – scoring 30.0 % –, which implies that more than half of the ST extracts containing puns, exactly 61.9 %, have been translated by means of textual fragments which also contain wordplay (See Table 2 and Graph 1).

Techniques which have preserved the semantic scenario at the expense of the pragmatic one, such as *Sacrifice of secondary of information* and *Non-selective translation*, have reached much lower percentages, respectively 17.8 % and 1.5 %.

Table 2 Distribution of translation techniques in the whole corpus

Translation technique	N	%
Punning correspondence	288	30.0
Change of pun	306	31.9
Sacrifice of secondary information	171	17.8
Non-selective translation	14	1.5
Diffuse paraphrase	90	9.4
Transference	58	6.1
Direct copy	7	0.7
Omission	9	0.9
Punoid	16	1.7
TOTAL	959	100



Graph 1 Use of translation techniques in the corpus

Meanwhile, techniques which have preserved none of the three scenarios, such as *Diffuse paraphrase* or *Omission*, have not scored high: 9.4 % and 0.9 % respectively. Source-culture oriented solutions, namely *Transference* – 6.1 % – and *Direct copy* – 0.7 % –, were not frequently used by the translators either.

Other solutions which are more oriented towards the preservation of the pragmatic scenario, as they try to recreate part of the cognitive effects derivable from the processing of wordplay, namely *Punoid*, were chosen by the translators only in 1.7 % of the cases.

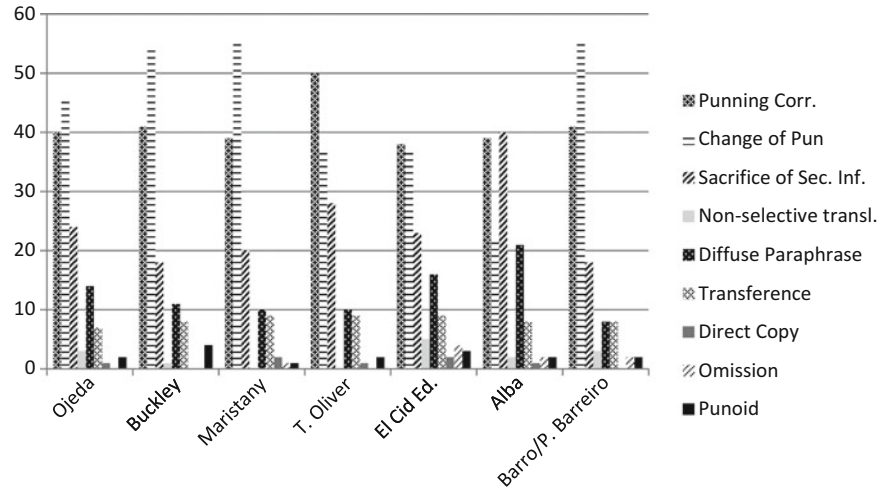
4.2 Variables

4.2.1 Version

There is a relation between the version variable and the choice of translation technique, as proved by the chi-square statistical test applied to the data. Thus, although the technique which ranks highest in the seven versions analysed involves the presence of wordplay in the TT, in four of them (those by Ojeda, Buckley, Maristany, and Barro & P. Barreiro) the first technique is *Change of pun*, whereas in the other three, the solution which occupies the first place is *Punning correspondence* (See Table 3 and Graph 2). In addition, if we add the percentages corresponding to those solutions which involve TTs containing puns (3.1 and 3.2.1), in Barro & P. Barreiro's translation the result amounts to 70.1 % whereas in Alba's it is only 44.5 %. In versions such as those by Buckley, Barro & P. Barreiro, Maristany, or Ojeda, then, there is a very clear tendency towards the preservation of the pragmatic scenario, which is much less clear in versions such as that by Alba.

Table 3 Distribution of translation techniques across versions

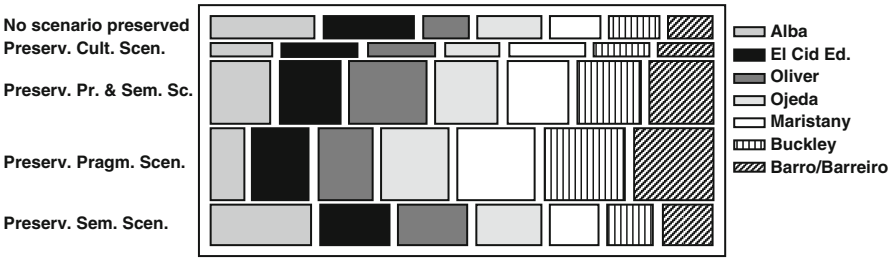
	Ojeda (Alianza)		El Cid editor		Buckley (Cátedra)		Maristany (Plaza and Janés)		T. Oliver (Akal)		Alba (Porrúa)		Barro & P. Barreiro (Xerais)	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%
P.C.	40	29.2	38	27.7	41	29.9	39	28.5	50	36.5	39	28.5	41	29.9
C.P.	46	33.6	37	27.0	54	39.4	55	40.1	37	27.0	22	16.0	55	40.2
S.S.I.	24	17.5	23	16.8	18	13.1	20	14.6	28	20.4	40	29.2	18	13.1
N.S.T.	3	2.2	5	3.6	1	0.7	0	0	0	0	2	1.5	3	2.2
D.P.	14	10.2	16	11.7	11	8.0	10	7.3	10	7.3	21	15.3	8	5.8
T.	7	5.1	9	6.6	8	5.8	9	6.6	9	6.6	8	5.8	8	5.8
D.C.	1	0.7	2	1.5	0	0	2	1.5	1	0.7	1	0.7	0	0
O.	0	0	4	2.9	0	0	1	0.7	0	0	2	1.5	2	1.5
P.	2	1.5	3	2.2	4	2.9	1	0.7	2	1.5	2	1.5	2	1.5
TOTAL	137	100	137	100	137	100	137	100	137	100	137	100	137	100



Graph 2 Distribution of translation techniques across versions

Table 4 Results of the chi-square test for translation technique by version

Test	Statistic	Df	P-value
Chi-square	49.963	24	0.0014



Graph 3 Mosaic plot for translation strategy by Version

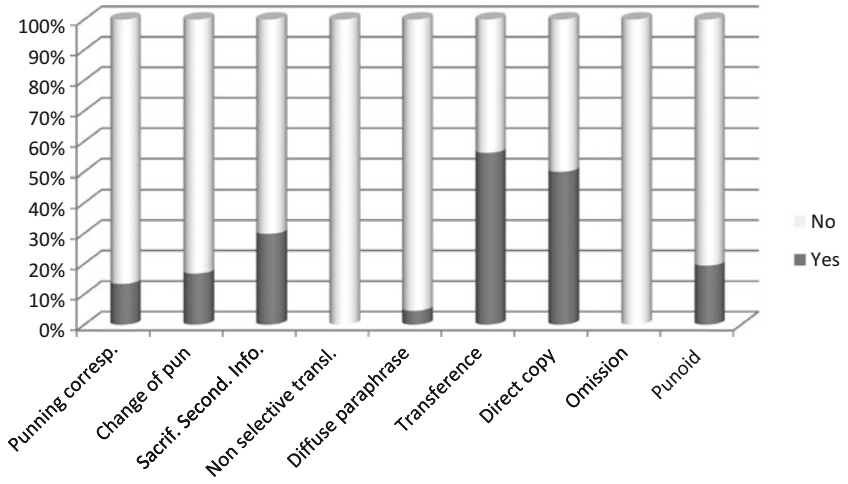
The chi-square statistical test reflects that there is a relation between choice of translation technique and version.[40] Since the P-value is less than 0.05, the hypothesis that choice of translation technique and version are independent – also called null hypothesis – can be rejected at the 95 % confidence level (See Table 4 and Graph 3).

[40] For this application of the chi-square test, as well as for the other two, translation techniques have been grouped under the categories corresponding to Sects. 3.1, 3.2, 3.3, 3.4, and 3.5, in order to endow the results of the test with more reliability.

4.2.2 Use of Editorial Means

As regards the use of editorial means, translators resorted to them in 107 instances, which accounts for 11.2 % in the whole corpus. However, two of the versions include no editorial means at all, whereas another one only includes 3 footnotes. If only the four versions which regularly include editorial techniques are considered, the percentage rises to 19.0 %. These editorial techniques are always used in combination with some other technique, and in this sense, a relation can be established between choice of translation technique and resort to editorial means. In other words, certain types of translation techniques seem to require the presence of an explanatory editorial means more than others. Thus, when *Transference* or *Direct copy* are used to deal with wordplay in the TT, the translation technique is accompanied by some type of editorial means in respectively 56.2 % and 50.0 % of the instances, whereas in the case of *Punning correspondence*, the percentage goes down to 13.4 % (See Graph 4).[41] This is logical, since if the translator decides to resort to a foreignizing technique, such as *Transference* or *Direct copy*, it is because he wants the target addressee to recover the cognitive effects intended by the original author, but as those techniques demand a certain background knowledge of the source culture and/or language by the target addressee, the translator often decides to provide some assumptions to ensure a comprehension as accurate as possible. In the case of *Punning correspondence*, more often than not the inclusion of an editorial technique would burden the target addressee with extra processing effort which would not be compensated for by additional cognitive effects.

As reflected in Table 5, the chi-square test also proves the interdependence between the two variables considered, namely use of editorial means and choice of



Graph 4 Use of editorial techniques across translation techniques

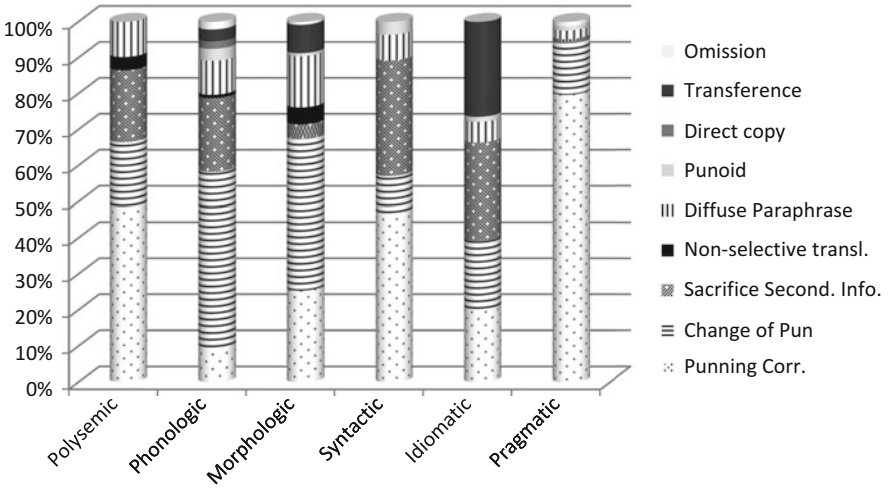
[41] Percentages correspond to those versions which regularly include editorial techniques.

Table 5 Results of the chi-square test for editorial technique by translation technique

Test	Statistic	Df	P-value
Chi-Square	44.766	4	0.0000

Table 6 Results of the chi-square test for translation technique by type of pun

Test	Statistic	Df	P-value
Chi-square	385.790	20	0.0000



Graph 5 Distribution of translation strategies across types of pun

translation technique. The null hypothesis can be rejected at the 95 % confidence level, as the P-value is less than 0.05.

4.2.3 Type of Pun

The linguistic device giving rise to the ST pun has also been found to have a clear effect on the choice of translation technique. In other words, the type of pun and the selection of translation technique variables have proved to be clearly related, as demonstrated by the chi-square test (See Table 6) and displayed in Graph 5. Thus, cross-linguistic differences demanded the recreation of a new pun in the TT for ST phonologic or morphologic puns. In the case of phonologic puns, two homophonous words in the SL, for instance, will not normally be homophonous in the TL, which requires a *Change of pun* solution if the pragmatic scenario is to be preserved. With regard to idiomatic puns, idiomatic expressions and proverbs are normally very closely linked to culture-specific aspects, and if translated literally into the TL will not normally keep the same meaning. It is precisely in this type of pun that *Transference* has been more widely used. As an exact equivalent of the original

idiom or proverb does not exist in the TL, translators very often decide to accompany this technique by some type of editorial means.

5 Conclusions

Within the framework of Relevance Theory and focusing on wordplay translation, the different techniques used by the translators of seven different versions of *Alice's Adventures in Wonderland* and *Through the Looking Glass and What Alice Found There* have been analysed throughout this paper. In general terms, it can be stated that the translators, guided by the principle of relevance, have tried to make the ST-writer-intended cognitive effects accessible to the target audience at minimal processing cost.

Those cases in which there was symmetry in the relation between form and content across SL and TL were normally taken advantage of by the translators. Thus, after metarepresenting the source writer's and target reader's cognitive environments, the translators normally decided to reproduce a congenial pun in the TT. However, more often than not there was lack of symmetry across SL and TL. The translator, then, had to decide whether prevalence should be given to the pragmatic scenario or to the semantic one. In general, they decided to maintain the pragmatic scenario and create a new pun in the TT, so that the cognitive effects derivable from the processing of wordplay could also be accessible to the target reader. This latter alternative was normally adhered to, to such an extent that *Change of pun* was the solution most frequently adopted in the whole corpus.

In spite of that general tendency, there were significant differences with respect to the choice of technique across the seven versions, as proved by the chi-square tests applied to the data. Whereas some of them were very clearly oriented towards the pragmatic scenario, in other versions – especially in that by de Alba – this orientation was not so clear.

Apart from the version variable, the type of pun one was also found to affect the choice of translation technique. Thus, for instance a polysemic pun is much more likely to be maintained in the TL if translated literally than a phonologic pun, which has favoured the use of *Punning Correspondence* for puns based on polysemy and the choice of *Change of pun* for puns rooted in phonology. *Transference*, in turn, is a translation technique mainly used to deal with idiomatic puns.

Moreover, in those versions in which editorial techniques were resorted to, their use was clearly related to translation technique selection. This finding can also be explained by means of the principle of relevance, as depending on the translation technique used, translators occasionally decide that the higher processing effort demanded by an editorial means may be outweighed by the additional cognitive effects derived from it.

Acknowledgements I would like to thank the anonymous reviewers of this article for their very valuable suggestions on earlier drafts of this paper. Needless to say, any remaining shortcomings are my only responsibility.

References

Alves, F., & Gonçalves, J. L. (2003). A relevance theory approach to the investigation of inferential processes in translation. In F. Alves (Ed.), *Triangulating translation: Perspectives in process oriented research* (pp. 3–24). Amsterdam/Philadelphia: John Benjamins.

Alves, F., & Gonçalves, J. L. (2007). Modelling translator's competence: Relevance and expertise under scrutiny. In Y. Gambier, M. Shlesinger, & R. Stolze (Eds.), *Translation studies: Doubts and directions* (pp. 41–55). Amsterdam/Philadelphia: John Benjamins.

Alves, F., & Gonçalves, J. L. (2010). Relevance and translation. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of translation studies* (Vol. 1, pp. 279–284). Amsterdam/Philadelphia: John Benjamins.

Asimakoulas, D. (2004). Towards a model of describing humour translation: A case study of the Greek subtitled versions of *Airplane!* and *Naked Gun. Meta, 49*(4), 822–842.

Cámara Aguilera, E. (1999). *Hacia una traducción de calidad. Técnicas de revisión y corrección de errores*. Granada: Grupo Editorial Universitario.

Carroll, L. (1976). *Alicia en el país de las maravillas* (J. de Ojeda, Trans., 4th ed.). Madrid: Alianza.

Carroll, L. (1982). *Alicia en el país de las maravillas. Al otro lado del espejo* (A. de Alba, Trans., 5th ed.). México D.F.: México.

Carroll, L.. (1984). *Alicia anotada. Alicia en el país de las maravillas & A través del espejo* (F. Torres Oliver, Trans.). Móstoles: Akal.

Carroll, L. (1985). *Do outro lado do espello e o que Alicia atopou aló* (T. Barro & F. P. Barreiro, Trans.). Vigo: Xerais.

Carroll, L. (1986). *Alicia en el país de las maravillas. Alicia a través del espejo* (L. Martistany, Trans.). Barcelona: Plaza & Janés.

Carroll, L. 2000. *The annotated Alice. Alice's adventures in wonderland and through the looking glass,* Ed. Martin Gardner, 1960. New York: Norton.

Carroll, L. (2002). *As aventuras de Alicia no país das marabillas* (T. Barro & F. P. Barreiro, Trans.). Vigo: Xerais.

Carroll, L. (2003). *A través del espejo y lo que Alicia encontró del otro lado*. Santa Fe: El Cid Editor.

Carroll, L. (2005). *A través del espejo y lo que Alicia encontró del otro lado*. (R. Buckley, Trans., 7th ed.). Madrid: Cátedra.

Carroll, L. (2009). *Alicia en el país de las maravillas*. Miami: El Cid Editor.

Carston, R. (2002a). *Thoughts and utterances*. Oxford: Blackwell.

Carston, R. (2002b). Linguistic meaning, communicated meaning and cognitive pragmatics. *Mind & Language, 17*, 127–148.

Delabastita, D. (1993). *There's a double tongue*. Amsterdam: Rodopi.

Delabastita, D. (1994). Focus on the pun: Wordplay as a special problem in translation studies. *Target, 6*(2), 223–243.

Delabastita, D. (1996). Introduction. In D. Delabastita (Ed.), *The Translator, 2*(2): *Essays on punning and translation* (pp. 127–139). Manchester: St. Jerome.

Delisle, J., Lee-Jahnke, H., & Cormier, M. C. (Eds.). (1999). *Terminologie de la Traduction/ Translation Terminology/Terminología de la Traducción/Terminologie der Übersetzung*. Amsterdam/Philadelphia: John Benjamins.

Díaz-Cintas, J., & Remael, A. (2014). *Audiovisual translation: Subtitling* (2nd ed.). London/New York: Routledge.

Díaz-Pérez, F. J. (2013). The translation of wordplay from the perspective of Relevance Theory: Translating sexual puns in two Shakespearian tragedies into Galician and Spanish. *Meta, 58*(2), 279–302.

Díaz-Pérez, F. J. (2014). Relevance Theory and translation: Translating puns in Spanish film titles into English. *Journal of Pragmatics, 70*, 108–129.

Diéguez, M. I. (2001). Aciertos y errores en la traducción automática: Metodología de la enseñanza-aprendizaje de la traducción humana. *Onomazein, 6*, 203–221.

Dynel, M. (2010). How do puns bear relevance? In M. Kisielewska-Krysiuk, A. Piskorska, et al. (Eds.), *Relevance studies in Poland: Exploring translation and communication problems* (Vol. 3, pp. 105–124). Warsaw: Warsaw University Press.

Franco Aixelá, J. (1996). Culture-specific items in translation. In R. Álvarez & M. C. A. Vidal (Eds.), *Translation, power, subversion* (pp. 52–78). Clevedon: Multilingual Matters.

González Davies, M., & Scott-Tennet, C. (2005). A problem-solving and student-centred approach to the translation of cultural references. *Meta, 50*(1), 160–179.

Gutt, E.-A. (1990). A theoretical account of translation—Without a translation theory. *Target, 2*(2), 135–164.

Gutt, E.-A. (1991). *Translation and relevance: Cognition and context*. Oxford: Blackwell.

Gutt, E.-A. (1998). Pragmatic aspects of translation: Some relevance-theory observations. In L. Hickey (Ed.), *The pragmatics of translation* (pp. 41–53). Clevedon: Multilingual Matters.

Gutt, E.-A. (2000). *Translation and relevance: Cognition and context*. Manchester: St. Jerome.

Gutt, E.-A. (2004). Challenges of metarepresentation to translation competence. In E. Fleishcmann, P. A. Schmitt, & G. Wotjak (Eds.), *Translationskompetenz: Proceedings of LICTRA 2001: VII. Leipziger Internationale Konferenz zu Grundfragen der Translatologie* (pp. 77–89). Tübingen: Stauffenburg.

Gutt, E.-A. (2005). On the significance of the cognitive core of translation. *The Translator, 11*(1), 25–49.

Haywood, L., Thomson, M., & Hervey, S. (2009). *Thinking Spanish translation. A course translation method: Spanish to English* (2nd ed.). London/New York: Routledge.

Higashimori, I. (2011). Jokes and metarepresentations: Definition jokes and metalinguistic jokes. In W. J. Sullivan & A. Lommel (Eds.), *LACUS Forum 36: Mechanisms of linguistic behavior* (pp. 139–150). Houston: LACUS.

Hrala, M. (1994). Criteria of translation evaluation. In FIT Committee for Translation Criticism (Ed.), *Miscellany on translation criticism* (pp. 67–76). Prague: Institute of Translation Studies, Charles University.

Hurtado Albir, A. (1996). La traductología: lingüística y traductología. *Trans, 1*, 151–160.

Hurtado Albir, A. (1999). *Enseñar a traducir*. Madrid: Edelsa.

Hurtado Albir, A. (2001). *Traducción y Traductología. Introducción a la Traductología*. Madrid: Cátedra.

Jaskanen, S. (1999). On the inside track to Loserville, USA: Strategies used in translating humour in two Finnish versions of Reality Bites. Unpublished M.A. thesis, University of Helsinki.

Jing, H. (2010). The translation of English and Chinese puns from the perspective of Relevance Theory. *The Journal of Specialised Translation, 13*, 81–99.

Kosińska, K. (2005). Puns in relevance. In A. Korzeniowska & M. Grzegorzewska (Eds.), *Relevance studies in Poland* (Vol. 2, pp. 75–80). Warsaw: The Institute of English Studies, University of Warsaw.

Levý, J. (1969). *Die literarische Übersetzung: Theorie einer Kunstgattung*. Frankfurt: Athenäum.

Marco, J. (2004). Les tècniques de traducció (dels referents culturals): retorn per a quedarnos-hi. *Quaderns: Revista de traducció, 11*, 129–149.

Marco, J. (2007). The terminology of translation: Epistemological, conceptual and intercultural problems and their social consequences. *Target, 19*(2), 255–269.

Marco, J. (2010). The translation of wordplay in literary texts. Typology, techniques and factors in a corpus of English-Catalan source text and target text segments. *Target, 22*(2), 264–297.

Martínez-Sierra, J. J. (2010). Using relevance as a tool for the analysis of the translation of humor in audiovisual texts. In J. L. Cifuentes, A. Gómez, A. Lillo, J. Mateo, & F. Yus (Eds.), *Los caminos de la lengua. Estudios en homenaje a Enrique Alcaraz Varó* (pp. 189–209). Alicante: Universidad de Alicante.

Molina, L., & Albir, A. H. (2002). Translation techniques revisited: A dynamic and functionalist approach. *Meta, 47*(4), 498–512.

Pedersen, J. (2011). *Subtitling norms for television*. Amsterdam/Philadelphia: John Benjamins.

Rosales Sequeiros, X. (2002). Interlingual pragmatic enrichment in translation. *Journal of Pragmatics, 34*(8), 1069–1089.

Rosales Sequeiros, X. (2005). *Effects of pragmatic interpretation on translation. Communicative gaps and textual discrepancies*. Munich: Lincom Europa.

Sanderson, J. D. (2009). Strategies for the dubbing of puns with one visual semantic layer. In J. Díaz-Cintas (Ed.), *New trends in audiovisual translation* (pp. 123–132). Bristol: Multiligual Matters.

Seewoester, S. (2011). The role of syllables and morphemes as mechanisms in humorous pun formation. In M. Dynel (Ed.), *The pragmatics of humour across discourse domains* (pp. 71–104). Amsterdam/Philadelphia: John Benjamins.

Solska, A. (2012). Relevance-theoretic comprehension procedure and processing multiple meanings in paradigmatic puns. In E. Walaszewska & A. Piskorska (Eds.), *Relevance Theory: More than understanding* (pp. 167–181). New Castle upon Tyne: Cambridge Scholars Publishing.

Sperber, D., & Wilson, D. (1986). *Relevance. Communication and cognition*. Oxford: Blackwell.

Sperber, D., & Wilson, D. (1995). *Relevance. Communication and cognition* (2nd ed.). Oxford: Blackwell.

Tanaka, K. (1992). The pun in advertising: A pragmatic approach. *Lingua, 87*, 91–102.

Tanaka, K. (1994). *Advertising language. A pragmatic approach to advertisements in Britain and Japan*. London/New York: Routledge.

Van Mulken, M., van Enschot-van Dijk, R., & Hoeken, H. (2005). Puns, relevance and appreciation in advertisements. *Journal of Pragmatics, 37*(5), 707–721.

Vázquez, E., Martínez, R. I., & Ortiz, J. (2011). *Errores de reproducción y trasmisión de sentido en traducción general y especializada (inglés/árabe-español): La experiencia en el aula de la universidad*. Granada: Universidad de Granada.

Venuti, L. (1995). *The translator's invisibility*. London/New York: Routledge.

Vinay, J.-P., & Dalbernet, J. (1958). *Stylistique Comparée du Français et l'Anglais*. Paris: Didier-Harrap.

Weissbrod, R. (1996). 'Curiouser and curiouser': Hebrew translations of wordplay in *Alice's Adventures in Wonderland.* In D. Delabastita (Ed.) *The Translator, 2*(2): *Essays on punning and translation*. (pp. 127-139). Manchester: St. Jerome.

Wilson, D. (2012). Metarepresentation in linguistic communication. In D. Wilson & D. Sperber (Eds.), *Meaning and relevance* (pp. 230–257). Cambridge: Cambridge University Press.

Wilson, D., & Carston, R. (2006). Metaphor, relevance and the 'emergent property' issue. *Mind & Language, 21*(3), 404–433.

Wilson, D., & Carston, R. (2007). A unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. In N. Burton-Roberts (Ed.), *Pragmatics* (pp. 230–259). Basingstoke: Palgrave Macmillan.

Wilson, D., & Sperber, D. (2004). Relevance Theory. In L. R. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 607–632). Oxford: Blackwell.

Yus, F. (2003). Humour and the search for relevance. *Journal of Pragmatics, 35*, 1295–1331.

Yus, F. (2008). A relevance-theoretic classification of jokes. *Lodz Papers in Pragmatics, 4*(1), 131–157.

Yus, F. (2012). Relevance, humour and translation. In E. Walaszewska & A. Piskorska (Eds.), *Relevance Theory: More than understanding* (pp. 117–145). New Castle upon Tyne: Cambridge Scholars Publishing.

Zabalbeascoa, P. (2004). Translating non-segmental features of textual communication: The case of metaphor within a binary-branch analysis. In D. Gile, G. Hansen, & K. Malmkjar (Eds.), *Claims, changes and challenges in translation studies* (pp. 99–111). Amsterdam/Philadelphia: John Benjamins.

Zabalbeascoa, P. (2005). Humor and translation – An interdiscipline. *International Journal of Humor Research, 18*(2), 185–207.

Zhonggang, S. (2006). A Relevance Theory perspective on translating the implicit information in literary texts. *Journal of Translation, 2*(2), 43–60.

Connective Items in Interpreting and Translation: Where Do They Come From?

Bart Defrancq, Koen Plevoets, and Cédric Magnifico

Abstract This paper presents corpus-based research into the use of connective items (*so, but, therefore,…*) by English and Dutch translators and interpreters with a view to determining (1) the relationship with connective items in the French source text that translators and interpreters are faced with; (2) the similarities and differences between translations and interpretations with regard to connective items and the way they relate to the source text; (3) the extent to which translations favour written features and interpretation spoken features of the target languages. The corpus data used in this study is drawn from a recently compiled corpus of interpretations and translations carried out at the European Parliament. The research shows that the approaches taken by interpreters and translators differ substantially: interpreters, regardless of the language they interpret into, use a broader range of translation options, omit more than translators, but – surprisingly - also add more items. A qualitative study of the additions reveals that interpreters use connective items to make clausal relations explicit, but also to connect on-coming clauses after substantial omissions or when facing processing difficulties.

Keywords Connectives • Parallel intermodal corpus • Interpreting • Translation • Additions

1 Introduction

This paper presents the results of a study on the use of connective markers by translators and simultaneous interpreters. Such a study is relevant for both translation and interpreting studies on a number of grounds. On the epistemological level, both disciplines can cross-fertilize, or as Shlesinger and Ordan (2012) put it:

> […] translation scholars can learn about the process and product of (written) translation by finding out more about interpreting – and interpreting scholars can infer about this

B. Defrancq (✉) • K. Plevoets • C. Magnifico
EQTIS, Department of Translation, Interpreting and Communication,
Ghent University, Groot-Brittanniëlaan, 45, 9000 Ghent, Belgium
e-mail: bart.defrancq@ugent.be

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_9

> high-pressure form of translation by observing the slower, more readily observable process and product of (written) translation; that one modality can teach us about the constraints, conventions and norms of the other; and that corpora of interpreted texts may teach us about the workings of oral vs. written discourse, both original and translated. (Shlesinger and Ordan 2012: 44, with reference to Chesterman 2004 and Pochhacker 2004)

The important questions in this regard are, of course: what properties do translation and interpreting share, and what are the specific properties of each discipline? And, how can we relate the discipline-specific properties to the circumstances in which both activities develop and, more specifically, the properties alluded to by Shlesinger and Ordan (2012)? We will try to answer these questions within the limits of our object of study and of a corpus-based methodology.

The discipline of translation studies has a long tradition of reflecting on the nature of translation. Corpus studies in the field of translation have contributed to that reflection by pointing at a number of textual properties which appear to be typical of translated text, as compared to non-translated text in the same language and to the source text from which it is translated. Translations were shown to be subject to source text interference (Mauranen 2004), to be more cohesive (Blum-Kulka 1986; Olohan and Baker 2000), to be lexically simpler (Laviosa 1998), etc. The field of interpreting studies has been less concerned with this research agenda, probably because the collection of interpreting corpora is infinitely more difficult than that of translation corpora (Shlesinger 1998). However, in recent years, a number of scholars have turned to interpreting corpora to check whether simultaneous interpreting presents the same tendencies as translation. For the time being, only collocations and lexical simplicity seem to have drawn their attention (Kajzer-Wietrzny 2012, Forthcoming; Bernardini 2014, Forthcoming). This study will take a look at a more pragmatic property, i.e. cohesion in the output of simultaneous interpreters.

Although translation and interpreting have a lot in common, the practitioners of both disciplines work in very different circumstances. There is of course the difference in modality (written vs. spoken), which has received much attention in corpus linguistics (Biber 1988 among many others), but has only just begun to be explored in corpus-based translation and interpreting research (Shlesinger 2009; Shlesinger and Ordan 2012). Research based on anecdotal evidence and experimental methods has shown that simultaneous interpreting presents typical features of spoken language, reflecting the limitations of on-line speech planning, such as disfluencies and self-repairs (Cecot 2001; Gile 1995). More importantly, the differences regarding modality are likely to be exacerbated within the translation-interpreting pair for two main reasons: (1) simultaneous interpreters are subject to a considerably higher cognitive load than spontaneous speakers (Gile 1997, 2008; Seeber 2011, 2013); (2) simultaneous interpreters are not entirely in control of their own production (Gile 1995): as they typically follow the speaker at a 2–3 s interval, their production cannot be planned more than 4–5 words ahead. This "short-sightedness" and the high cognitive load are likely to have an impact on the cohesion of interpreters' output, as compared with translators' output.

The paper is structured in the following way: Sect. 2 presents a broad overview of the relevant literature on connective items. Sections 3 and 4 describe the data, methodology and results of the quantitative study, while Sect. 5 reports on a limited qualitative study of additions. The conclusions of the present study are presented in Section 6.

2 Review of the Literature

In Halliday and Hasan's (1976) terms, cohesion refers to "relations of meaning that exist within the text and that define it as a text" (Halliday and Hasan 1976: 4). The concept encompasses various linguistic phenomena ("cohesion ties"), whose common property is that they include units whose interpretation depends on other units in the discourse: pronouns, ellipsis, substitution, lexical items and conjunctions. We will focus our attention on the last category, more commonly known as "connective devices", such as *but, so, therefore, consequently*…

There are various reasons why the study of connective devices is relevant in translation and interpreting research: first of all, there is a vast amount of linguistic research on connective devices, including corpus-based contrastive research. Many of these items are fairly well described, including their frequencies in spoken and written corpora, showing that some items are more typical of spoken than of written registers and *vice versa* (see Sect. 3). For a study on spoken and written features of interpreting and translation, these data are of course particularly relevant.

Secondly, there is corpus-based translation research in the area of connective devices. Connective devices are among the first items to be considered instrumental in explicitation processes undertaken by translators. Blum-Kulka (1986), in fact, observes that in translations done by graduate research assistants, connectives are added to the translation. Other studies confirm Blum-Kulka's observation (Øverås 1998) and the related claim that some connectives are more frequent in translations than in original texts of the same language (Olohan and Baker 2000; Mauranen 2000; Puurtinen 2004). However, in more recent years, explicitation has come under fire, especially in the area of connectives. Saldanha (2008) and Becher (2010) point to conceptual and methodological problems in the research on explicitation, while illustrating their point with corpus data on connectives. Finally, in their paper on causal connectives in literary translation from Dutch into English and French, Vandepitte et al. (2013) observe a substantial amount (up to 25 %) of implicitation, i.e. omission, of connectives. So, although the evidence in the literature seems to support diverging claims, the least we can say is that connectives are key evidence in the debate on the nature of translation processes.

Thirdly, the interest of translation scholars in connectives has spilled over into interpreting studies: Shlesinger (2009) and Shlesinger and Ordan (2012) point out the frequent use of some adverbs – mainly connective adverbs – in simultaneous interpreting (versus written translation). The higher frequencies, they claim, prove that simultaneous interpreting has in fact more in common with spontaneous speech than with written translation.

Finally, there is fairly widespread belief in interpreting studies that pragmatic markers, including connective items are vulnerable in the interpretation process. Indeed, they do not, by definition, belong to the propositional content of the clauses they interrelate and are thought to be among the first victims of cognitive overload in the mind of the interpreter. Analysing dialogue interpretations carried out in courts, Hale (2004) and Mason (2008) observe that connective items are frequently omitted by interpreters, especially when interpreting long turns which require more memory capacity. As early as 1975, in the framework of an experimental study on errors and omissions in simultaneous interpreting, Barik (1975) decides to ignore omissions of connective and other pragmatic markers, because he considers them not to cause loss of meaning. Moreover, it has been claimed that simultaneous interpreters lack a structural overview of the text they are interpreting as they are expected to follow the source at an interval of only a couple of seconds (Gile 1995). As connective items are specifically geared towards discourse structure, the lack of structural overview may have a detrimental effect on their frequency. Shlesinger (1995) concludes that cohesive shifts in simultaneous interpreting occur and mainly consist of omissions of non-essential items in the area of connective items.

In sum, there is both monolingual and parallel corpus evidence from translation and interpreting that suggests that connective items are a promising field for the comparative study of translation and interpreting. Our research question is as follows: do translators and interpreters display convergent or divergent patterns of use in the area of clausal connectives? Drawing on previous findings, we will work with the following hypotheses:

- if explicitation is a defining feature of translation, it should also be found in interpreting;
- however, due to cognitive management problems and a lack of structural overview, interpreters may also be expected to omit quite a number of connectives;
- the first and second expectation are not contradictory in principle as the locus of explicitation and omission may be different;
- if simultaneous interpreting really is a spoken translation register, shifts should be observed towards the use of connectives that are more typical of spoken registers.

The connective items used in this study were selected from the causal and the concessive domains, as these are the two domains on which quantitative data from monolingual spoken and written corpora are available. We decided to focus on coordinate conjunctions and adverbial connectives only, as bringing in other causal or concessive devices (subordinators, particular verb forms) would have required us to take into account additional explanatory factors for translators' and interpreters' strategies (for instance *saucissonage* in the case of subordinating conjunctions, see Ilg 1978). In the next paragraphs, we briefly discuss the different items and their frequencies in spoken and written corpora, as reported in relevant studies. What should be kept in mind, though, is that spoken data usually include more connective devices than written data (Soria 2005).

2.1 French

There are quantitative data for the following causal items: *alors* ('then, so'), *donc* ('so'), *par conséquent* ('consequently'); and for the following concessive items: *quand même* ('yet'), *tout de même* ('yet'), *malgré tout* ('yet'), *cependant* ('however'), *toutefois* ('however'), *néanmoins* ('*nevertheless*') and *or* ('*however*'). In a comparative study on spoken and written registers, Teston and Véronis (2004) find that *alors, donc, quand même, malgré tout* are significantly more frequent in oral data than in written and that *cependant* and *par conséquent* are significantly more frequent in written registers. Schlamberger Brezar (2012) observes that in corpora of spoken data the frequencies of *mais* ('but') and *alors* are higher in the more spontaneous register (TV debate) than in official speeches and literary dialogues. Based on Gougenheim et al. (1964), Gettrup and Nølke (1984) state that *quand même, tout de même, pourtant* ('yet') are the only concessive markers that are frequent in spoken French, excluding *néanmoins, cependant* and *toutefois*. All of these items have been found in the corpus we used for this study. Other causal and concessive connectives that occur quite regularly are: *c'est/voilà pourquoi* ('that is why'), *c'est pour cela/ça que* ('that is why'), *dès lors* ('therefore, thus') and *c'est la raison pour laquelle* ('that is the reason why'). As no empirical study seems to have covered these items, we have decided not to classify them as typically spoken or written. An overview of the French connective items included in this study is given in Table 1.

2.2 English

Chafe (1987) reports that "[*a*]*nd, so* and *but* appear to be the three most common connectives in spoken English" (p. 42). Based on a comparison of the London Lund Corpus of Spoken English and the London Oslo Berger corpus of written data, Altenberg (1984) concludes that the causal connective *so* is considerably more frequent in spoken data than in written, while *therefore* and *thus* are typical of written registers, the latter not occurring once in the oral data. Regarding concessive items, Barth (2000) and Taboada and Gomez Gonzalez (2013) reach identical conclusions: while *but* is frequent in both spoken and written material, it almost has a monopoly on concessions in oral data. Barth (2000) also finds *yet* and *nevertheless* to be

Table 1 Connective items in French

	Oral	Written	?
Causal	*alors, donc*	*par conséquent*	*dès lors, c'est pourquoi, c'est pour cela, c'est la raison pour laquelle*
Concessive	*mais, pourtant, malgré tout, quand même, tout de même*	*néanmoins, cependant, toutefois, or*	

Table 2 Connective items in English

	Oral	Written	?
Causal	*so*	*therefore, thus*	*that is why, consequently, as a result*
Concessive	*but, though*	*however, yet, still*	*nevertheless, nonetheless*

marginally used in spoken data, while Taboada and Gomez Gonzalez (2013) report a higher relative frequency of *though* in oral than in spoken data. Other items, such as *however* or *still* are only found in written data. Adopting a more fine-grained approach to register, Biber et al. (1999) report that *so*, *though* and *anyway* predominate in conversation, while *therefore, thus* and *however* are among the most frequent connective items in academic prose. With the exception of *anyway*, all the items mentioned occur in our corpus. Additional causal and concessive items were found that could not be classified according to register. An overview of the English connective items is shown in Table 2.

2.3 Dutch

The data on Dutch are scarce. There is a considerable amount of corpus research into causal connectives, but focused on the semantic and pragmatic features of the items (for an overview, see Sanders et al. 2012) and not on their frequencies in spoken and written data. De Sutter (2010) observes that the relative frequency of *dus* ('so') is twice as high in spoken Dutch as in written Dutch. Based on a corpus study of legal Dutch, Van Noortwijk (1995) concludes that *derhalve* ('consequently') is typical of legal language and barely occurs in standard Dutch. Other items can only be classified according to normative sources. Van Belle (1996), for instance states that causal *bijgevolg* and concessive *echter* and *nochtans* are typical of written Dutch, while oral registers use *dus* ('so'), *maar* ('but') and *toch* ('though') more frequently. On the basis of these data, a tentative classification is proposed in Table 3.

Tables 1, 2, and 3 will be used as reference sets for the comparison of source and target texts in the corpus.

3 Data and Methodology

To analyse the use of connective items by translators and interpreters, we compiled a so-called parallel intermodal corpus (PIC), i.e. a corpus of authentic source and target texts, both spoken and written (interpreted and translated) and delivered in comparable contexts (Bernardini 2014). The PIC methodology is very recent: results from similar corpora are reported in Shlesinger 2009; Shlesinger and Ordan 2012 (oral/written); Kajzer-Wietrzny (2012, Forthcoming), Bernardini (2014,

Table 3 Connective items in Dutch

	Oral	Written	?
Causal	*dus*	*bijgevolg, derhalve*	*daardoor, daarom, dan ook*
Concessive	*maar, toch*	*echter, nochtans*	*(desal)niettemin,*

Table 4 Size of the different subcorpora

FR source speeches	FR verbatim reports	NL interpretations	NL translations	EN interpretations	EN translations	Total
31,471	30,456	26,606	29,604	28,196	28,816	175,149

Forthcoming). The corpus we used is based on data retrieved from sittings of the European Parliament held between 1st September 2008 and 21st October 2008. The original speeches and their interpretations are drawn from EPICG (*European Parliament Interpreting Corpus Ghent*), a larger corpus consisting of transcriptions of speeches held during plenary sessions of the European Parliament and of their interpretations. In its current shape, it comprises approximately 190,000 words, including mainly French and Spanish source texts and Dutch and English target texts. Source and target texts are transcribed according to the Valibel instructions (Bachy et al. 2007). The data used here are 39 French source texts and interpretations into Dutch and English. We opted for two target languages, in order to be able to assess whether data vary more cross-modally (between translations and interpretations) or cross-linguistically (between different languages).

For the purpose of this study, written data were collected from the verbatim reports of the same French source texts and their translations in Dutch and English. In all, the different subcomponents used for this study amount to 175,149 words distributed as shown in Table 4.

It is important to analyse the text cycle within the European Parliament to better understand how the different components relate to one another. Members of the European Parliament (MEPs) mostly prepare a written version of their speech, which they read out at the plenary. Video recordings of the speeches show MEPs standing up, holding a text which they look at most of the time while they speak. As they are offered little time to make their point, speeches are read at an excessively high delivery rate, which in some cases, goes up to 200 words per minute. Speeches delivered by members of the European Commission and the Council of the EU tend to be longer and slower. Impromptu speeches (i.e. speeches that are not read from a text) are extremely rare: in our sample of 39 speeches, only one can be considered an impromptu speech, delivered by a member of the European Parliament who can be seen in the video recording to speak without a text.

Speeches are interpreted into all the official languages of the EU. In 2008, the EU counted 23 official languages. French speeches were therefore interpreted at that time into 22 other languages. In most cases, interpreters work directly from the source speech. Relay interpreting is however also practiced, as booths can no longer

cover all the combinations of 22 C languages with the A language. However, relay interpreting is usually not necessary between the biggest languages (German, English, French, Italian and Spanish), neither between the historical languages, which, apart from some of the already mentioned languages, also include Dutch. It is therefore safe to say that the interpreters in the Dutch and English booths work completely independently and that the English and Dutch data used for this study are not in any way connected to one another. This is confirmed by the Ear-Voice-Span (EVS) practiced by the Dutch and English interpreters: in both booths EVSs are comparable and completely in line with EVSs observed in the literature on interpreting. If one of the booths worked in relay, the EVS would of course be substantially longer. According to Bernardini (forthcoming), who interviewed some EP interpreters, the written versions of the speeches are not available to the interpreters. Interpreters work on the basis of the oral input alone (apart from the general background knowledge they may have collected from written documents).

After the oral phase that takes place in the hemicycle, the texts go through a written phase. The speeches themselves are transformed into verbatim reports that are translated into the other official languages. Recordings of interpretations are not used in that process. Cross-influences from other translations are unlikely for the same reasons as the ones previously invoked: the language units covering the big languages and the historical languages normally translate directly from French. We cannot, however, completely rule out that translators also use the work of translators in other units as source texts in their work.

Schematically, the text cycle looks as shown in Fig. 1.

The use of data from the European Parliament has clear advantages and disadvantages. The main advantage lies in the near-laboratory conditions in which the data are produced: as the written and spoken source texts are very similar, the former being a verbatim report of the latter, translation and interpreting strategies are likely to stand out more clearly against this common background. On the other hand, it is important to note that the oral source data in the corpus are of a particular nature: as pointed out before, speeches held during plenary sessions of the

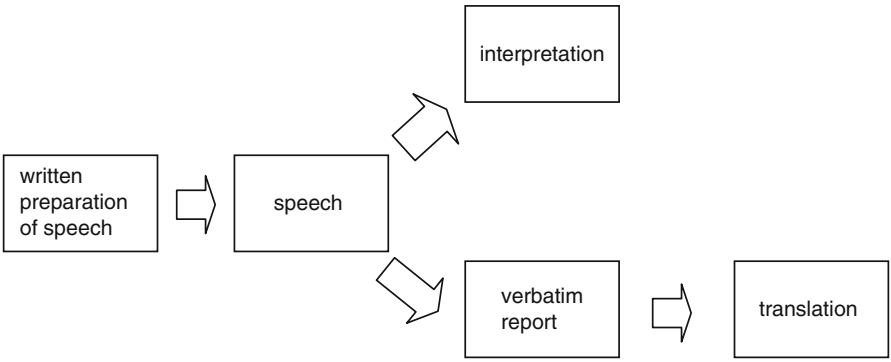


Fig. 1 Text cycle in the European Parliament

European Parliament are rarely impromptu, which is due to have an important impact on the pragmatic features of the source data. Source data features, including those of the oral version, are likely to be more typical of written language. Consequently, we considered it crucial to include a comparative analysis of both types of source texts in our comparative analysis of translation and interpreting. In addition, speeches are relatively short and delivered at particularly high delivery rates, forcing interpreters to work in ways that may not be entirely representative of standard interpreting practice.

Occurrences of the items described in Sect. 2 (Tables 1, 2, and 3) were identified in all subcorpora, source and target corpora alike. We made use of a monolingual concordancer (Wconcord), as the data in the different subcorpora are not (yet) aligned. For each occurrence, its context was retrieved, including the corresponding context in the source or target text. Occurrences of the items that do not have a connective function were removed from the data set. This was, for instance, the case of instances of restrictive *but* in combination with *nothing* and of the modal, attenuating use of *maar* in combination with imperatives: *doe maar* ('please do'). We then checked whether the corresponding source or target text contained any kind of marker that could be related to the occurrence of the relevant connective item that was retrieved. Even semantically different markers were identified as potential sources of translated items or as potential translations of a source item. Seven categories of translation options were distinguished:

1. A source item from Table 1 corresponds to a target item from either Table 2 or Table 3. In other words, one of the causal or concessive adverbs mentioned in Section 2 for French is translated by means of a causal or concessive adverb mentioned in Section 2 for English and for Dutch. We will call this category 'in-group equivalent'. Example (1), drawn from the speech and interpretation corpora, illustrates the use of *nonetheless* as an in-group equivalent of *toutefois* ('however'):

   (1a) il est / difficile de prendre la parole devant votre assemblée // **toutefois** / il nous faut en venir aux réalités
   [EPICG_08.10.08_preparationoftheeuropeancouncil2_jouyet_fr]

   (1b) it is very difficult to now get back down to business and to take the floor / in front of the chamber **nonetheless** we do have to come back down to solid ground
   [EPICG_08.10.08_preparationoftheeuropeancouncil2_jouyet_I_en]

2. A target item from Table 2 or 3 corresponds to a source item which is not mentioned in Table 1, but has a causal or concessive meaning. This category will be called 'equivalent in source'. Example (2) is drawn from the verbatim report and translation corpora. In French, initial *aussi* followed by subject verb inversion (*devrait-elle*) has a causal meaning:

   (2a) **Aussi,** en plus de l'envoi sur le terrain d'observateurs européens, dans le cadre de l' OSCE, la priorité de l'Union devrait-elle être d'enrayer l'escalade à tout prix […].
   [EPICG_01.09.08_situation en georgie_franciswurtz_fr]

(2b) In addition to sending in European observers under the aegis of the OSCE, the EU's priority should **therefore** be to prevent any escalation at any cost […]
[EPICG_01.09.08_situation en georgie_franciswurtz_V_en]

3. A target item from Table 2 or 3 corresponds to a source item which is neither mentioned in Table 1 nor has a causal or concessive meaning. This category will be called 'non-equivalent in source'. In example (3), which is drawn from the verbatim report and translation corpora, the French source item is *et*, which corresponds functionally to *and*.

(3a) On l'écrit dans nos règlements, on en parle dans nos discours **et** rien n'avance beaucoup.
[EPICG_20.10.08_gouvernance et partenariat aux niveaux national et régional_jeanmariebeaupuy_O_fr]

(3b) It is incorporated into our regulations, it is discussed in our debates, **but** not much progress is made.
[EPICG_20.10.08_gouvernance et partenariat aux niveaux national et régional_jeanmariebeaupuy_V_en]

4. A target item from Table 2 or 3 has no equivalent whatsoever in the source. Obviously we will call these cases 'additions'. In example (4), drawn from the speech and interpretation corpora, *so* has no stimulus in the source text; its position corresponds to the position indicated with square brackets in the speech:

(4a) enfin nous devrions réfléchir / aussi au fait de savoir si l' imposition de sanctions euh dans euh ce cas euh serait dans l' intérêt de la Géorgie [] je demande à chacun d' y euh / réfléchir
[EPICG_03.09.2008_évaluationdessanctions_jeanpierrejouyet_fr]

(4b) I mean we should be aware of the fact that if we impose such sanctions euh we ha/ we have to ensure that is in the interest of Georgia **so** before we take any steps everyone should think very carefully about it
[EPICG_03.09.2008_évaluationdessanctions_jeanpierrejouyet_I_en]

5. A source item from Table 1 corresponds to a target item which is not mentioned in Table 2 or Table 3, but has a causal or concessive meaning. This category will be called 'equivalent in target'. Example (5) illustrates an occurrence of *donc* ('so') in the source speech, which is translated by *that means that* in the interpretation:

(5a) et puis ils ont un fusil à l'épaule / ils ont **donc** acquis un statut de respectabilité dans la région
[EPICG_08.10.08_formalsitting_betancourt_fr]

(5b) and then they have a machine gun or a rifle slung over their shoulders and **that means that** they've got a respectable status in that region
[EPICG_08.10.08_formalsitting_betancourt_I_en]

6. A source item from Table 1 corresponds to a target item which is neither mentioned in Table 2 or Table 3 nor has a causal or concessive meaning. This category will be called 'non-equivalent in target' and is illustrated by example (6), which is drawn from the speech and interpretation corpora. The French speech contains the complex item *voilà pourquoi*, roughly equivalent with 'that is why'.

   (6a) si un système voit lui échapper ainsi ses propres créatures c'est qu'il est dans une crise existentielle **voilà pourquoi** si on veut éviter d' autres effondrements toujours plus douloureux il faut oser des ruptures
   [EPICG_24.09.08_situationsystèmefinanciermondial_franciswurtz_fr]

   (6b) now // er / clearly his own creatures have escaped his er grasp and he's in a some sort of existential crisis // **basically** we have to be brave and er / er break away from the current mould
   [EPICG_24.09.08_situationsystèmefinanciermondial_franciswurtz_fr]

7. A source item from Table 1 has no equivalent whatsoever in the target. These cases will be called 'omissions' and are illustrated in example (7), drawn from the verbatim report and translation corpora.

   (7a) **Donc** leur combat continue.
   [EPICG_08.10.08_formalsitting_betancourt_O_fr]

   (7b) Their fight goes on.
   [EPICG_08.10.08_formalsitting_betancourt_V_en]

4 Results

Figure 2 gives an overview of the normalised frequencies of the connective items listed in Sect. 3 as observed in the six subcomponents of the corpus (all raw and normalised frequency data are included in the Annex to this paper). The most striking feature is

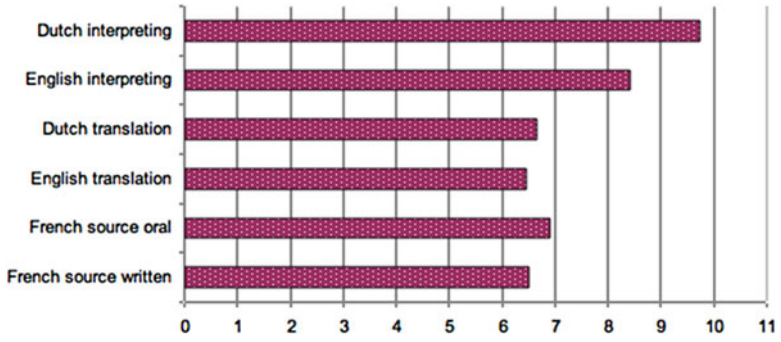


Fig. 2 Normalised frequencies (/1,000 words) of causal and concessive items in source texts and target texts

that the frequencies mainly differ on the spoken-written dimension: connective devices are more frequent in all oral subcomponents of the corpus, be it in source texts or in target texts. Interpreters are thus no exception to the general tendency observed in spoken language (Soria 2005). On the contrary, interpreting seems to exacerbate that tendency: while the frequency differences between the oral and written source texts are not significant (X-squared = 0.3611, df = 1, p = 0.5479), the differences between the oral and written target texts (i.e. interpretations and translations) are highly significant (X-squared = 23.8777, df = 1, p < 0.0001)

At first sight, there seems to be no trace of explicitation in translations, quite on the contrary: translations contain fewer instances of the listed connective items. However, the presented frequencies do not include translation options other than the ones covered by the items lists in Sect. 2. The picture they present of the degree of connectivity is therefore incomplete. To obtain a more complete picture, we need an overview of all the relations between source and target texts. Figure 3 provides such an overview. It covers the seven categories of translation options listed in Section 3.

The frequencies of the relevant connective items in target texts are represented by the columns to the right of the category axis (positive values). Relevant source items that do not have a relevant item in the target text are shown to the left of the category axis (negative values). By 'relevant' we mean belonging to one of the groups listed in Tables 2, 3 and 4. We chose to represent absolute frequencies in this case, as Fig. 3 aggregates data on both source and target texts. What Fig. 3 shows is that interpreters use a broader spectrum of translation options than translators (X-squared = 115.1544, df = 6, p < 0.0001), regardless of the language into which they interpret: translations and interpretations into English do not differ significantly from translations and interpretations into Dutch with respect to the frequencies of connective items (X-squared = 8.6405, df = 6, p-value = 0.1948). The modal divide is clearly more relevant in this context than the language divide.

Interpreters appear to use fewer in-group equivalents than translators; they omit and add more items and, except for the Dutch interpreters, they use other connective items more often to translate relevant source items. Both groups of interpreters also

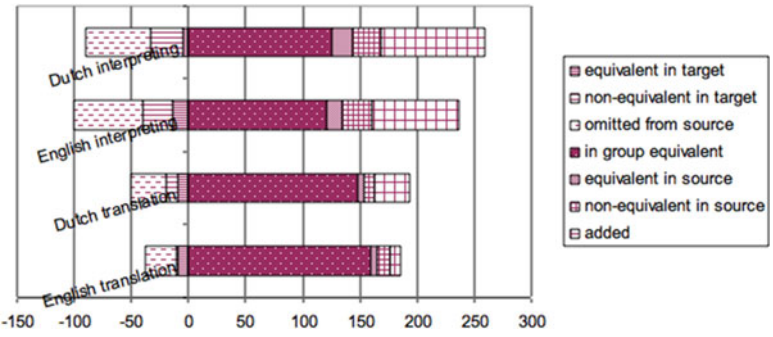


Fig. 3 Frequencies of translation options found in English and Dutch translations and interpretations

use relevant target items more often to translate source items that are not listed among the relevant items for this study. Surprisingly, if we add up all the semantically equivalent items in target texts produced by interpreters (in-group equivalent, equivalent in target and equivalent in source), it appears that there is only 40–45 % semantic overlap between the source and the target texts, whereas translators reach 70 %. In other words, as far as causal and concessive relations are concerned, interpreters in the European Parliament appear to drastically reshape the discourse structure of the source text. While reshaping, interpreters appear to omit quite a lot of connective items, which confirms previous research (Shlesinger 1995) and they appear to omit them more frequently than translators, which seems perfectly logical given the higher cognitive load they experience and the lack of structural overview. However, interpreters also appear to add quite a lot of connective items. In fact, they add more items than they omit and they add them to a significantly higher extent than translators (X-squared = 59.5566, df = 1, $p < 0.0001$). These findings are unexpected and will be looked into in detail in the qualitative analysis of this study (Sect. 5).

Not only do interpreters modify the discourse structure more than translators, they do so at different points of the text. Figure 4 gives an overview of omissions and additions shared or not by translations into different languages, by interpretations into different languages and by translations and interpretations in the same language.

As Fig. 4 aggregates data on both source and target texts, it presents absolute frequencies only. It appears that for every two subcorpora that we compared, there are at least some overlapping omissions. The overlap seems to be greater within one mode and across languages than within one language and across modes. However, neither the differences between modes (X-squared = 1.7184, df = 1, p-value = 0.1899), nor between the languages (X-squared = 0.9915, df = 1, p-value = 0.3194) appear to be significant. As far as additions are concerned, the modal dimension yields significant differences (X-squared = 20.1974, df = 1, Fisher-Exact: p-value = 0.0001, the

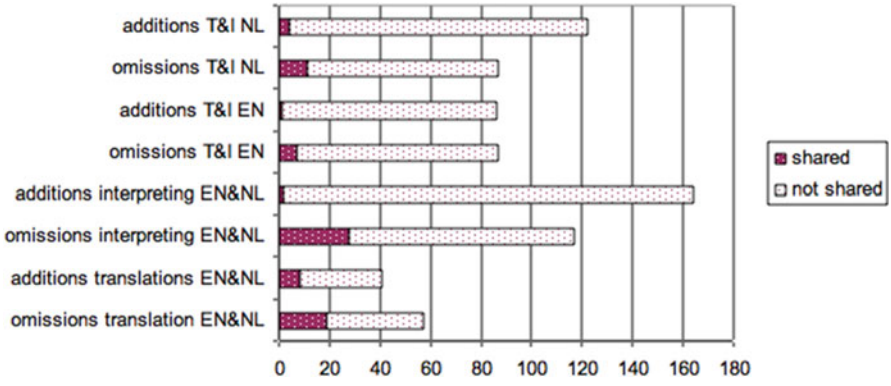


Fig. 4 Frequencies of omissions and additions in translation and interpreting

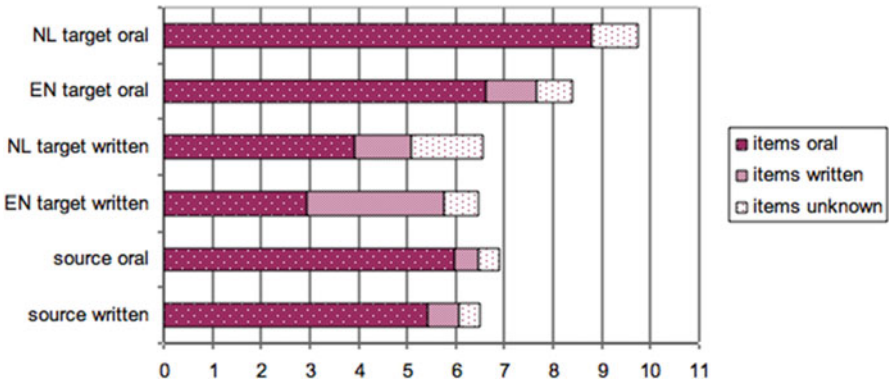


Fig. 5 Normalised frequencies and nature (spoken vs. written, /1,000 words) of connective items in source and target texts

language dimension does not (X-squared = 0.9626, df = 1, Fisher-Exact: p-value = 0.6509). With respect to additions, translations into different languages are thus more alike than interpretations and translations into the same language. Interpretations into different languages are also more alike than interpretations and translations into the same language. This might be an indication that a factor traditionally held responsible for additions in translation, i.e. explicitation, is perhaps not responsible for all additions in interpretations. The qualitative analysis reported in Sect. 5 will shed some light on this issue.

Regarding the spoken or written features of interpreting and translation, we also performed separate frequency counts for connective items typical of spoken and typical of written language. The results are shown in Fig. 5.

The picture that emerges is not as clear-cut as expected. First of all, regarding the source texts, the written verbatim reports use fewer typically oral connective items than the speeches read by the MPs. The difference is not significant though in statistical terms (X-squared = 1.0754, df = 2, p-value = 0.5841). This result probably reflects the fact that the speeches held by the members of the European Parliament are prepared in written form and read out at the plenary.

Translations in English and Dutch both display high frequencies of typically written items and low frequencies of typically oral items, as expected. They differ significantly in this respect from both the interpretations (X-squared = 118.4937, df = 2, p-value < 0.0001) and the written source texts (X-squared = 51.6204, df = 2, p < 0.0001). It is therefore safe to conclude that the translation process appears to produce texts that are more written than the written source texts.

Interpreting, on the other hand, does not seem to favour necessarily items that are more typical of spoken registers. Only Dutch interpretations present more of these items and fewer items typical of written registers than their source texts, as expected. Surprisingly, the English interpretations present slightly more typically written items than the French oral source. Taken together, English and Dutch interpretations,

therefore, do not differ significantly from their oral sources with respect to the frequencies of typically written or spoken connective items (X-squared = 2.5893, df = 2, p = 0.274).

5 Qualitative Analysis

As pointed out in Sect. 4, translators and interpreters show significantly diverging patterns regarding omissions and additions. Instances of additions and omissions shared by translators and interpreters are found, but the numbers are low. In this section we will carry out a limited qualitative analysis of additions with a view to determining whether the different patterns of addition are to be explained by fundamentally different underlying practices or priorities. We will focus on additions, as these have been studied more thoroughly in translation research, offering a good basis for comparison with interpretation. The two most frequently added markers in English and Dutch will be studied: *so, but* and *dus*, *maar*.

In translation studies, additions are traditionally analysed as explicitations (Blum-Kulka 1986; Olohan and Baker 2000): they are believed to make explicit in the target text certain aspects of meaning which can only be inferred from the source text. It is well known that hearers often infer relationships between clauses, even though no lexical item instructs them to do so (Blakemore 2002). In the examples (8) and (9), taken from Blakemore (2002: 78–79), the second clause can be interpreted as a conclusion based on the first clause. The relationship is left implicit in (8) and is made explicit in (9).

(8) Tom can open Ben's safe. He knows the combination.
(9) Tom can open Ben's safe. *So* he knows the combination.

Explicitation narrows down the range of possible interpretations: while (9) only conveys the described interpretation, (8) allows other relationships to be inferred.

For all 161 additions of *so, but, dus* and *maar*, we verified whether they could be considered to explicitate an implicit but inferable relationship in the source text or not. To identify the possible relationships, we used the typologies of Müller (2005) for *so* and Bell (1998) for *but*. For the Dutch items, we used Evers-Vermeul (2010) for *dus* and Foolen (1993) for *maar* in so far as the uses they describe differ from the uses of the English items. Our purpose is of course not to provide an in-depth analysis and reasoned categorisation of all the examples, but just to check whether the use of a connective item could be based on an inference which is authorised by the source text.

Table 5 provides an overview of the frequencies of additions for each item. These frequencies appear to vary cross-modally in both languages: in translation, additions of *but* and *maar* outnumber additions of *so* and *dus*. The opposite appears to hold in interpreting.

Table 5 Frequencies of additions of *so, dus, but* and *maar* in translations and interpretations

	so added	*dus* added	*but* added	*maar* added
Translation	1	4	8	14
Interpreting	40	42	26	26

5.1 Translation

In translation, all additions but one are typical instances of explicitation. *But* and *maar* are frequently added to explicitate the cancellation of ideational information. In most cases, the relationship is inferable from the opposite polarities of the clauses, as in (10):

(10a) Nous sommes en liberté, quelques-uns, pas tous.
[EPICG_08.10.08_formalsitting_ingridbettancourt_O_fr]
(10b) We are free, some of us, **but** not all.
[EPICG_08.10.08_formalsitting_ingridbettancourt_V_en]
(10c) Wij zijn vrij, dat wil zeggen sommigen van ons, **maar** niet allemaal.
[EPICG_08.10.08_formalsitting_ingridbettancourt_V_nl]

It is worth noting that there is significant overlap between English and Dutch translations on this particular point: all cases where both English and Dutch translators chose to explicitate the same relationship, are cases with opposite polarities.

The other additions of *but/maar* involve cancellations at the rhetorical level, where possible conclusions drawn on the basis of disclosed information are countered. In (11), for instance, the conclusion that a particular state of affairs only applies to three areas in the world is countered by the addition of a fourth area. In the source text, this cancellation can be inferred; in the target text it is explicitated.

(11a) Cette stratégie est désastreuse pour la Géorgie, pour le Caucase et pour l'Europe. La leçon vaut pareillement pour la direction russe.
'This is a disastrous strategy for Georgia, for the Caucasus and for Europe. There is also a lesson to be learnt for the Russian leadership.'
[EPICG_01.09.08_situation en georgie_franciswurtz_O_fr]
(11b) Het bleek een rampzalige strategie voor Georgië, de Kaukasus en Europa. **Maar** de les treft zeker ook Rusland.
[EPICG_01.09.08_situation en georgie_franciswurtz_V_nl]

Regarding *so* and *dus*, additions are rare in translations. In English translations, *so* is added only once in a clear case of explicitation (example 12). The causal relationship which is inferable from the use of the French verb form ending in *–ant*, is made explicit by the addition of the connective item:

(12a) assiégés par toutes sortes de monstres qui les poursuivent sans répit faisant de leur corps le siège de la douleur.
[EPICG_08.10.08_formalsitting_ingridbettancourt_O_fr]
(12b) besieged by all sorts of monsters that relentlessly pursue them, **so** their bodies are racked with pain.
[EPICG_08.10.08_formalsitting_ingridbettancourt_V_en]

In Dutch translations, *dus* is added four times: twice to explicitate a causal relationship, as in (13):

(13a) Néanmoins, il n'est pas souhaitable que l'Europe de la santé ne soit pas bâtie par les deux colégislateurs, c'est-à-dire vous et nous, et à l'issue d'un dialogue politique et d'un processus démocratique.
[EPICG_25.09.08_paquetsocial_roselynebachelotnarquin_O_fr]

(13b) Het is evenmin wenselijk het Europa van de gezondheid op te bouwen buiten de twee medewetgevers om, dat wil zeggen zonder het Parlement en de Raad, en **dus** zonder een politieke dialoog en een democratisch proces.
'It is undesirable to build a health care Europe without the colegislators, i.e. without Parliament and Council and [dus] without a political dialogue and a democratic process'
[EPICG_25.09.08_paquetsocial_roselynebachelotnarquin_O_fr]

In both other examples *dus* is used to reactivate previously given information (Evers-Vermeul 2010). In example (14), the speaker refers in the preceding context to a model for the fight against VAT fraud and refers to it again in this sentence. The translator chooses to add *dus*, in order to explicitate the fact that the reference is repeated:

(14a) Je crois que la Commission devrait analyser ces propositions, car les modèles sont là.
[EPICG_02.09.08_explicationdevote_astridlulling_O_fr]

(14b) Ik denk dat de Commissie deze voorstellen moet analyseren, want er bestaan **dus** weldegelijk geschikte modellen.
'I think that the Commission should analyse these proposals, for there are [dus] no doubt suitable models'
[EPICG_02.09.08_explicationdevote_astridlulling_V_en]

Examples such as (14) could also be considered instances of explicitation, as the information status of the clause can be inferred from the source text, but is explicitated in the target text.

Unsurprisingly, there is no indication whatsoever in translations that one of the four items is used to with another purpose than to explicitate inferable clausal relations. The only example that cannot be accounted for in terms of explicitation is a Dutch example, in which the use of *maar* is ambiguous: in (15) *maar* could either be a connective marker or a modal particle reinforcing the adverbial expression of degree, *al te*:

(15) Hoewel het […] tegen China afgekondigde embargo volledig legitiem is, mag het ons niet verbazen als het geen enkel positief effect sorteert, aangezien de Europese Unie de opheffing van het wapenembargo niet afhankelijk heeft gemaakt van concrete eisen. **Maar** al te vaak is het sanctiebeleid vaag en rekbaar, beheerst door de politieke grillen van de meest invloedrijke lidstaten of door het commerciële of geopolitieke belang van de entiteit in kwestie.
'Although the embargo imposed upon China is absolutely legitimate, it should not come as a surprise that it did not produce any positive result, since the European Union did not make the withdrawal of the embargo dependent

on concrete requirements. [Maar] all too often the sanction policy is vague and flexible, swayed by the political whims of the most influential Member States or by the commercial or geopolitical importance of the entity involved'
[EPICG_03.09.08_evaluationdessanctions_hélèneflautre_V_nl]

5.2 Interpreting

Nearly three out of four examples of added items in interpretations can be accounted for in terms of explicitation. It is worth noting that there is more diversity in the explicitated relationships.

With respect to *but* and *maar*, English and Dutch interpretations present the same types of explicitations as were found in translations: cancellations on the ideational and the rhetorical level. On the ideational level, the Dutch interpreter in (16) adds *maar* at exactly the same position as the English and Dutch translators in examples (10b) and (10c):

(16) we zijn vrij / ten minste een aantal van ons **maar** niet allemaal
[EPICG_08.10.08_formalsitting_ingridbettancourt_fr]

On the rhetorical level, example (17b) explicitates the cancellation of an inference intended but left implicit by the speaker of (17a). The first clause insists on the fairly long period of time that has passed since the beginning of the subprime crisis, creating the inference that central bankers, such as mister Trichet, should have been able to work out that this was not just a market correction. Nevertheless, they still claimed that it was one:

(17a) je rappelle que cinq mois après le déclenchement de la crise des subprimes monsieur Trichet au nom des dix principales banques centrales mondiales ne parlait encore que de simples je le cite corrections de marché
[EPICG_25.09.08_situationsystèmefinanciermondial_franciswurtz_fr]

(17b) er now we're talking about er months after the sub-prime crisis broke out **but** mister Trichet and others / were simply told about a market correction
[EPICG_25.09.08_situationsystèmefinanciermondial_franciswurtz_I_en]

English interpretations furthermore contain a couple of added instances of what Bell (1998: 527) calls "sequential" *but*, used by the interpreter to signal the return to the main topic of discourse, as in (18), where the source speaker picks up the thread of his speech again after making a joke about the French president:

(18a) il pourrait prendre une chambre au Kremlin et y rester indéfiniment / euh ça serait aussi une possibilité / moi je crois que // la chose suivant / premièrement monsieur Daul / s'il y a quelque chose à ne pas discuter / c'est l'intégration de la Géorgie et de l'Ukraine dans l'OTAN
[EPICG_01.09.08_situation en georgie_daniel cohn-bendit_fr]

(18b) he could might as well take a room in the Kremlin and stay there at this rate that is a possibility / **but** what I think is first of all mister Daul if there's something we shouldn't be discussing it's the integration of Georgia and Ukraine into NATO
[EPICG_01.09.08_situation en georgie_daniel cohn-bendit_fr]

Regarding *so* and *dus,* explicitation of causal relationships is found in interpretations, as witnessed by example (19).

(19a) il est urgent de rassurer les déposants et d'irriguer le marché interbancaire c'est de cette manière que nous restaurerons la confiance
[EPICG_08.10.08_preparationoftheeuropeancouncil2_jouyet_fr]
(19b) it's urgent to reassure depositors and to to to suc/ stabilise the interbank markets and **so** in in this way we will restore confidence
[EPICG_08.10.08_preparationoftheeuropeancouncil2_jouyet_I_en]

So and *dus* are also added by interpreters to explicitate a recapitulation or paraphrase of preceding arguments or to provide an example, as in (20):

(20a) les premières leçons que je tire de l/ cette première partie de la présidence française c'est que aucune crise n'efface les autres / la crise financière n'efface pas la crise extérieure
[EPICG_08.10.08_preparationoftheeuropeancouncil_jouyet_fr]
(20b) the first lessons I draw from the this first part of the French presidency is that // no crisis is a makes the other go away **so** the financial crisis doesn't mean the external crises have gone away
[EPICG_08.10.08_preparationoftheeuropeancouncil_jouyet_I_nl]

In Dutch, reactivation of previously given information is explicitated by *dus* in a number of cases, such as (21). The related use of *so* explicitating the main idea unit in English also occurs.

(21a) la Cour n'interdit pas au Conseil de prendre de nouvelles mesures de gel de fonds / à condition que les personnes concernées / puissent avoir des informations sur la raison pour laquelle / elles sont visées par de telles mesures
[EPICG_03.09.2008_évaluation_jeanpierrejouyet_fr]
(21b) het Hof verbiedt het niet om nieuwe maatregelen tegen deze lieden te nemen mits zij **dus** op de hoogte gebracht worden van de redenen waarom deze maatregelen tegen hen genomen worden
'the Court does not forbid to take new measures against these individuals provided they [dus] receive information on why they are targeted by such measures'
[EPICG_03.09.2008_évaluation_jeanpierrejouyet_I_nl]

Finally, interpreters working into English add instances of *so* to introduce a request or a question in English, as in (22):

(22a) ils sont la cible facile de leur énervement // permettez-moi de prononcer devant vous chacun de leurs noms
[EPICG_08.10.08_formalsitting_ingridbettancourt_fr]

(22b) they are the easy targets of their anger and their annoyance // **so** if I may I'd just like to read out all of their names
[EPICG_08.10.08_formalsitting_ingridbettancourt_I_nl]

In sum, the following numbers of interpreter-added instances can be described as instances of explicitations: 23 out of 26 additions of *but*, 18 out of 24 additions of *maar*, 34 out of 40 additions of *so* and 31 out of 41 additions of *dus*. In the other cases, an analysis in terms of explicitation is excluded, as the target text differs so much from the source text that no inference based on the source text can be the basis for an explicitation in the target text. Example (23) illustrates these cases:

(23a) je vous conseille aussi comme l'a fait mon excellente euh collègue euh libérale de lire une biographie de la comtesse de Ségur [notamment celle de madame Strich aux excellentes éditions euh Bartillat et vous verrez que tout le poids qu'il faut donner] euh au mot que vous avez prononcé à deux reprises si je vous ai bien écouté le mot interdépendance
[EPICG_21.10.08_relationUERussie_paulmariecouteaux_fr]

(23b) I note er the our liberal colleague's er recommendation to read the biography of Madame Ségur **but //** (inspiration) on d/ a number of casions the same speaker mentioned the word interdependent
[EPICG_21.10.08_relationUERussie_paulmariecouteaux_I_en]

Example (23b) shows a substantial omission corresponding to the part of the sentence between square brackets in (23a). It is precisely at this point that the interpreter inserts *but*, which does not have any corresponding source item and does not seem to be motivated by any inference that could be drawn from (23a). A similar example in a Dutch interpretation is given in (24):

(24a) la compréhension de tous est tout à fait essentiel / car ce projet de cadre commun de référence qui vous a été remis à la fin de l'année dernière / euh naturellement / il faut que / il soit pris en considération
'it is vital that everyone understands this for this draft common reference framework that was submitted to you at the end of last year of course should be taken into consideration'
[EPICG_01.09.08_cadrecommun_jacquestoubon_fr]

(24b) dat betekent dat eenieder euh moet begrijpen waar het om euh gaat / wat **dus** euh vorig jaar als conferentie heeft plaats gehad // dat moet we ook euh mm mee kunnen wegen
'what we are talking about what [dus] took place as a conference last year that should also be taken into consideration'
[EPICG_01.09.08_cadrecommun_jacquestoubon_I_nl]

In (24b) the interpreter misunderstands the source item *référence* ('reference') rendering it by *conferentie* ('conference'). *Dus* is added to the text just

before the mistake is produced and is immediately followed by a hesitation marker. It does not correspond to any item in the source text and as the interpreter is referring to a conference, it cannot be interpreted either as reactivating information, because the speaker makes no mention of a conference which took place.

These examples are fairly typical of additions in target texts that cannot be explained by explicitation. They raise interesting questions about the specific use of connective items by interpreters, as they seem to originate in both cases by the interpreter's struggle with the input text. Additions such as the ones in (23) and (24) obviously do not occur in translations, as translators are not expected to leave any traces of the difficulties they might have had with the source text, in the target text they produce. In the next Section we will offer complementary explanatory hypotheses regarding the addition of connective items, which take into account both examples of what we could call non-explicitating additions and examples we previously analysed as instances of explicitation.

5.3 Discussion

It is important to point out from the start that the very fact that additions occur in interpretations is surprising *per se*: adding items is counter-productive in interpretation as it requires cognitive resources that are already scarce. Moreover, interpreters appear to add substantially more connective items than translators, although translators do not face the cognitive limitations that interpreters face. The question thus arises: what benefit do interpreters draw from using connective items when speakers do not use them? The answer to this question will be twofold: we will first consider possible explanations in the area of explicitation and, afterwards, discuss the cases that fall outside the scope of explicitation.

In translation, the benefit of explicitation is frequently explained in terms of the mediating role of translators (Blum-Kulka 1986): translators add information when they consider their readership unable to draw similar amounts and similar kinds of information from translations as the readership of source texts draws from source texts. In a relevance-theoretic framework, translators are believed to make some sort of cost-benefit analysis. The overall cognitive outcome of the target text for the target readership is to be more or less the same as the outcome of the source text for the source readership. Therefore, additional efforts in the shape of items added to the target text are required whenever there is a risk that the outcome is lower (Gutt 1991). Applied to the addition of connectives, this would mean, following Blakemore (2002), that translators explicitate clausal relations when they feel that there is a risk that the readership is unable to identify the relationship on the sole basis of inference.

As interpreters add many more connective items than translators, even though the source texts are nearly identical, it seems logical to conclude that the former seem to

assess this risk to be much higher than the latter. The interesting question that arises in this context is: is there a basis for such different assessments? One obvious reason has to do with the audience of the texts produced by translators and interpreters: readers of translations have more time to identify clausal relationships than the audience of interpretations, who are faced with an ongoing stream of information. This seems to be a good reason for interpreters to explicitate more than translators.

Risk avoidance can also account for the addition of connective devices whose relationship is not inferable from the source text. Interpreters are indeed also aware of the inherent risks of the interpreting process itself (Monacelli 2009). Due to the high cognitive load interpreters experience, their productions inevitably reflect the source text less accurately than translations. Greater awareness of this risk could also lead to more explicit clausal relationships: connective items are after all a "cheap" means of solving cohesion problems due to errors and omissions. It is worth noting in this respect that both *so* and *but* have been described in the literature as "chaining" connectives (de Cock 2007), i.e. connective items allowing speakers to loosely connect independent clauses, while only giving rudimentary semantic information about the relationship between these clauses. In other words, in terms of a cost-benefit analysis, adding *so* or *but* to the target text increases cohesion at a low semantic and syntactic cost. Speakers do not commit themselves to a precise semantic profile nor do they engage in complex syntactic planning. Chaining connectives are therefore typical of spontaneous spoken registers, as on-line planning of speech requires cost-effective solutions. Due to the cognitive load they experience, interpreters are even more likely to use the cost-effective chaining strategies. As a consequence, additions of *so* and *but* (and of their Dutch equivalents) are expected in interpretations, but unlikely in translations.

While the cognitive load is indeed likely to promote chaining strategies and, therefore, the use and addition of chaining connectives, it is also at the heart of an alternative explanation. Cognitive load is indeed claimed by some scholars to prompt connective items directly. In a study on the use of *so* in different registers of Hong Kong English, Lam (2009) points out that *so* is used to signal that the speaker faces processing problems and needs more time to initiate the turn, a function Buysse (2012) also finds in British English and different varieties of learner English. Lam's description seems to fit very well with cases like (25b), where the interpreter uses *so* after a moment of poor translation and a long pause (2+ seconds, signalled by the double slash):

(25a) elle négocie son crédit douanier sur le marché boursier ou en banque et il est bonifiable si nous voulons aider des pays en voie développement l'exportat/ l'importateur peut offrir un montant de crédit douanier supérieur au montant de droits de douane

'it negociates its customs credit on the stock market or with a bank and it is transferable: to help developing countries, the export/ the importer can offer a customs credit that is higher than the amount of customs duties'

[EPICG_08.10.08_suspensionofthewtodoharound(debate)_martinez_fr]

(25b) it has to be sorted out in the commodity exchanges and elsewhere // **so** there is the whole issue of the customs credit euh possibly being more than the customs duty
[EPICG_08.10.08_suspensionofthewtodoharound(debate)_martinez_I_en]

The quality of the interpretation and the hesitation clearly point to processing difficulties, caused by a variety of factors: the technical nature of the speech, the extremely intricate reasoning developed by the speaker and the possible confusion created by the speaker's repair (*importateur*). Some of the features of *so* in (25b) also parallel features of delaying *so* described by Buysse (2012): the item is preceded by a pause and is pronounced with a rising intonation. It seems therefore reasonable to assume that some added instances of *so* in interpretations are directly prompted by processing difficulties.

However, there is a notable difference regarding genre: the speeches held at the European Parliament and their simultaneous interpretations are monologic texts, while the use of a delaying *so* is reported in dialogues. This is not a minor issue: Buysse (2012) claims that delaying *so* has a floor-holding function. While verbalising the processing difficulties they experience, speakers signal that they want to keep the turn. Obviously, there is little reason for interpreters to try to hold the floor, as the conventions of the simultaneous interpreting activity grant them a monopoly on speech. Therefore, if interpreters use *so* to signal processing difficulties at all, it cannot be the case that they do so to hold the floor. It also remains to be seen whether this kind of hypothesis can also account for the additions of English *but* and Dutch *dus* and *maar*.

Further research will be necessary to determine which of the hypotheses, i.e. chaining strategies or delaying strategies, is the most plausible one. Both hypotheses explain the difference between translation and interpreting equally well, as they both rely on cognitive load and online planning of speech which are plausibly higher on the interpreters' than on the translators' agenda.

6 Conclusions

The aim of this paper was to describe the use of connective items by translators and interpreters, with a focus on interpreters. As interpreters and translators perform the same basic activity, i.e. translate source texts, but in very different circumstances and for different kinds of public (audience vs. readership), we expected the use of connectives by interpreters to present both similarities and differences with respect to their use by translators.

The corpus data we used for the study yielded mixed results: we were able to confirm that omissions and additions occur both in translation and in interpreting and that omission is more frequent in interpretations. This is expected, as interpreters work under a heavier cognitive load than translators and are more likely to omit parts of the source speech, either because they do not have time to process or even

to hear it, or because they apply a conscious strategy of omitting parts of the source text that are not vital to the meaning. However, we also found that interpreters add more connective items than translators, which was unexpected, as adding information is counterproductive: it increases the cognitive load which interpreters face. The qualitative analysis showed that additions made by translators are all cases of explicitation of clausal relationships which can be inferred from the source text, whereas interpreters also appear to add connective items at places where no such relationships exist. Different hypotheses were put forward to explain this peculiarity of interpreting. Additions and omissions were also found to be partially modally determined: interpretations and translations into different languages share more additions than translations and interpretations into the same language. We were able to demonstrate that the whole range of translation options chosen by interpreters and translators is different.

As far as the register features of interpreting are concerned, it appeared that interpretations contain more connective items than both translations and their source texts, which places them closer to the spoken end of the register spectrum. However, a closer look at the data revealed that interpreting does not necessarily promote the use of items that are more typical of spoken registers.

Finally, the research described here showed what results a corpus-based approach can yield. It also calls for more research to be done in the same vein. The prevalence of mode as an explanatory factor for translation options, for instance, can easily be verified taking into account interpretations into other languages than the ones analysed here. Other types of connective items should be analysed, especially the ones that are likely to be added by interpreters to connect unrelated clauses. Additive markers are of course among the most likely candidates. Finally, interpreters should also be compared to spontaneous speakers of the same language, in order to see whether the frequencies of connective items are comparable.

7 Annex

Frequency data included in Figures 2, 3, 4, and 5

Figure 2. Frequencies of causal and concessive items in source texts and target texts

	Absolute frequencies	Normalised frequencies /1,000 words
French source written	198	6.50
French source oral	217	6.90
English translation	186	6.45
Dutch translation	194	6.65
English interpreting	237	8.41
Dutch interpreting	259	9.73

Figure 3. Frequencies of translation options found in English and Dutch translations and interpretations

	Omitted from source	Non-equiv. in target	Equiv. in target	In group equivalent	Equiv. in source	Non-equiv. in source	Added
English translation	27	2	9	160	5	11	10
Dutch translation	30	10	10	148	6	9	31
English interpreting	60	26	14	121	14	26	76
Dutch interpreting	57	28	5	127	19	25	88

Figure 4. Frequencies of omissions and additions in translation and interpreting

	Shared	Not shared
Omissions translation EN&NL	19	38
Additions translations EN&NL	8	33
Omissions interpreting EN&NL	28	89
Additions interpreting EN&NL	3	161
Omissions T&I EN	7	80
Additions T&I EN	1	85
Omissions T&I NL	11	76
Additions T&I NL	4	118

Figure 5. Frequencies and nature (spoken vs. written) of connective items in source and target texts

	Source written		Source oral		EN target written		NL target written		EN target oral		NL target oral	
	N	/1,000	N	/1,000	N	/1,000	N	/1,000	N	/1,000	N	/1,000
Items oral	165	5.42	188	5.97	85	2.95	116	3.92	187	6.63	234	8.80
Items written	20	0.66	16	0.51	81	2.81	35	1.18	29	1.03	0	0.00
Items unknown	13	0.43	13	0.41	20	0.69	43	1.45	21	0.74	25	0.94

References

Altenberg, B. (1984). Causal linking in spoken and written English. *Studia Linguistica, 38*(1), 20–69.

Bachy, S., Dister, A., Francard, M., Geron, G., Giroul, V., Hambye, P., Simon, A.-C., & Wilmet, R. (2007). *Conventions de Transcription Régissant les Corpus de la Banque de*

Données VALIBEL. https://www.uclouvain.be/cps/ucl/doc/valibel/documents/conventions_valibel_2004.PDF
Barik, H. (1975). Simultaneous interpretation: Qualitative and linguistic data. *Language and Speech, 18*(2), 272–298.
Barth, D. (2000). That's true, although not really, but still: Expressing concession in spoken English. In E. Couper-Kuhlen & Bernd Kortmann (Eds.), *Cause, condition, concession, contrast: Cognitive and discourse perspectives* (pp. 411–438). Berlin: Mouton/de Gruyter.
Becher, V. (2010). Abandoning the notion of translation-inherent explicitation: Against a dogma of translation studies. *Across Languages and Cultures, 11*(1), 1–25.
Bell, D. (1998). Cancellative discourse markers: A core/periphery approach. *Pragmatics, 8*(4), 387–403.
Bernardini, S. (2014). Intermodal corpora in contrastive and translation studies. Paper read at *Using Corpora in Contrastive and Translation Studies 4*. University of Lancaster.
Bernardini, S. (Forthcoming). From EPIC to EPTIC: Exploring simplification in interpreting and translation from an intermodal perspective. *Target*.
Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.
Blakemore, D. (2002). *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers*. Cambridge: Cambridge University Press.
Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In J. House & S. Blum-Kulka (Eds.), *Interlingual and intercultural communication* (pp. 17–35). Tübingen: Narr.
Buysse, L. (2012). *So* as a multifunctional discourse marker in native and learner speech. *Journal of Pragmatics, 44*, 1764–1782.
Cecot, M. (2001). Pauses in simultaneous interpretation: A contrastive analysis of professional interpreters' performances. *The Interpreters' Newsletter, 11*, 63–85.
Chafe, W. (1987). Cognitive constraints on information flow. In R. Tomlin (Ed.), *Coherence and grounding in discourse* (pp. 21–51). Amsterdam: Benjamins.
Chesterman, A. (2004). Paradigm problems? In C. Schaffner (Ed.), *Translation research and interpreting research: Traditions, gaps and synergies* (pp. 52–56). Clevedon/Buffalo/Toronto: Multilingual Matters.
de Cock, S. (2007). Routinized building blocks in native speaker and learner speech. In M. C. Campoy & M. J. Luzón (Eds.), *Spoken Corpora in applied linguistics. Linguistic insights, 51* (pp. 217–234). Bern: Peter Lang.
De Sutter, G. (2010). Dus (dus) is (dus) een accessibility marker (dus)? Reactie op: J. Evers-Vermeul 'Dus vooraan of in het midden? *Nederlandse Taalkunde, 15*(2), 198–203.
Evers-Vermeul, J. (2010). 'Dus' vooraan of in het midden? Over vorm-functierelaties in het gebruik van connectieven. *Nederlandse Taalkunde, 15*(2), 149–175.
Foolen, Ad. (1993). *De betekenis van partikels: een dokumentatie van de stand van het onderzoek, met bijzondere aandacht voor "maar"*. PhD dissertation, Radboud University. http://repository.ubn.ru.nl/bitstream/handle/2066/93646/93646.pdf?sequence=1
Gettrup, H., & Nølke, H. (1984). Stratégies concessives: Une étude de six adverbes français. *Revue Romane, 19*(1), 3–47.
Gile, D. (1995). *Regards sur la recherche en interprétation de conférence*. Lille: PUL.
Gile, D. (1997). Conference interpreting as a cognitive management problem. In J. H. Danks, G. M. Shreve, S. B. Fountain, & M. McBeath (Eds.), *Cognitive processes in translation and interpreting* (pp. 196–214). Thousand Oaks/London/New Delhi: SAGE Publications.
Gile, D. (2008). Local cognitive load in simultaneous interpreting and its implications for empirical research. *Forum, 6*, 59–77.
Gougenheim, G., Michea, R., Rivenc, P., & Sauvageot, A. (1964). *L'élaboration du français fondamental: étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris: Didier.
Gutt, E.-A. (1991). *Translation and relevance: Cognition and context*. Oxford: Basil Blackwell.
Hale, S. (2004). *The discourse of court interpreting*. Amsterdam: Benjamins.

Halliday, M. A. K., & Ruqayia Hasan, R. (1976). *Cohesion in English*. London: Longman.
Ilg, G. (1978). L'apprentissage de l'interprétation simultanée. *Parallèles, 1*, 69–99.
Kajzer-Wietrzny, M. (2012). *Interpreting universals and interpreting style*. PhD University of Poznan. https://repozytorium.amu.edu.pl/jspui/bitstream/10593/2425/1/Paca%20doktorska%20Marty%20Kajzer-Wietrzny.pdf
Kajzer-Wietrzny, M. (Forthcoming). Simplification in interpreting and translation. *Across Languages and Cultures, 16*.
Lam, P. (2009). The effect of text type on the use of so as a discourse particle. *Discourse Studies, 11*(3), 353–372.
Laviosa, S. (1998). Core patterns of Lexical use in a comparable corpus of English narrative prose. *Meta, 43*(4), 557–570.
Mason, M. (2008). *Courtroom interpreting*. Lanham: University Press of America.
Mauranen, A. (2000). Strange strings in translated language: A study on corpora. In M. Olohan (Ed.), *Intercultural Faultlines. Research models in translation studies 1: Textual and cognitive aspects* (pp. 119–141). Manchester: St. Jerome Publishing.
Mauranen, A. (2004). Corpora, universals and interference. In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (pp. 65–82). Amsterdam: Benjamins.
Monacelli, C. (2009). *Self-preservation in simultaneous interpreting*. Amsterdam: Benjamins.
Müller, S. (2005). *Discourse markers in native and non-native English*. Amsterdam: Benjamins.
Olohan, M., & Baker, M. (2000). Reporting that in translated English. Evidence for subconscious processes of explicitation? *Across Languages and Cultures, 1*(2), 141–158.
Øverås, L. (1998). In search of the third code: An investigation of norms in literary translation. *Meta, 43*(4), 571–588.
Pochhacker, F. (2004). I in TS: On partnership in translation studies. In C. Schaffner (Ed.), *Translation research and interpreting research: Traditions, gaps and synergies* (pp. 104–115). Clevedon/Buffalo/Toronto: Multilingual Matters.
Puurtinen, T. (2004). Explicitations of clausal relations. A corpus-based analysis of clause connectives in translated and non-translated Finnish children's literature. In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (pp. 65–82). Amsterdam: Benjamins.
Saldanha, G. (2008). Explicitation revisited: Bringing the reader into the picture. *Trans-kom, 1*(1), 20–35.
Sanders, T., Vis, K., & Broeder, D. (2012). Project notes of CLARIN project DiscAn: Towards a discourse annotation system for Dutch language corpora. In *Proceedings of the Eighth workshop on interoperable semantic annotation (isa-8)*. http://www.clarin.nl/sites/default/files/isa8_submission_7-2.pdf
Schlamberger Brezar, M. (2012). Les marqueurs discursifs « *mais* » et « *alors* » en tant qu'indicateurs du degré de l'oralité dans les discours officiels, les débats télévisés et les dialogues littéraires. *Linguistica, 52*(1), 213–224.
Seeber, K. (2011). Cognitive load in simultaneous interpreting: Existing theories – New models. *Interpreting, 13*(2), 176–204.
Seeber, K. (2013). Cognitive load in simultaneous interpreting: Measures and methods. *Target, 25*(1), 18–35.
Shlesinger, M. (1995). Shifts in cohesion in simultaneous interpreting. *The Translator, 1*(2), 193–214.
Shlesinger, M. (1998). Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta, 43*(4), 486–493.
Shlesinger, M. (2009). Towards a definition of interpretese: An intermodal, corpus-based study. In G. Hansen, A. Chesterman, & H. Gerzymisch-Arbogast (Eds.), *Efforts and models in interpreting and translation research: A tribute to Daniel Gile* (pp. 237–253). Amsterdam: John Benjamins.
Shlesinger, M., & Ordan, N. (2012). More spoken or more translated? Exploring a known unknown of simultaneous interpreting. *Target, 24*(1), 43–60.
Soria, C. (2005). Constraints on the use of connectives in discourse. In *Proceedings of the first international symposium on the exploration and modelling of meaning*. Biarritz. http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=bras&subURL=sem05/proceedings-final/15-Soria.pdf

Taboada, M, & María de los Ángeles Gómez-González. (2013). Discourse markers and coherence relations: Comparison across markers, languages and modalities. In M. Taboada, S. Doval Suárez, & E. González (Eds.), *Contrastive discourse analysis. Functional and corpus perspectives* (pp. 17–40). Alvarez Sheffield: Equinox.

Teston, S., & Véronis, J. (2004). Recherche de critères formels pour l'identification automatique des particules discursives. In *Journée ATALA "Modéliser et décrire l'organisation discursive à l'heure du document numérique"* http://archivesic.ccsd.cnrs.fr/docs/00/06/25/15/PDF/sic_00001231.pdf

Van Belle, W. (1996). Over Belgisch Nederlands. In S. Ted & S. Peter (Eds.), *Schrijfwijsheden. Visies op taal en taaladvies* (pp. 209–216). Den Haag: Sdu Uitgevers.

Van Noortwijk, C. (1995). *Het woordgebruik meester. Een vergelijking van enkele kwantitatieve aspecten van het woordgebruik in juridische en algemeen Nederlandse teksten*. PhD. Universiteit van Rotterdam. http://repub.eur.nl/pub/1433/Woordgebruik_meester.pdf

Vandepitte, S., Denturck, K., & Willems, D. (2013). Translator respect for source text information structure: A parallel investigation of causal connectors. *Across Languages and Cultures, 14*(1), 47–73.

Corpus Perspectives on Russian Discursive Units: Semantics, Pragmatics, and Contrastive Analysis

Dmitrij Dobrovol'skij and Ludmila Pöppel

Abstract The present study analyzes a group of Russian discursive units with focus-sensitive semantics such as *imenno* (just/precisely), *kak raz* (just/precisely), *to-to i ono* (that's just it/the point/problem), *to-to i est'* (that's just it/the point/problem) and *to-to i delo* (that's just it/the point/problem). They are important elements of communication but have not yet been adequately described. Some of the analyzed lexical units – for example, *imenno* and *kak raz* or *to-to i ono, to-to i est'* and *to-to i delo* – are near synonyms. Others, such as *kak raz* and *to-to i ono*, are not near synonyms, but they nevertheless belong to the semantic class of focus-sensitive elements. Thus they can all be put into a single group according to the principle of family resemblance. The material itself suggests the logic of the analysis – on the basis of pairs or groups of the semantically closest near synonyms: (1) *imenno* vs. *kak raz*; (2) *imenno* vs. *to-to i ono,* (3) *to-to i ono* vs. *to-to i est'* vs. *to-to i delo.*

Near-synonyms within these groups can be distinguished from each other on the basis of semantics, pragmatics, and usage preferences. Identifying differences of various types requires a good corpus with numerous examples, for they can be present simultaneously on several levels: semantic and pragmatic, pragmatic and usual, etc. Often, although not always, pragmatic and/or usual differences are semantically motivated. Syntactic distinctions among near-synonyms, including those in certain syntactic patterns, are also generally motivated by differences in their semantics. In a number of cases the problem is solved through the use of translational equivalents, that is, not on the level of individual lexical units (words and phrasemes) but on that of the entire utterance. Using relevant lexicographic information, text corpora, including parallel corpora, and works of fiction, we shall:

D. Dobrovol'skij (✉)
Russian Language Institute, Russian Academy of Sciences, Moscow, Russia
e-mail: dm-dbrv@yandex.ru

L. Pöppel
Department of Slavic and Baltic Studies, Finnish, Dutch and German,
Stockholm University, Stockholm, Sweden
e-mail: ludmila.poppel@slav.su.se

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_10

(a) clarify semantic and pragmatic properties as well as usage peculiarities of the focus sensitive discursive units *imenno, kak raz, to-to i ono, to-to i est'* and *to-to i delo*;
(b) analyze their systemic and translational equivalents in English and Swedish.

Keywords Discursive units • Focus sensitive items • Synonymy • Cross-linguistic equivalence • Systemic equivalents • Translational equivalents

1 Research Goals and Data

There is a group of discursive units in Russian[1] which have a certain semantic resemblance and common pragmatic features. All of these units are focus-sensitive.

The group of units considered in the present study includes the particles and constructions *imenno* (just/precisely), *kak raz* (just/precisely), *to-to i ono* (that's just it/the point/problem), *to-to i est'* (that's just it/the point/problem) and *to-to i delo* (that's just it/the point/problem). Some of them are traditionally described as synonyms, for example, *imenno* and *kak raz* in MAS (1985–1988), BTS (2002); *to-to i ono* and *to-to i est'* in Molotkov (1967). Intuitively one senses that despite the similarity in meaning of those described as synonyms, they are not interchangeable in all contexts because each of them has individual characteristics.

The pragmatic function of the group of units under consideration depends on the dialogic situation and can consist in the expression of agreement, disagreement, doubt, etc. In certain contexts some of these units are interchangeable. Elsewhere, however, they cannot be easily substituted for each other, since the semantic structure of each of the units contains features that the semantics of the others lacks, and there are also other reasons of a pragmatic and stylistic nature. Obviously, in contexts in which individual semantic features are being profiled, substituting a unit for a near-synonym is impossible.

The pragmatic limitations derive from the specific functional preferences of each of these units. For some of them expressing agreement is more typical, whereas for others it is disagreement. Certain functional peculiarities and distinguishing semantic features have already been described in the literature, especially with reference to *imenno*, *kak raz* and *to-to i ono* on the basis of Russian materials (Dobrovol'skij and Levontina 2012, 2014; Levontina 2004; Paillard 1998a, b) and *eben*, *gerade*, *ausgerechnet* on the basis of a Russian-German and German-Russian parallel corpus (Dobrovol'skij and Šarandin 2013). Using corpus data, including materials of parallel corpora, the present study aims to identify and describe the distinguishing

[1] For descriptions of Russian discursive words see especially Baranov et al. (1993), Kiseleva and Paillard (1998, 2003), Kobozeva and Zakharov (2004), Kobozeva (2006, 2007), Paukkeri (2006), Šaronov (2009) which are specifically devoted to this layer of the lexicon. Such lexical units in other languages are studied, for example, in Fischer (2000), Sorjonen (2001), Travis (2005), Romero-Trillo (2009).

characteristics of the discursive units *imenno, kak raz, to-to i ono, to-to i est'* and *to-to i delo* and their translational equivalents in English and Swedish.

Our working hypotheses are as follows:

(a) differences among near-synonyms are determined not only by semantics, but can be motivated by pragmatics and usage as well;
(b) syntactic differences between the synonyms considered here are motivated by semantic differences;
(c) because different languages lack semantic equivalents, they solve the problem on the level of the utterance, where they encounter not systemic equivalents but entirely different parallels –translational equivalents, which are determined by pragmatics to a greater degree than by semantics.

The analysis is corpus-based. The empirical data has been collected from the Russian National Corpus (RNC) and Språkbanken (the Swedish Language Bank). The RNC is thus far the most comprehensive corpus of Russian. It makes it possible to study the usage of the discourse units in both fictional and journalistic Russian texts starting from the eighteenth century, and to distinguish obsolete and currently used discourse units. The RNC consists of a collection of corpora, one of which (not very big yet but constantly developing) is the Parallel Corpus, which enables the researcher to thoroughly analyze English translations of the discursive units (in fiction) and vice versa. The search for Swedish translations was done in the parallel corpus Språkbanken (the Swedish Language Bank). Because at present the Russian-Swedish corpus is not adequate to this purpose, we also conducted this search manually. Språkbanken was additionally used to find Swedish examples which use the discursive units of interest to us and which have Russian translations outside this corpus.

Because our goals did not include statistical analysis but merely the confirmation of the basic hypothesis on the discrepancy in the semantic structures of comparable units in different languages, we think that the empirical data is adequate in scope. We have collected some 200 Swedish and 250 English examples. While not enough for a full-fledged statistical analysis, they are entirely sufficient for identifying basic tendencies. All examples were collected between 1 April and 1 June 2014.

The words and constructions under analysis are discussed in groups arranged according to semantic and/or pragmatic proximity.

2 *Imenno* vs. *kak raz*

To begin with, we will compare two discursive units – *imenno* and *kak raz* – in the function of focus sensitive particles in order to determine their specific semantic and pragmatic properties.[2] As was noted in Sect. 1, in a number of dictionaries such as

[2] We are interested only in the semantic, pragmatic (and to some extent syntactic) differences between *imenno* and *kak raz* and between other discursive units treated in the article. Prosodic differences, the importance of which is discussed on the basis of other discursive units in Kobozeva (2006, 2007), Kobozeva and Zakharov (2004), are a topic for a separate study.

MAS (1985–1988: I, 661; II, 18), BTS (2002: 389, 410) *imenno* and *kak raz* are understood to be mutual full synonyms. Following Levontina (2004), however, we consider that between these discursive units there are substantial semantic and pragmatic differences, and these will be described in the present section.

2.1 Semantics

As a focus particle *imenno* can be described as follows: 'among a certain number of objects, events, etc. some particular one of them is singled out and focused upon according to the feature that the speaker considers to be decisive, i.e. the most important element of the situation'. Cf. (1).

(1) Nexljudov vspomnil, čto slyšal, kak ėtot Šenbok *imenno* potomu, čto on prožil vse svoe sostojanie i nadelal neoplatnyx dolgov, byl po kakoj-to osobennoj protekcii naznačen opekunom nad sostojaniem starogo bogača, promatyvavšego svoe sostojanie, i teper', očevidno, žil ėtoj opekoj. [L. N. Tolstoj. Voskresenie (1899)]

Nekhludoff remembered having heard that this Schonbock, *just* because, he had spent all he had, had attained by some special influence the post of guardian to a rich old man who was squandering his property – and was now evidently living by this guardianship. [Leo Tolstoy. The Awakening (William E. Smith, 1900)][3]

In (1) *imenno* simply performs a focusing function, singling out one component of the utterance, namely the reason for this particular state of affairs.

The core meaning of *kak raz* is 'to point out an often random coincidence of two values or parameters'. Unlike *imenno*, *kak raz* focuses on the fact that the choice of an object is random or leads to unpredictable results.[4] Cf. (2).

(2) – Zdravstvuj, Rėd. A ja *kak raz* tebja išču. – Znaju, – govorju. [A. N. Strugackij, B. N. Strugackij. Piknik na obočine (1971)]

"Hello, Red. I was *just* coming to see you." "I know." [Arkady Strugatsky, Boris Strugatsky. Roadside Picnic (Antonina W. Bouis, 1977)]

Often what is in the scope of *kak raz* is an event rather than an object. Cf. (3).

(3) *Kak raz* togda, kogda Varenuxa, derža v rukax trubku, razdumyval o tom, kuda by emu ešče pozvonit', vošla ta samaja ženščina, čto prinesla i pervuju molniju, i vručila Varenuxe novyj konvertik. [M. A. Bulgakov. Master i Margarita (1929–1940)]

Just as Varenukha, receiver in hand, was pondering where else he might call, the same woman who had brought the first telegram came in and handed

[3] Here and in similar examples the English translation is included for the sake of understanding. Unless otherwise indicated, translations of Russian examples are by Charles Rougle. For a contrastive analysis cf. 2.2, 3.2 and 4.2.

[4] For this reason, *imenno* is often used in contexts of identification (on contexts of this type cf. in more detail Padučeva 2014). As for *kak raz*, this discursive unit is more seldom used in such contexts, since the notion of chance coincidence is emphasized in its semantics.

Varenukha a new envelope. [Mikhail Bulgakov. Master and Margarita (Richard Pevear and Larissa Volokhonsky, 1979)]

Here two events happen to coincide in time, so that *kak raz* is more appropriate than *imenno* (cf. Levontina 2004).

In contexts where the distinctive features are weakened or are not the focus of attention, of course, interchanges are possible. Cf. (4).

(4) Na vsjakij slučaj slonenok rastopyril uši, kak protivouragannye ščity. *Kak raz* ėto i okazalos' samoj bol'šoj ošibkoj. [Alexandr Dorofeev. Ėle-Fantik // «Murzilka», 2003]

To be on the safe side, the little elephant spread its ears like tornado shields. *Precisely* this proved to be its biggest mistake.

In (4) *kak raz* singles out an event (the elephant spreading its ears) from among all other possible events and points out a correlation between this event and another event that failed because the first event (spreading its ears) was the wrong thing to do. The first event was a deliberate action, but it was a gamble and led to unpredictable results. Therefore, the use of *kak raz* in the same context with *na vsjakij slučaj/ to be on the safe side,* meaning 'to take safety precautions in case something happens,' is natural if the story is told from the little elephant's perspective.

It is possible to replace *kak raz* with *imenno,* but in that case the interchange would mean that the story is being told from the perspective of a narrator standing outside the inner world of the text.

At the same time, there are contexts in which no such interchange is possible. This applies in particular to the special questions *kto imenno?* (who, exactly?); *čto imenno?* (what, exactly?); *kogda imenno?* (when, exactly?), etc., where **kto kak raz; *čto kak raz; *kogda kak raz,* etc. are clearly impossible.[5] The prohibition against the use of *kak raz* is understandable, since the gist of such questions is to focus the interrogative word, and it is not possible to take other values into consideration. The exception among special questions is those introduced by *počemu* (why) and its synonyms. Cf. (5).

(5) – Pogodi, – skazal Žixar'. – Ty daveča pro osinu govoril, na kotoroj mne, predatelju, povesit'sja. Ty otkuda ėtu osinu vzjal? I počemu *kak raz* osinu, a ne berezu i ne dub, u kotoryx vetki pokrepče budut? [Mixail Uspenskij. Tam, gde nas net (1995)]

"Wait a minute," said Žixar'. "A while ago you were talking about an aspen tree on which to hang me, a traitor. Where did this aspen come from? And why *precisely* an aspen rather than a birch or an oak, which have stronger branches?"

The reason this is possible is obvious – why-questions presume an underlying comparison of two different values. That is, the question *počemu X?* (why X?) can also be construed as *počemu X, a ne Y?* (why X but not Y).

[5] According to Levontina 2004, we are dealing with a special reading of *imenno* in this case, namely "*imenno* 1" ≈ 'exactly'.

Table 1 Separate utterances: *imenno* vs. *kak raz*

	imenno	*kak raz*
Ending with a full stop	534	0
Ending with an exclamation point	397	0
Ending with a dash	557	0
Total	1,488	0

The difference in the semantics of *imenno* and *kak raz* also motivates their syntactic behavior. Unlike *kak raz*, *imenno* can function not only as a focus particle but also as a separate utterance. Cf. (5).

(5) – Ne Eleny li Stanislavovny budete synok? – Da. *Imenno.* (Cf. **Kak raz.*) [I. A. Il'f, E. P. Petrov. Dvenadcat' stul'ev (1927)]
"Not by any chance Elena Stanislavovna's son?" "*Right!*" [Ilya Ilf, Evgeny Petrov. The Twelve Chairs (John Richardson, 1961)]

As a separate utterance *imenno* focuses what was stated in the preceding utterance. *Kak raz*, on the other hand, always compares two different values, of which at least one must be expressed explicitly. This is confirmed by the corpus data of the RNC. Cf. Table 1.

As can be seen in Table 1, out of a total of more than 10,500 contexts in the RNC (without homonym disambiguation) we found 1,488 in which *imenno* is used as a separate utterance. The corresponding results for *kak raz* are 0 contexts as a separate utterance out of a total of more than 24,000 without homonym disambiguation. The corpora data indicate that these discursive units display non-random differences in syntactic behavior, and they corroborate our hypothesis that these syntactic features have a semantic basis.

When *imenno* functions as a separate utterance, it displays the variant *vot imenno*, in which the focusing function is strengthened by the deictic element *vot*.[6] Cf. (6).

(6) – Net, ja dumaju, bez šutok, čto dlja togo, čtob uznat' ljubov', nado ošibit'sja i potom popravit'sja, – skazala knjaginja Betsi. – Daže posle braka? – šutlivo skazala žena poslannika. – Nikogda ne pozdno raskajat'sja, – skazal diplomat anglijskuju poslovicu. – *Vot imenno*, – podxvatila Betsi, – nado ošibit'sja i popravit'sja. [L. N. Tolstoj. Anna Karenina (1878)]
"No; I imagine, joking apart, that to know love, one must make mistakes and then correct them," said Princess Betsy. "Even after marriage?" said the ambassador's wife playfully. "It's never too late to mend." The attaché repeated the English proverb. "*Just so*," Betsy agreed; "one must make mistakes and correct them." [Leo Tolstoy. Anna Karenina (Constance Garnett, 1911)]

The pragmatics of *imenno* as a separate utterance is to express agreement with a previously stated or expressed hypothesis or opinion. When used as a separate utterance it often confirms an opinion expressed by the interlocutor, and in such cases the semantic valency normally filled by the focusing element is left unfilled. This

[6] The variant *vot imenno* occurs not only as a separate utterance, but – albeit more seldom – in the position of a focus particle within the utterance.

unfilled valency is by default filled by an element from the preceding utterance, which in a dialogue is usually the speech of the interlocutor. Thus *imenno* focuses the central element of the interlocutor's utterance, namely the element of the situation that is critical for it to be understood correctly.[7]

2.2 Contrastive Analysis

We checked equivalents in bilingual Russian-English (Axmanova and Smirnickij 1985; Wheeler et al. 1997; Ermolovič 2011) and Russian-Swedish dictionaries (Birgegård and Sharapova Marklund 2010; Davidsson 1976) and looked for translational equivalents in corpora and works of fiction. In none of these bilingual dictionaries was *imenno* clearly distinguished as two lexemes – as an independent utterance and as a focus particle. The English near-equivalents of *imenno* are *just, exactly, particular, in particular, indeed, precisely, specifically, obviously, actually, it is… that/who.* It is obvious that *just, particular, in particular, obviously,* and *actually* can only be translational equivalents of *imenno* as a focus particle. The equivalents of *imenno* as a separate utterance are *just it, exactly, precisely, indeed, specifically*.[8] Cf. (7).

(7) – Doktor Pil'man, možet byt', Vy skažete svoim zemljakam neskol'ko slov po ėtomu povodu? – Čto *imenno* ix interesuet? [A. N. Strigackij, B. N. Strigackij. Piknik na obočine (1971)]
"Dr Pilman, would you care to say a few words to your fellow townsmen on the subject?" "What *in particular* interests you?" [9] [Arkady Strugatsky, Boris Strugatsky. Roadside Picnic (Antonina W. Bouis, 1977)]

Kak raz has the following translational equivalents: *just, exactly, right, right away, surely, precisely, directly, actually*. Cf. (8).

(8) – Sdajte ob"avlenie Paše. Ona sejčas *kak raz* edet v nočnuju. Sekretar' sel čitat' peredovuju. [I. A. Il'f, E. P. Petrov. Dvenadcat' stul'ev (1927)]
Give the advertisements to Pasha. He's[10] *just* going over there. The editor sat down to read the editorial. [Ilya Ilf, Evgeny Petrov. The Twelve Chairs (John Richardson, 1961)]

The Swedish equivalents of *imenno* are *just, exakt, just det*, and of *kak raz – precis [så].* In translations (besides those considered in dictionaries) we have found the following equivalents: *imenno – precis; just precis; förresten; ja, inte sant?; korrekt*; *jadå.* As in English, in Swedish there are various equivalents corresponding to

[7] Analysis shows that *imenno* as a separate utterance is very similar semantically to *to-to i ono.* The differences between them will be discussed in Sect. 3.

[8] For reasons of space, in Sects. 2.2, 3.2 and 4.2 only one illustrative example is provided for each analyzed Russian discursive unit.

[9] Mistranslation – *them* in the original.

[10] Mistranslation – *she* in the original.

imenno as a focus particle and as an independent utterance – *just, förresten* can only be translational equivalents of *imenno* as a focus particle, while *just det; ja, inte sant?; jadå* are equivalents as a separate utterance. Cf. (9).

(9) Kak ėto ni stranno možet pokazat'sja, no Konstantin Levin byl vljublen *imenno* v dom, v sem'ju, v osobennosti v ženskuju polovinu sem'i Ščerbackix. [L. N. Tolstoj. Anna Karenina (1878)]

Hur underligt det än låter, så hade Konstantin Levin blivit förälskad *just* i själva hemmet, i hela familjen, särskilt i dess kvinnliga hälft. [Leo Tolstoj. Anna Karenina. (Sigurd Agrell, 1927)]

Strange as it may appear, *it was* with the household, the family, *that* Konstantin Levin was in love, especially with the feminine half of the household. [Leo Tolstoy. Anna Karenina (Constance Garnett, 1911)]

Ideally, corpus analysis also presumes analysis of the conditions for the use of the relevant elements of the source language – in this case Russian – in reverse translations; that is, in the translation of English and Swedish texts into Russian. Cf. (10).

(10) "Och vid elvatiden på kvällen efter att han blivit skjuten saknades datorn i hans bostad"."*Korrekt*". [Stieg Larsson. Flickan som lekte med elden (2006)]

– A v odinnadcat' čacov večera, kogda ego zastrelili, noutbuka v kvartire ne okazalos'. – *Vot imenno.* [Stieg Larsson. Devuška, kotoraja igrala s ognem (Inna Streblova, 2009)]

"And by 11:00 that night – when the police arrived at his apartment – the computer was gone." "*Correct*." [Stieg Larsson. The Girl Who Played With Fire (Reg Keland, 2009)]

Analysis of even these isolated examples enables us to identify additional correspondences: in (10), for instance, where *korrekt* is the equivalent of *imenno* as an independent utterance.

Such words as *just, exactly, indeed, actually* in English and *just, exakt, precis* in Swedish can be considered near-synonyms, but they also display different semantic features. These semantic features do not coincide with the semantic features that distinguish *imenno* from *kak raz*. This explains the fact that both *imenno* and *kak raz* can be translated with the help of the same lexical units.

(11) I *imenno* tam osobenno živo vspominaetsja Rossija, i *imenno* derevnja. [L. N. Tolstoj. Anna Karenina (1878)]

(11a) And it's *just* there that Russia comes back to me most vividly, and *especially* the country. [Leo Tolstoy. Anna Karenina (Constance Garnett, 1911)]

(11b) Och *det är* på sådana platser *som* minnesbilderna från Ryssland, och då *just* landsbygden, är som starkast. [Leo Tolstoj. Anna Karenina (Ulla Roseen, 2007)]

In (11) the first *imenno* is translated into English (11a) as *just*, the second as *especially.* As for the Swedish translations in (11b), the first translational equivalent is the construction *det är… som*, the second is *just*. In other words, there are many ways of expressing the same pragmatic function of *imenno* in English and Swedish, but there is no exact semantic equivalent on the lexical level. This is also true of *kak raz*.

We have examined the semantic oppositions and pragmatic similarities and differences between the focusing units *imenno* and *kak raz*, and we have also noted the syntactic and pragmatic ability of *imenno* to function as a separate utterance. This characteristic is semantically motivated and is confirmed by the corpus data of the RNC. *Kak raz* cannot function as an independent utterance, and this as well is based on semantics and is confirmed by the corpus data. For this reason, *kak raz* and *to-to i ono* cannot be substituted for each other, and the opposition is therefore irrelevant.

If *kak raz* is a near-synonym of *imenno* as a focus particle, *imenno* as an independent utterance has other near-synonyms such as *to-to i ono, to-to i est', to-to i delo, v tom-to i delo, to-to že*, etc. In turn here we will examine *to-to i ono, to-to i est',* and *to-to i delo*, and Sect. 3 will discuss the opposition between *imenno* and *to-to i ono* as separate utterances.

3 *Imenno* vs. *to-to i ono*

3.1 Semantics, Pragmatics

MAS (1985–1988: I, 661; IV, 391) interprets *imenno* in the relevant meaning as an "affirmative word", and *to-to i ono* as a unit that is "commonly used for affirmation or emphasis of something said". From these interpretations it is difficult to understand just what constitutes the semantic and pragmatic differences between these discursive units. *To-to i ono* is described in general and more fully in phraseological dictionaries such as Molotkov (1967) and Lubensky (2013). For more detail see Sect. 4.1.

To-to i ono as a separate utterance can be defined as: 'what the speaker determines to be the most important element in the interlocutor's utterance, that which is crucial to an adequate understanding of a given situation.' Cf. (12).

(12) – Skažite, tovarišč Gavrilov, odnoj li perevozkoj kontrabandy delo ograničivaetsja? Kombat otvetil ugrjumo: – *Vot to-to i ono-to!* [N. A. Ostrovskij. Kak zakaljalas' stal' (č. 2) (1930–1934)]

"Probably there's something more serious than smuggling going on. What do you say, Comrade Gavrilov?" *"That's just the trouble,"* the Battalion Commander replied gloomily. [Nikolai Ostrovsky. How the Steel was Tempered (pt 2) (R. Prokofieva, 1952)]

In (12) *to-to i ono* serves to single out the most important element in the utterance of the interlocutor from the perspective of the speaker, namely "there's something more serious than smuggling going on."

In cases where the main element of the situation coincides with the element in the interlocutor's utterance that the speaker singles out as most important, *imenno* and *to-to i ono* are interchangeable, since the subtle differences between them are neutralized. Cf. (13).

(13) – Zapomnite, mamaša, – buržua. Žujut oni nas, žujut i vsasyvajut. – Bogatye, značit? – sprosila mat'. – Vot *imenno!* V ėtom ix nesčastie. [Maksim Gor'kij. Mat' (1906)]

"Remember that word, dear granny – bourgeois! Brr! How they chew us and grind us and suck the life out of us!" "The rich, you mean?" "*Yes, the rich*. And that's their misfortune." [Maxime Gorky. Mother (D. J. Hogarth, 1921)]

Otherwise such an interchange is either undesirable or causes a shift of accent. Cf. (14).

(14) – Ponjatno. Nado bylo takie dela drugim poručat'. Ėto tebe ne na mašinke stukat'. – Net, ja sam dolžen byl. – A esli by Bati sidel? – Nu, esli by sidel. Vot *to-to i ono*! [Mixail Gigolašvili. Čertovo koleso (2007)]

"I see. These things should have been entrusted to someone else. It's not something you can just bat out on a typewriter." "No, I had to do it myself." "And if Bati had been in jail?" "Well, if he'd been in jail. *That's just the thing!*"

In (14) *to-to i ono* singles out only the main element in the utterance of the interlocutor. Since that sentence is in the subjunctive, *to-to i ono* here refers to a hypothetic situation. In such cases substitution with *imenno* is undesirable.

Another constraint that makes the interchange undesirable is constructional in nature. *To-to i ono* can combine with the clitic *-to*, which is not the case with *imenno.* Cf. (15).

(15) – A ty ego klass vedeš' teper'… U nego vse otličniki oni byli… ran'še, – on podmignul mne, – a teper'? U tebja?.. – Kakie že oni otličniki, kogda ėlementarnyx veščej?.. – Vot *to-to i ono-to*… – I on zasmejalsja. [Bulat Okudžava. Noven'kij kak s igoločki (1962)]

"And you are teaching his class now… With him they were all top students… earlier," he winked at me, "but now? With you?" "What sort of top students are they, when it's just a question of elementary things?…" "*That's just what I mean*…" And he laughed.

The ability to combine with the clitic *-to* is evidently connected with the polemical element contained in the pragmatic potential of *to-to i ono*. In contrast to *imenno,* the basic pragmatic feature of *to-to i ono* is argumentativeness, which presumes disagreement – both with one of the participants in the situation and with the hypothetical supporters of an opinion that does not coincide with that of the speaker. Cf. (16) and (17).

(16) Otvet: Gospodi, čego tol'ko ne byvaet! Ee podmenili vo dvorce, a ego v roddome. Vopros: V kakom dvorce? Otvet: A počemy vy ne sprašivaete, v kakom roddome? *To-to i ono!* Vse xotjat uznat', kak podmenivajut vo dvorcax […]. I nikomu ne interesno pro rajonnuju bol'nicu […]. [Mixail Šiškin. Venerin volos (2004) // «Znamja», 2005]

Answer: Lord, the things that happen! Her they switched in the palace, and him in the birthing home. Question: in which palace? Answer: Why don't you ask in which birthing home? *That's the whole point!* Everyone wants to know how babies get switched in the places… And nobody cares about a district hospital […].

Table 2 Separate utterance: *imenno* vs. *to-to i ono*

	imenno		*to-to i ono*	
	imenno	*vot imenno*	*to-to i ono*	*vot to-to i ono*
Ending with a full stop	228	306	59	36
Ending with an exclamation point	167	230	60	24
Ending with a dash	242	315	25	21
Total	637	851	144	81

(17) – Tam ros reliktovyj granatovyj les, – govorit Abbas. – Ty videl kogda-nibud' dikij granat, kotoromu tri veka v korne? Vot *to-to i ono*. [Aleksandr Iličevskij. Pers (2009)]

"There used to be an old-growth pomegranate forest there," says Abbas. "Have you ever seen a three-hundred-year-old wild pomegranate? *That's the whole problem*."

The polemical aspect is particularly obvious in examples where *to-to i ono* is a reaction to a rhetorical question, as in (16) and (17). In these cases substituting *imenno* is undesirable. When substitution is possible, the contexts are still not identical pragmatically, since *imenno* has a different semantic prosody. Here *to-to i ono* is always potentially argumentative.

We have also checked the frequency occurrences in the RNC for *imenno* and *to-to i ono* as an independent utterance both with *vot* and without. Cf. Table 2.

In the RNC we found 1,488 examples with *imenno* (of a total of more than 10,500 without preliminary homonym disambiguation) and 225 with *to-to i ono* (out of a total of 316).

Evidently this is because the construction *to-to i ono* has additional meanings that demand more specific contexts.

3.2 Contrastive Analysis

According to Lubensky (2013: 637–238) the English equivalents of *to-to i ono* are *that's just it /the thing/ the point); that's the whole point; (and) that's the problem/ the trouble; you've put your finger on it; my point exactly; that's just my point; the thing is…; the (whole) point is…; my point is…; the problem/the trouble is*…. The Swedish equivalents considered in Birgegård and Sharapova Markund (2010: 711) are *just [så är] det, precis*. In addition, we have found the following equivalents in works of fiction: *jag vill mena det; jo; det är just det; jo, jo… det är just det, så är det med den saken, exakt*.

The Swedish equivalents *precis, just det* and *exakt* are possible only in contexts where the polemical element is neutralized, and they are therefore used to translate *imenno* and *to-to i ono*. The same applies to English equivalents such as *that's just it/the thing/the point; that's the whole point*. Other equivalents are required to translate *to-to i ono* in contexts where the polemical element is focused.

4 *To-to i ono* vs. *to-to i est'* and *to-to i delo*

Molotkov (1967: 479) describes the discursive units *to-to i ono* and *to-to i est'* as synonyms, and these units are interpreted in two verbatim entries with mutual references.

TO-TO •<**VOT**>**TO-TO I EST'** *Coll.* Expression, usually didactic or reproachful in nature, acknowledging the correctness of what has been said. Cf. < vot>to-to i ono.

< **VOT**>**TO-TO I ONO** *Coll.* Expression, usually didactic or reproachful in nature, acknowledging the correctness of what has been said. Cf. < vot>to-to i est'.[11]

Lubensky (2013: 637–638) defines the two units in one entry as synonymous and identical in function:

(VOT) TO-TO I ONO<ONO-TO>; (VOT) TO-TO (ONO) I EST' *all coll* [Interj; used as indep. sent or main clause in a complex sent (usu. foll. by a čto-clause); these forms only; fixed WO] this/that is the important factor, the essential thing (used to emphasize that what has just been said or is about to be said is the central issue, the most important aspect of the matter in question): **that's just it<the thing, the point>; that's the whole point;** [lim.] **(and) that's the problem<the trouble>; you've put your finger on it; my point exactly; that's just my point;** [when foll. by a čto-clause] **the thing is…; the (whole) point is…; my point is…;** [lim.] **the problem<the trouble>is…**

Lubensky considers these two constructions both in the function of an independent sentence and in the function of a main clause in a complex sentence followed by a *čto*-clause. In the present analysis we consider the constructions only as independent sentences.

To-to i delo it is not addressed in any of the dictionaries that we have checked. Our analysis of the corpus data indicates that there are significant non-semantic differences between the three units.

4.1 Semantic and Pragmatic Analysis

As we are going to show using corpus evidence, *to-to i est'* and *to-to i delo* are hardly ever used in present-day Russian, while *to-to i ono* occurs frequently. The semantic and pragmatic characteristics of these first two constructions, therefore, can only be described in general terms. Context analysis shows that between *to-to i ono*, on the one hand, and *to-to i est'* and *to-to i delo*, on the other, there is a clear semantic and pragmatic similarity. Cf. (18) and (19).

[11] To facilitate understanding, the entry is translated into English

(18) Oleg edva sderžival sebja i izbegal smotret' na Staxoviča. – K-kak tvoe mnenie, Sereža? – Lučše by napast', – skazal Serežka, smutivšis'. – *To-to i est*'… [A. A. Fadeev. Molodaja gvardija (1943–1951)]

Oleg could hardly restrain himself and avoided looking at Stakhovich. "W-what do you think, Sergei?" "I think we'd better make the attack," Sergei said in some confusion. "*That's it, then.*" [Alexander Fadeev. The Young Guard (Violet Dutt, 1958)]

(19) – Imja, sudar', imja! Ėto vsego nužnee v našej knižnoj kommercii. – Da gde ž mne prikažeš' ego vzjat'? – Vot *to-to i delo!* [M. N. Zagoskin. Moskva i moskviči (1842–1850)]

"A name, sir, a name! That's what's needed most of all in our book trade." "But where am I supposed to find it?" "*That's the whole problem!*"

In (18) and (19) the three constructions are interchangeable, at least from the perspective of present-day usage.

The construction *to-to i ono* and the constructions *to-to i est'* and *to-to i delo* all have a polemical potential. Cf. (20) and (21).

(20) – Interesno, čemu ix tam v gorodax učat? – Izvestno čemu, – soobrazil Čonkin. – Salo derevenskoe žrat'. – *To-to i est'*, – soglasilas' Njura. [Vladimir Vojnovič. Žizn' i neobyčajnye priključenija soldata Ivana Čonkina (1969–1975)]

"I wonder what they teach them there in the city." "That's easy," announced Chonkin. "To live off the fat of the countryside." "*That's for sure*," agreed Nyura. [Vladimir Voinovich. The Life and Extraordinary Adventures of Private Ivan Chonkin (Richard Lourie, 1977)]

In (20) Njura expresses agreement with Čonkin that town-dwellers learn only one thing about the countryside – how to live off the fat of the countryside.

(21) A otkeda, skaži, iz zaviruxi burannoj krov'? Veter ved' ėto, vozdux, snegovaja pyl'. A *to-to i est'*, kuma, ne veter ėto buran, a razvedenka-oborotenka detenyša-ved'menočka svoego poterjala, iščet v pole, plačet, ne možet najtit'. [B. L. Pasternak. Doktor Živago (1945–1955)]

And how is it, tell me, that blood can come from a stormy whirl? Isn't it just wind, air, snowy powder? *But the fact is*, my pet, that the storm is not wind, it's a changeling she-werewolf that's lost her young one, and searches for him in the field, and weeps because she can't find him. [Boris Pasternak. Doctor Zhivago (Richard Pevear and Larissa Volokhonsky, 2010)]

In (21) *to-to i est'* is not only polemical but also expresses disagreement with the question asked in the speaker's preceding sentence. From the viewpoint of present-day usage it is difficult to judge whether replacing *to-to i i est'* and *to-to delo* with *to-to i ono* would add any extra element to the utterance.

In order to find possible usage differences between *to-to i ono, to-to i est'* and *to-to i delo,* we analyzed all contexts from the RNC containing these constructions. The search yielded 393 contexts with *to-to i ono,* 449 with *to-to i est',* and 41 with *to-to i delo*. The frequency of occurrence is shown in Table 3.

Table 3 *To-to i ono, to-to i est'* and *to-to i delo*: frequency of occurrence in the RNC

Number of occurrences	*to-to i ono*	*to-to i est'*	*to-to i delo*
Total	393	449	41
From 1970s	259	4	3
1960s	33	1	1
1950s	24	13	1
1940s	9	11	0
1930s	19	25	0
1920s	20	17	8
1910s	11	10	2
1900s	11	26	3
XIX century	6	342	23

The constructions *to-to i est'* and *to-to i delo* are not used in present-day Russian. They were frequent in the nineteenth century, however, and therefore should be labeled as archaic. *To-to i ono*, in contrast, is frequent in modern usage but was practically not used in the nineteenth century. From the 1920s through the 1940s *to-to i ono* and *to-to i est'* were equally frequent. This balance was disturbed in the 1950s, leading to the disappearance of *to-to i est'*. As for *to-to i delo*, it is approximately ten times less frequent in the RNC in comparison with *to-to i ono* and *to-to i est'* and was practically not used after the 1920s. *To-to i ono-to* occurs in contexts throughout the whole period, beginning in the nineteenth century.

It should also be pointed out that because the RNC is relatively small (the main corpus includes about 200 million running words), the results can be considered only preliminary and must be verified on the basis of larger corpora.

We have also observed some differences in combinatorics. *To-to i ono* combines easily with *-to*, while combinations of *to-to i est'* and *to-to i delo* with *-to* occur very seldom.[12] We have found 54 contexts with *to-to i ono-to,* 2 contexts with *to-to i est'-to* and no examples with *to-to i delo-to*; cf. (22) and (23).

(22) Kazalos' by, čto ešče čeloveku nužno? No čelovek, osobenno čelovek tvorčeskij, kak izvestno, nikogda ne ostanavlivaetsja na dostignutom. Postavil odin unitaz, xočetsja postavit' vtoroj, a kuda? Vot *to-to i ono-to*… [Vladimir Vojnovič. Ivan'kiada, ili Rasskaz o vselenii pisatelja Vojnoviča v novuju kvartiru (1976)]

It would seem a man could want no more. But, as is well-known, a man, especially a creative man, never rests on his accomplishments. He puts in one toilet; then he wants to put in another. But where? *Ah, that's the problem!* [Vladimir Voinovich. The Ivankiad (David Lapeza, 1976)]

(23) – Čtoby ėtu knigu pravil'no opredelit', ee vsju pročitat' nužno, – obratilsja on nakonec k Lapinu tonom upreka. – Ėto dolgo, – skazal doktor vinovatym

[12] A search in Google Books yielded 3 examples with *to-to i est'-to* and 3 examples with *to-to i delo-to*.

golosom […]. – *To-to i est'-to!* [F. D. Krjukov. Obysk (1906–1915) // «Sovetskaja Rossija», 1990]

"To classify this book correctly it will be necessary to read it in its entirety," he finally addressed Lapin in a reproachful tone. "That will take a long time," said the doctor in a guilty voice. "*Precisely!*"

To-to i est' combines more often with vocatives than *to-to i ono*: a search in the RNC gave 43 contexts with *to-to i est'* and 12 with *to-to i ono*. Cf. (24).

(24) – Pojdem-ka. Posmotriš', kak ja živu i rabotaju. – Večerom spektakl', – vozrazil Lik, – i zavtra ja uezžaju! – *To-to i ono,* milyj, *to-to i ono*. Xvataj! Pol'zujsja! Drugogo šansa nikogda ne budet. [V. V. Nabokov. Lik (1938)]

"Come on, let's go. You'll see how I live and work." "I have a performance tonight", Lik objected, "and I'm leaving tomorrow." "*That's just the point*, my friend, *that's just the point*. Seize the opportunity! Take advantage of it! There will never be another chance." [Vladimir Nabokov. Lik (Vladimir Nabokov, 1966)]

Most of the contexts with *to-to i est'* are from the nineteenth century. To explain this phenomenon, we would at a minimum need to have information on the statistical distribution of vocatives in the speech of nineteenth-century literary characters as compared with present-day literature.

4.2 Contrastive Analysis

Both Lubensky (2013) and Birgegård and Sharapova Markund (2010) consider English and Swedish equivalents of *to-to i est'* together with *to-to i ono.* Cf. 3.2. As for *to-to i delo*, we have found two examples in the parallel corpus with *to-to i est'* in the RNC. Cf. (25).

(25) – Da začem že ty tak odelsja? Ty smotriš' kakim-to ploxim gorodskim meščaninom … ili raznoščikom… ili otstavnym dvorovym. Otčego ėtot kaftan, a ne poddevka ili prosto krest'janskij armjak? – *To-to i est'*, – načal Neždanov, kotoryj v svoem kostjume dejstvitel'no smaxival na melkogo prasola iz meščan […]. [I. S. Turgenev. Nov' (1877)]

"But why did you get yourself up like this? You look like some sort of shopkeeper, or pedlar, or a retired servant. Why this long coat? Why not simply like a peasant?" "*Why*?" Nejdanov began. He certainly did look like some sort of fishmonger in that garb […]. [Ivan Turgenev. Virgin Soil (Rochelle S. Townsend, 1929)]

As often happens in translations of literary works, in both contexts the pragmatic thrust of the dialogue is captured, but there are no lexical equivalents of *to-to i est'*. The Swedish equivalents behave similarly. Cf. (26).

(26) – Oni otvergajut spravedlivost' sobstvennosti, kapitala, nasledstvennosti, a ja, ne otricaja ėtogo glavnogo stimula […], xoču tol'ko regulirovat' trud. – *To-to*

i est', ty vzjal čužuju mysl', otrezal ot nee vse, čto sostavljaet ee silu, i xočeš' uverit', čto èto čto-to novoe, – skazal Nikolaj, […]. [L. N. Tolstoj. Anna Karenina (1878)]

"De förkastar rättmätigheten i ägandet, kapitalet, arvet, medan jag inte alls bestrider sådana viktiga stimuli […] utan bara vill reglera arbetet." "*Det är just det* jag menar, du har tagit någon annans tanke, skurit bort allting som ger den styrka och sedan envisas du med att påstå att du har kommit på något nytt", sa Nikolaj […]. [Lev Tolstoj. Anna Karenina (Ulla Roseen, 2007)]

"They deny the justice of property, of capital, of inheritance, while I do not deny this chief stimulus." […] "All I want is to regulate labor." "*Which means*, you've borrowed an idea, stripped it of all that gave it its force, and want to make believe that it's something new," said Nikolay, […]. [Leo Tolstoy. Anna Karenina (Constance Garnett, 1911)]

In the Swedish translation of this passage, the construction *det är just det* is an adequate pragmatic equivalent of *to-to i est'*. In the dictionaries we consulted and in the parallel corpus there were no equivalents of the construction *to-to i delo.*

5 Conclusion

Our analysis shows that in the discursive examples in Russian considered here, synonymy is not as complete as it appears at first glance. Using examples of synonymous particles and phrasemes, we have demonstrated that seemingly fully synonymous particles and constructions such as *to-to i ono, to-to i est'* and *to-to i delo* or *imenno* and *kak raz* differ with respect to their syntactic use and certain semantic and pragmatic features, as well as from the perspective of diachrony, style, and frequency.

Our work with the corpus indicates that there are almost no cases in which the use of one near-synonym is correct and another entirely impossible. Interchanges are practically always permissible, albeit with different frequency and different degrees of cognitive entrenchment. Sometimes substitutions produce slight shifts that are always apparent to a sensitive native speaker. The following assumptions were advanced as working hypotheses of the study:

(a) differences between near-synonyms are determined not only by semantics, but can also be motivated by pragmatics and usage;
(b) syntactic differences between the synonyms under consideration are motivated by distinctions in their semantics;
(c) different languages, when they do not have good semantic equivalents, solve the problem on the level of the utterance, where they use not systematic equivalents but entirely different parallels – translational equivalents, which are determined more by pragmatics than by semantics.

The corpus data have confirmed the three hypotheses. As for the first hypothesis, the differences between *imenno* and *kak raz* have primarily to do with semantics. If

imenno as a focus particle singles out the element of the utterance that is within its sphere of influence, the semantics of *kak raz* presumes the explicit or implicit comparison of two values. All other differences – pragmatic, syntactic, and usual – derive from this basic semantic distinction. The differences between *imenno* and *to-to i ono* depend in equal measure upon semantics and pragmatics. From the semantic point of view, *imenno* as an independent utterance underscores what the speaker regards as the main element of the situation, whereas *to-to i ono* focuses the main element of the preceding utterance. This binds *to-to i ono* more closely to the verbal form. In most contexts, of course, this difference is neutralized. The principal pragmatic difference between *to-to i ono* and *imenno* is that *to-to i ono* contains a polemical element, which in part explains the semantic difference noted above. The differences between *to-to i ono, to-to i est'* and *to-to i delo* are based solely on usage. Of the three constructions, only *to-to i ono* is typical of modern usage, while *to-to i est'* and *to-to i delo* are perceived as archaic, which is indeed confirmed by the corpus data. The data also confirm that *to-to i delo* is used much more seldom than *to-to i est'* and can be considered practically obsolete.

The second hypothesis on the semantic motivation of syntactic differences has been confirmed on the basis of the opposition between *imenno* and *kak raz*. The fact that *imenno* can function both as a focus particle and as an independent utterance, whereas *kak raz* can only serve as a focus particle, derives from the semantic differences between these units discussed in Sect. 2.1.

The third hypothesis is also confirmed. The analysis (see Sects. 2.2, 3.2 and 4.2) has shown that in order to find a functional equivalent that can be adequately used in the translation of a given context it is not at all necessary to have an equivalent in the language system.

Acknowledgements This paper is based on work supported by the RFFI under Grant 13-06-00403. Thanks also go to Pierre-Yves Modicom (Université Paris-Sorbonne), who read a draft version of the present article, for an interesting discussion of theoretical issues raised by us, and to the anonymous reviewers for a number of valuable comments that we have attempted to take into account.

References

Axmanova, O. S., & Smirnickij, A. I. (1985). *Russko-anglijskij slovar'* [Russian-English Dictionary]. Moskva: Russkij jazyk.

Baranov, A. N., Plungjan, V. A., & Raxilina, E. V. (1993). *Putevoditel' po diskursivnym slovam russkogo jazyka* [Guide to Russian discursive words]. Moskva: Pomovskij i partnery.

Birgegård, U., & Sharapova Marklund, E. (Eds.). (2010). *Norstedts ryska ordbok: rysk-svensk, svensk-rysk* [Norstedt's Russian dictionary: Russian-Swedish, Swedish-Russian]. Stockholm: Norstedts akademiska förlag.

BTS. (2002). *Bol'šoj tolkovyj slovar' russkogo jazyka* [Comprehensive explanatory dictionary of Russian], (Ed.). Sergej A. Kuznecov. Sankt-Peterburg: Norint.

Davidsson, K. (Ed.). (1976). *Russko-švedskij slovar'* [Russian-Swedish dictionary]. Moskva: Russkij jazyk.

Dobrovol'skij, D. O., & Levontina, I. B. (2012). O sinonimii fokusirujuščix častic (na materiale nemeckogo i russkogo jazykov) [Synonymous focus particles in German and Russian]. In *Computational linguistics and intellectual technologies. Papers from the annual international conference "Dialogue 2012"*. Issue 11 (18), (Vol. 1, pp. 138–149). Moskva: RGGU.

Dobrovol'skij, D. O., & Levontina, I. B. (2014). Timiologičeskij komponent v semantike diskursivnyx slov [The timiological component in the semantics of discursive words]. In A. D. Šmelev (Ed.), *Trudy Instituta russkogo jazyka RAN II* (pp. 334–343). Moskva: Institut russkogo jazyka.

Dobrovol'skij, D., & Šarandin, A. (2013). Die Fokuspartikel EBEN und ihre Quasisynonyme in deutsch-russischer lexikographischer Perspektive. In E. Breindl & A. Klosa (Eds.), *Germanistische Linguistik, 221–222* (19–57). Hidesheim/Zürich/New York: Georg Olms Verlag.

Ermolovič, D. I. (2011). *Anglo-russkij i russko-anglijskij slovar'* [English-Russian and Russian-English dictionary]. Moskva: AST, Astrel', Xarvest.

Fischer, K. (2000). *From cognitive semantics to lexical pragmatics: The functional polysemy of discourse particles*. Berlin: Mouton de Gruyter.

Kiseleva, K. L., & Paillard, D. (Eds.). (1998). *Diskursivnyje slova russkogo jazyka: opyt kontekstno-semantičeskogo opisanija* [Russian discursive words: an attempt at a context-semantic description]. Moskva: Metatekst.

Kiseleva, K. L., & Paillard, D. (Eds.). (2003). *Diskursivnyje slova russkogo jazyka: kontekstnoe var'irovanie i semantičeskoe edinstvo* [Russian discursive words: contextual variation and semantic invariance]. Moskva: Azbukovnik.

Kobozeva, I. M. (2006). Opisanie označajuščego diskursivnyx slov v slovare: nerealizovannye vozmožnosti [Describing the signifier of discursive words in the dictionary: Unrealized possibilities]. In *Vestnik MGU*. Serija 9, 2. Filologija.

Kobozeva, I. M. (2007). Polisemija diskursivnyx slov i vozmožnosti ee razrešenija v kontekste predloženija (na primere slova *vot*) [Ambiguity of discourse markers – Can it be resolved in clausal context? (the case of *vot*).] In *Computational linguistics and intellectual technologies. Papers from the annual international conference "Dialogue 2007"*. Vypusk 6 (13), 250–255. Moskva: RGGU.

Kobozeva, I. M., & Zakharov, L. M. (2004). Types of information for the multimedia dictionary of Russian discourse markers. In *Proceedings of the 9th international conference "Speech and computer"*. St-Petersburg: St-Petersburg University.

Levontina, I. B. (2004). Imenno 2, kak raz 1. In J. D. Apresjan (Ed.), *Novyj ob"jasnitel'nyj slovar' sinonimov russkogo jazyka*. Izd. 2 ispr. i dop. Moskva; Wien: Jazyki slavjanskoj kul'tury, Wiener Slawistischer Almanach.

Lubensky, S. (2013). *Russian-English dictionary of idioms*. New Haven: Yale University Press.

MAS – Malyj akademičeskij slovar'. (1985–1988). *Slovar' russkogo jazyka v uetyrex tomax* [Dictionary of Russian in four volumes]. 3-e, stereotip. izd. Moskva: Russkij jazyk.

Molotkov, A. I. (Ed.). (1967). *Fraseologičeskij slovar' russkogo jazyka* [Phraseological dictionary of Russian]. Moskva: Sovetskaja ėnciklopedija.

Padučeva, E. V. (2014). Nestandartnye otricanija v russkom jazyke: vnešnee, smeščennoe, global'noe, radikal'noe [Nonstandard negations in Russian: external, shifted, global, radical]. In *Voprosy jazykoznanija, 5*, 3–23.

Paillard, D. (1998a). *Kak raz* ili Mirom pravit slučaj [*Kak raz*, or The world is ruled by chance]. In Ksenija Kiseleva & Denis Paillard (Eds.), *Diskursivnyje slova russkogo jazyka: opyt kontekstno-semantičeskogo opisanija* (pp. 278–284). Moskva: Metatekst.

Paillard, D. (1998b). *Imenno* ili Kak nazyvat' vešči svoimi imenami. [*Imenno,* or How to call things by their names.]. In K. Kiseleva & D. Paillard (Eds.), *Diskursivnyje slova russkogo jazyka: opyt kontekstno-semantičeskogo opisanija* (pp. 285–293). Moskva: Metatekst.

Paukkeri, P. (2006). *Recipient v russkom razgovore: o raspredelenii funkcij meždu otvetami* da, nu *i* tak [The recipient in Russian conversation: On the distribution of functions between the answers *da, nu,* and *tak*]. Helsinki: Helsinki University.

Romero-Trillo, J. (2009). Discourse markers. In J. Mey (Ed.), *Concise encyclopedia of pragmatics* (2nd ed., pp. 191–194). Amsterdam: John Benjamins.

Šaronov, I. A. (2009). Kommunikativy i metody ix opisanija [Communicative units and methods of their description]. In *Computational linguistics and intellectual technologies. Papers from the annual international conference "Dialogue 2009"*. Vypusk 8 (15), 543–548. Moskva: RGGU.

Sorjonen, M.-L. (2001). *Responding in conversation. A study of response particles in Finnish*. Philadelphia: John Benjamins.

Travis, C. E. (2005). *Discourse markers in Colombian Spanish: A study in polysemy*. Berlin: Mouton de Gruyter.

Wheeler, M., Unbegaun, B., & Falla, P. (Eds.). (1997). *The Oxford Russian dictionary* (Revised and updated Colin Howlett). Oxford: Oxford University Press.

On Concluders and Other Discourse Markers in the Concluding Moves of English and Italian Historical Research Articles

Silvia Cacchiani

Abstract Starting from the assumption that local and disciplinary cultures have an impact on the rhetorical organization of the text and on identity construction within a genre, this paper takes a corpus-assisted approach to genre variation across English and Italian research articles in history. Specifically, the main emphasis lies on *'conclu*'* and its lemmatizations, or, more precisely, on second-level Summarizers and Concluders and with metadiscourse across moves. As will be seen, second-level discourse markers (SLDMs) represent a marked option, in that they add extra meaning to their more general, more transparent, more frequent, and less specific counterparts. Whereas variation within the unit or pattern results from combinations with discourse markers from the same or other categories, variation across English and Italian is better accounted for within an interpersonal model of metadiscourse, in terms of different strategies on the interactional level.

Keywords English • Italian • History research articles • Second-level discourse markers • Conclusions

S. Cacchiani (✉)
Department of Language and Cultural Studies, University of Modena and Reggio Emilia, Modena, MO, Italy
e-mail: silvia.cacchiani@unimore.it

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_11

1 Introduction[1]

The study of research articles (RAs) has long been a major concern of research in English for Academic Purposes (for one, Swales 1990). Recent developments into corpus compilation and the development of query tools, however, have led to efforts to explore other genres, and, importantly, to investigate cross-linguistic and cross-cultural variation.[2] Whereas EAP and register studies alike have thus recently looked at language variation across genres and disciplines (e.g. Hyland and Bondi 2006), our discussion is typical in its focus on cross-linguistic and cross-cultural variation in English and Italian historical RAs. More specifically, we will focus attention on the rhetorical features of their concluding moves. We thus provide a qualitative investigation into *'conclu*'* and its lemmatizations, in an apparent reaction to the general orientation of the field to concentrate on RA introductions (Swales 1990) or abstracts (cf. López-Arroyo 2004, on the move structure of English and Spanish medical RA abstracts, and, more recently, the essays collected in Bondi and Lorés Sanz 2013).

Research on the role played by local and disciplinary cultures and work on the rhetorical organization of the text provide the rationale for this study. EAP research (Fløttum et al. 2006) suggests that what shapes identity within a genre are factors such as the author's national native language culture, the world of the academia – which provides the author with a general academic identity –, the author's discipline and disciplinary identity, features of the genre, and the discourse community. In the same way, we can expect cultural variation for the same genre in different languages.

When we turn to contrastive rhetoric and studies in L2 writing, an immediate issue is that L2 writers tend to reproduce L1 patterns of text organization. Particularly, corpus based studies in lexical research for translation, text production and reception, have shown that the treatment of specific words in monolingual learner's dictionaries and of their translation equivalents in bilingual dictionaries do not always provide a comprehensive account of differences in meaning and use (Siepmann 2005).[3] In the case of connectors, dictionary equivalents may be used differently across languages and genres. In principle, recourse to dictionary equivalents may result in unusual writing, with particular connectors being over- or underrepresented.

[1] I would like to thank two anonymous reviewers, and the special issue editor, Jesús Romero-Trillo, for their invaluable feedback on earlier versions of this paper. Needless to say, the usual disclaimers apply.

[2] Admittedly, as a reviewer rightly points out, much of the recent impetus has come from research into news discourse and the language of economics in particular. See, among many others, Murphy (2005), for a contrastive study of markers of attribution in English and Italian opinion articles, or Musacchio and Ahmad (2009) and Musacchio (2011), for English and Italian economics metaphors.

[3] More particularly, Siepmann (2005: 241–326) provides extensive discussion of the inclusion and treatment of English, French and German second-level discourse markers in the macro- and micro-structures of mono- and bilingual dictionaries.

In practice, if this is certainly true of one-word units, a more pertinent challenge is posed by multi-word units with different degrees of fixedness. Connectors may indeed be seen as a learning, translation and writing problem. When we consider native and non-native writing, phraseological competence is shown to be a feature of native speakers (Howarth 1996). Conversely, fairly proficient non-native speakers transform, misuse, under-represent, over-generalize, or extend specific L2 patterns, which makes their writing less effective (cf. De Cock 1998; Granger 1998; Siepmann 2005, among others; and Ädel 2006, on metadiscourse in L1 and L2 English).

Turning to English-Italian cross-linguistic studies, in their reference grammar of modern Italian Maiden and Robustelli (2000) observe that the same connectors are used differently across the two languages. To take one example, while frequent recourse to connectors such as *'invece'* 'instead' and *'infatti'* 'indeed, but, sure enough' would be a feature of Italian, the underlying coherence relation is more often left implicit in English. Possibly as a consequence of the lack of large comparable and parallel corpora, contrastive and translation studies of English and Italian connectors focus on lexicalized and relatively frequent one-word units of the type listed in bilingual desk dictionaries (cf. Bruti 1999, on *'infatti'* and *'in fact'*; Palumbo and Musacchio 2010, on *'infatti'* and *'invece'*). Our analysis takes the first steps towards redressing the research imbalance between functionally equivalent one-word and multi-word connectors in English and Italian RAs in history. For this purpose, we shall integrate the mainly qualitative results of a preliminary corpus-based and corpus-driven analysis (Sinclair 1991) with a more genre-oriented perspective.

Specifically, we address the issue of identifying a rationale behind the uses, functions and behaviour of *second-level discourse markers* (SLDMs), i.e. cohesive devices which seem to be especially infrequent in the text (Siepmann 2005; see Sect. 2.1).[4] The main emphasis lies on the way Summarizers and Concluders interact with the partially overlapping category of Reformulators and Resumers, and with Inferrers and other categories, within the concluding moves (Swales 1990, 2004) of English and Italian historical RAs. In doing so, we proceed on the assumption that SLDMs introduce more specialized and precise meanings than their more frequent counterparts (usually one-word or lexicalized units), and that these meanings point to an overlap between elements of interactive and interactional metadiscourse. This can be shown by shifting the focus from an initial and much needed overview of the above-mentioned categories, to *'conclu*'*, its lemmatizations and their interplay with other metadiscourse.

[4] To the extent that the analysis we present places the main emphasis on coherence relations and metadiscourse, we use the terms *connector* (or *connective*, cf. Bondi 2013) and *discourse marker* (Siepmann 2005). This enables us to better position ourselves within descriptive approaches to discourse that concentrate on the encoding of structural relationships between segments of text and discourse. For terminological issues, see, among others, Shourop's (1999) tutorial overview of discourse markers and functionally related expressions, Aijmer and Simon-Vendenbergen's delimitation of the terms *pragmatic markers* and *discourse markers* (2006: 3–4), or Bondi's (2013) encyclopaedic entry on connectives and cognate terms.

The major advantage of a corpus-assisted approach that uses insights from contrastive rhetoric, descriptive work on discourse markers, and genre-oriented research on academic metadiscourse and disciplinary cultures, rests on the interesting reflection that it will offer on RA Conclusions. First, because this is a section most often neglected in genre-based EAP studies and, second, because of its contribution to research into the whys and wherefores of English and Italian multi-word units as expressions of specific local and disciplinary cultures.

2 Materials and Methods

The data for this study comes from two corpora, *HEM-History_EN* and *HEM-History_IT*. The *HEM-History_EN* was built and is currently held at the University of Modena and Reggio Emilia. It comprises approximately 2,700,000 tokens. The articles were downloaded electronically from peer-reviewed academic journals nominated by disciplinary experts among the leading international publications in history. The journals in the English corpus are: *American Historical Review* (AHR), *American Quarterly* (AQ), *Gender & History* (GH), *Historical Research* (HR), *Journal of European Ideas* (JEI), *Journal of Interdisciplinary History* (JIH), *Journal of Medieval History* (JMH), *Journal of Social History* (JSH), *Labour History Review* (LHR), *Studies in History* (SH). They span the years 1999–2000.

Not surprisingly, the Italian corpus addresses a more restricted, national readership. It covers a parallel range of disciplines in history for the years 1999–2001. All the journals have been nominated by leading Italian historians. The journals comprising the Italian corpus to date are: *Dimensioni e problemi della ricerca storica* (DPRS), *Meridiana* (MER), *Passato e presente* (PeP) – approximately equivalent to AHR, AQ, HR, SH; *Il pensiero politico* (PP) – approximately equivalent to JEI and LHR; *Intersezioni* (INT) – a close counterpart of JIH; *Quaderni medievali* (QM) and *Studi Medievali* (SES) – both corresponding to JMH; *Società e storia* (SES) – with GH as its closer counterpart.

Since the journals are not available electronically, they have been scanned from printed sources following corpus design methodology. Only approximately 1,000,000 tokens have already reached the final proofreading stage. The investigation is therefore restricted to this initial sample, from the journals *Il pensiero politico* (PP), *Intersezioni* (INT), *Meridiana* (MER), *Passato e presente* (PeP), *Quaderni medievali* (QM), and to their closest English counterparts: *Journal of Social History* (JSH), *Labour History Review* (LHR), *Journal of Interdisciplinary History* (JIH), *Historical Research* (HR), *Journal of Medieval History* (JMH). The English and Italian used in the aforementioned papers are taken to be representative of the language standard accepted for publication by leading journals in the relevant disciplines.

The focus is on Summarizers and Concluders, Reformulators and Resumers, and Inferrers, and on the way they are found to interact in the text, within multi-word units or extended collocations. Whereas this amounts to taking into account

variability within a string, the relatively small size of our corpus and the inflectional nature of Italian, a pro-drop language, do not make our data a sufficient basis for extensive generalization and practical applications (e.g. in bilingual lexicography and the teaching of L2 academic writing). At this initial stage of research we therefore set out to test whether and to what extent previous observations on the above categories can be extended from other genres and disciplines to historical RAs and from English to Italian.

Specifically, after introducing a working definition of the items under discussion against the background of current debate on phraseology (Sect. 2.1), and with the use Mike Scott's (2012) *WordSmith Tools*, we provide a list of Summarizers and Concluders, Inferrers, Reformulators and Resumers (Sect. 2.2) based on a combination of corpus-based and corpus-driven procedures. Following a brief discussion of the marked nature of SLDMs (Sect. 2.3), the second part of the study (Sect. 3) qualifies as a more genre-oriented investigation. Focusing on *'conclu*'* and its lemmatizations within the relevant concordance lines and extended text in the Viewer, we subsequently look into Summarizers and Concluders, Reformulators and Resumers, and Inferrers, with a view to understanding the rationale behind their uses and functions in the concluding moves (Swales 1990, 2004) of English and Italian historical RAs.

2.1 One-Word and Multi-Word Units

The context of this analysis is provided by previous work in contrastive rhetoric, phraseology and cultural and disciplinary variation in metadiscourse. More specifically, we bank heavily on Siepmann's (2005) corpus-based taxonomy of *second-level discourse markers* (cf. Table 1), which also incorporates studies on metadiscourse (Vande Kopple 1985; Hyland 2005), research on the pragmatics of discourse markers (Fraser 1988), and work in Rhetorical Structure Theory (Mann and Thompson 1988; Mann 1999).

Whereas *first-level discourse markers* (FLDMs) are especially frequent units traditionally recorded in the dictionary, *second-level discourse markers* (SLDMs) are "medium-frequency fixed expressions or collocations composed of two or more printed words acting as a single unit. Their function is to facilitate the process of interpreting coherence relation(s) between elements, sequences or text segments and/or aspects of the communicative situation" (Siepmann 2005: 52). They are relatively infrequent fixed-expressions and collocations (less than 200 tokens per million words) and, we may want to add, combinations of one-word units. They allow for variation of at least one element within the recurring pattern, and they are *cue phrases* in the sense of Knott and Dale (1994) and Knott and Sanders (1998). Although the units gathered from our corpora are highly infrequent and cannot be viewed as SLDMs at least in this respect, we still retain the label for lack of a better term.

SLDMs may result from accumulation of markers (e.g. *'(First) we should consider'*; *'To paint an extreme example, consider'*) and are not restricted to *lexical bundles* or

Table 1 Siepmann's (2005) taxonomy of second-level discourse markers

	Category	SLDM
1	Comparison and Contrast markers	*The same can be said for*; *Analogously*; *It is one thing … It is another*
2	Concession markers	*It would be a mistake to* INF; *[Although] It could be argued …, it is also worth remembering that*
3	Exemplifiers	*As with*; *To paint an extreme example, consider*
4	Explainers	*This is because*; *The explanation seems*
5	Definers	*An X is a Y such that*; *Narrowly defined,*
6	Enumerators	*(First) We should consider*; *Beyond this*
7	Summarizers and Concluders	*A final point:*; *It remains for me to* INF
8	Inferrers	*So it turns out that*; *This is not to imply that*
9	Cause and Reason markers	*A number of factors account for this.*; *There are two main reasons for this.*
10	Announcers	*I will now briefly describe*; *Consideration of … must be left until*
11	Topic initiators (or Topic shifters)	*It is often said that*; *Now consider*
12	Excluders	*Space limitations preclude*; *This is not the place*
13	Digression markers	*It should be mentioned in passing that*; *Incidentally,*
14	Question and Answer markers	*The question then arises:*; *The next obvious question is*
15	Emphasizers	*It must be emphasized that*; *Note that/Note NP*
16	Informers	*It should be recognized that*; *A first point is that*
17	Clarification markers	*But that is not the point.*; *The key point is that*
18	Suggestors	*One thing is certain:*; *It will be readily seen that*
19	Hypothesis and Model markers	*It is a fair guess that*; *Let us imagine that*
20	Restrictors	*To further confound the picture*; *A further problem is that*
21	Referrers and Attributors	*[Name] argues*; *…, it has been seen that*
22	Reformulators and Resumers	*Put another way,*; *In other words,*

clusters (Scott 2012). That is, word strings that appear in a genre more frequently than expected by chance and occur in multiple texts in that genre (Biber et al. 1999; Biber 2006). Siepmann's (2005) work on SLDMs broadens the picture and shifts the focus from recurrent word strings to variability within the string itself, as in *'To give/take/paint an (extreme) example, (let's) consider/take/turn to'*. Table 1 also reveals that SLDMs can be realized as structurally complete set expressions (e.g. *'But this is not the point.'*) and structurally incomplete ones (e.g. *'Put another way,'*), sentence fragments (anticipatory/dummy-*it* constructions, as in *'It has been seen that'*), and sentence-integrated markers (*'As with'*). To put it with Granger and Paquot (2008), they are phraseological units that serve a textual function: complex conjunctions (*'Given that'*), linking adverbials (*'In other words,'*), textual sentence stems (*'The final point is'*). Additionally, communicative and attitudinal formulae can be found (*'It is clear that'*) and may interact with textual phrasemes.

The determining factor for distinguishing SLDMs is their textual function, which can be identified on the basis of the coherence relation(s) signalled by the corresponding FLDM(s). Within Hyland's (2005, 2008) interpersonal model of metadiscourse, signals of coherence relations typically belong to *interactive metadiscourse*, which helps orient the reader through the text. A second dimension, the *interactional* one, concerns the way writers involve the reader in the text. SLDMs cross-cut both categories. Consider, in this respect, the Emphasizer *note that*, an *engagement marker* in Hyland's (2005) model, which explicitly builds the writer's relationship with the reader. Another example is *'It is clear that'*, which can categorize as an Inferrer, and a *booster*, in that it emphasizes certainty. Conversely, *'It is a fair guess that'*, a Hypothesis marker, also qualifies as a *hedge*. That is, it withholds complete commitment to a proposition. Likewise, *self-mentions*, which refer to the degree of explicit authorial presence in the text measured by the use of first-person pronouns and possessive adjectives and pronouns, introduce a dimension of variation in SLDMs (*'(First) we should consider'*; *'…, it remains for me to'*). Finally, *attitude markers*, which express the writer's attitude to the proposition, occur in diverse combinations with and within SLDMs, as in the Concession markers *'It would be a mistake to'* and *'It is also worth remembering that'*.

Attitudes are forms of *evaluation* on the part of the speaker. Following Thompson and Hunston (2000: 5), by evaluation we mean "the expression of the speaker's or writer's attitude or stance [(Conrad and Biber 2000)] towards, viewpoint on, or feelings about the entities or propositions that he or she is talking about. That attitude may relate to certainty [(epistemic modality)], obligation", the good/bad dimension,[5] relevance/significance, and expectedness. Evaluation has a threefold function: besides revealing the value system of the writer and his community and helping compose a shared value-system with his/her reader, it may have a role in organizing the discourse, and, third, it may help construct and maintain writer-reader relations (Thompson and Hunston 2000; see also Hunston 2010). This brings us back to Hyland's (2005) interactive model of metadiscourse and the growing interest in *participant-oriented metadiscourse* (next to *research-oriented* and *text-oriented* metadiscourse, cf. Hyland 2008). *Participant-oriented metadiscourse* comprises both *stance* features, which convey the writer's attitudes and evaluations (*'are likely to be'*), and *engagement* features, which address readers directly (*'Note that'*).

If, besides developing a sound argument and producing compelling evidence for one's claims, the persuasive force of an academic text also derives from the writer's ability to engage in a convincing dialogue with the reader, interactional metadiscourse

[5] One anonymous reviewer recommends substituting *evil* for *bad*, probably based on Martin and White (2005). In their Appraisal Framework, adjectives such as *'bad'*, *'immoral'*, *'evil'* group together in that they convey a judgment of moral sanction and describe the negative dimension of social praise (as expressed by *'good'*, *'moral'*, *'ethical'*). However, we take sides with Thompson and Hunston (2000) and posit a *good/bad* dimension where *'good'* and *'bad'* (though not *'evil'*) respectively express an evaluation of desirability (positive) as opposed to undesirability (negative). This is most often an accidental quality of the entity, which overlays with its basic referential meaning. Following this view, what is useful can be seen as not only important but also desirable and good in terms of goal-achievement (as in 14a, §75 or 14a, §83).

and evaluation cannot be discounted from our treatment of SLDMs. While we adopt Siepmann's (2005) multilingual, corpus-based taxonomy, we also integrate it with insights from Hyland's (2005, 2008) work on metadiscourse[6] and work on the transmission of evaluation. This shall enable us to identify a preliminary list of *prima facie* functionally equivalent Summarizers and Concluders, Inferrers, Reformulators and Resumers, and then concentrate on their diverse interaction in RA Conclusions.[7]

2.2 *Summarizers and Concluders, Reformulators and Resumers, Inferrers*

Summarizers and Concluders (Quirk et al. 1985: *summatives*) may signal the last element in a list (*'finally'*) or be used to sum up (English: *'altogether'*, *'then'*, *'therefore'*, and more formal expressions like *'to conclude'*, *'in conclusion'*; Italian: *'In/in breve'*, *'Allo scopo di sintetizzare'*). Besides introducing the final point in an enumeration, they can introduce a short *summary* of the preceding text, often serving what Siepmann (2005) calls a *solutionhood* function. Summarizers and Concluders partly overlap with Reformulators and Resumers, which reword the lexical content of a text span while also providing additional illustrative, explanatory material. In their turn, both Summarizers and Concluders and Reformulators and Resumers tend to combine with Inferrers and also serve as Inferrers. Inferrers (Quirk et al. 1985: *resultives*) indicate that the truth of one statement follows from the truth of the former. The relevant FLDMs are English *'thus'*, *'therefore'* and Italian *'dunque'*, *'pertanto'*, *'quindi'*.

In this section we provide lists of functionally equivalent English and Italian Summarizers and Concluders, Reformulators and Resumers, and Inferrers. Tables 2, 3, and 4 summarize the results of a number of corpus-based and corpus-driven searches (Sinclair 1991). After running five-, four-, three-, and two-token WordLists to get a preliminary list of items, we moved on to a manual selection of possible candidates for analysis on the basis of their concordances and, accordingly,

[6] For recent developments along similar lines, see Ghezzi (2014). Following Traugott (2003), Ghezzi (2014: 16) defines *intersubjectivity* as encoding the addresser's attention towards the addressee's cognitive stances and social identities. She then categorizes discourse and pragmatic markers into four intersubjective types and functions: *responsive*, *attitudinal*, *textual interactional*, and *textual interactive*. While clearly intended to address subjectivity and intersubjectivity in the diachronic development of discourse and pragmatic markers, this four-way classification is highly reminiscent of Hyland (2005).

[7] Admittedly, based on procedural encoding and encoding of constituents of conceptual representations (Blakemore 2002), work on discourse markers, metadiscourse and evaluation in EAP would benefit enormously from synchronically-oriented reflection on subtle meaning differences across functionally related units. Also, casting first- and second-level discourse markers as metadiscourse in Cognitive Grammar terms (for instance along the lines of Verhagen 2005), we might gain considerable insights into the extent to which individual units broadly serving the same role may differ as to their ability to manage intersubjective coordination relations. This, however, is matter for future research.

Table 2 Summarizers and Concluders

HEM-History_EN	HEM-History_IT
We may conclude by -ing	*È possibile concludere che* (One can conclude that; It is appropriate to conclude that; It can/must be concluded that)
I'd like/I would like to conclude by –ing	*Come considerazione conclusiva,* (As a final consideration,)
This leads to a further conclusion.	None
(So) NP/DET provides us with grounds for concluding that	*Concludendo,* (By way of concluding,); *È possibile concludere che* (One can conclude that; It is appropriate to conclude that; It can/must be concluded that)
In conclusion,	*In conclusione,*
A final point:	None
Let us now turn to our final point.	*Veniamo ora alle conclusioni (che è possibile ricavare dal nostro lavoro).* (Let us turn now to the conclusions (that can be drawn from our work).)
To conclude,	*Per concludere,*
To sum up,	*In sintesi,* (In summary,)
None	*Allo scopo di sintetizzare (con maggior precisione),* (To sum up (more specifically,))
What I conclude is that; *I conclude that*	*Come considerazione conclusiva,* (As a final conclusion,)

Table 3 Reformulators and resumers

HEM-History_EN	HEM-History_IT
In a word,	*In breve,* (In short:); *In estrema sintesi,* (Summarizing briefly,; Lit. As a very brief summary,; Very briefly,)
(and) (More/more) specifically,	*(e) (Più/più) in particolare,*; *Con maggior precisione,*
…, to be specific,	*Mi riferisco, in particolare, a*: (Specifically, I am referring to/ talking about)
We might call this	*Si tratta di* (This is)
Another way … is to	*Detto altrimenti,* (Put differently,)
…, also called	*…, altrimenti definito*; *…, detto altrimenti*
In another way,	*In altre parole,* (In other words,)
To put it differently/another way,	*In altri termini,* (In other words,)
Put another way,	*Detto altrimenti* (Put differently,)
As discussed above,	*Come accennato sopra,* (As suggested above,)
To conclude/sum up,	*Si può (quindi) concludere che* (It can/must be (thus) concluded that; One can/must conclude that); *Per concludere,* (To conclude,); *Concludendo,* (By way of concluding,); *In conclusion,* (In conclusion,)
NP/DET can be summarized as follows.	*Si può sintetizzare sottolineando* (One can summarize by highlighting; A summary can be made by highlighting)
NP/DET can be summarized by the following table.	*La tavola riassume/sintetizza i dati.* (The Table shows/ summarizes the data.)
To summarize; *Summarizing:*; *In sum*	*In (estrema) sintesi,* (Summarizing briefly,; Lit. As a very brief summary,; Very briefly,); *Concludendo,* (By way of concluding,); *In conclusione,* (In conclusion,)
None	*Proviamo a riassumere (NP/embedded clause)* (Let us try to sum up (NP/embedded clause))
None	*Se dovessimo riassumere schematicamente gli elementi salienti,* cond II. (If we had to outline crucial facts, NP would INF.)

Table 4 Inferrers

HEM-History_EN	HEM-History_IT
The corollary (to such/to this/of this) was/ is that	*Questo ha rilevanti implicazioni per* (This has important implications for)
(Clearly) the implication (here/of this) is that	*Ciò/esso implica che/NP* (This/It implies that/NP); *Le implicazioni di ciò/esso VP* (Its implications VP)
The (simplest) conclusion is (thus) that	*Si osserva chiaramente che* (One can clearly observe that; It can be clearly observed that; It is clearly evident/shown that)
From which/this it follows that	*Da NP/DET appare evidente che* (It is clear from NP/DET)
It follows from this (therefore) that	*Da cui* (Hence,); *Da queste considerazioni risulta che* (It follows from these considerations that)
It (therefore) comes as no surprise that	None
It is obvious/evident that; *What is obvious is that*	*Si osserva chiaramente che* (One can clearly observe that; It can be clearly observed that; It is clearly evident/shown that)
Hence, NP/DET are likely to affect	*Questi dati confermano che* (The data demonstrate/confirm that)
It (therefore) seems likely (therefore)/ appears that	*Ciò indica probabilmente che* (This probably suggests/indicates/shows that)
This is not, of course, to imply that	None
ADJ by implication,	None
That this is the case is (further) suggested by; *That this is not the case is clear/ evident/obvious from*	None
As a result/as a consequence,	*Questi risultati indicano (dunque) che* (These results (thus) indicate/show that)

of their function(s) in context. Whereas cross-linguistic equivalents are matched in the table on the basis of meaning, function and (where possible) structure, a closer investigation into their frequency of occurrence across the two corpora is matter for future research. As is only natural, the shorter the unit, the more frequent its use, and, similarly, the less variable the unit, the more frequent its use. Optional items are given in round brackets and a slash separates alternative options. They are more often FLDMs (English *'thus'*, *'therefore'*; Italian *'dunque'*, *'quindi'*) or stance features and speech act modifiers (cf. Searle and Vandervecken 1985; Merlini Barbaresi 1997), e.g. English *'More specifically'* and adjective selection (e.g. *'It is clear/ evident/obvious from'*); Italian *'Più in particolare'* or *'Con maggior precisione'*.

For the sake of clarity, Italian discourse markers are rendered literally when the corpus does not return any English counterpart, e.g. Italian *'Allo scopo di sintetizzare (con maggior precisione)'* 'To sum up (more specifically)' – that is, an *elegant variation* (Siepmann 2005) of *'In sintesi'* (the Italian analogue of *'To summarize'* and *'To sum up'*). Literal translations are also given in the common case of lack of structural equivalents, e.g. Italian *'Come considerazione conclusiva'* 'As a final conclusion' and English *'What I conclude is that'* or *'I would like to conclude*

by -ing'. Third, another important set comprises broadly functional equivalents that differ along the micro-pragmatic dimension. To take one example, consider *'Come considerazione conclusiva'* 'As a final conclusion': unlike the self-mentions *'What I conclude is that'* and *'I would like to conclude by -ing'*, *'Come considerazione conclusiva'* works towards conciseness, depersonalization and objectivation (Gotti 2008). Also, compare *'È possibile concludere che'* 'One can conclude that; It is appropriate to conclude that; It can/must be concluded that' and *'We may conclude by -ING'*. The third-person impersonal dummy-*it* construction *'È possibile* INF*'* in *'È possibile concludere che'* expresses the writer's attitude to the proposition by pointing to the strength of the immediately following inferential conclusions. Conversely, inclusive-*we* in *'We may conclude by -ing'* mentions the writer, builds and strengthens his/her relationship with the reader and, crucially, construes the latter as a participant in the communicative situation.

2.3 Why Second-Level Discourse Markers?

In this section we address the issue of recourse to SLDMs where more frequent FLDMs are available for selection. Assessing their use against the parameters put forth within different approaches to markedness/unmarkedness suggests that they represent the marked member of the opposition.

First, SLDMs show medium to low frequency of use. This is perfectly in line with Greenberg's (1966) *principle of distribution*, according to which the number of unmarked members is always greater than that of marked members. The unmarked member of an opposition is the dominant and most common one, whereas the marked member shows higher specificity and complexity in many respects, thus occurring less frequently (Battistella 1990). Specificity must therefore play a role in motivating recourse to SLDMs. The other way round, cf. Waugh and Lafford's (1994) discussion on the *principle of dependency*, specificity would imply that the unmarked element has an enveloping general meaning (set) while the marked one depends on it (subset). If the unmarked category is always presupposed, then the unmarked member remains the only representative of one category when some specific features of the other members are neutralized (cf. Trubetzkoy 1939/1969; Jakobson 1936/1971; Lyons 1977 and the discussion on the *principle of neutralization*). What this argument boils down to is the marked nature of SLDMs. Turning now to Tables 2, 3, and 4, the data suggests that SLDMs can be variously realized as set expressions, sentence fragments and sentence-integrated markers. Highly infrequent one-word items or lexicalized units have also been included. It is clear that variation can result from introducing a second function or a metadiscursive feature within a unit. Some examples here are: English *'further'*, a Summarizer, in *'That this is the case is further suggested by'*, altogether an Inferrer, or *'of course'*, a Suggestor which clearly marks speaker's stance in *'This is not, of course, to imply that'*, which serves as an Inferrer. By the same token, Italian *'probabilmente'* 'probably' modulates – or, better, downgrades – degree of certainty in *'Ciò indica*

probabilmente che' 'Lit. This probably suggests/indicates/shows that' (as against, e.g., *'Ciò/Esso implica che/NP'* 'Lit. This/It implies that/NP'). Another example is *'Si può sintetizzare sottolineando'* 'Lit. One can summarize by highlighting; A summary can be made by highlighting', which comprises a Resumer and an Emphasizer.

Second, SLDMs may also combine and interact with their FLDMs, e.g. English *'therefore'* in *'It (therefore) comes as no surprise that'*, or Italian *'quindi'* 'thus, therefore' in *'Si può quindi concludere che'* 'Lit. It can/must be thus concluded that; One can/must conclude that', or *'dunque'* 'thus' in *'Questi risultati indicano dunque che'* 'These results (thus) indicate/show that'. In this case SLDMs specify the meaning and function of FLDMs, most often giving a more precise meaning (e.g. Italian *'In estrema sintesi'* 'very briefly, ultimately; Lit. In a very brief summary'). Together with the FLDM, they can be seen as a special type of lexical focus markers (in the sense of König 1991), which contribute communicative dynamism and point to new/relevant information in the sentence.

Having said this, we are now able to turn to *'conclu*'* and its lemmatizations in the Conclusions of English and Italian historical RAs. As suggested above, central to the analysis is the idea that SLDMs are the more specific counterparts of FLDMs and may interact with interactional metadiscourse and evaluation to different extents. Zooming in on Concluders, we shall see, they are often found to serve a double purpose and thus overlap with Inferrers, combine with Reformulators and Resumers, or interact with other discourse markers.

3 Results and Discussion

To better characterize the role played by Summarizers and Concluders and the way they overlap and interact with both Reformulators and Resumers and Inferrers, we now turn to the more genre-oriented part of our investigation and concentrate on the use of *'conclu*'* and its lemmatizations (English *'conclu*'*: *'conclude(d)'*, *'conclusion(s)'*; Italian *'conclu*'*: *'conclusivo/a'* 'concluding, Sg', *'conclusivi/e'* 'concluding, Pl', *'conclusione'* 'conclusion', *'conclusioni'* 'conclusions', *'concludere'* 'to conclude' and its inflected forms) in the rhetorical-argumentative structure of the text and in its concluding *moves* (that is, the "discoursal and rhetorical unit[s] that perform a coherent communicative function in […] discourse", Swales 2004: 228).

To address this issue, for each corpus we proceeded as follows: as a first step, we downloaded the concordances for *'conclu*'* and its lemmatizations. Using the Viewer tool and the Concordancer, we were then able to take a closer comparative look at its uses in the Conclusions. After dealing with sections introduced by an illocution signal (e.g. *'Conclusions'*; *'Conclusioni'*), the remaining part of the analysis is devoted to *'conclu*'* and its lemmatizations in the Conclusions. Our starting point is Bondi and Mazzi's (2008: 164) characterization of historical RA Conclusions as *inferential conclusions*. Though the Conclusions are not always nor exclusively labelled as such, they encapsulate (Sinclair 1993), re-state and evaluate (previous) findings. Four moves can be identified: (a) Re-stating findings; (b) Signalling inferential conclu-

sions; (c) Establishing links between writer's contribution and broad disciplinary debate; (d) Speculating about future/practical implications.

If SLDMs add extra-meaning to more general, more frequent, and less specific options (Sect. 2.3), their use can be accounted for in terms of different choices with respect to types and degrees of evaluation and interactional elements. Our final analysis thus regards: (a) how *'conclu*'* interacts with other discourse markers to mark coherence relations; (b) how it assists the writer interact with the reader; (c) how it combines with evaluation across rhetorical moves. To enter more specifically into the analysis, within the examples selected we adopt the following conventions: single underlining is used for discourse markers and italics to signal participant-oriented metadiscourse. Square brackets are used to label the category of the discourse marker and to add comments on dialogic/monologic positioning, epistemic commitment, evaluation and move structure.

Excluded from the investigation are: (a) examples which situate *'conclu*'* and its lemmatizations in the Introduction and indicate research article structure (Swales 1990), as shown in (1a) and (1b); (b) examples which situate the lemmatizations of *'conclu*'* in the Results section, where the author details sequences of events (2a, 2b); (c) instances in which *'conclu*'* signals Reference and Attribution (3a, 3b):

(1a) 179 sense or cosmological in a dualist one. In conclusion **[**Concluder; *narrative discourse***]**, *I* shall address **[**Announcer**]** some of these
[hem-hi\jmh\264(20~1.txt 62]

(1b) 116 un forte sfondo comune. *Proveremo* dunque **[**Inferrer**]** in conclusione **[**Concluder/Enumerator; *narrative discourse***]** a ipotizzare **[**Announcer; Concluder**]** –
[rastor\mer\37(200~3.txt 69]
[116 a widely shared background. We shall thus conclude by drawing some tentative hypothesis –]

(2a) 82 de deux reiterating the warnings. It **concluded [***narrative discourse***]** with Senator Humphrey asking
[hem-hi\jsh\332(19~1.txt]

(2b) 79 Antonio di Bernardo de' Medici, **a conclusione di** una lunga lettera inviata **[***narrative discourse***]**
[rastor\qm\47(199~4.txt]
[79 Antonio by Bernardo de' Medici, to conclude a lengthy letter that he sent]

(3a) 90 degenerate hybrids." "Who" Stout concluded **[**Referrer and attributor**]**, "shall form the families of the
[hem-hi\jsh\336b6b~1.txt]

(3b) 267 umanesimo. In sintesi **[**Summarizer**]**, conclude **[**Referrer and attributor**]** Garin, Gentile
[rastor\pep\51(200~2.txt]
[267 humanism. In brief, Garin concludes, Gentile]

3.1 'Conclu' in English RAs*

The corpus returns 277 concordance lines for *'conclu*'* and its lemmatizations. Only 70 instances, however, are used as Concluders in the concluding moves. As a heading, *'Conclu*'* serves a prospective function (Sinclair 1993) in sections labelled *'Conclusions'/'Conclusion'* (3 hits each), *'Conclusions and implications'* (1 hit). *'Conclu*'* is an illocution marker which signals the underlying speech act. It is a general noun that indicates the communicative goal of the immediately following paragraphs.

In the first example we examine (4a), the writer starts off introducing his counterargument, based on *'variable attestation'* (§102) as against *'conventional assumption'* (§96).

(4a) 96 **CONCLUSION** The conventional assumption that women's identity (unlike that of men) is intrinsically defined in terms of marital status, [...] flows logically from the assumption that women are either customarily or legally under the guardianship of men. [...]
102 But **[FLDM: Restrictor]** the variable attestation of other types of appositives upsets this logic **[***Re-stating principal findings and introducing counterargument*; INANIMATE SUBJECTS (SHELL NOUNS)**]**.
[hem-hi\jmh\253(19~4.txt]

The Conclusions then link to the interpretation of historical events via recourse to personal evaluation (*'unrealistic'*) in a second-level Concluder (4b, §103: *'It would be unrealistic to conclude, for example, that'*): the writer introduces his/her inferential conclusions, which follow logically from the data. Here, Inferrers represent the most frequent discourse marker (e.g. 4b, §103; 4c, §105). When embedded in this type of Conclusions, *'conclude'* (4b, §103) links up to the argumentative discourse. *'As just noted'* (4b, §104) jumps back along the narrative discourse line to briefly summarize events. Third, a particular line of reasoning or action is recommended as logically following from the data (4c, §105: *'The great variety … suggests that'*) by making recourse to *should*-conditionals and the passive voice (4c, §105: *'should be used very cautiously, if at all'*). Notice, however, that *'very cautiously'* and *'if at all'* turn the speech act into an act of cautioning.

(4b) 103 *It would be unrealistic to* **[Concession marker]** conclude, for example **[FLDM: Exemplifier]**, that **[Concluder/Informer/Inferrer;** DUMMY-IT**]** a woman who lacks any appositive specifications was not a citizen or did not work for a living **[***Inferential conclusions; Interpreting events***]**.
104 As just noted **[Summarizer]**, it is only in the case of designations of high social rank that the absence of a relevant epithet invariably signifies that the person in question was indeed **[Emphasizer]** not invested with that social status **[***Re-stating findings***]**.
[hem-hi\jmh\253(19~4.txt]

(4c) 105 The great variety of phrases used to identify women in Douai **[*Re-stating findings*]** *suggests that* **[**Inferrer; *Inferential conclusions*; *INANIMATE SUBJECT*] this particular piece of conventional wisdom *should be used very cautiously, if at all* **[**Concession marker; *Recommendation for action; argumentative discourse*; *DEONTIC AUX; PASSIVE VOICE*]**.**
[hem-hi\jmh\253(19~4.txt]

In a similar manner, Inferrers play a major role in (5). The example illustrates the case of inferential conclusions that follow logically from the data (5a, §74: *'It would be consistent with the evidence to suppose that'*) and link to the broad disciplinary debate via Attribution markers (5a, §74: *'as William of Poitiers notes …, 'he was …"*). The writer then proceeds to signal his/her contribution to the debate (5b, §75: *'The considerations lead me to conclude that'*; 5c, §76: *'The point I wish to make … is that'*), also by making recourse to Suggestors (5b, §75: *'He may well be'*; 5c, §76: *'… does not obviously suggest that'*).

(5a) 74 were so prominent in their support for Eustace, for *as William of Poitiers notes* in an apparent reference to the skirmish of 1051, 'he was æformerly their bitter enemy' **[**Reference and Attribution marker; *Establishing link between writer's contribution and broad disciplinary debate*] (the use of the word æformerly' *should be noted* **[**Emphasizer; *DEONTIC AUX; PASSIVE VOICE*]) and Kent was traditionally a stronghold of the Godwin family. […] *It would be consistent with the evidence to suppose that* **[**Suggestor; *DUMMY-IT*] Eustace was the patron of the Tapestry but **[**FLDM: Contrast marker; *Introducing counterargument*] that it was designed and made on his behalf by English elements who had been favourable to his attack on Dover and who remained favourable to his cause. […]
[hem-hi\jmh\253(19~1.txt]

(5b) 75 These considerations **[***Link to findings*] *lead me* **[***Link to findings; Signalling writer's contribution*] *to conclude that* **[**Concluder/Inferrer; *SELF-MENTION*] Eustace *cannot be dismissed as a less likely candidate than* Odo purely on the basis of the political content of the Tapestry and *he may well* **[**Suggestor; *Emphasizer* well] *be a more likely one. I have also* **[**FLDM: Enumerator] *suggested that* **[**Resumer; *Highlighting and pointing to writer's contribution*; *SELF-MENTION*] the Tapestry was intended as a gift to Odo. […]
[hem-hi\jmh\253(19~1.txt]

(5c) 76 The point *I wish to make* **[**Informer; Concluder], however **[**FLDM: Restrictor; *Introducing counterargument; Highlighting and pointing to writer's contribution*; *SELF-MENTION*], *is that* **[**Concluder/Enumerator; Emphasizer] the content of the Tapestry *does not obviously* **[***Emphasizer*, cf. Hyland (2005)] *suggest* **[**Inferrer; *INANIMATE SUBJECT*] that Odo had a directive or guiding influence over its design […].
[hem-hi\jmh\253(19~1.txt]

Examples (4) and (5) illustrate distinctive features of English historical RAs and allow us to move to a broad discussion of discourse markers, metadiscourse and evaluation. A first point to be made is that evaluation in dummy-*it* constructions is systematically used to signal the legitimacy of data analysis, interpretation and conclusions (4b, §103: *'It would be unrealistic to'*; 5b, §74: *'It would be consistent with the evidence to conclude that'*). Conversely, the writer turns to self-mentions (5c, §75: *'I have also suggested that'*) to take responsibility for his/her claims. Second, SLDMs can be ambiguous between different readings, as in *'These considerations lead me to conclude that'* (5b, §75). Here, *'conclude'* serves as an Inferrer rather than a Concluder, which would simply introduce the last item in a list. This is apparent when *'conclusion(s)'* combines with first-level Inferrers such as *'hence'* or *'thus'* (examples 6 and 7):

(6) 1 also not going to be correlated with R1) **[***Re-stating findings***]**. Hence **[**FLDM: Inferrer**]**, *that* no *substantive* conclusions *ought to be drawn* from the result that T and R1 are not correlated follows *immediately* from the *procedure* **[**Concluder/Inferrer; *DEONTIC AUX; PASSIVE VOICE***]**
[hem-hi\jih\1(1999~4.txt]

(7) 1 The *simplest* conclusion is thus **[**FLDM: Inferrer**]** that **[**Concluder/Inferrer**]** the idea of the Four Highways *is nothing more than a twelfth-century myth*: it was invented by Henry of Huntingdon around 1130 and ***thus*** **[**FLDM: Inferrer**]** had no Anglo-Saxon origins. Those who, like Pollock, try to derive legal principles from it, *fall into error*. Nevertheless **[**FLDM: Restrictor**]**, no matter how fanciful **[**Concession marker**]** the development of the story, the inclusion of the Four Highways in law codes implies that **[**Summarizer/Inferrer; *INANIMATE SUBJECT*] they *should* play a part in *our understanding* of the legal culture of the twelfth century; *only unreconstructed Whiggism would lead one to think otherwise*. **[***Speculating about practical implications***]**
[hem-hi\jmh\264(20~2.txt]

'Conclusion' is frequent in the '*One/the* ADJ/superlative degree of ADJ *conclusion is that*' pattern, as in *'The simplest conclusion is thus that'* (7) and *'One simple, though correct conclusion is that'* (8). In general, adjectives point to the conclusiveness of the argument (*'clear'*, *'categorical'*, *'inescapable'*, *'substantive'*) or characterize the conclusions as legitimate and logically compelling (*'minimal'*, *'general'*, *'simple'*, *'correct'*, *'safe'*). One exception is (9), where *'important'* expresses evaluation for relevance, and the strength of the conclusions is highlighted by bringing to the fore the logical link to the *'evidence'*:

(8) 10 One *simple, though correct,* conclusion is that this represents a degree of
[hem-hi\lhr\1(2000~2.txt]

(9) 24 *The most important* conclusions to be drawn from the evidence relating to vagabondage concern land. Land was [hem-hi\hr\18c016~1.txt]

Example (9) can be seen as an elegant variation (Siepmann 2005) of SLDMs of the type 'General noun *shows/demonstrates/implies that*'. This type, however, is not found in the Conclusions, where it is replaced by the type 'Re-statement of findings *indicates/shows/demonstrates/implies that*'. Consider example (10), which provides a continuation to the inferential conclusions in (4) above. In *'The diversity in phrases … implies that'* (10, §106), *'implies'* reinforces epistemic certainty by pointing to the logical strength of the Conclusions. The writer's commitment to the truth of the proposition is thus reinforced, and the underlying speech act intensified. Consider also *'The combination of this variety … not only indicates … but also suggests that'* (10, §107): though weaker, *'indicate'* and *'suggest'* can be interpreted along the same lines. Inanimate subjects, re-statements of findings and discourse-oriented verbs help characterize the Conclusions as a logical consequence of the research.

(10) 106 The diversity in phrases which are appended to personal names of women (what we have called **[**Reformulator; *Narrative discourse*; INCLUSIVE-WE**]** 'appositives') *implies* that **[**Inferrer; INANIMATE SUBJECT**]** family status was not a rigid standard in terms of which Douaisian society was customarily organized **[***Re-stating findings***]**.

107 The combination of this variety in appositives with the high incidence of women's names unaccompanied by any identifying information at all **[***Re-stating findings***]** not only *indicates* **[**Inferrer; INANIMATE SUBJECT**]** that formulas for identification were unstable, but also **[**FLDM: Contrast marker/Enumerator**]** *suggests* **[**Inferrer; INANIMATE SUBJECT**]** that the nature of women's identity itself was in flux and not yet fully socially determined. **[**§§106-107: *Re-stating findings in inferential conclusions*; cf. example (4)**]**
[hem-hi\jmh\253(19~4.txt]

This is perfectly in line with the writer's withdrawal from the text, also a feature of third-person passive constructions. One example here is *'That no substantive conclusion ought to be drawn from the results … follows immediately from the procedure'* (6), where *'immediately'* points directly to the legitimate and conclusive nature of the research.

Once the writer takes responsibility for his/her conclusions, these are presented as true and consensually given. Accordingly, *'obviously'* signals the assumption of pre-existing shared knowledge in *'The point I wish to make, however, is that the content of the Tapestry does not obviously suggest'* (5c, §76). Example (5c, §76) also illustrates the case of self-mention: while suggesting the efficacy of the relationship between data analysis, interpretation of events, and writer's claims, the writer may

recur to self-mention to point to, and take responsibility for, his/her interpretation and contribution, as in *'These considerations lead me to conclude that'* (5a, §75).

This seems to be a feature of the type of Conclusions in which the writer summarizes his counterargument against widely-accepted claims or conventional assumptions (11). If this is the case, the writer is more likely to also recur to hedges, which mark a statement as plausible rather than certain, e.g. *'would'*, as in *'It would be unrealistic to conclude that'* (4b, §103). In like manner, hedges are at work in *'Perhaps the safest conclusion would be to say that'* (12), where *'perhaps'* and *'would'* clearly downgrade the writer's commitment to his proposition, and *'say'* does not qualify as a strong assertive.

(11) *I* conclude by *suggesting* that [*SELF-MENTION*], even if [hem-hi\jmh\264(20~1.txt]

(12) 46 cline would be ill-judged. *Perhaps* **[***Hedge***]** *the safest* **[***Emphasizer/Booster***]** conclusion *would be* **[***Hedge***]** *to say* that [*INANINATE SUBJECT; SHELL-NOUN*] [Brockworth [hem-hi\hr\177(19~2.txt]

3.2 'Conclu' in Italian RAs*

In Italian RAs, *'Conclu*'/'conclu*'* can be used as a heading and serves as an illocution marker in sections labelled *'Conclusioni'* 'Conclusions' (15 hits), *'Conclusione'* 'Conclusion' (1 hit), *'Conclusioni miste'* 'Mixed conclusions' (1 hit), *'Considerazioni conclusive'* 'Concluding remarks' (1 hit), *'Osservazioni conclusive'* 'Concluding observations' (1 hit), *'Qualche riflessione conclusiva'* 'Some concluding reflections' (1 hit). Given our contrastive focus on discourse markers, each and every Italian example comes with an English translation. Where the relatively more culture-specific counterpart and the literal rendering of the discourse marker are not coextensive, both options are given for the sake of comparison. In an attempt to provide translations that are adequate to the culture-specific target-genre conventions, target-text translations are based on subsequent corpus searches for the frequency and use of specific words and expressions, also with stop-words in their (extended) left and right context. Particularly, search nodes comprised lemmatizations of lexical words (e.g. *'concl*'*, *'observ*'*, *'appea*'*), and verbs were further tested for uses in the active and passive voice, with third-person singular inflection (as in *'it is'*, *'it appears'*) or inanimate subject (as in *'the results suggest'*).

Although the overall move structure of the concluding sections does not radically differ from the English Conclusions, Italian Conclusions unfold in slightly different manners. Specifically, Italian Conclusions do not appear to establish links between the writer's contribution and the disciplinary debate. Instead, they highlight the writer's interpretation of the findings and, at times, speculate about practical implications (14). The type of interpretation given is presented as legitimate and not falsifiable via recourse to third-person singular impersonal constructions of the type *'Da queste considerazioni risulta che'* 'Lit. It follows from these considerations that'

(13a, §67) or to impersonal-*si* constructions, which work towards depersonalization (Gotti 2008) at the detriment of dialogic positioning and reader engagement. If we now turn to discourse markers, though various types are found to interact in longer units, there seems to be a pronounced preference for Inferrers over Concluders or other types. For instance, the abovementioned *'Da queste considerazioni risulta che'* (13a, §67) could be seen as a discourse marker serving the dual function of Inferrer/Concluder. Importantly, frequent recourse to discourse markers with dual functions helps the writer re-state and evaluate findings in inferential conclusions (examples 13 and 14). Thus, *'si è potuto verificare NP'* 'We were able to probe; Lit. It was possible to probe NP; NP could be probed' (13a, §68) has a double reading as an Inferrer and Summarizer. *'Considerando, infine, che'* 'Lastly, considering that' (13a, §72) features *'Considerando'*, a Topic initiator, and *'infine'* 'finally, lastly', a Summarizer and Concluder, or Enumerator. Yet another example is *'si può constatare'* 'It is easy to see; Lit. One can observe NP; NP can be observed' (13a, §72), which may suggest and inform at the same time.

(13a) 67 **CONCLUSIONI** Da queste considerazioni *risulta* che **[**Inferrer/Concluder**]** i monasteri che *con certezza sono da ascrivere* **[***DEONTIC EXPRESSION*; *PASSIVE VOICE***]** all'opera fondatrice di Domenico sono San Salvatore di Scandrigli [...] e Santa Maria a Sora, mentre la fondazione di Sant'Angelo sul monte Caccume riguarda *probabilmente* una ecclesia castri**.**

68 [...] All'origine di queste istituzioni *si è potuto verificare* **[**Informer/Summarizer**]** l'intervento di famiglie aristocratiche come i conti di Sabina o quelli di Valva [...]**. [***Re-stating findings*; *IMPERSONAL-SI***]**

72 Considerando **[**Topic intiator**]**, infine **[**FLDM: Summarizer and Concluder/Enumerator**]**, che le famiglie di maggiore rilievo facevano accogliere i loro membri nel monastero o cercavano di entrare nella clientela vassallatica dell'abate, *si può constatare* **[**Suggestor/Informer**]** l'emergere di una gerarchia al vertice della quale vi era la famiglia fondatrice**. [***Re-stating findings in inferential conclusions*; *IMPERSONAL-SI***]**

[rastor\ss\199~3.txt]

[67 CONCLUSIONS *This would lead us to conclude that / The simple conclusion that could be drawn from this is that* {DA QUESTE CONSIDERAZIONI RISULTA CHE: it follows from these observations that} while San Salvatore di Scandrigli […] and Santa Maria a Sora are among the monasteries that can be *clearly* {CON CERTEZZA: with certainty; without any doubts} linked to Domenico's founding work, the founding of Sant'Angelo sul Monte Caccume is probably related to an ecclesia castri.

68 […] *We were able to* {SI È POTUTO INF: it was possible to INF; NP could INF, PASS} probe the intervention of aristocratic families such as the Counts of Sabia or those of Valva […] as leading to the founding of these institutions.

72 *Lastly, considering that* {CONSIDERANDO INFINE CHE} the most prominent families ensured that family members be granted access to the monastery or made efforts to become vassals and be granted an estate by the lord abbot, *it is easy to see* {SI PUÒ CONSTATARE: One can observe; NP can be observed} the emergence of a hierarchical organization, with the founding family at its head.]

Consider also (13b, §73): Definers (*'si presenta come'* 'qualifies as'; *'non si può parlare di'* 'it is not a matter of; Lit. one cannot talk about') interact with an Inferrer/Informer (*'Appare chiaro che'* 'It is clear that') to re-state and evaluate findings:

(13b) 72 Il monastero di San Bartolomeo di Trisulti solleva altre problematiche **[Restrictor]**.

73 Da un lato **[Comparison and Contrast marker]** si presenta come **[Definer]** una fondazione privata, sul tipo di quelle analizzate [...], dall'altro [Comparison and contrast marker] *non si può parlare* di **[Definer; *IMPERSONAL-SI*]** una famiglia in cerca di affermazione all'interno di un determinato territorio. *Appare chiaro che* **[Inferrer/Informer]** le modalità dell'Eigenkloster vengono fatte proprie dai ceti emergenti [...] **[*Re-stating findings*; *3SG IMP*]**.
[rastor\ss\199~3.txt]
[72 San Bartolomeo di Trisulti raises other issues.

73 *On the one hand* {DA UN LATO}, *it qualifies as* {SI PRESENTA COME} a private foundation, much like the ones described above […]. *On the other hand* {DALL'ALTRO}, *it is not a matter of* {NON SI PUÒ PARLARE DI: one cannot talk about} a family seeking to gain prominence and prestige in a certain area. *It is clear that* {APPARE CHIARO CHE NP: it is clearly shown that NP; NP is clearly shown to INF} the new emerging class is now endorsing the system known as Eigenkloster [proprietary monastery] […].]

Once again, second-level discourse markers in (14a) are ambiguous between two different though related readings, e.g. Inferrer and Concluder in (14a, §75): *'Proviamo a tirare le fila dei ragionamenti sviluppati nelle pagine precedenti e a trarre qualche utile implicazione'* 'Let us try to wrap up and explore some important economic implications; Lit. Let us try to bring together the arguments made in the previous pages and explore some useful economic implications'.

A slightly different example is *'È, infatti, evidente che'* 'Indeed, it is clear that' (14b, §81), which combines a first level Explainer (*'Infatti'*) and an Inferrer. In like manner, a first-level Emphasizer or Restrictor (*'In particolare'*) interacts with a Summarizer in *'In particolare, nel testo si è sostenuto che'* 'Specifically, we have shown that; Lit. Particularly, it was argued in the text that' (14a, §83). Example (14a) thus re-states and evaluates findings which allow the writer to speculate about practical implications (14a, §83) and recommend for action (14b, §§88–89):

(14a) 75 **CONCLUSIONI:** Cosa c'entra il Mezzogiorno? **[Question marker]** *Proviamo* a [*IMPERATIVE*; *INCLUSIVE-WE*] tirare le fila dei ragionamenti sviluppati nelle pagine precedenti **[Concluder]**, e a trarre qualche *utile* implicazione **[Concluder/Inferrer]** per l'economia del Mezzogiorno. [...]

77 *Sembra* **[Informer]** invece **[FLDM: Contrast marker]** *utile* **[***3SG IMP***]** richiamare **[Announcer/Emphasizer]** le maggiori difficoltà che emergono […].

81 È, infatti **[FLDM: Explainer]**, *evidente che* **[Inferrer; *3SG IMP*]**, nel calcolo complessivo sarà - a parità di altre condizioni - più rilevante il peso di coloro che, disponendo di redditi e ricchezze più elevate, daranno una valutazione maggiore ai danni subiti o ai benefici ottenuti.

82 A questi limiti è *possibile* porre rimedio.

83 In particolare **[FLDM: Emphasizer/Restrictor]**, nel testo *si è sostenuto* che **[Summarizer]** […] **[*Re-stating findings*; *IMPERSONAL-SI*]**. Ecco **[Topic initiator]**, dunque **[FLDM: Inferrer]**, in che senso quanto precede è *particolarmente rilevante* per il Mezzogiorno. **[*Evaluating findings and speculating about practical implications*]**

[rastor\mer\379f73~1.txt]

[75 CONCLUSIONS: Where does the Mezzo Giorno come into it? *Let us try to wrap up* {PROVIAMO A TIRARE LE FILA DEI RAGIONAMENTI SVILUPPATI NELLE PAGINE PRECEDENTI: let us try to bring together the arguments made in the previous pages} and explore some *important* {UTILE: useful} economic implications for the Mezzo Giorno. […]

77 *By contrast, it seems appropriate to* {SEMBRA, INVECE, UTILE INF: Instead, it seems useful to INF} recall the markedly increased difficulties that we are deemed to encounter […].

81 *Indeed, it is clear that* {È, INFATTI, EVIDENTE CHE} all things being equal people with higher incomes and greater riches carry a greater weight in the overall calculation, in that they place greater value in damages and benefits.

82 There are ways to overcome these drawbacks.

83 *Specifically,* {IN PARTICOLARE,: (More) particularly,} *we have shown that* {NEL TESTO SI È SOSTENUTO CHE: it was argued in the text that} […]. *Thus, this is* {ECCO, DUNQUE,: thus, here is} the sense in which our discussion is especially relevant to the Mezzo Giorno.]

(14b) 87 Il problema *menzionato in precedenza* **[*narrative discourse*]** rischia, dunque **[FLDM: Inferrer]**, di **[Inferrer]** essere *particolarmente severo* nel Mezzogiorno **[*Evaluating and re-stating findings*]**.

88 L'alternativa *sta* nel **[Definer; Contrast marker]** complesso rafforzamento istituzionale di cui si è detto **[*narrative discourse*; *IMPERSONAL-SI*]** e che, non soltanto per questioni legate all'ambiente, appare necessario.

89 Invocare forme di federalismo […], *non appare sufficiente* **[Informer; *3SG IMP*]**. […] **[§ 88-89: *Speculating about future/ practical implications; Recommending for action*]**

[rastor\mer\379f73~1.txt]

[87 The above mentioned issue *may thus* {RISCHIA DUNQUE DI: thus risks to} turn into an extremely serious problem in the Mezzo Giorno.

88 *As suggested above, the way out is to* {L'ALTERNATIVA …, DI CUI SI È DETTO} reinforce the institutions. However complex, this move *seems to be required* {APPARE NECESSARIO: is shown to be required) to solve environmental issues and other problems.

89 Promoting recourse to federalism […] *does not seem to be a viable option* {NON APPARE ESSERE SUFFICIENTE: is shown not to be enough} […].]

Within the Conclusions, *'conclu*'* is used as a Concluder and Summarizer in 78 out of 288 concordance lines, its most frequent lemmatization being *'In/in conclusione'* (17 hits). Examples (15) to (17) illustrate how *'conclu*'* may combine with first-level Inferrers (*'quindi'*, *'sicché'*) which seem to bring to the fore its dual use as a Concluder and an Inferrer:

(15) 61 preponderanza femminile». *Si può* quindi [FLDM: Inferrer] concludere [*IMPERSONAL-SI*] che nel Croce dei primi anni
[rastor\pep\47(199~4.txt 62]
[61 […] women's importance." *We may thus conclude that* {SI PUÒ QUINDI CONCLUDERE CHE: it can thus/therefore be concluded that; one can thus/therefore conclude that} in Croce's early work […].]

(16) 39 alcuni nodi irrisolti - ha prodotto conseguenze disastrose. Sicché [FLDM: Inferrer], in conclusione, senza lasciarsi andare per questo [Cause and Reason marker; *IMPERSONAL-SI*] a *fuorvianti* profezie apocalittiche, *c'è da supporre che* [Inferrer; *3SG IMP*]
[rastor\mer\382dab~1.txt 95]
[39 […] some unresolved issues – had dire consequences. *To conclude,/In conclusion,* {IN CONCLUSIONE: In conclusion} while there is no need to become carried away in delivering most misleading apocalyptic prophecies on these grounds alone, *it is thus safe to presume that* {SICCHÉ … C'È DA SUPPORRE CHE: It must thus be assumed that} […].]

(17) 11 Dunque [FLDM: Inferrer], se [Hypothesis marker] prestiamo fede ai testimoni, *non possiamo che* concludere che [Inferrer; Concluder; *DEONTIC EXPRESSION*; *INCLUSIVE-WE*] Trencavelli aveva una buona cultura ed era in grado di leggere e commentare l'Olivi in latino e in volgare.
[rastor\qm\47(199~3.txt 75]
[11 Therefore, giving credence to the historical sources *we cannot but conclude that* {NON POSSIAMO CHE CONCLUDERE CHE} Trencavelli was quite knowledgeable and could read and discuss Olivi in Latin and Vulgar […].]

Though present, content disjuncts which specify degree of truth (Quirk et al. 1985) and adjectives which express different degrees of certainty in dummy-*it* and copular constructions, are not a favourite choice. Consider, in this respect, *'chiaro'* 'clear' in *'appare chiaro che'* 'it is clear that; Lit. It is clearly shown that NP; NP is clearly shown to INF' (14a, §73). Another example is *'evidente'* – meaning 'which does not

leave room for doubts and alternative interpretations' (DISC: Sabatini and Coletti 2007); my translation) –, as in *'È infatti evidente che'* 'Indeed, it is clear that' (14a, §81).

When signalling practical and future implications of one's own research in the relevant move, pointing to that move as part of the Conclusions, or recommending for action, the tendency is to express evaluations along the moral and social dimensions, e.g. *'appare necessario'* 'it is required; it seems to be required; Lit. it is shown to be required' (14b, §88), or *'non appare sufficiente'* 'is not a viable option; does not appear to be enough; Lit. is shown not to be enough; is not enough'(14b, §89). Also likely are combinations of the categories of usefulness and importance, e.g. *'qualche utile implicazione per l'economia'* 'some important practical economic implications; Lit. some useful economic implications' (14a, §75), or *'particolarmente rilevante'* 'especially relevant' (14a, §83), where importance is assessed in terms of expected practical advantages.

Signalling the conclusiveness of the results is more often the job of other types of comments on the validity of the propositions, and, specifically, of depersonalization strategies that come about with discourse-oriented verbs in dummy-*it* constructions, impersonal-*si* constructions, and directives realized by strong deontic modals. Take *appare* in *'risulta'*, as in *'Risulta che'* 'This would lead us to conclude; The simple conclusion is that; Lit. It follows that' (13a, §67), where *'risulta'* illustrates the case of a discourse-oriented third-person singular impersonal verb meaning 'to be shown that, to be obvious/clear that' (Sabatini and Coletti 2007). Likewise, impersonal constructions with copular *'appare'* and inanimate subject are synonymous with 'to be shown to be' (Sabatini and Coletti 2007). Hence, *'appare necessario'* 'is required; Lit. is shown to be required' (14b, §88) and *'non appare sufficiente* 'does not appear to be a viable option; Lit. is shown not to be enough; is not enough' (14b, §89).

For impersonal-*si*, let us turn now to *'In particolare, nel testo si è sostenuto che'* 'Specifically, we have shown that; Lit. Particularly, it was argued in the text that' (14a, §83). If, on a generic interpretation, an impersonal-*si* or third-person singular impersonal sentence is true for everybody (the non-person comprises both 'I' and 'you'), it is also true for the writer (the non-person refers to 'I'). Or, the other way round, if a sentence is true for the writer (hence, inclusive-*si*, which takes the writer as a referent), it is also true for all readers (quasi-universal, generic-*si*). Both participate in the same process of knowledge construction. Hence, reference to *'nel testo'*, which does not have any counterpart in English: *'nel testo'* metonymically describes a shared communicative situation and, more precisely, an objective line of reasoning which can be repeated and verified, from data through analysis/interpretation to conclusions. In brief, impersonal-*si* points to a communicative event (arguing for a given thesis) that is construed as necessarily shared between the writer and each and every intended reader: based on selected historical documents, facts and data, the interpretations and inferential conclusions drawn by the reader (each and every intended reader) cannot be at variance with the writer's conclusions. Impersonal constructions qualify the argumentation as scientific and objectively grounded.

This is also true for *'si può constatare'* 'it is easy to see; Lit. one can observe; NP can be observed' (13a, §72), which can be interpreted as a relatively strong

claim based on the association with data and findings in the underlying frame. Therefore, the findings and the mechanism of knowledge construction do not leave room for alternative interpretations. Since *'constatare'* can passivize, *'Si può constatare'* 'lends itself to be interpreted as a passive' (Sabatini and Coletti 2007), which points to the ability to get to identical conclusions following from identical premises.

When we adopt impersonal deontic constructions that express inexhaustive possibility (Kaufmann 2012), the type of interpretation given is clearly presented as not falsifiable. One example here is *'c'è da supporre che'* 'it is safe to presume that; Lit. It must/is to be assumed that' (16). The difference between *'c'è da supporre che'* and deontic expressions with first-person plural inclusive-*we* such as *'non possiamo che'* 'we cannot but', is that the latter denotes the writer's interest in explicitly engaging with the reader.

The expression of dialogic orientation, however, represents a significant departure from standard practice in Italian RAs. When present, it comes about with a metonymic reference to the entire knowledge construction process, which characterizes the act as objective, valid and legitimate, e.g. *'Proviamo a tirare le fila dei ragionamenti sviluppati nelle pagine precedenti e a trarre qualche utile implicazione'* 'Lit. Let us try to bring together the arguments made in the previous pages and explore some useful economic implications' (14a, §75). Importantly, based on subsequent corpus searches it is easy to anticipate that *'dei ragionamenti sviluppati nelle pagine precedenti'* would not have a counterpart in English RAs (hence, 'Let us try to wrap up and explore some important economic implications').

Dialogic positioning is also a feature of questions and rhetorical questions. The corpus returns two examples, from the same text: *'Cosa c'entra il Mezzogiorno?'* 'Where does the Mezzo Giorno come into it?' (14a, §75) and *'Potrebbe mai aversi sviluppo economico […]?'* 'Could we ever have economic development […]?' (18, §3). Example (18) is a good illustration of how reduced epistemic certainty in the rare case of tentative conclusions – more precisely, recommendation(s) for future action –, requires major departures from depersonalization and objectivation strategies. Objectivation strategies, we have argued, are particular to the type of inferential conclusions that come with a strong degree of confidence and are therefore presented as compelling, repeatable, and non-falsifiable. More precisely, it is not only the strength of the conclusions, but also the ability to engage with the reader that constitute the main focus of the rhetorical question. Reasoning by asking questions, however, helps develop the argument and bring home a major point, further presented as the only logically possible answer via recourse to a depersonalized form in *'L'obbligatoria risposta è'* 'the only appropriate answer is; Lit. The inevitable answer is'. This provides a transition to a plausible assertion in a set of statements that are hedged via recourse to downgrading adjectives, verbal mood and degree words. More precisely, the findings are given 'rapid' considerations (*'rapide considerazioni'*). Since only careful and detailed considerations would lead to valid conclusions and generalizations, the ensuing recommendation for action is presented as a plausible option rather than as a strong and far-reaching solution: *'Se queste rapide considerazioni sono fondate, viene da concludere che l'espressione "sviluppo locale" dovrebbe essere abbandonata'* 'Lit. If these rapid considerations

are well-grounded, it might be concluded that the expression "local development" should be abandoned' (18, §3). As can be seen, the mechanisms of impersonal knowledge construction that are particular to Italian RAs are at play: inanimate (abstract) subjects (*'considerazioni'*; *'l'espressione "sviluppo locale"'*), impersonal constructions (*'viene da concludere che'*) and passive voice (*'dovrebbe essere abbandonato'*). Nevertheless, 'rapid' (*'rapide considerazioni'*) comes about with another hedging device in order to downgrade the underlying speech act: while embedded in a Pattern I *if*-conditional, recourse to the conditional mood downgrades the strength of deontic *'dovere'* 'must' in *'dovrebbe essere abbandonato'*, a directive which expresses non-exhaustive possibility. As a final transition, however, this conditional directive is further modulated and reinforced to some degree by the background (cf. Smith 2003) argument in *'perché un po' vaga e piuttosto sovraccarica'* 'as slightly vague and rather overcharged', where hedges are at play in the form of degree words (Quirk et al. 1985): while diminisher *'un po"* 'slightly' indicates that the quality described by the adjective is present to a low degree, compromiser *'piuttosto'* 'rather' has a slightly lowering effect and negative overtones.

(18) 3 Potrebbe mai aversi sviluppo economico in un'area senza il contributo di almeno qualche risorsa proveniente da quell'area? **[**Question answer pattern; *rhetorical question***]**. L'*obbligatoria* risposta negativa a questa domanda *spinge* a concludere che **[**Concluder; *INANIMATE SUBJ***]** lo sviluppo è sempre, almeno un po', locale. D'altro canto **[**Comparison and Contrast marker**]**, se **[**Hypothesis**]** in un'area lo sviluppo manca *viene da* pensare che **[**Suggestor; *3SG IMP***]** la causa sia il difetto, in quell'area, di almeno qualcuna delle risorse (intese in senso lato) necessarie. Dunque **[**FLDM: Inferrer**]**, lo sviluppo è anche, almeno un po', non-locale. Se queste *rapide* **[***Hedge***]** considerazioni sono fondate *viene da* concludere che **[**Concluder/Inferrer; *3SG IMP***]** l'espressione «sviluppo locale» dovrebbe **[***Hedge***]** essere abbandonata perché **[**Cause and reason marker**]** un po' vaga **[***Hedge***]** e *piuttosto* **[***Hedge***]** sovraccarica. **[***Conditional prediction based on empirical hypothesis***]**
[rastor\mer\34-35(~7.txt 2]
[3 Could there ever be economic developments in an area without any contribution from at least one local resource? *'No' is the only appropriate answer to the question. This leads us to conclude that* {L'OBBLIGATORIA RISPOSTA NEGATIVA A QUESTA DOMANDA SPINGE A CONCLUDERE CHE: the inevitable negative answer to this question leads one to conclude that} economic development is always local to some degree. *On the other hand* {D'ALTRO CANTO,}, *if* {SE} an area is not developed *we might assume that* {VIENE DA PENSARE CHE: it might be suggested that}, broadly, some required resources are missing. *Thus* {DUNQUE}, economic development is also to some extent minimally non-local. *If* {SE} these rapid considerations are well-grounded, *one safe conclusion is that* {VIENE DA CONCLUDERE CHE: it might be concluded that} the expression "local development" should be abandoned *as* {PERCHÉ: because it is} slightly vague and rather overcharged.]

4 Conclusions

In this paper we concentrated on the lemmatizations of *'conclu*'* and their uses as Summarizers and Concluders. More specifically, besides discussing their interaction with the partially overlapping categories of Reformulators and Resumers, and Inferrers, we investigated in what ways they are found to combine with other categories and, more generally, with other metadiscourse in the Conclusions of English and Italian historical research articles. This enabled us to look into the reasons behind their use while also offering some reflections on the move structure of RA Conclusions.

Given the need for scholars in non-English medium institutions to publish in international high-impact journals, this brings us to the question of what the observed (dis-)similarities mean for effective L2 *writing-for-publication* (Hyland 2013).[8] With Hyland (2013: 68), we argue for the need of small elective writing-for-publication programs that concentrate on the development of genre awareness, disciplinary and genre-specific grammar ability, use of lexico-grammar and knowledge of rhetorical structures. We would not claim that we have provided a thorough description of these categories,[9] but we believe that we can conclude by offering some initial suggestions which might prove useful in creating consciousness rising (Svalberg 2007) teaching materials, activities and tasks (Ur 2012) targeted at the needs of PhD students and scholars in history based at Italian institutions.

When we compare the move structure of the Conclusions of English and Italian research articles in history, the differences are not very striking. Importantly, both qualify as inferential conclusions that flow logically from data selection and analysis. English Conclusions proceed through four rhetorical moves: (a) Re-stating and evaluating findings; (b) Signalling inferential conclusions; (c) Establishing links between writer's contribution and broad disciplinary debate; (d) Speculating about future/practical implications (a clearly optional move). As small as our corpus might be, the data suggests that move 'c' is considerably more likely to occur in English than Italian. Though more research is needed to explore whether this observed preference can be substantiated, in general we do not expect to find links to the broad disciplinary debate in the concluding moves of Italian historical research articles.

[8] Scholars specializing in Italian history at non-Italian medium institutions also need to gain full command of the genre features and conventions set by nationally recognized scholars. For reasons of space, however, we restrict discussion to the more frequent case of Italian researchers working in non-English medium institutions.

[9] This study can be seen as a contribution to the vast area of studies in the rhetorical organization of the text, but also to the growing literature on local and disciplinary cultures. Having only sought to shed some light into the uses and internal variability of a restricted set of discourse markers, however, it is clear that future descriptive research must consistently take into account the quantitative dimension and concentrate on (dis-)similarities in the lexicalization of coherence relations across English and Italian RAs. This amounts to concentrating on position and frequency of syndetic and asyndetic coordination and subordination within specific moves, as well as variability in the lexicalization of coherence relations, within an interpersonal model of metadiscourse.

Researchers who join a specific writing-for-publication course exhibit suitable to high disciplinary knowledge and expertise, and adequate to full mastery of the research article genre in their native language. They have full awareness of the centrality of inferential conclusions in historical research articles. This can provide a starting point for in-class discussion of move structure and discourse markers across languages and local disciplinary cultures. Ability to join in discussions along these lines is to be seen as assisting in performing genre-awareness tasks and activities that engage learners in comparing and contrasting enriched input (Ellis 2009) based on adapted Conclusions.

Another important point concerns second-level discourse markers. Researchers that join a small elective writing-for-publication course are fully engaged (Svalberg 2007, 2009), upper-intermediate or advanced L2 learners. This means that following recycling and consolidation of previously acquired structures (connectors and linking phrases) along the lines of traditional student grammars, more genre- and disciplinary-specific tasks and activities should be created to encourage in-class discussion about the types and categories of first- and second-level discourse markers, and, importantly, about their interaction in the text. In-class discussion would be based on learner experience in the native language and culture and on an array of authentic corpus examples that provide enhanced input for the target of study.

It is apparent from the analysis that SLDMs are marked options, which add extra meaning to their less specific and more general, more transparent, and more frequent first-level counterparts. Variation within the unit results from the insertion of FLDMs and from combinations within the extended concordance line with discourse markers from other categories. Within the Conclusions, English *'conclusion(s)'* and Italian *'conclusione/i'* take on a dual reading – as Concluders and Inferrers –, which is brought to the fore in combination with first-level Inferrers (e.g. *'The simplest conclusion is thus'*; *'Sicché, in conclusione* 'Lit. Thus, to conclude; Thus, in conclusion'). While we acknowledge that further corpus-based and corpus-driven investigation of first- and second-level discourse markers is necessary to learn more about their use, it is worth considering that a major mismatch concerns different interactional concerns across English and Italian concluding moves. Particularly, English *'conclusion'* (rather than *'conclusions'*) is frequent in the '*One/The* ADJ/(superlative degree of) ADJ *conclusion is that*' pattern, where the adjective points to the conclusiveness of the argument (*'clear'*, *'categorical'*, *'inescapable'*, *'substantive'*), or characterizes the conclusions as legitimate and logically compelling (*'minimal'*, *'general'*, *'simple'*, *'correct'*, *'safe'*). On the contrary, Italian *'conclusioni'* 'conclusions' (rather than *'conclusione'* 'conclusion') is not found to combine frequently with epistemic adjectives. This might represent a first potential writing and translation problem.

In this context, if we want to put culture-specific genre conventions at the centre of the discussion, it is interesting to consider the expression of the author's voice in participant-oriented metadiscourse. Broadly, inanimate subjects, restatements of findings and discourse-oriented verbs help characterize the conclusions as the logical consequence of the research in English and Italian. We can suggest the following pattern for English: 'Re-statement of findings *indicates/shows/demonstrates/implies*

that'. In Italian, however, suggesting the conclusiveness of the results is more likely to be the job of the following devices:

- third-person singular impersonal forms (13b, §73: *Appare chiaro che* 'It is clear that; Lit. It is clearly shown that NP; NP is clearly shown to');
- impersonal-*si* constructions (13a, §72: *'si può constatare'* 'it is easy to see; Lit. one can observe; NP can be observed');
- passives (14a, §83: *'si è sostenuto che'* 'we have shown that; Lit. it was argued that');
- strong modals that constrain the reader to follow the writer's line of reasoning (16: *'c'è da supporre che'* 'it is safe to presume that; Lit. it must/is to be assumed that').

This does not mean that we should exclude any reference to sources and findings (if rare) or to specific steps in the process of knowledge construction, e.g. *'Da queste considerazioni risulta che'* 'This would lead us to conclude that' (and its elegant variation 'The simple/safest conclusion that could be drawn from this is that; Lit. It follows from these considerations that') (13a, §67), or *'In particolare, nel testo si è sostenuto che'* 'Specifically, we have shown that; Lit. Particularly, it was argued in the text that' (14a, §83).

The function of the abovementioned strategies is to let data and knowledge construction process speak for themselves. In Italian, impersonal-*si* in particular is key in construing the writer (inclusive-*si*) and each and every intended reader (generic-*si*) as following the same steps in the argument. If rare, possible alternates are examples which depend at least as much on dialogic positioning as on legitimate and logically compelling data selection, analysis and interpretation: *'se prestiamo fede ai testimoni'* 'giving credence to the historical sources; Lit. If we give credence to the historical sources' (17), or *'Proviamo a tirare le fila dei ragionamenti sviluppati nelle pagine precedenti'* 'Let us wrap up; Lit. Let us try to bring together the arguments made in the previous pages' (14, §75).

Compared to Italian, English shows a stronger preference for dialogic positioning, most notably via recourse to inclusive-*we*. While suggesting the legitimacy of the link between data analysis, interpretation and writer claims, however, the writer may recur to self-mention when summarizing his/her counterargument against widely-accepted claims or conventional assumptions. Whereas this option is not available to Italian research article conclusions, it is adopted in English to highlight the writer's responsibility and focus on his/her contribution to the debate, as in *'These considerations lead me to conclude that'* (5a, §75) or *'I conclude by suggesting that'* (7). Possible alternates do away with self-mentions, modulate epistemic certainty via recourse to the conditional mood, other hedges in dummy-*it* constructions, and general nouns that designate steps in the process of knowledge construction (respectively, 4b, §103: *'It would be unrealistic to conclude that'*, and 12: *'Perhaps the safest conclusion would be to say that'*). In summary, though there are (approximate) structural and functional-pragmatic equivalents – e.g. Inferrers such as English *'It is clear that'* and Italian *'Appare chiaro che'* –, it is not hard to see how English and Italian historical research articles shape author's identity

differently. This is immediately apparent from the Italian-to-English translations provided in Sects. 3 and 3.1.

In order to put a more applied perspective on the above characterization of identity construction and participant positioning and engagement in English and Italian Conclusions, it is obvious that a combination of contrastive analysis of enhanced input with translation and (back-)translation tasks (Cook 2010) can promote genre and language awareness and increase knowledge of cross-linguistic and cross-cultural differences. Broadly, under contrastive analysis we group tasks ranging from noticing and cognitive comparison to direct proactive explicit instruction and L2 practice for deductive explicit learning (Ellis 2009) and consolidation. More precisely, learners should engage in tasks that promote discussion of selected input based on specific rules of thumb (Ur 2012) that translate the data interpretation above into intelligible and transparent explanations.

A possible collaborative task could be adapted from the following:

(a) *Instruction*: The following Concluders (b) are taken from English and Italian research articles in history. Read the rule of thumb (c) carefully and answer the following questions (d).
(b) *Examples*: bi. *'Come considerazione conclusiva'*, *'Si può concludere che'*, or *'È possibile concludere'*; bii. functionally adequate counterparts: *'What I conclude is that'*, *'I would like to conclude by -ing'* or *'We may conclude by -ing'* (cf. Table 2).
(c) *Rules of thumb*: ci. Italian Conclusions let the data speak for themselves; writer and readers necessarily follow the same steps in the knowledge construction process; cii. English Conclusions rely on data analysis and interpretation. Additionally, they put the focus on the author's responsibility and on engaging with the intended reader directly.
(d) *Potential questions*: di. Do English Concluders have structural equivalents in Italian?; diia. Which options are available to English?; diib. Does the writer engage with the reader?; diic. Why does English make recourse to first-person plural *'we'*?; diiia. Which choices are available to Italian?; diiib. What's the reason behind using nouns such as *'considerazione'* in *'considerazione conclusiva'*?; diiic. Why does Italian make systematic recourse to impersonal dummy-*it* constructions and impersonal-*si* constructions?; diiid. Considering that, in principle, first-person plural *we* and passive voice are possible alternates of impersonal-*si* constructions, would they represent a viable option in Italian Conclusions?; div. Would you use inclusive-*we* in Italian? If yes/no, why?; dv. Would you use impersonal constructions or the passive voice in English? If yes/no, why?

This type of task clearly involves metalinguistic reflection on identity construction in the concluding moves of English and Italian research articles in history. Recourse to translation and back-translation – to be carried out without time constraints –, would then follow along with subsequent in-class discussion in the form of acceptability judgments and metalinguistic corrective feedback. While the emphasis lies on knowledge about language, this can also contribute to implicit language learning.

Given any of the examples above and the relevant translation (Sects. 3.1 and 3.2), we would thus expect a move from literal renderings to more genre-specific and functionally adequate translations. To take one final example, *'viene da pensare che'* (18) would be rendered literally as 'it might be suggested that', but it would translate into *'we might assume that'* following noticing activities along the lines suggested above. The other way round, *'one safe/legitimate conclusion is that'* (18) would back-translate into *'si può concludere che'* or *'viene da concludere che'* via 'può essere concluso che', where the expression of epistemic modality moves from adjectival modification to auxiliary selection.

References

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins.

Aijmer, K., & Simon-Vendenbergen, A.-M. (Eds.). (2006). *Pragmatic markers in contrast*. Amsterdam: Elsevier.

Battistella, E. L. (1990). *Markedness: The evaluative superstructure of language*. Albany: State University of New York Press.

Biber, D. (2006). *University language. A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson Education.

Blakemore, D. (2002). *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers*. Cambridge: Cambridge University Press.

Bondi, M. (2013). Connectives. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (2nd ed., pp. 891–897). Chichester: Wiley-Blackwell.

Bondi, M., & Mazzi, D. (2008). *Per concludere veramente*: Signalling conclusions in historical research articles in Italian and English. *La Torre di Babele. Rivista di letteratura e linguistica, 5*, 159–171.

Bondi, M., & Sanz, R. L. (Eds.). (2013). *Abstracts in academic discourse. Variation and change*. Bern: Peter Lang.

Bruti, S. (1999). *In fact* and *infatti*: Same, similar or different? *Pragmatics, 9*(4), 519–534.

Conrad, S., & Biber, D. (2000). Adverbial marking of stance in speech and writing. In S. Hunston & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 56–73). Oxford: Oxford University Press.

Cook, G. (2010). *Translation in language teaching. An argument for reassessment*. Oxford: Oxford University Press.

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics, 3*, 59–80.

Ellis, R. (2009). Implicit and explicit learning, knowledge and instruction. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 3–25). Bristol: Multilingual Matters.

Fløttum, K., Dahl, T., & Kinn, T. (2006). *Academic voices across languages and disciplines*. Amsterdam: John Benjamins.

Fraser, B. (1988). Types of English discourse markers. *Acta Linguistica Hungarica, 38*, 19–33.

Ghezzi, C. (2014). The development of discourse and pragmatic markers. In C. Ghezzi & P. Molinelli (Eds.), *Discourse and pragmatic markers from Latin to Romance languages* (pp. 10–26). Oxford: Oxford University Press.

Gotti, M. (2008). *Investigating specialized discourse* (2nd Rev. ed.). Bern: Peter Lang.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 145–160). Oxford: Oxford University Press.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective* (pp. 27–50). Amsterdam: John Benjamins.

Greenberg, J. (1966). *Language universals*. The Hague: Mouton de Gruyter.

Howarth, P. A. (1996). *Phraseology in English academic writing*. Tübingen: Niemeyer Verlag.

Hunston, S. (2010). *Corpus approaches to evaluation. Phraseology and evaluative language*. New York: Routledge.

Hyland, K. (2005). *Metadiscourse. Exploring interaction in writing*. London: Continuum.

Hyland, K. (2008). *As can be seen*. Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*, 4–21.

Hyland, K. (2013). Writing in the university: Education, knowledge and reputation. *Language Teaching, 46*(1), 53–70.

Hyland, K., & Bondi, M. (Eds.). (2006). *Academic discourse across disciplines*. Bern: Peter Lang.

Jakobson, R. (1936/1971). Beitrag zur allgemeinen Kasuslehre: Gesamtbedeutungen der russischen Kasus. In *Roman Jakobson. Selected writings 1*, 23–71. The Hague: Mouton de Gruyter.

Kaufmann, M. (2012). *Interpreting imperatives*. Dordrecht: Springer.

Knott, A., & Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes, 18*(1), 35–62.

Knott, A., & Sanders, T. (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics, 30*, 135–175.

König, E. (1991). *The meaning of focus particles: A comparative perspective*. London: Routledge.

Vande Kopple, W. J. (1985). Some exploratory discourse on metadiscourse. *College Composition and Communication, 36*(1), 82–93.

López-Arroyo, B. (2004). English and Spanish medical research papers and abstracts: How differently are they structured? In J. M. Bravo (Ed.), *A new spectrum of translation studies* (pp. 175–193). Valladolid: Universidad de Valladolid.

Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

Maiden, M., & Robustelli, C. (2000). *A reference grammar of Modern Italian*. London: Arnold.

Mann, W. C. (1999). *Introduction to rhetorical structure theory*. http://www.sil.org/linguistics/RST/rintro99.htm. Retrieved 1 June 2013.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8*(3), 243–281.

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation. Appraisal in English*. Basingstoke: Palgrave MacMillan.

Merlini Barbaresi, L. (1997). Modification of speech acts. Aggravation and mitigation. In *Proceedings of the 16th international congress of linguistics* (Paper No 0353). Oxford: Pergamon Press.

Murphy, A. (2005). Markers of attribution in English and Italian opinion articles: A comparative corpus-based study. *ICAME, 29*, 131–150.

Musacchio, T. (2011). Metaphors and metaphor-like processes across languages: Notes on English and Italian language of economics. In K. Ahmad (Ed.), *Affective computing and sentiment analysis. Metaphor, ontology and terminology* (pp. 89–98). Berlin: Springer.

Musacchio, T., & Ahmad, K. (2009). Variation and variability of economics metaphors in an English-Italian corpus of reports, newspaper and magazine articles. In A. Wallington, J. Barnden, M. Lee, R. Moon, G. Phillip, & J. Littlemore (Eds.), *Corpus-based approaches to figurative language* (Cognitive science research papers, pp. 115–122). Birmingham: University of Birmingham.

Palumbo, G., & Musacchio, M. T. (2010). When a clue is not a clue. A corpus-driven study of explicit vs. implicit signalling of sentence links in popular economics translation. *Rivista Internazionale di Tecnica della Traduzione, 12*, 63–76.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English syntax*. London: Longman.
Sabatini, F., & Coletti, V. (2007). *Il Sabatini Coletti. Dizionario della lingua italiana 2008*. Milano: Rizzoli Larousse.
Scott, M. (2012). *WordSmith Tools 6.0*. Oxford: Oxford University Press.
Searle, J. R., & Vandervecken, D. (1985). *Foundations of illocutionary logic*. Cambridge: Cambridge University Press.
Shourop, L. (1999). Discourse markers. *Lingua, 107*, 227–265.
Siepmann, D. (2005). *Discourse markers across languages*. London: Routledge.
Sinclair, J. M. (1991). *Corpus concordance collocation*. Oxford: Oxford University Press.
Sinclair, J. M. (1993). Written discourse structure. In J. M. Sinclair, M. Hoey, & G. Fox (Eds.), *Techniques of description: Spoken and written discourse. A Festschrift for Malcolm Coulthard* (pp. 6–31). London: Routledge.
Smith, C. S. (2003). *Modes of discourse. The local structure of texts*. Cambridge: Cambridge University Press.
Svalberg, A. M.-L. (2007). Language awareness and language learning. *Language Teaching, 40*(4), 287–308.
Svalberg, A. M.-L. (2009). Engagement with language. Developing a construct. *Language Awareness, 18*(3–4), 242–258.
Swales, J. M. (1990). *Genre analysis. English in academic and research settings*. Cambridge: Cambridge University Press.
Swales, J. M. (2004). *Research genres. Explorations and applications*. Cambridge: Cambridge University Press.
Thompson, G., & Hunston, S. (2000). Evaluation: An introduction. In S. Hunston & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 1–27). Oxford: Oxford University Press.
Traugott, E.-C. (2003). From subjectification to intersubjectification. In R. Hickey (Ed.), *Motives for language change* (pp. 624–647). Cambridge: Cambridge University Press.
Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. Göttingen: Vandenhoeck & Ruprecht. English edition: Trubetzkoy, N. S. (1969). *Principles of phonology* (C. A. M. Baltaxe, Trans.). Berkeley: University of California Press.
Ur, P. (2012). *A course in English language teaching*. Cambridge: Cambridge University Press.
Verhagen, A. (2005). *Constructions of intersubjectivity. Discourse, syntax and cognition*. Oxford: Oxford University Press.
Waugh, L. R., & Lafford, B. A. (1994). Markedness. In R. E. Asher & J. M. Y. Simpson (Eds.), *The encyclopaedia of language and linguistics 5* (pp. 2378–2383). Oxford: Pergamon Press.

Corpus-Based Interpreting Studies and Public Service Interpreting and Translation Training Programs: The Case of Interpreters Working in Gender Violence Contexts

Raquel Lázaro Gutiérrez and María del Mar Sánchez Ramos

Abstract The growing popularity of Public Service Interpreting and Translation (PSIT) in different fields, such as healthcare or legal environments, has highlighted the need for interlingual and intercultural communication between public service providers and users who do not have any or sufficient command of the official language of the public authorities. Training is essential in those settings if we want to successfully achieve the appropriate communication. Interpreting and translation training programs are especially useful in the cases of gender violence victims from other countries, with different pragmatic communication strategies. This article explores the use of Corpus-based Interpreting Studies (CIS) as a methodology to train interpreters in gender-based violence context. After a theoretical introduction on CIS, PSIT and interpreting in gender violence contexts a particular emphasis is placed on the design, compilation process and use of a monolingual corpus and concordance software.

Keywords Public Service Interpreting and Translation • Corpus-based Interpreting Studies • Gender-based violence • Interpreting training • Concordance software

1 Introduction

The necessity of communication links between public service providers and the users who do not have a command of the official language of public authorities has developed what is known as a new academic and professional discipline within Translation Studies, namely Public Service Interpreting and Translation (PSIT) or Community Interpreting and Translation, which covers a wide range of fields including, among others, healthcare, educational, legal and administrative settings.

R. Lázaro Gutiérrez (✉) • M.d.M. Sánchez Ramos
Departamento de Filología Moderna, Facultad de Filosofía y Letras,
University of Alcalá/Group FITISPos-UAH, Alcalá de Henares (Madrid), Spain
e-mail: raquel.lazaro@uah.es; mar.sanchezr@uah.es

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_12

Due to the growing demand for professional PSIT translators and interpreters specialised in particular fields, different research projects have emerged, such as InterMed,[1] funded by the Spanish Ministry of Economy and Competitiveness and focused on interpreting in medical contexts, Interpreting and Translation in Prison Settings,[2] funded by the University of Alcalá, or SOS-VICS (Speaking Out for Support, co-financed by the EU's Directorate General Justice and nine Spanish universities[3]), whose main goal is to train interpreters to work in gender violence contexts, thus facilitating the assistance to gender violence victims of foreign origin who may not fluently speak the language of the interaction, and, at the same time to contribute to raising awareness of the need to hire professional interpreters during linguistic mediation in such cases. Despite increasing research on PSIT, research based on interpreters working with gender violence victims is still scarce. In this specific case, interpreters need to understand protocols involved in gender violence, as well as key applicable legal terminology and procedures, and the definition of concepts related to gender violence. Apart from that, the pragmatic meaning of the language of the victims is usually hard to render, as people from distant cultures may have different communication styles and may use a great variety of mechanisms to convey a particular meaning.

The Spanish central and regional administrations are putting a lot of effort into combatting gender violence and dealing with victims, who are local. However, foreign victims receive little assistance. For them to be able to access the services provided, an interpreter is essential. If the interpreters are not qualified and specialised enough, the interpretation may not reach the desired quality standards resulting in inaccurate and inefficient communication between victims and service providers. Without qualified and specialised interpreters, the rights of the victims may be violated and their risk of exclusion may thus increase.

Driven by this increasingly demanding necessity of training interpreters involved in gender violence cases, we highlight how Corpus-based Translation Studies (CTS) can be the groundwork for our study to build a monolingual corpus (Spanish) (original oral discourses delivered in similar settings, written texts, such as protocols, leaflets and guides and written spontaneous discourse) to obtain valuable information about specific features of gender violence spoken language and derive pedagogical applications. To this end, we will describe how we have built, compiled and analysed this corpus and discuss the main obstacles related to spoken speech (i.e how to tackle with pragmatics, paralinguistic dimensions of language, copyright issues and ethical implications) that we have overcome. As part of our main tasks, our research investigates the genre of gender violence speech. We will centre our attention on the numerous advantages in the area of interpreting training that our *ad hoc* corpus provides in terms of discourse patterns, pragmatic conventions, lexical clusters and semantic prosody, among others. We will conclude our paper with

[1] Ref.: FFI2011-25500.

[2] Ref.: CCG2013/HUM-010.

[3] These are the current research projects in which the Group FITISPos-UAH is currently working. SOS-VICS (Ref.: JUST/2011/JPEN/2912) co-ordinator is the University of Vigo.

some specific examples on the exploitation of our corpus and draw conclusions concerning the advantages of CTS in PSIT training.

This study started in 2013 and is still under development. The authors of this piece of research belong to the Research Group FITISPos-UAH, based in Madrid, and specialised in training and research in public service translation and interpreting. The methodology of this study, if proven useful, will constitute suggestions for trainers in the field, and the final outcome, both scientific and practical, will be particularly useful to train interpreters of the Master's Degree in Public Service Translation and Interpreting at the University of Alcalá, and others wishing to or in need of specialising in gender violence contexts.

2 Corpus-Based Interpreting Studies

The initial steps of corpus linguistics can be traced back to the pre-Chomskian period (McEnery et al. 2006: 3), where followers of the structuralist tradition used a corpus-based methodology to generate empirical results based on observed data. This area of research, defined as "the study of language based on examples of 'real life' language use" (McEnery and Wilson 2004: 1) has opened many possibilities for the study of language. Similarly, corpora have attracted increasing attention in translation studies over the last years. Depending on the nature of the work carried out, researchers have used corpora to investigate the features of translated texts (Baker 1995, 1996; Kenny 2001; Saldanha 2004), or the possibilities of using corpora as translation and terminology resources (Bowker 2003; Zanettin et al. 2003; Zhu and Wang 2011). There is no doubt that the study of real language in its context can provide valuable information. Calzada Pérez (2007: 216) highlights that:

> […] por su flexibilidad y capacidad de adaptación, los CTS aúnan metodologías descriptivas y lingüísticas; análisis del proceso y producto; exégesis de detalles o amplios patrones de comportamiento de interés tanto por cuestiones formales como por facetas culturales, ideológicas e incluso literarias (Calzada Pérez 2007: 216)
>
> […] due to their flexibility and adaptation possibilities, corpus-based translation studies merge descriptive and linguistic methodologies, process and product analysis, display of details or wide behaviour patterns, which are interesting both because of formal issues and cultural, ideological and even literary aspects. (Calzada Pérez 2007: 216, our translation)

Corpus-based translation studies (CTS) has proved to be a reliable way of collecting data to generalise about the so-called translated linguistic features or universals of translation (Baker 1993). The benefits and the pedagogical implications of using corpora within translation studies have been shown by various researchers. We can take Bowker and Pearson (2002), Corpas Pastor and Seghiri (2009), Lee and Swales (2006), or Sánchez Ramos and Vigier Moreno (forthcoming) as examples of authors that consider using corpus to research and teach specialised translation.

Nevertheless, CIS has not enjoyed the same popularity as CTS. Schlesinger's seminal paper (1998) set the groundwork for a CIS methodology. The use of corpora as a methodology within interpreting studies poses a number of challenges and

opportunities, and a number of difficulties. The main difficulty when dealing with a corpus-driven methodology in the study of interpreting is due to the obstacles involved in the analysis of oral discourse translation (i.e. transcribed speech): "The recording and transcription of unscripted speech events is highly labour intensive in comparison to the work involved in collecting quantities of written text for analysis (Thompson 2005: 254). Additional difficulty can be found in the compilation stage, as stated by Pöchhacker (2008), as it is difficult to obtain data and consent from speakers or service providers. Other types of difficulties to be taken into account have to do with the interpreter-mediated event, the different speakers and their roles, the interpreting mode, and the target audience (Shlesinger 1998).

Despite this challenging background, corpora may constitute the future of interpreting studies (Luzón et al. 2008). Although CIS are small in number if compared with CTS, there is a growing number of interpreting studies based on corpus data. We can take some works as examples. Ryu et al. (2003) focus on the use of corpora and simultaneous interpreting, as well as on the compilation of a bilingual corpus for linguistic and contrastive purposes. Other studies within the field of simultaneous interpreting explore the interpreter's speeches using an aligned simultaneous monologue interpreting corpus in order to research the interpreter's speaking speed and the difference between the beginning time of the speaker's utterance and that of the interpreter's utterance (Takagi et al. 2002). Taking into account different variables (the recording time, the number of utterance units, the speaking time, among others), these authors carry out an exhaustive statistical analysis of their corpus. Other authors like Van Beisen (1999) have used a corpus-driven methodology to study the different techniques involved in interpreting (i.e. anticipation). Lázaro Gutiérrez (2012) carries out a discourse analysis study from a corpus of 75 transcriptions of real doctor-patient conversations. This corpus was not tagged, but was manually processed to find out features of the asymmetry of the encounter.

Apart from studies based on manual corpora, there are few examples of projects based on machine-readable corpora. The University of Bolognia – European Parliamentary Interpreting Corpus (EPIC), 2005 – constitutes one of the first examples of interpreting corpus compilation using a machine-readable methodology. EPIC is the first large-scale interpreting corpus aimed at collecting a "large quantity of authentic simultaneous interpreting data to produce much-needed empirical research on the characteristics of interpreted speeches and to inform and improve training practices" (Russo et al. 2012: 53). EPIC is a trilingual open corpus (English, Italian and Spanish), including source speeches in those three languages and interpreted speeches in all possible combinations and directions (Russo et al. 2012: 53). It consists of nine sub-corpora (177,295 words in total). Recently, community interpreting has been looking into new ways of researching different relationships involved in social discourse. Thus, Angermeyer et al. (2012) offer to the academic community what is called *ComInDat Pilot Corpus*, a collection of two corpora of interpreted doctor-patient communication and a second corpus of interpreted court proceedings. Other authors stimulate the corpus-based interpreting methodology by providing different types of corpora – CoSi, a corpus of consecutive and simultaneous interpreting – in order to encourage the research community to use corpora in inter-

preting studies (House et al. 2012). This corpus was created by using the EXMARaLDA software (Schmidt and Wörner 2009, Schmidt and Kai 2012), which includes the EXMARaLDA Partitur-Editor, a tool for editing transcriptions in musical score notation. These are just some examples of the growing interest of corpora and interpreting and their valuable source for research.

Following in the footsteps of all these studies, our research is meant to contribute to the area of CIS by compiling a monolingual multimodal corpus on gender violence, whose ultimate goal is to train interpreters in this area. Interpreters dealing with gender violence cases have to perform their work in many different public service interpreting settings, such as police offices, medical practices, courts, social work and psychology practices. In what follows, a brief description of public service interpreting will be provided.

3 Public Service Interpreting

Public service interpreting, also called community interpreting, is performed at institutions which offer public services for the general population, as is the case in courts, hospitals, police offices, healthcare centres, schools, public administration offices, and the like. Public service interpreters bridge communication gaps between service providers (lawyers, doctors, teachers, police officers, social workers…) and users. One of the first definitions of public service interpreting is given by Wadensjö (1998: 49):

> Interpreting carried out in face-to-face encounters between officials and laypeople, meeting for a particular purpose at a public institution is (in English-speaking countries) often termed community interpreting.

The areas where public service interpreting is performed are multiple and include a great variety of settings, such as legal (considered by some authors as a distinct variety, apart from public service interpreting (Phelan 2001)), healthcare, educational, administrative, social, police setting, amongst others. Public service interpreters use the bi-directional modality both onsite, over the phone, or through videoconferencing technologies. This area of interpreting has specific characteristics that differentiate it from others. Here we include a compilation of features taken from different sources:

– Interpreters must have a deep knowledge both of the languages they interpret into and from and of the cultures their clients belong to (Valero Garcés 2006). The understanding and expression of concepts related to gender-based violence may vary amongst cultures. Thus, foreign victims may have ways of expressing their problems which may result exotic or even incomprehensible to members of the host culture. Interpreters must not only be familiar with these culturally marked pragmatics patterns of victims, but also with cultural and institutional constructs belonging to the host culture, which will condition the development of interactions between service providers and users.

- The asymmetry between the participants in the conversations (Lázaro Gutiérrez 2012). The characteristic asymmetry of institutional encounters increases when the user of public services does not speak the institutional language (the language used by the service provider, not only at a semantic level, but also at a pragmatic one).
- The tense situations in which these interpreted conversations sometimes take place (Phelan 2001; Valero Garcés 2006). Victims of gender-based violence find themselves in a complicated personal situation. Interpreters usually suffer from the stress generated by having to re-verbalise traumatic events.
- The scarce (though growing) acknowledgement of the profession, which usually leads to the fact that non-professional interpreters undertake this task, or results in poor working conditions for professional interpreters working in this field, who receive low salaries, are assigned tasks other than interpreting, have little support and resources (not receiving previous information about the topic of the conversations to be interpreted or the peculiarities of the interactants, or being called very shortly before the assignment starts (Lázaro Gutiérrez 2014).
- The performance by the interpreter of a much broader task than that of simply interpreting, which includes, among other issues, the weight and responsibility of co-ordinating the turns in the conversation (Wadensjö 1998).

Taking into account the specific characteristics of public service interpreting, according to Inglis (quoted in Iliescu 2001) a public service interpreter should:

- Master a sufficient number of general and specialised terms.
- Be able to remember communicative pragmatic patterns such as greetings, farewells, questions and other ways to obtain information; know how to ask for explanations and repetitions, spell, make remarks about a particular aspect of the conversation, control the sequence of the interaction, express agreement and disagreement, self-repair, apologise; and have the ability to repair a negative impression on the listener or perform other phatic patterns such as compliments and good wishes.
- Be aware of the nature and characteristics of discourse.
- Recognise and transfer register and tenor.
- Have a good command of syntax.
- Have a good command of discursive strategies.
- Be able to recognise and transfer the illocutionary force of the original message.
- Be able to perceive interactants' opinions and their degree of knowledge about the topic of conversation.
- Be able to grasp and transfer the interactants' points of view.
- Be able to improve the structure of the discourse.
- Notice the interactants' cultural differences and have expert knowledge about them.

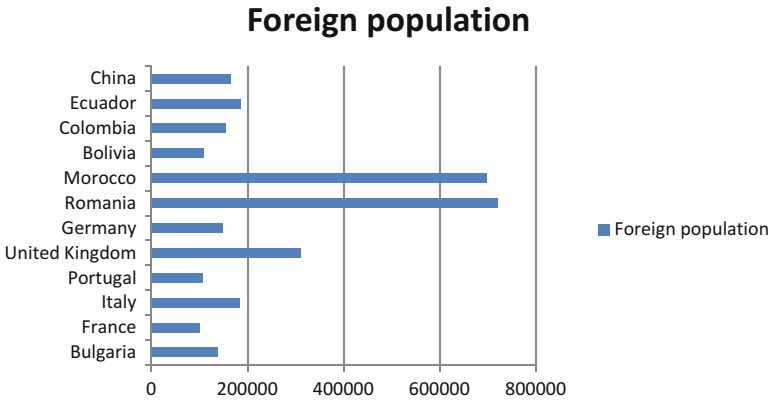
Thus, interpreters willing to specialise in gender violence contexts should acquire specific vocabulary about this topic, knowledge about the most frequent structures

of this kind of interactions (questions, narratives, explanations, and the like), the nature of the encounters, the pragmatic peculiarities of the discourse (metaphors, use of empathy), amongst many other abilities mentioned by Inglis (Iliescu 2001), such as a good use of syntax, grammar and pragmatics and a deep cultural knowledge. We consider that a multimodal corpus of real conversations, natural oral or written discourse, and written documents will be useful in order to spot and systematise these features. From the findings obtained after the analysis of such a multimodal corpus, it will be possible to elaborate useful training materials for interpreters willing to specialise in gender violence contexts, that will contribute to the acquisition of what has been called "pragmatic competence", which consists of the "the ability to use language effectively in order to achieve a specific purpose and to understand language in context (Thomas 1983: 92).

4 Interpreting in Gender Violence Contexts

The research program which is presented has been developed in Spain, although we believe that its methodology could be extrapolated to many different contexts. According to the most recent data of the European Commission (Eurostat), Spain received the 5th largest number of immigrants in the European Union in 2012 (304,100). Germany reported the largest number of immigrants (592,200), followed by the United Kingdom (498,000), Italy (350,800) and France (327,400).

According to the Spanish Institute of Statistics (*Instituto Nacional de Estadística*), in 2014, the highest number of migrants came from Romania, Morocco and the United Kingdom. These data give us a clue about how necessary interpreters are for foreign languages such as Romanian and Arabic in Spain so that the population from these countries can successfully access public services.



Immigration to Spain according to nationality in 2014. Spanish Institute of Statistics

On the other hand, focusing on gender violence issues, it is worth mentioning that in 2012, the European Institute for Gender Equality (2012) reviewed the implementation by the member states of the measures agreed to at the Beijing Declaration and Platform for Action against violence against women and support to victims and alerted about the fact that no less than 33 % of European women had ever suffered gender violence.

The *Directive 2012/29/EU of the European Parliament and of the Council of 25 October, 2012, establishing minimum standards on the rights, support and protection of victims of crime, and replacing Council Framework Decision 2001/220/JHA* is worth highlighting as it makes it clear that victims of crime (such as victims of gender-based violence), must be provided with the necessary tools to grant them access to legal services, such as the assistance of an interpreter. This is a legal instrument that reinforces national legislation of the member states and is based on other similar proposals, especially the *Directive 2011/36/EU of the European Parliament and of the Council of 5 April 2011 on preventing and combating trafficking in human beings and protecting its victims, and replacing Council Framework Decision 2002/629/JHA,* the *Directive 2011/92/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision 2004/68/JHA,* or the *Council Framework Decision 2001/220/JHA of 15 March 2001 on the standing of victims in criminal proceedings.* Apart from these, Art. 6 from the *Directive 2012/29/EU of the European Parliament and of the Council of 25 October, 2012, establishing minimum standards on the rights, support and protection of victims of crime, and replacing Council Framework Decision 2001/220/JHA* displays along 6 sub-articles the right to translation and interpreting, which is also recognised in the *Directive 2010/64/EU of the European Parliament and of the Council of 20 October 2010 on the right to interpretation and translation in criminal proceedings.*

In Spain, national authorities have put into practice a set of measures aimed at softening this gender-based violence phenomenon and at contributing to establish a national strategy for the elimination of this social cancer through personalised attention to the victims. The main aim of this strategy is to unite all the institutions and organisations of the country into a national network, and even to duplicate staff and resources in order to fight against gender violence. Amongst these measures is the *Basic Law 1/2004 of 28 December on Integrated Protection Measures Against Gender Violence* (*Ley Orgánica 1/2004 de Medidas de Protección Integral contra la Violencia de Género)* and the *Royal Decree 233/2005 on the Creation and Constitution of Courts on Violence against Women (Real Decreto 233/2005 a través del cual se establece la creación y la constitución de juzgados de violencia sobre la mujer).*

Other measures were the launching of the Observatory against Domestic and Gender Violence (*Observatorio contra la Violencia Doméstica y de Género*) of the General Council of the Judiciary as an instrument of analysis and action within the Spanish Justice Administration that promotes initiatives and measures oriented

towards the elimination of the social problem posed by domestic and gender violence. The National Observatory on Violence against Women (*Observatorio Estatal de Violencia sobre la Mujer*) of the Ministry of Healthcare, Social Affairs and Equality was launched in 2006 and is the institution responsible for the elaboration of reports and research about gender violence, as well as for the evaluation of the impact of the adopted policies and measures through the compilation, analysis and dissemination of materials about gender violence. In the last decade, the Network of Local Points of the Regional Observatory on Gender Violence and the hotline 016, for the assistance of women suffering from gender violence, were launched amongst many other mechanisms and resources.

However, the Macrosurvey on Gender Violence carried out in 2011 by the Centre for Sociological Research (*Centro de Investigaciones Sociológicas*) found out that the prevalence of gender violence against foreign women doubles that against Spanish women. According to this same source, 469,317 foreign women had suffered gender violence at some time in their lives, and 130,241 had suffered it in 2010. On the other hand, data published by the Spanish General Council of the Judiciary (*Consejo General del Poder Judicial*) indicates that in 2010 12 % of the gender violence victims who attended Spanish courts were of foreign origin. Only 3 years later, in 2013, this figure mounted to 35 %. If linguistic assistance is not provided, all these measures and resources are out of reach for these foreign gender-based violence victims who do not speak Spanish (fluently).

These new European and national regulations are reflected on the development of research projects such as "Speak Out for Support – SOS-VICS" (JUST/2011/JPEN/2912), co-ordinated by the University of Vigo and with the participation of nine Spanish universities. This is a pilot project whose main objective is to train interpreters who want to specialise in gender violence contexts, interpreting for foreign victims. These interpreters, apart from a few lessons they may receive if they follow some Spanish postgraduate studies in public service interpreting and translation, do not usually get any specific training about interpreting in gender-based violence contexts, contrary to what happens with other agents (doctors, police officers, social workers, and so on) who assist victims of gender violence.

We think that, in order to perform accurately, interpreters should acquire thematic knowledge about gender violence, knowledge about institutional procedures and usual communicative events, as well as linguistic, pragmatic, cultural and, particularly, terminological command. Besides, they should be able to manage the stress experimented both by themselves (interpreters) and the other participants in the interaction, following Inglis (1984, quoted in Iliescu 2001). Interpreters working in these contexts need specialisation in linguistic mediation with foreign victims, apart from the command of the languages involved and knowledge about institutional protocols on the one hand, and emotions management on the other. With our proposal we intend to contribute to their training through a corpus-based methodology and, in what follows, we will describe the methodological phases of our project.

5 A Corpus for Interpreters Working in Gender Violence Contexts: Design, Compilation and Use

The main purpose of our project is to compile a monolingual electronic archive (Bowker 2003) on gender violence for pedagogical purposes. Although the potential benefits of applying a systematic corpus-based methodology for research on interpreting have been sufficiently acknowledged in the last few years (Straniero Sergio and Falbo 2012), they are mainly focused on studying interpreting through corpora. Hence our archive has been designed with a pedagogical purpose in mind. This archive is divided into three different, but related, corpora. As stated in Sect. 4, it is of paramount importance to offer high quality training to our students so that they can provide a successful interpreting service in gender violence contexts. We believe that this training gap can be filled with the design and compilation of an archive focused on gender violence to analyse and research the gender violence genre (terminology, spoken and written language, register, pragmatic patterns, etc).

Our final corpus will provide:

1. a comprehensive knowledge of the gender violence genre
2. real language and authentic material to design our interpreting training sessions
3. fairly accurate statistics of word occurrences (essential to design vocabulary acquisition activities)
4. examples of discursive patterns corresponding to different cultural pragmatics
5. quick access to large texts.

From a pedagogical perspective, a corpus-based methodology has proved to be the most adequate. Based on a deductive approach, this methodology will enable us to analyse patterns of use for pre-defined linguistic features (i.e. word frequency, linguistic and register variations of a given category, frequency of pragmatic patterns, and so on).

The different advantages of working with a corpus-based methodology for pedagogical purposes were highlighted by authors such as Flowerdew (1993: 91). According to this author, working with corpus and concordance programs has three main applications for trainers: (i) as a linguistic informant, (ii) as a source of input for training, and, finally, (iii) as input for materials developments. Firstly, as a linguistic informant, the trainer has the possibility to access the corpus in order to corroborate both grammatical and lexical choices, as well as expressions and other pragmatic patterns. Secondly, as a source of input for training, the trainer can use the corpus to generate authentic examples of usage, which would reflect all the levels of language (including the pragmatic level) and the communicative situation. Finally, applying corpus as input for material development can be successful if the following conditions are fulfilled:

1. the trainer is aware of the students' strengths and weaknesses and knows which linguistic points (lexis, grammar, pragmatics…) need to be improved
2. the trainer wishes to design his or her own material

3. the trainer is computer-literate and has the proper software and concordance tools.

Our corpus is made up of three main sources:

1. Texts: manuals and protocols, scientific documents
2. Videos: simulated videos and real conversations
3. Texts: written spontaneous discourse.

It comes as no surprise that our corpus could cover many areas related to gender violence due to the fact that interpreters can work in different settings. As a starting point for our research, we have decided to compile a monolingual corpus on gender violence context in medical and social settings.

5.1 Corpus Design

We were aware that just a compilation of texts (both scientific documents and practical documents such as manuals and guides of practice) was not enough to cater to the training needs of the interpreters. Public service interpreters have to deal with spontaneous oral conversations and interpret both the service provider's and the victim's discourse, each of them with different characteristics, most of them have to do with the pragmatic level of language. Bearing this in mind, our corpus includes spoken discourse (both real and simulated conversations), and written discourse spontaneously produced by victims.

Our corpus contains data from different sources compiled between 1998 and 2014. As stated before, it consists of three sub-corpora. We describe our corpus and the relevant background information contained in our corpus data in the next paragraphs.

(a) Texts: manuals and protocols, scientific documents

Corpus data was collected mainly from the Internet. Official websites of public administrations and NGO associations were especially useful. The usefulness of the project SOS-VICS, which offers a great repository of documents on this field in its public webpage, has to be highlighted. In order to follow a well-designed compilation protocol, we used specific software to automate the downloading process (i.e. HTTrack, GNU Wget or Jdownloader). These tools allow the downloading of websites at one go. Texts can be downloaded automatically. Once the documents had been found and downloaded, the texts had to be converted to .txt files in order to be processed by corpus analysis software. This task is especially necessary in the case of texts retrieved in .pdf format. Finally, all documents were stored in different files.

(b) Videos: simulated videos and real conversations

Another important part of our corpus was a set of videos of medical consultations with victims of gender violence. Here we can distinguish two different

resources: simulated videos, which are accessible via internet and have been published by universities or healthcare organisms; and real conversations, which were recorded in Spanish general practitioner (GP) consultations and hospitals and belong to the FITISPos-UAH Group (1998–2004). Another source of material comes from the InterMed project, which records and analyses real conversations with foreign patients in GP consultations.

Visits to several organisations and associations, such as local points of the Regional Observatory on Gender Violence and several women's associations, were carried out.

Copyright issue was the main problem at this stage. Many associations have shown interest in the project and have contributed to the corpus compilation.

As a future task, and in order to get the most out of this multimodal material, we believe it is important to tag our corpus so that we can analyse it. Tagging will allow us to clearly differentiate the different actors taking part in the discourse (i.e. healthcare providers and patients or victims). Based upon the work done in other projects – European Parliamentary Comparable and Paralles Corpora, ECPC – (Calzada Pérez et al. 2007), our corpus will be annotated in XML (*Extensible Mark-up Language*) with different types of data: linguistic, contextual, and metalinguistic information. XML-tagging will be carried out semi-automatically with the use of regular expressions.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE maltrato SYSTEM "maltrato.dtd">
<maltrato>
<doctor>D<text num="1"> a ver Fátima ¿eres alérgica a algo? ¿alergia?</text></doctor>
<paciente>P<text num="2"> no</text></paciente>
<doctor>D<text num="3"> ¿alguna enfermedad importante a parte de lo de la piel?</text></doctor>
<paciente>P<text num="4"> sí pues como... no sé cómo se llama [eh]</text></paciente>
<doctor>D<text num="5"> [esa] enfer[medad] </text></doctor>
<paciente>P<text num="6"> [sí]</text></paciente>
<doctor>D ¿alguna otra?</doctor>
<paciente>P<text num="8"> pues no</text></paciente>
<doctor>D<text num="9"> de corazón [pulmón]</text></doctor>
<paciente>P<text num="10"> [no]     </text></paciente>
<doctor>D<text num="11"> estómago                    </text></doctor>
<paciente>P<text num="12"> no</text></paciente>
<doctor>D</doctor>
<paciente>P<text num="14"> no</text></paciente>
<doctor>D<text num="15"> ¿depresión o algo?</text></paciente>
<paciente>P<text num="16"> no [creo]</text></paciente>
<doctor>D<text num="17"> [de] tristeza o</text></doctor>
<paciente>P<text num="18"> no</text></paciente>
<doctor>D<text num="19"> estás bien ¿verdad? ¿tomas algún tratamiento? Ahora lo que te echas para la
<paciente>P<text num="20"> sí</text></paciente>
<doctor>D<text num="21"> ¿pero aparte tomas algo por la boca todos los [días]?</text></doctor>
<paciente>P<text num="20"> sí</text></paciente>
<doctor>D<text num="23"> ¿qué tomas?</text></doctor>
<paciente>P<text num="24"> pues la primera vez es de hoy no sé cómo se llama</text></paciente>
<doctor>D<text num="25"> ah la antibiótico</text></doctor>
<paciente>P<text num="26"> sí</text></paciente>
<doctor>D<text num="27"> vale por la mañana y por la noche ¿no?</text></doctor>
<paciente>P<text num="28"> sí</text></paciente>
<doctor>D<text num="29"> vale entonces esto comenzó ayer ¿no?</text></doctor>
<paciente>P<text num="30"> sí</text></paciente>
<doctor>D<text num="31"> ¿qué pasó ayer?</text></doctor>
<paciente>P<text num="32"> pues ayer están discutiendo mi (xxx) mi hermanos los dos eso sí</text></p
<doctor>D<text num="33"> sus hermanos empezaron a discutir</text></doctor>
<paciente>P<text num="34"> sí</text></paciente>
<doctor>D<text num="35"> ¿y qué pasó?</text></doctor>
<paciente>P<text num="36"> pues me empezó mi mano así me</text></paciente>
<doctor>D<text num="37"> la mano</text></doctor>
<paciente>P<text num="38"> sí</text></paciente>
<doctor>D<text num="39"> ¿pasó algo entre sus hermanos a parte de la de las voces o</text></doctor>
```

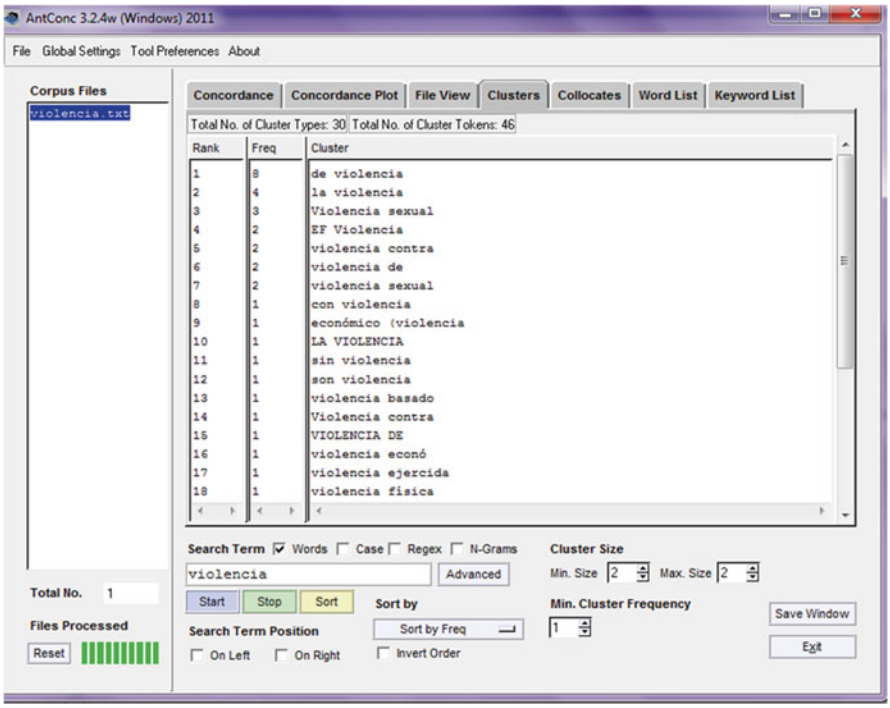
XML-tagging

(c) Texts: written spontaneous discourse

In spite of every effort to obtain recordings from which samples of discourse uttered by victims could be extracted, the number of real conversations we were able to gather was scarce. Aware of the great importance of the compilation of spontaneous discourse, it was decided to include in our corpus the contributions that women themselves made to specialised consultation fori about their situations as victims. These contributions are either accessible online, such as the ones sent to the forum of the Region of Castilla-La Mancha, or have been provided by women's associations.

In terms of corpus analysis, there is no doubt that working with corpus requires an efficient use of software. Although we are in a very initial stage of our research, we believe it is relevant to know the tools to be used in order to accomplish our main goal: training interpreters working in gender violence contexts. For instance, concordance software programs will provide the most frequent words service providers and victims use in their communicative interaction. It will enable us to identify appropriate (and inappropriate) terminology, collocations, phraseology, pragmatic patterns, style and register of the gender violence discourse. There are different programs available. One of the most popular is Wordsmith, designed by Mike Scott. It offers a wide range of possibilities for analysing corpora, such as XML reading facilities, wordlists, keywords and concordances. It incorporates follow-up concordance searches, file viewer utility, a corpus corruption detector or a concgram facility. XML reading facilities can be very useful if we want to analyse our second sub-corpus (videos) separately so that we can compare the different speakers' speeches in terms of terminology, for example. This analysis will retrieve the most frequent words to elaborate a wordlist in order to design activities to train our interpreting students. Other functions are particularly useful to grasp the pragmatics of discourse, as they can be used to retrieve utterances in context or collocational and contextual information (concord).

Wordsmith is an indispensable tool but there are other programs that provide reliable information. We can take AntConc as an example, a freeware multiplatform tool created by Laurence Anthony in 2004. Functions are very similar. Wordlist and word frequency proved useful for the three sub-corpora as well as collocates and clusters function.



AntConc collocates function

5.2 *Some Examples of Practical Application*

As previously stated, one of the main challenges for interpreters working in fields related to gender-based violence consists on the rendition of the pragmatic content of messages. They struggle every day to be able to render pragmatically appropriate utterances, matching both the situational context and the intents of the people whose discourse they are interpreting. Interpreter-mediated natural conversations are an example of communication across languages and cultures, which, as Kecskes and Romero-Trillo (2013: 1) point out, has become the new challenge for pragmatics research in the twenty-first century.

Interpreters usually receive training about interpreting techniques (including memorising, diction, making notes, and the like), terminology, and advice on how to prepare for an assignment. One of the most difficult aspects is to obtain contextual information, including details about the communicative event, which is usually structured under institutional constraints, or the culture of the participants, which may determine their communicative styles and their use of language and pragmatic patterns. Training and information gathering about these specific issues must not be taken for granted because, as Kecskes and Romero-Trillo (2013: 1) state, "our individual comprehension of language is dependent upon our biographical socio-cultural experience". Vital experiences might not be enough for interpreters to grasp the

meaning of the utterances of the participants in interpreter-mediated interactions. This is so because of interpreters might belong to the cultural group of one of the interactants or to none of them, making it obvious that they will lack pragmatic information from at least one of the parties.

Once our corpus is fully compiled, we will be able to analyse it taking into account a discourse level, that is, our corpus will allow us to analyse it according to the particular setting of conversations (social, medical, legal, etc. and the different cultures of the victims. Our point of departure is that, as Baider (2013: 8) suggests "words are not culturally neutral and bring with them certain culture-specific ways of thinking". We agree with Wierzbicka (2006) in that we cannot take for granted equivalence between two languages, and go a step further and add that we cannot take for granted equivalence between two cultures, even if people belonging to both of them speak the same language.

Particularly in the field of gender-based violence and precisely when it comes to interpret the discourse of the victims, it is especially challenging for the interpreter to grasp and transmit the meaning of taboo concepts, as they are usually expressed in a very creative way, so as to conceal the taboo part of them. Although it is still soon to advance results of the analysis of our corpus, which is still in a compilation phase, a good example of this can be obtained from Torruella Valverde (2013), when she reports about the experiences of interpreters and tells and gives the following example. One interpreter of Arabic had to render once the meaning of a gesture performed by a gender-based violence victim who was declaring in court. She brought her hand close to her ear, as if she was speaking on the phone, but without touching her head. The interpreter, who shared the culture of the victim, could understand that she meant that her husband had threatened her for death. Imagine how difficult it is for an interpreter who does not know the culture of their interlocutors to grasp the meaning of certain metaphors and uses of language that are the result of the speakers' creativity to avoid taboos.

6 Conclusions

The emergence of new projects focused on CIS has made it necessary to reflect on the need for more research in corpus linguistics and interpreting. Public Service interpreters in particular often lack both general and specialised training. Furthermore, many scholars support the idea that interpreters require a precise understanding of specific discourses, such as the one produced in gender violence contexts, as there are many specific features (particularly pragmatic ones) that are characteristic of specific contexts.

However, training interpreters in specific contexts is challenging because training materials must be consistent with real data. The analysis through CIS of the gender violence discourse produced both by victims and by service providers provides the researcher with valuable linguistic material (i.e. terminology, phraseology, metaphors, pragmatic patterns) that can be transformed into useful training resources.

Although we are aware that CIS has challenges and obstacles that need to be sorted out, we believe that this area of research is the key to accessing real data. We are aware that many methodological questions still remain open and are only likely to be answered once our research has been completed. A closer collaboration between academics, service providers and interpreters could be one of the potential answers, as it may allow us to enlarge our corpus with pieces of real oral discourse (recordings). Although gathering collaboration for this purpose is usually a hard task, it is hoped that, in the future, we will be able to reach our objectives and enlarge our corpus to gradually become a useful resource for the academic community, thus representing what gender violence discourse is and how it should be taught to interpreters in the classroom.

References

Angermeyer, P. S., Meyer, B., & Schmidt, T. (2012). Sharing community interpreting corpora: A pilot study. In T. Schmidt & K. Wörner (Eds.), *Multilingual corpora and multilingual corpus analysis* (pp. 275–294). Amsterdam/Philadelphia: John Benjamins Publishing.

AntConc website http://www.laurenceanthony.net/software.html. Accessed 20, 22 Sept 2014.

Baider, F. (2013). Hate: Saliency features in cross-cultural semantics. In K. Istvan & J. Romero-Trillo (Eds.), *Research trends in intercultural pragmatics* (pp. 7–28). Boston/Berlin: De Gruyter.

Baker, M. (1993). Corpus linguistics and translation studies – Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–253). Amsterdam/Philadelphia: John Benjamins.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target, 7*(2), 223–243.

Baker, M. (1996). Corpus-based translation-studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP, and translation: Studies in language engineering in honor of Juan C. Sager* (pp. 175–186). Amsterdam/Philadelphia: John Benjamins.

Bowker, L. (2003). Corpus-based applications for translator training: Exploring possibilities. In S. Granger, J. Lerot, & S. Petch-Tyson (Eds.), *Corpus-based approaches to contrastive linguistics and translation studies* (pp. 169–183). Amsterdam/Philadelphia: Rodopi.

Bowker, L., & Pearson, J. (2002). *Working with specialized language. A practical guide to using corpora*. London/New York: Routledge.

Calzada Pérez, M. (2007). *El espejo traductológico. Teorías y didácticas para la formación del traductor*. Madrid: Editorial Octaedro.

Calzada Pérez, M., Cucala, N. M., & Martínez, J. M. (2007). ECPC: European Parliamentary Comparable and Parallel Corpora. *Procesamiento del Lenguaje Natural, 37*, 349–351.

Centro de Investigaciones Sociológicas. (2011). *Macroencuesta sobre Violencia de Género.* Madrid: Ministerio de Sanidad, Ciencias Sociales e Igualdad.

Corpas Pastor, G., & Seghiri, M. (2009). Virtual corpora as documentation resources: Translating travel insurance documents (English–Spanish). In A. Beeby, P. Rodríguez Inés, & P. Sánchez-Gijón (Eds.), *Corpus use and translating. Corpus use for learning to translate and learning corpus use to translate* (pp. 75–107). Amsterdam/Philadelphia: John Benjamins.

EPIC website http://sslmitdev-online.sslmit.unibo.it/corpora/corpora.php. Accessed 20, 22 Sept 2014.

Flowerdew, J. (1993). Concordancing as a tool in course design. *System, 21*(2), 231–244.

House, J., Meyer, B., & Schmidt, T. (2012). CoSi – A corpus of consecutive and simultaneous interpreting. In T. Schmidt & K. Wörner (Eds.), *Multilingual corpora and multilingual corpus analysis* (pp. 295–304). Amsterdam/Philadelphia: John Benjamins Publishing.

Iliescu Gheorghiu, C. (2001). *Introducción a la Interpretación: la modalidad consecutiva*. Alicante: Universidad de Alicante.

Inglis, M. (1984). *A preparatory course in bilateral interpreting*. Edinburgh: University of Edinburgh, Department of Linguistics.

Instituto Asturiano de la Mujer. Disponible en. http://institutoasturianodelamujer.com/iam/category/biblioteca-digital/. Consultado el 15 May 2014.

Instituto Europeo de Igualdad de Género (European Institute for Gender Equality) (2012). *Examen de la aplicación de la Plataforma de Acción de Beijing en los Estados miembros de la Unión Europea sobre violencia contra la mujer y apoyo a las víctimas*. Brussels: EIGE.

Kecskes, I., & Romero-Trillo, J. (2013). Introduction. In I. Kecskes & J. Romero-Trillo (Eds.), *Research trends in intercultural pragmatics* (pp. 1–3). Boston/Berlin: De Gruyter.

Kenny, D. (2001). *Lexis and creativity in translation. A corpus-based study*. Manchester: St. Jerome.

Lázaro Gutiérrez, R. (2012). *La interpretación en el ámbito sanitario. Estudio de la asimetría en consultas médicas*. Saarbrücken: Editorial Académica Española.

Lázaro Gutiérrez, R. (2014). Use and abuse of an interpreter. In C. Valero-Garcés (Ed.), *(RE) Considerando ética e ideología en situaciones de conflicto*. Alcalá de Henares: Servicio de Publicaciones de la Universidad de Alcalá.

Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes, 25*, 56–75.

Linkterpreting, Universidad de Vigo. http://linkterpreting.uvigo.es/interpretacion-social/material-didactico/. Accessed 15 May 2014.

Luzón, M. J., Campoy, M. C., del Mar Sánchez Ramos, M., & Salazar, P. (2008). Spoken corpora: New perspectives in oral language use and teaching. In M. C. Campoy & M. J. Luzón (Eds.), *Spoken corpora in applied linguistics* (pp. 3–30). Bern: Peter Lang.

McEnery, T., & Wilson, A. (2004). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies. An advanced resource book*. London/New York: Routledge.

Phelan, M. (2001). *The interpreter's resource*. Manchester: Multilingual Matters.

Pöchhacker, F. (2008). Inside the 'black box'. Can interpreting studies help the profession if access to real life settings is denied? *The Linguist, 48*(2), 22–23.

Russo, M., Bendazzoli, C., Sandrelli, A., & Spinolo, N. (2012). The European parliament interpreting corpus (EPIC): Implementation and developments. In F. S. Straniero & C. Falbo (Eds.), *Breaking the ground in corpus-based interpreting studies* (pp. 53–90). Bern: Peter Lang.

Ryu, K., Matsubara, S., Kawaguchi, N., & Yasuyoshi, I. (2003). Bilingual speech dialogue corpus for simultaneous machine interpretation research. In *Proceedings of the 6th Oriental COCOSDA workshop in Singapore* on Oct 1–3 2003 (pp. 164–168). http://ir2.nul.nagoya-u.ac.jp/jspui/handle/2237/15076. Accessed 15 May 2014.

Saldanha, G. (2004). Accounting for the exception to the norm: A study of split infinitives in translated English. *Language Matters, Studies in the Languages of Africa, 35*(1), 39–53.

Sánchez Ramos, María del Mar., & Javier Vigier Moreno, F. (in press). Using corpus management tools in public service translator training: An example of their application in the translation of judgments. In C. C. Antonio Pareja-Lora & P. Rodríguez Arancón (Eds.), *Technological innovation for specialized linguistics domains*. Oxford: Oxford University Press.

Schmidt, T., & Wörner, K. (2009). EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics, 19*, 565–582.

Schmidt, T., & Kai, W. (2012). *Multilingual corpora and multilingual corpus analysis*. Amsterdam/Philadelphia: John Benjamins Publishing.

Straniero Sergio, F., & Falbo, C. (Eds.). (2012). *Breaking ground in corpus-based interpreting studies*. Bern: Peter Lang.

Shlesinger, M. (1998). Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta, 43*(4), 486–493.

Takagi, A., Matsubara, S., & Inagaki, Y. (2002). A corpus-based analysis of simultaneous interpretation. In *Proceedings of international joint conference of the 5th symposium on natural language processing* (SNLP-2002) (pp. 167–174). http://slp.itc.nagoya-u.ac.jp/web/papers/2002/snlp2002_takagi.pdf. Accessed 15 May 2014.

Thomas, J. (1983). Cross-cultural pragmatic failure. *Applied Linguistics, 4*(2), 91–112.

Thompson, P. (2005). Spoken language corpora. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 59–70). Oxford: Oxbow Books.

Torruella Valverde, J. (2013). *La adecuación en la interpretación en los SSPP. Comunicación no verbal e implicaciones culturales (árabe-español Alcalá de Henares).* Unpublished master's dissertation.

Valero Garcés, C. (2006/2008). *Formas de Mediación Intercultural e Interpretación en los Servicios Públicos. Conceptos, Datos, Situaciones y Práctica*. Granada: Comares.

Van Besien, F. (1999). Anticipation in simultaneous interpretation. *Meta, 44*(2), 250–259.

Wadensjö, C. (1998). *Interpreting as interaction*. Londres y Nueva York: Longman.

Wierzbicka, A. (2006). *English: Meaning and culture*. Oxford: Oxford University Press.

Wordsmith website http://www.laurenceanthony.net/software.html Accessed 20 Sept 2014.

Zanettin, F., Bernardini, S., & Stewart, D. (Eds.). (2003). *Corpora in translator education*. Manchester: St. Jerome.

Zhu, C., & Hui, W. (2011). A corpus-based, machine-aided mode of translator training. *The Interpreter and the Translator Trainer, 5*(2), 269–291.

Part III
Book Reviews

Zappavigna, M. (2012). *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. London: Bloomsbury

Rachelle Vessey

Abstract The book *Discourse of Twitter and Social Media* provides readers with an accessible and engaging introduction to pioneering research in the field of linguistics and discourse studies. The research combines systemic functional grammar (more specifically, appraisal theory) and corpus linguistics in the study of Twitter data and it draws on a wealth of literature from the field of media and communication studies. The book usefully builds on the innovative theorisation of "ambient affiliation" in Twitter – a concept introduced in the author's (2011) article. Although the book unfortunately does not provide detail on the methods used for analysis and the analysis excludes multilingual data, ultimately it presents new and innovative ways of approaching the discourse of Twitter, a type of data that had yet to be examined from a linguistic perspective.

Keywords Corpus linguistics • Discourse analysis • Social media • Twitter

This book is the first of its kind and serves as an excellent introduction to the discourse of Twitter and social media more generally. While there has been considerable research on Twitter in other fields, there has been little within the field of linguistics or within discourse studies more specifically. The book is, therefore, an important addition to these fields because, as argued by the author, the advent of social media has placed "new and interesting semiotic pressure" (p. 2) on language. Throughout the book, Zappavigna provides numerous examples and case studies to highlight the kinds of linguistic innovations taking place within social media and her research paves the way for further work in this area.

At the heart of the book is the study of the interpersonal nature of communication in social media. The author primarily uses appraisal theory to illustrate the evaluative and interpersonal nature of social media discourse. Indeed, one of the most

R. Vessey (✉)
School of Education, Communication and Language Sciences,
Newcastle University, Newcastle upon Tyne, UK
e-mail: Rachelle.Vessey@ncl.ac.uk

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_13

notable features of the research undertaken within this book is the combination of approaches from systemic functional grammar (more specifically, appraisal theory) with corpus linguistics, while at the same time drawing on a wealth of literature from the field of media and communication studies. While innovative and ambitious, this combination poses some issues, which are discussed below.

The book begins with an introduction to the topics covered in the book and an outline of the history and features of Twitter. Most notable in this chapter are the introductions to hashtags and "ambient affiliation", which are expanded upon in later chapters. The author also provides an overview of the methodological approaches used in the analysis, namely, a qualitative and quantitative approach to interpersonal and ideational meaning drawing on systemic functional linguistics and corpus linguistics (p. 12). However, this brief overview is unfortunately not expanded upon in later chapters.

Chapter Two outlines the general features of social media language and the challenges it poses when used as data. The author highlights issues such as representativeness, noise, time sensitivity, and size of data as potential roadblocks in analysis. The author also introduces readers to the primary dataset under examination throughout the book – a 100,000-word corpus of tweets called the HERMES corpus. Chapter Three explores this corpus, showing some of the unique linguistic features of Twitter such as hyperlinks, forms of address, and retweets. The author lists some of the most frequent linguistic patterns in the HERMES corpus, such as the 3-gram "Thanks for the", and suggests the social functions of these patterns.

Chapter Four presents findings from a study of evaluation within a subset of 100 tweets from the HERMES corpus. After briefly overviewing appraisal theory, the author explains how different appraisal systems can be applied to the HERMES corpus in order to ascertain the kinds of interpersonal relationships that are being fostered in new and innovative ways in social media language. The author also addresses the evaluative role of emoticons, albeit to a lesser extent, and provides a table containing articulation systems for emoticons.

In Chapter Five, Zappavigna elaborates on a theorisation of "ambient affiliation" in Twitter – a concept introduced in the author's (2011) article. Arguably, the theorisation about the role of hashtags is one of the most valuable parts of this book. The author argues that hashtags mark meanings and hypercharge them with an additional semiotic pull akin to a "gravitational field" (p. 95). It is this "field" that creates the opportunity for ambient affiliation – that is, a form of virtual bonding around a topic of shared interest (p. 96). In this way, affiliation is continually evolving because groups shift as the hashtags change depending on what people are talking about at a given time (p. 98). Importantly, the notion of "community" evolves in this context, because affiliation is not fixed and the author draws on a "semiotic rather than purely interactional definition of community" (p. 99). Zappavigna lists some of the general affordances (p. 86) and functions (p. 87) of hashtags before exploring specific frequent hashtags in the HERMES corpus.

In Chapter Six, the author addresses the topic of internet memes, which figure significantly in the HERMES corpus. Memes pertain to the copying of multimedia

that is quickly generated and transmitted; this is relevant to social media language because Zappavigna notes that electronic texts are easily remixed through image, verbiage, audio and video manipulation "to produce many derivatives of an original concept" (p. 100). These memes can take shape as "phrasal templates", which serve as a formulaic scaffolding in which lexical items are customisable in individual slots (p. 106). The author gives several examples of phrasal templates for memes, including, for example, the popular "im in ur [noun] [present infinitive verb] [noun]" template. Zappavigna explains that these memes are deployed for social bonding rather than sharing information, and humour is a common strategy that supports bonding.

In Chapters Seven, Eight, and Nine, the author addresses specific functions and features of social media language: slang, humour, and politics, respectively. In chapter Seven, Zappavigna explores the interpersonal function of creating and maintaining solidarity through the use of slang. More specifically, Zappavigna addresses the importance of "geek identity" and how it becomes indexed by specific lexical items in the dataset; technology, it would seem, is a major motivation for affiliation and identification. Chapter Eight addresses an important factor in social media – humour, and in particular the humorous uses of the hashtag #fail. Finally, before turning to the conclusion in chapter Ten, the author addresses the issue of political discourse in chapter Nine. Here, Zappavigna summarises findings from the "Obama win" corpus – a specialised corpus of tweets collected in the 24 h following Barack Obama's victory in the 2008 US presidential elections.

As can be seen from the range of topics covered, this book is an ambitious and important addition to the field. Nevertheless, there are some shortfalls. These all perhaps stem from the fact that the book amalgamates theory, literature, and methods from different fields; unfortunately as a result, there is some lack of clarity with regard to theoretical concepts, methodology, and focus.

Firstly, although this book is entitled "Discourse of Twitter and Social Media", it notably fails to provide a definition and theorisation of "discourse". As a result, it is sometimes unclear if the focus of the book is structural features of the language of Twitter (e.g. hashtags, forms of address) or the discourses contained within a dataset drawn from Twitter (e.g. political discourse, construction of "geek identity") (cf. Barton and Lee 2013: 4–6). Since the author examines both, there is some confusion as to the distinction between what constitutes "language" and what constitutes "discourse" within this context. Although the author draws on emergent theory in order to avoid imposing predetermined structures and hierarchies on patterns (p. 13), the theorisation of a core concept such as *discourse* seems rather crucial in a book with this title.

Secondly, there are some issues pertaining to the consistency of application of corpus linguistic terminology (e.g. the use of "semantic prosody" and "prosodies of evaluation" on p. 181) and methods. These issues begin with the data under investigation. Although the author provides some details about the HERMES corpus and its collection and content, there are other corpora, such as the "Obama win" corpus (p. 177), where very few details are provided about size and parameters. There are issues throughout the book concerning quantification and it is often unclear how

salience is established within datasets – especially if details about the corpora are not provided. There is little quantification of memes and slang, and in one particular example the author discusses a "common 4-gram" (p. 169) but does not explain what "common" means, nor does Zappavigna provide raw or normalised frequencies. Frequency was also a concern with reference to the issue of "rebroadcast" material (retweets). Although the author explores the primary social functions of retweeting (p. 36), Zappavigna provides no insight into the methodological implications for using such repetitive data. Zappavigna states (p. 22) that (some?) automatically generated retweets were discarded, but does not specify how it was established that tweets were automated or if *all* automated tweets were eliminated. Notably, in other cases retweets were judged to be important "since they give important insight into what is considered significant enough to republish within Twitter communities" (p. 22). However, it seems rather subjective to deem some retweets significant while others are excluded altogether. Furthermore, the repetitive nature of retweets has important implications for studies of frequency – a core concept in corpus linguistics. However, the author provides no insight into how frequency was addressed methodologically within the study. Indeed, since this book is the first of its kind and ambitious in its approach, it might have been useful if the author had elaborated on how the challenges associated with using social media as data were addressed. More methodological detail would have been informative for future research on Twitter and social media.

Finally, another potential shortfall of this book is its streamlining of cultural and linguistic features. For example, it is not entirely clear why Asian-style emoticons were not considered within the emoticon system network in Chapter Four. Also, the author specifically labels HERMES as "an English-language corpus", despite the fact that "many tweets are in languages other than English" (p. 20). The author notes that "posts with non-English characters were removed from the corpus, and then a statistical language model written in pearl was used to further filter out non-English tweets" (p. 24). Certainly, no single researcher could be expected to analyse a corpus containing numerous different languages; nevertheless, it seems an oversight to avoid discussing the inherently multilingual nature of social media (see e.g. Barton and Lee 2013; Crystal 2011; Danet and Herring 2007). In other words, it is more difficult to assert that a dataset is "English" in a superdiverse environment (Vertovec 2007; Blommaert and Rampton 2011), and more reflection on the role of this international language as opposed to minoritised languages would have strengthened this book.

In summary, in ten short chapters, this accessible and engaging book presents new and innovative ways of approaching the discourse of Twitter, a type of data that had yet to be examined from a linguistic perspective. This book will be indispensable as an introductory volume for students and researchers following in Zappavigna's footsteps in this evolving area of research.

References

Barton, D., & Lee, C. (2013). *Language online: Investigating digital texts and practices*. Abingdon: Routledge.

Blommaert, J., & Rampton, B. (2011). Language and superdiversity. *Diversities, 13*(2), 1–21.

Crystal, D. (2011). *Internet linguistics: A student guide*. London: Routledge.

Danet, B., & Herring, S. (Eds.). (2007). *The multilingual internet: Language, culture, and communication online*. Oxford: Oxford University Press.

Vertovec, S. (2007). Super-diversity and its implications. *Ethnic and Racial Studies, 30*(6), 1024–1054.

Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on Twitter. *New Media & Society, 13*(5), 788–806.

Aijmer, K. and Altenberg, B. (eds.) (2013). *Advances in Corpus-Based Contrastive Linguistics. Studies in Honour of Stig Johansson*. Amsterdam: John Benjamins

Elaine Vaughan

Abstract Karin Aijmer and Bengt Altenberg's edited volume brings together 12 papers which represent the state-of-the-art in corpus-based contrastive analysis. Contrastive analysis, and corpus-based contrastive analysis in particular, appear to be enjoying something of a revival, with the emphasis on 'appear': the former has a long history and, from at least the 1990s on, there is a consistent line of enquiry evident in the literature relating to the latter. This volume is based on a workshop on 'corpus-based contrastive analysis' convened during the 2011 ICAME conference and, like the conference, in honour of Stig Johansson. Within the 12 individual papers by 22 international contributors that make up this volume, Dutch, English, French, German, Norwegian, Spanish and Swedish are discussed. It will be of use to the reader interested more broadly in how this 'conversation' between different disciplinary paradigms and corpus methodology is developing, as well as its primary audience: readers interested in contrastive linguistics (corpus-based or otherwise), translation studies and foreign language pedagogy.

Keywords Contrastive analysis • Corpus-based contrastive analysis • Corpus-based approach • Multilingual corpora • Language comparison

Contrastive analysis, and corpus-based contrastive analysis in particular, appear to be enjoying something of a revival, with the emphasis on 'appear': the former has a long history and, from at least the nineties on, there is a consistent line of enquiry evident in the literature relating to the latter. Krzeszowski (e.g. 1985) traces contrastive analysis itself back as far as the fifteenth century, with contrastive theories proper appearing at the beginning of the seventeenth century. The search for links between language families dominated the nineteenth century with investigations

E. Vaughan (✉)
School of Modern Languages & Applied Linguistics,
University of Limerick, Ireland
e-mail: Elaine.Vaughan@ul.ie

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_14

which were mainly based on empirical and historical methodologies (cf. Gómez-González and Doval-Suárez 2005). After the Second World War and into the 1970s, contrastive linguistics as a discipline was often deployed in the service of practical, pedagogical purposes, the goal being to understand the differences between languages in order to teach a target language more effectively. For example, many practising English language teachers' first brush with contrastive analysis is via Swan and Smith's (e.g. 2001) *Learner English*, which contrasts English with 22 other languages or groups of languages in terms of phonological systems, lexical, grammatical and syntactic features, and punctuation in the written language. This orientation in the research was unidirectional in the sense that one of the languages being compared (usually English) was adopted as a frame of reference (Gómez-González and Doval-Suárez 2005: 21) for the other language/s. The foundations of the approach in much of the contrastive studies at that point were, as Aijmer and Altenberg point out in the introduction to this collection of papers, very much 'intuitive and limited to comparing abstract systems (or subsystems) rather than exploring languages in use' (2014: 1).

The research agenda in contrastive linguistics is now quite firmly underpinned by the comparison of different languages – rather than adopting only one language as a frame of reference – and as a basis for this comparison the use of computer corpora of different languages, with the implications this has for theories and methods of contemporary contrastive linguistics. Marzo et al. (2012: 1) have stressed "…the necessity of a fully-fledged corpus-driven approach in contrastive linguistics and its indispensable interaction with theoretical findings". This edited volume not only pursues this agenda but is also interesting in the way it illustrates research paradigms (such as contrastive linguistics, in this case) driving forward developments in software for corpus compilation, viewing/storing and analysis. Corpus linguistics itself, of course, was made possible by and develops in tandem with technological advances – to an extent at least; but the reciprocity visible in different disciplines' theoretical and methodological interactions with corpora, corpus linguistic methods and corpus linguistics as a broader area, in terms of how they adapt to one another is an interesting study. This book will be of use to the reader interested more broadly in how this 'conversation' between different disciplinary paradigms and corpus methodology is developing, as well as its primary audience: readers interested in contrastive linguistics (corpus-based or otherwise), translation studies and foreign language pedagogy.

The emergence of a 'new era' in contrastive linguistics, instituted in the early nineties, or the corpus-based approach to contrastive linguistics, can be credited in no small part to the work of Stig Johansson and his team on the English-Norwegian Parallel Corpus, amongst other projects and research. This volume is based on a workshop on 'corpus-based contrastive analysis' convened during the 2011 ICAME conference and is, like the conference, in honour of Stig Johansson. Within the 12 individual papers by 22 international contributors that make up this volume, Dutch, English, French, German, Norwegian, Spanish and Swedish are discussed.

The introduction gives Stig Johansson the prominence in corpus-based contrastive linguistics he so richly deserves and, *inter alia*, mentions why Johansson called the English-Norwegian Parallel Corpus a 'parallel' corpus – inspired, apparently,

partly by the Rosetta Stone's and partly by the Anglo-Saxon translation of the Vulgate version of the Bible's interlinear, parallel presentation. The distinction made between types of corpora most frequently consulted by contrastive linguists is obviously highly relevant to the current volume, and made clear by Johansson himself (1998): within the field of multilingual corpora, he bracketed off *comparable corpora*, with original texts in the same language; *translation corpora*, with original texts in a language and their translations into other languages; and *parallel corpora*, with original texts aligned with translations, where corresponding units of one type or another are linked. A key issue for the more general reader is the basis upon which languages can be compared at all, and whether this be functional, communicative or pragmatic, the question of equivalencies, or *tertia comparationis*, needs to be established. Due to the level of rigour and detail in all of the chapters that make up the volume, it is only possible to give a brief outline of the main focus of each, but what follows should provide a flavour of the research territory of the volume.

Thomas Egan's chapter addresses the issue of *tertia comparationis*, and focusses on the concept of 'betweenness' in English and French via translations of the Norwegian preposition *mellom*. He outlines his semantic/pragmatic (rather than syntactic) basis for comparison, isolates seven senses of 'betweenness' and posits an eighth, idiomatic, which captures the remainder. He finds a considerable degree of similarity between English and French encoding of 'betweenness' and highlights the benefits of a 3-text approach in comparison to 2-text approaches, suggesting that further insights may be gained from comparing both with a language other than Norwegian. Subsequent chapters take different approaches to establishing bases for comparison, which as they are various, will be interesting to readers: Åke Vikberg's chapter offers a typological perspective and reviews a specific feature of Swedish, its range of motion verbs, with reference to these verbs' correspondents in English, German, French and Finnish. The corpus-based approach in this study allows for a far more fine-grained analysis, which brings into focus different languages' perspectives on motion depending on different types of situation. The chapter by Sylvie de Cock and Diane Goosens illustrates the use of corpus tools to retrieve comparable units of language for a study of approximators in English and French business news, with their investigation positioned around analysis of types of approximation and preferences in their use. They employ part-of-speech tagging to extract numbers and then a collocation programme to identify approximators in their vicinity, and through this approach find that there is less approximation around numbers in French. They suggest that this may be due to differing levels of formality in business reportage in the corpora they use, approximation being often a feature of less formal discourse. Rabadán and Izquierdo focus on affixal negation by investigating how English is translated into Spanish in this respect, and comparing these translations with non-translated Spanish texts. They illustrate how typological correspondence between forms may not actually match up to actual distribution and use of the forms. This chapter gives a detailed outline of the composition of the corpora consulted, which, as is acknowledged here and elsewhere in the volume, an essential prerequisite to interpretation of results.

Chapters by Anne-Marie Simon-Vandenbergen and Kate Beeching deal with the rather slippery fish that are pragmatic markers. Simon-Vandenbergen focuses on the adverbs *basically*, *essentially* and *fundamentally*, items that have been studied from a monolingual and contrastive point of view previously. Her point of departure is to investigate how exploring their translations into French and Dutch can shed further light onto their "subtle and contextual shades of meaning" (p. 136) in English. Beeching looks at how parallel corpora can be used to provide evidence of semantic change and focuses on the French pragmatic marker, *quand même*. She notes that pragmatic markers require recourse to the examples of authentic use corpus data provide, as it is far more difficult to have a broad or even conscious enough intuitive sense of the many ways in which they operate. While showing that existing corpora can be used to great complementary effect, she notes the lack of translated spontaneous spoken data in this regard. Stenström's chapter on the markers *vale* in Spanish and *okay* in English pursues a number of lines of investigation including what both markers have in common functionally, whether they occur with the same frequency, and who is using them and how. She uses corpora designed to represent the speech of teenagers in Spain and the UK, and finds that *okay* and *vale* are indeed characteristic of teenage talk, with *okay*'s functional versatility matched by *vale*'s frequency. Where it was supposed that there might be a level of pragmatic borrowing of *okay* in the Spanish data, this was not the case, suggesting that *vale* "does the job" (p. 136) it needs to.

Sylviane Granger and Marie-Aude Lefer take as their starting point Johansson's assertion of how valuable the application of cross-linguistic corpus research is to bilingual lexicography. They absolutely concur, but note a lack of research which refers to multilingual corpora in this regard. They investigate lexical bundles (also referred to as chunks or n-grams more generally) related to *yet* and *encore* in English and French bilingual dictionaries, comparing them with corpora of English and French (with the caveat that French does not yet have a representative national corpus along the lines of the British National Corpus, which they consult). They find that in one direction – French to English – the phraseological coverage of *encore* is limited in the lexicographic data they use, while the English to French entries on *yet* appear to reflect naturally occurring language use. Ebeling, Oksefjell Ebeling and Hasselgård's chapter also explores phraseological units, in the sense that they study what they define as "recurrent word-combinations" which function as semantic units (so not, for example, non-phraseological n-grams such as *he said and*, p. 179). They explore phraseological differences between English and Norwegian using the English-Norwegian parallel corpus. One of the interesting aspects of this chapter is their discussion of the methodological issues in defining and retrieving comparable units for analysis – given that where in English a compound like *mobile phone* is represented in two orthographic words, whereas in Norwegian two stems are joined together (*mobiltelfon*) – and the corpus-driven approach they take to solving this conundrum. They present three case studies which explore in depth the 3-word combinations that they noted especially as being more frequent in translated than original texts.

Kunz and Steiner's chapter opens out from previous chapters' analyses of single items and lexical bundles to look at the broader sphere of cohesion in texts and cohesive substitution in German and English specifically. They clarify their approach to cohesion as a concept, and the relationship between cohesive reference, substitution/ellipsis and lexical cohesion, before moving on to specifically contrast English and German realisations of substitution, noting a greater variety of devices in German than English. Herriman's chapter investigates the extraposition of clausal subjects in English and Swedish. Both languages are characterised by the late positioning of 'newsworthy' elements, or Hallidayan Rheme, and the thematic variation which allows finite and non-finite clausal subjects to be postponed until after their predicate, replaced by anticipatory subject pronouns, or extraposition. Herriman uses the English-Swedish Parallel Corpus as a source of data, and finds that extraposition is more frequent in the Swedish data, explaining, perhaps, observed 'overuse' of extraposition by Swedish learners of English. The final chapter by Lavid, Arús and Moratón explores thematic variation in specific genres of English and Spanish (news reporting and commentaries). The availability of journalistic texts make them an attractive proposition in terms of data collection; journalistic discourse is not homogenous, however, and is characterised by "considerable generic variation both within and across media, languages and cultures" (p. 261). This chapter also takes its units of analysis and theoretical grounding from Hallidayan Systemic-Functional Linguistics. It is the generic characteristics of news reporting and commentaries that prove the most significant factor in thematic variation, and the chapter fleshes out Theme, an "elusive linguistic category" (p. 282). As with the all the chapters in this volume, while the findings are important and interesting, the method by which the findings are arrived at, and the transparency with which the chapters describe their data and methodology, are equally so.

Many different corpora were consulted in the papers that make up this volume, and though it is perhaps unusual for an edited collection to contain appendices, it would have been useful to have the corpora consulted uniformly described – each chapter does indeed outline the corpus or corpora consulted, but some in more detail than others. This is a small quibble, will be mostly irrelevant for the vast majority of the audience for the volume (who will most likely have more than a passing knowledge of the corpora in question), and it certainly does not detract from the volume. The volume is carefully sequenced, and though readers will most likely be targeting a particular chapter, it is worth reading in sequence for the clear overview this gives of the 'state of the art' in corpus-based contrastive linguistics.

The value of corpora in "opening up new fronts" to use an interesting metaphor (Gómez-González and Doval-Suárez 2005) can be seen very clearly for the research agenda at the heart of this volume. Perhaps in the history of contrastive analysis we can also see parts of the history of language study in and of itself – the move from intuition to empiricism, prescriptivism to descriptivism, a unitary approach to characterising language use (syntactic, grammatical, lexical) to a more unified approach (the lexico-grammatical turn, the pragmatic turn) as well as the blending of schools of thought on language use and methods and units of analysis. This edited volume is a valuable compendium of current research using multilingual corpora.

References

Gómez-González, M., & Doval-Suárez, S. M. (2005). On contrastive linguistics: Trends, challenges, problems. In C. Butler, M. Gómez-González, & S. M. Doval-Suárez (Eds.), *The dynamics of language use* (pp. 19–45). Amsterdam: John Benjamins.

Johansson, S. (1998). On the role of corpora in cross-linguistic research. In S. Johansson & S. Oksefjell (Eds.), *Corpora and cross-linguistic research* (pp. 1–24). Amsterdam: Rodopi.

Krzeszowski, T.P. (1985). The so-called 'sign theory' as the first method in contrastive linguistics. In U. Pieper & G. Stickel (Eds.), *Studia Linguistica. Diachronica et Synchronica* (pp. 485–501). Berlin: Mouton de Gruyter.

Marzo, S., Heylen, K., & De Sutter, G. (2012). Developments in corpus-based contrastive linguistics. In S. Marzo, K. Heylen, & G. De Sutter (Eds.), *Corpus studies in contrastive linguistics* (pp. 1–6). Amsterdam: John Benjamins.

Swan, M., & Smith, B. (Eds.). (2001). *Learner English: A teacher's guide to interference and other problems*. Cambridge: Cambridge University Press.

Adolphs, S. and Carter, R. (2013). *Spoken Corpus Linguistics: From Monomodal to Multimodal*. London: Routledge

Keiko Tsuchiya

Abstract Svenja Adolphs and Ronald Carter's 'Spoken Corpus Linguistics: From Monomodal to Multimodal' (2013, Routledge) is one of the innovative and advanced volumes in the fields of corpus linguistics and pragmatics. It illuminates the emergent areas of spoken corpus linguistics and covers a variety of issues from practical guidelines for designing spoken and multimodal corpora to some pedagogic implications derived from the analyses using these corpora. It also offers several case studies on discursive practices, prosody, listener responses, and gestures in talk. With these areas of focus, this volume makes a distinct contribution to the series, Routledge Advances in Corpus Linguistics. The book is divided into two parts: monomodal spoken corpus analysis and multimodal spoken corpus analysis. Monomodal spoken corpus analysis focuses on one mode, spoken language, in other words, 'textual dimension of communication' (p. 1), while multimodal spoken corpus analysis deals with plural and diverse aspects in spoken interaction which include 'textual, prosodic and gestural representations' (ibid). The former introduces a practical framework for design and development of spoken corpora, and includes case studies of monomodal spoken corpus analysis on multi-word units and discourse markers. In the latter part, the book moves from monomodal corpus analysis to multimodal corpus analysis, where studies on prosody and gestures are presented. This cutting-edge work will stimulate its readers' ingenuity, and take corpus research forward into its next stage.

Keywords Multimodal corpus • Monomodal corpus • Spoken corpus linguistics • Multimodal discourse analysis • Phraseology

Interdisciplinary and multimodal approaches are recent trends in corpus linguistics (O'Keeffe and McCarthy 2010). 'Spoken Corpus Linguistics: From Monomodal to Multimodal' is one of such innovative and advanced volumes in the field. It illuminates the emergent areas of spoken corpus linguistics and covers a variety of issues

K. Tsuchiya (✉)
Foreign Language Centre, Tokai University, Tokyo, Japan
e-mail: ktsuchiya@tokai-u.jp

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3_15

from practical guidelines for designing spoken and multimodal corpora to some pedagogic implications derived from the analyses using these corpora. It also offers several case studies on discursive practices, prosody, listener responses, and gestures in talk. With these areas of focus, this volume makes a distinct contribution to the series, Routledge Advances in Corpus Linguistics.

While it addresses a wide range of topics in spoken corpus linguistics inclusively, it also provides sound and detailed descriptions on each topic. Furthermore, it also keeps a good balance between theoretical and practical aspects in spoken corpus construction and development. This book is perhaps not designed as an entry level beginners' guide for those who are new to corpus linguistics since there are no detailed instructions on, for example, what we can do with search engines of existing corpora such as the British National Corpus and Xaira. Yet despite this, the well-structured organisation makes the volume readily accessible to a broad range of readers. It will be particularly useful for graduate students and researchers in corpus linguistics or/and discourse analysis, who attempt to design and develop spoken corpora for their own research purposes. It will also be beneficial for computer scientists and engineers doing research on spoken and multimodal interaction (Chapter 7), and practitioners in language education wanting to conduct spoken corpus analysis of learner English beyond the level of concordances and key word lists (Chapter 4: Case Study 1).

The volume is divided into two parts: monomodal spoken corpus analysis and multimodal spoken corpus analysis. Monomodal spoken corpus analysis focuses on one mode, spoken language, in other words, 'textual dimension of communication' (p. 1), while multimodal spoken corpus analysis deals with plural and diverse aspects in spoken interaction which include 'textual, prosodic and gestural representations' (ibid). The former introduces a practical framework for design and development of spoken corpora (Chapter 1), and includes case studies of monomodal spoken corpus analysis on multi-word units (Chapter 2 and Chapter 4: Case Study 2) and on discourse markers (Chapter 3 and Chapter 4: Case Study 1). In the latter part, the book moves from monomodal corpus analysis to multimodal corpus analysis, where studies on prosody (Chapter 5) and gestures (Chapters 6 and 7) are presented. Then, readers will enjoy the forward-looking discussion on a future direction in spoken corpora at the end of this book (Chapter 8). After a concise review of the history of spoken corpora, a framework for spoken corpus design is presented, which has three phases: recording, transcribing (make-up and coding) and analysis (Chapter 1). The issues researchers need to consider at each stage are described precisely in this chapter. To build a spoken corpus, researchers should decide what kind of data set and how much they need, and in what condition they can record them. Researchers also need to deal with ethical issues when recording interactions among participants, and to make a decision on what forms of annotation and what level of detail are required, and which transcription format is utilised. The chapter provides several examples of transcription formats for spoken corpora such as column type or musical notation type (Chapter 1), which will enable its

readers to have some images of what spoken corpora look like, and stimulate their ideas for corpus design. This volume updates a corpus design guideline specifically for spoken and multimodal corpora, which differentiates the work from previous practical guidelines for corpus building (McEnery et al. 2006).

For researchers who work on lexical and discourse features of spoken interaction and wish to analyse them not as a single word but as clusters, the monomodal spoken corpus analyses on multi-word units (Chapter 2 and Chapter 4: Case Study 2) might be the first chapters to read. Chapter 2 starts with a review of existing studies on multi-word units and moves to an analysis of two- to six word clusters in the spoken data extracted from the British National Corpus. The study shows inspiring findings in the use of these clusters in spoken interaction (i.e. *I think that*, *It seems to me that*) in relation to theories in discourse analysis, for example, face saving and politeness (Brown and Levinson 1987) (Chapter 2). The second case study (Chapter 4: Case Study 2) also looks at the use of multi-word units, but this time in a particular academic setting, namely lectures (i.e. *well I'm saying*, *which again*). The analysis highlights the close relationship between the use of particular multi-word units, or lexical bundles, and the context. The relationship between language and context is one of the important themes underlying this volume and the authors' previous works (Adolphs 2008; Carter and McCarthy 1997).

Two comparative studies of English variations, including learner English, are presented in this volume, focusing on listener response (Chapter 3 and Chapter 4: Case Study 1). An in-depth review of previous studies in listenership and response tokens is provided in Chapter 3, which is particularly invaluable for those who are interested in behaviours of listeners in spoken interaction. Referring to preceding research, response tokens are categorised according to their forms (minimal response tokens, non-minimal response tokens, clusters) and functions (continuer tokens, convergence tokens, engagement tokens, information receipt tokens) in the study (O'Keeffe and Adolphs 2008). The first study compares the use of listener responses between British English and Irish English based on spoken data sets in the two sub-corpora derived from CANCODE (the Cambridge and Nottingham Corpus of Discourse in English) and LCIE (the Limerick Corpus of Irish English) (Chapter 3). The findings from the analysis are clearly and sufficiently presented with sample extracts, tables and graphs. By reading this chapter, early career researchers will also have the benefit of learning how to develop a comparative study using two or multiple spoken corpora, and how to present the results of quantitative and qualitative analyses.

The other case study on discursive features (Chapter 4: Case Study 1) will attract attention from not only researchers but also practitioners in language education, particularly those who are teaching in Eastern Asian countries, since the study compares the use of discourse markers in pedagogic settings between native speakers of British English and learners of English in Hong Kong. Using data sets in two spoken corpora, CANCODE and the Hong Kong Learner Corpus, the use of discourse markers is analysed in terms of their positions (utterance initial, medial, final) and

functions (interpersonal, cognitive, topical, and textual). Pedagogical implications from the analysis are provided, which will also be profitable for teachers of English in secondary and higher education. These case studies demonstrate practical implementations of spoken corpora in discourse studies, where quantitative corpus-based approaches are 'happily married' with qualitative discourse analysis.

The authors take their readers to the frontier of applied linguistics in the latter half of this book: multimodal corpus analysis. After reading this section, readers will reaffirm the fact that multimodal analysis means not just studies on gestures but on any facets in human interaction. Two case studies on phraseology are offered in this volume (Chapter 5): the first one examines the use of the cluster, *I don't know why*, in Chinese learners of English using the Nottingham International Corpus of Learner English (NICLEs-CHN), focusing on boundaries of intonation units in the cluster. In the second case study, the use of *I think* in the English Native Speaker Interview Corpus (ENSIC) is analysed in relation to pauses. Through these two case studies, the authors construe the feasibility and potential of the innovative research approach with multimodal spoken corpora, which enable researchers to analyse clusters with other elements (i.e. intonation units or pause). This chapter will strongly be recommended for researchers whose expertise falls in interdisciplinary fields relating to phraseology, prosody and discourse in spoken interaction.

Multimodal discourse analysis (Kress 2011) is a close 'neighbour' to multimodal corpus analysis. The authors review theories and practices in multimodal discourse analysis, which integrate well into their studies. An advanced study on use of head nods together with verbal response tokens (continuer *mhm* and agreement *that's right*) is presented (Chapter 6). The audio-visual data was recorded and stored in the Nottingham Multimodal Corpus (NMMC) for the analysis. This chapter describes the detailed process of building the multimodal corpus, from how to set the digital video cameras for recording, to what coding schemes are applied to visual and verbal data. A multimodal annotation software called the Digital Replay System (DRS) is used for the analysis, which was developed at the University of Nottingham for the project. As the authors claim, the DRS is still at the development stage, but the reader can see how advanced this analysis tool is from the description and the captured images of the DRS although the DRS is not available currently (see Knight 2011 for more information).

Head and hand movements are central interests in this example of multimodal corpus analysis (see also Knight 2011 for related work). The book also presents an interesting multimodal corpus study on hand movements (Chapter 7). When capturing hand movements, the authors adopt not the traditional method of marker-tracking, but the new method of video-tracking, which utilises the locational information stored in the video recordings. Again, the image captures of the video-tracking appear in this chapter, which helps its readers to understand the methodology visually. The case study compares the use of hand movements in relation to pauses in verbal interaction between two different academic contexts, lectures and supervisions. The findings indicate that the frequencies of filled pause + gesture combination are similar among all of the participants (supervisors, students and

lectures) while the frequency of beat gestures, which are used to 'emphasise their co-occurring speech' (p. 173, also see Goldin-Meadow 1999), in lectures are higher than the others.

As a concluding remark, the book outlines the future with regard to spoken corpus linguistics (Chapter 8), where the potential usage of ubiquitous computing environments (i.e. GPS, Wifi) is suggested (Adolphs et al. 2011). What I found interesting and persuasive in their discussion is that the authors see the potential in the creative and efficient use of system logs. As presented in this volume, for example, researchers can track hand movements extracting locational information from video-recordings (see the case study in Chapter 7). In more theoretical aspects, research on language and context is highlighted as one of the future areas of research, and in order to explore the field, interdisciplinary approaches will be necessary. Researchers are expected to lead further integration of corpus linguistics with theories and practices in discourse and conversation analysis, and systemic functional linguistics.

On a more practical side, the book briefly introduces several software applications, which are used in their studies. WordSmith is used for the analysis of multiword units, Praat and Adobe Audition for prosody, and Transana and the Digital Replay System for capturing body movements. Some of these are free software, and the others are available for purchase. This book does not spare many pages for providing step-by-step instructions on how to use these devices, but focuses its attention on the practical implementations of these tools, in other words, what can be done with these application packages, from the content and case studies presented in this book. This is likely to be a particularly attractive feature of this book for potential readers.

This volume provides innovative approaches and insightful discussions on spoken corpus linguistics, and its content is presented orderly and logically. However, I noticed three things that could be improved. First, in terms of formatting, it could be more reader friendly if there were more detailed information in the table of contents such as section titles and pages of tables and figures. The reader with specific interests may need to browse through the book to find the particular section relevant for their own interests. Second, it could be more helpful especially for younger researchers if lists of further readings were provided at the end of each chapter. Third, it would add extra value to this volume if it has deeper analyses on the multimodal corpus studies and includes some case studies relating to ubiquitous computing environments. However, all of these are minor issues and readers can always consult the notes and the index to look for specific topics and related works.

Overall, this book breaks new ground in corpus linguistics and will attract a broad spectrum of readers in applied linguistics, including corpus linguistics, multimodal discourse analysis, phraseology and related interdisciplinary fields. For lecturers in corpus linguistics in higher education, this would be a must-buy book to purchase when planning to update and revise their module. Researchers in computer science and engineering and practitioners in language education will also find this volume useful to generate creative ideas for their research on spoken interaction. This cutting-edge work will stimulate its readers' ingenuity, and take corpus research

forward into its next stage. Therefore, I would recommend this book especially for early career researchers who have already experienced some corpus studies and want to progress their research further with multimodal analysis.

References

Adolphs, S. (2008). *Corpus and context: Investigating pragmatic functions in spoken discourse*. Amsterdam: John Benjamins Publishing Company.

Adolphs, S., Knight, D., & Carter, R. (2011). Capturing context for heterogeneous corpus analysis. *International Journal of Corpus Linguistics, 16*, 305–324.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Carter, R., & McCarthy, M. (1997). *Exploring spoken English*. Cambridge: Cambridge University Press.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences, 3*(11), 419–429.

Knight, D. (2011). *Multimodality and active listenership*. London: Continuum.

Kress, G. (2011). *Multimodal discourse: The modes and media of contemporary communication*. London: Bloomsbury Academic.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

O'Keeffe, A., & Adolphs, S. (2008). Response tokens in British and Irish discourse. In K. P. Schneider & A. Barron (Eds.), *Variational pragmatics*. Amsterdam: John Benjamins Publishing Company.

O'Keeffe, A., & McCarthy, M. (Eds.). (2010). *The Routledge handbook of corpus linguistics*. Abingdon: Routledge.

Author Index

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3

Subject Index

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*,
Yearbook of Corpus Linguistics and Pragmatics 3, DOI 10.1007/978-3-319-17948-3