Yearbook of Corpus Linguistics and Pragmatics

Jesús Romero-Trillo *Editor*

Yearbook of Corpus Linguistics and Pragmatics 2013

New Domains and Methodologies

Springer

Yearbook of Corpus Linguistics and Pragmatics 2013

YEARBOOK OF CORPUS LINGUISTICS AND PRAGMATICS

2013

For further volumes:
http://www.springer.com/series/11559

Jesús Romero-Trillo
Editor

Yearbook of Corpus Linguistics and Pragmatics 2013

New Domains and Methodologies

Springer

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

ISSN 2213-6819 ISSN 2213-6827 (electronic)
ISBN 978-94-007-6249-7 ISBN 978-94-007-6250-3 (eBook)
DOI 10.1007/978-94-007-6250-3
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013940618

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

New Domains and Methodologies in Corpus Linguistics and Pragmatics Research, an Introduction

Jesús Romero-Trillo

The present volume, *New Domains and Methodologies in Corpus Linguistics and Pragmatics Research*, marks the launch of the new Springer series *Yearbook of Corpus Linguistics and Pragmatics*. The series intends to address the interface between Corpus Linguistics and Pragmatics and is conceived to offer a platform to scholars who combine both disciplines. The rationale behind the series, which follows a peer-reviewed editorial process, is to publish research that aims at the pragmatic analysis of language in real contexts through the use of rigorous corpus analysis techniques.

The editor of the series published a volume some years ago (Romero-Trillo 2008) that represented a provocative stir in the mutualistic, though sometimes excluded and excluding, relationship between Pragmatics and Corpus Linguistics. The volume convened scholars who belonged to different generations of linguistics but shared the intuition that the only way to understand the pragmatic meaning of interaction was through the analysis of a representative volume of data, sieved through pragmatics theories. In fact, the volume intended to bridge the gap between two ways of looking at language: corpus linguistics as a method of analysis primarily informed by mathematics and statistics with the aid of an excellent and meticulous methodology; and pragmatics, on the other hand, which was perceived to have an indefinite methodology when it accounted for the interpretation of the pervasive distance between sentence and intended meaning in communication.

Since then, the scenario has changed and many scholars have trodden the narrow path between corpus linguistics and pragmatics trying to justify the theoretical pragmatics density of their descriptions, while at the same time showing a corpus linguistics aptitude that goes from technological to statistical expertise. This ambivalent orientation has sometimes created incertitude and a *niche-less* ambit for research progress, especially in its application to new linguistic domains.

J. Romero-Trillo (✉)
Departamento de Filología Inglesa, Facultad de Filosofía y Letras,
Universidad Autónoma de Madrid, C/Francisco Tomás y Valiente 1,
28049 Madrid, Spain
e-mail: jesus.romero@uam.es

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_1,

In sum, this new series is aimed at researchers who want to unite the delicacy of pragmatics analysis with the guaranteed representativeness of corpus linguistics. In fact, the series volumes will pay special attention to the recently universalised corpus compilation capacity of all scholars via the ubiquitous access to digital, computerised and visual data. I am sure that the near future will bring very interesting and surprising linguistic data coming from established social networks, such *youtube, facebook, twitter, google+, foursquare, flikr,* etc., and others in the offing for the general public such as *delicious, pinterest* or *paper.li*, *inter alia*. The vibrant combination of language (spoken, written and mixed –as in chat rooms) and visual information in these online networks, with the availability and mobility of technological gadgets that provide these services, will surely influence language and communication in the near future and will be the source of research for corpus linguistics, pragmatics, sociolinguistics, language education, psycholinguistics, etc.

One of the first results of this multifaceted data mining possibilities is that researchers have started to use tailor-made corpora, aside with the classic reference corpora. The design and pragmatic analysis of these often ad hoc and usually smaller corpora is also one of the main topics in this volume, as they tend to focus on the pragmatics of well-defined situations in such a way that can illustrate the specific features of communication in new and under-researched contexts. In fact, the present volume approaches some the 'trending topics' that have been mentioned above, in combination with some theoretical issues that are currently discussed in the synergic practice of the two disciplines.

The volume is structured in four sections. The first three contain chapters that investigate the following topics: first, 'Current theoretical issues in pragmatics and corpus linguistics research'; second, 'New domains for corpus linguistics and pragmatics'; and third, 'New methodologies for the pragmatic analysis of speech through corpora'. The fourth part reviews two books that will surely be of great interest to the readers.

The opening chapter of the first part, *'Current Theoretical Issues in Pragmatics and Corpus Linguistics Research'*, is authored by Marcus Callies and the title is '*Advancing the Research Agenda of Interlanguage Pragmatics: The Role of Learner Corpora*'. The chapter reviews the role of pragmatics in Second Language Acquisition research and defends a broader role of the discipline in Interlanguage Pragmatics (ILP). The author argues that pragmatic knowledge in a foreign/second language (L2) includes more than the sociopragmatic and pragmalinguistic abilities for understanding and performing speech acts. In his opinion, learner corpora can override the limitations posed by the dominance of data elicitation techniques in ILP. By way of illustration, the chapter shows empirical results in French and German learners of English of the pragmalinguistic component of L2 pragmatic knowledge through the study of information organization in discourse, and the use of lexico-grammatical means of information highlighting for intensification and contrast.

The second chapter, 'Corpus Linguistics and Conversation Analysis at the Interface: Theoretical Perspectives, Practical Outcomes', authored by Steve Walsh, offers a theoretical perspective on the pros and cons associated with a

combined Corpus Linguistics (CL) and Conversation Analytic (CA) approach to the study of language. For the author, Corpus Linguistics and Conversation Analysis have different origins and research foci, and often some scholars believe that they incompatible because Corpus Linguistics is mainly quantitative, while Conversation Analysis focuses on the study of talk-in-interaction. The author compares the various arguments in favour of one or the other approach, with especial reference to their possible combination in the co-construction of meaning in an educational context.

The last chapter of this section is entitled 'Small Corpora and Pragmatics' and written by Elaine Vaughan and Brian Clancy. Their contribution describes the growing interest in the study of pragmatics based on small context-specific corpora, both spoken and written. According to the authors, the advantage of the analysis of language at this scale is that fine-grained distinctions that pertain to contextual or genre-based features can be better studied when the corpus collection has been carefully controlled. The authors provide evidence with two corpus case studies that illustrate the symbiosis of contextual control and corpus linguistics for pragmatics research.

The second part, 'New Domains for Corpus Linguistics and Pragmatics', opens with the chapter by Britt Erman, Margareta Lewis and Lars Fant 'Multiword Structures in Different Materials, and with Different Goals and Methodologies'. The chapter delves into the patterns of word combinations in Second Language Acquisition in spoken and written corpora. The authors describe different paradigms to assess the production of these structures according to some variables: medium, size, control of task, topic and discipline. Two methods, the lexical bundle and the 'comprehensive' method, are applied to the analysis of spoken language of English and Spanish native and non-native students. The authors argue that the two methods can be combined to broaden the conception of these structures, but also to understand what being native-like means for the design of materials and the analysis of production in language teaching contexts.

The next chapter, 'Discourse Functions of Recurrent Multi-word Sequences in Online and Spoken Intercultural Communication' by Yen-Liang Lin, also approaches multi-word sequences. In this case the study investigates the discourse functions of multi-word sequences comparing computer-mediated communication (CMC) and face-to-face (FTF) interaction. The author concentrates on some recurrent multi-word (three-word) sequences firstly over time, and then focuses on the 50 most common three-word sequences. The chapter compares the online and spoken datasets and also two reference corpora. The sequences are classified according to three categories – social interaction, necessary topics and discourse devices – with regard to the primary discourse function they realise. The chapter concludes with the explanation of the functional differences present in both types of communication, face-to-face and computer-mediated.

Dawn Knight, Svenja Adolphs and Ronald Carter author the chapter 'Formality in Digital Discourse: A Study of Hedging in CANELC'. Their study presents a corpus-based analysis of formality in e-language and compares the levels of formality in e-language with spoken and written discourse in the BNC. The chapter focuses

on common indicators of formality in discourse, especially in hedging. The data for the analysis comes from the recent one-million-word Cambridge and Nottingham e-language Corpus (CANELC), which contains language from online discussion boards, blogs, tweets, emails and SMS messages.

The last chapter in this section is 'A Corpus-Based Classification of Commitments in Business English', by Rachele De Felice. The study presents a corpus-based analysis of commitments in Business English emails. The author uses a speech act-annotated corpus of emails that departs from a detailed analysis of the lexicon and phraseology of commitments. The chapter proposes a new classification of this speech act with a clear illustration of the contribution of corpus linguistics to the pragmatic description of workplace communication.

The third part, 'New Methodologies for the Pragmatic Analysis of Speech Through Corpora', starts with the chapter written by Carolina Amador Moreno, Michael McCarthy and Anne O'Keeffe, entitled 'Can English Provide a Framework for Spanish Response Tokens?' Their study investigates if response items in Spanish can be analysed using frameworks developed for the study of related items in English. They base their research on the Spanish corpus COREC, the Corpus Oral de Referencia del Español Contemporáneo, and for English the British English corpus, CANCODE, and the Cambridge and Nottingham Corpus of Discourse in English. The authors try to assess the possibility and appropriateness of using English-based frameworks for the analysis of Spanish and, as a second step, to develop the notion of 'good listenership'. Specifically, their study concentrates on (a) formal aspects of response items in Spanish, (b) the pragmatic coverage of the items and their translatability and transferability, and (c) some insights into potential cross-cultural misunderstandings with English as the comparison language. Their conclusion supports the idea that response tokens are essential elements for active listenership, and that fluency can only be really appraised in dialogic contexts.

The second chapter is written by Jesús Romero-Trillo and is entitled 'The Corpus of Language and Nature (CLAN Project)®: A Tool for the Study of the Relationship Between Cognition and Emotions in Language'. This chapter makes a description of the theoretical tenets of the structure and compilation of the Corpus of Language and Nature. The chapter describes the foundations of the analysis of natural landscapes from an ecological perspective and makes the link with the theory of the cognitive appraisal of natural landscapes developed by Romero-Trillo and Espigares (2012). The chapter presents the design of the corpus compilation step by step as a way to illustrate how to use modern technological and computer resources for linguistic analysis and data archiving. The objective of the chapter is also to illustrate the process of design and collection for readers intending to start a corpus with a sound experimental design.

The third chapter, 'System Networks as a Tool for the Pragmatic Analysis in an EFL Spoken Corpus', is authored by Silvia Riesco-Bernier. It presents the design of a network to operationalize the study of regulatory functions in EFL pre-school teacher talk. Based on Systemic Functional Linguistics and Classroom Discourse Analysis applied to a corpus of EFL data, the study defines the variables that portray

the discourse-semantic options in the instantiation of regulatory functions in teacher talk. The elaboration of the system network is supported by statistical tests and the results can certainly constitute a valid instrument to systematize the study of pragmatic content in classroom discourse.

The last chapter of this section, written by Anna Gladkova, is 'A Cultural Semantic and Ethnopragmatic Analysis of the Russian Praise Words *Molodec* and *Umnica* (with Reference to English and Chinese)'. Using data from the Russian National Corpus the author explores the semantics and ethnopragmatics of two Russian praise words *molodec* and *umnica*. The methodology used in the study is Natural Semantic Metalanguage (NSM), which formulates semantic explications and cultural scripts as a reflection of the underlying cultural ideas expressed by these concepts. The cultural specificity of the terms is compared with other Russian cultural key words and ideas and in comparison with pragmatic equivalents in English (good boy/girl) and in Chinese (*guāi*).

The last part reviews two books of great interest for pragmatics and corpus linguistics: *Corpus Linguistics: Methods, Theory and Practice* by Tony McEnery and Andrew Hardie, *and Cyberpragmatics. Internet-Mediated Communication in Context* by Francisco Yus. Undoubtedly, the description of the two volumes complements the rich and diverse topics dealt with in the book chapters, which as mentioned above, range from the more theoretical aspects of the intersection of corpus linguistics and pragmatics, to the more applied, pedagogic and computer-based domains that result from their synergy.

To conclude, I believe that this first volume of the new series 'Yearbook of Corpus Linguistics and Pragmatics' will offer an indispensible source of expertise to both experienced and novice scholars in these disciplines, and will also contribute to the understanding of language and communication with up-to-date methodologies that will cover a broad spectrum of topics.

References

Romero-Trillo, J. 2008. *Pragmatics and corpus linguistics, a mutualistic entente*. Berlin: Mouton de Gruyter.

Romero-Trillo, J., and T. Espigares. 2012. The cognitive representation of natural landscapes in language. *Pragmatics and Cognition* 20: 168–185.

Part I
Current Theoretical Issues in Pragmatics and Corpus Linguistics Research

Advancing the Research Agenda of Interlanguage Pragmatics: The Role of Learner Corpora

Marcus Callies

1 Pragmatics in Second Language Acquisition Research: A Critical Assessment

1.1 Interlanguage Pragmatics and Its Scope of Inquiry

Broadly defined, pragmatics as a discipline can be conceived of as "the study of language from the point of view of the users, especially of the choices they make, the constraints they encounter in using language in social interaction, and the effects their use of language has on the other participants in an act of communication" (Crystal 2003: 364). Leech (1983: 10f.) distinguishes between two components of general pragmatics. First, he defines socio-pragmatics as "the sociological interface of pragmatics" that focuses on the conditions of language use which derive from the social situation, i.e. the social setting of language use, including variables such as cultural context, social status or social distance of speakers. Second, pragmalinguistics is "the more linguistic end of pragmatics", considering the particular linguistic resources which a given language provides for conveying particular illocutions, i.e. the range of structural resources from which speakers can choose when using language in a specific communicative situation, e.g. speech act verbs, imperatives, politeness markers, pragmatic markers etc.

The study of pragmatics as a field of inquiry within Second Language Acquisition (SLA) research is usually referred to as Interlanguage Pragmatics (ILP). ILP is commonly defined as "the study of nonnative speakers' comprehension, production, and acquisition of linguistic action in L2" (Kasper 2010: 141). While this suggests a relatively broad range of research topics as in pragmatics in general, ILP to date

M. Callies (✉)
English-Speaking Cultures, University of Bremen,
FB 10, Bibliothekstr.1, GW2, Bremen 28359, Germany
e-mail: callies@uni-bremen.de

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_2,

has operated on a fairly narrow understanding of what constitutes linguistic action in L2. One of the main reasons for this is that traditionally, ILP has been heavily influenced by and largely modeled on cross-cultural pragmatics, adopting its research topics, theories and methodologies (Kasper 2010: 141). Thus, it has predominantly been concerned with politeness phenomena by investigating foreign/second language (L2) learners' comprehension and production of a variety of speech act types such as requests, apologies, refusals, complaints, compliments and compliment responses, and the use of internal and external modification to these speech acts. The findings of these investigations have subsequently been compared with native speaker performance.

In their review of research methods in ILP, Kasper and Dahl (1991) define the field "in a narrow sense, referring to nonnative speakers' (NNSs') comprehension and production of speech acts, and how their L2-related speech act knowledge is acquired" (1991: 216). Studies addressing topics like conversational management, discourse organization, or sociolinguistic aspects of language, e.g. address forms, were explicitly left outside of the scope of this article. This narrow view has been taken over in many overview articles and book chapters on ILP that have been published since. For example, Ellis (2008: 160), explicitly referring to Kasper and Dahl (1991), also adopts the narrow sense of ILP arguing that this aspect of pragmatics has received the greatest attention in SLA research. Ellis even maintains that the scope of pragmatics in ILP is "relatively well-defined. Researchers have investigated what speakers accomplish when they perform utterances in terms of: (1) interactional acts and (2) speech acts" (2008: 159). In sum, this perspective has led to a narrow research focus and sociopragmatic bias in ILP where the dominant area of investigation has been the speech act.

Almost 20 years after Kasper and Dahl's review paper, Bardovi-Harlig (2010) provided a state-of-the-art meta-analysis of published research in ILP. Noting that "the study of interlanguage pragmatics has not typically been as broad as the areas outlined by the definition of pragmatics used in the handbook",[1] she states that "within second language studies, work in pragmatics has often been narrower than in the field of pragmatics at large" and that "there seems to be less agreement in the field about the scope of *pragmatics*" (2010: 219f.; emphasis in original). Her meta-analysis of a sample of 152 research articles published between 1979 and 2008 reveals that in 99 out of the 152 studies reviewed (65.1 %), pragmatic competence was operationalized in terms of speech acts. This leads her to conclude that "the dominant area of investigation within interlanguage pragmatics has been the speech act" (2010: 219). Only few studies have investigated other pragmatic phenomena, e.g. turn structure (sequencing of turns, repair, alignment, greeting and leave taking), pragmalinguistic devices, i.e. grammatical and lexical devices

[1] Bardovi-Harlig refers to the *Handbooks of Pragmatics* series published with DeGruyter Mouton. In the general preface to the series, the editors state that all the handbooks in the series share the same wide understanding of pragmatics as the scientific study of all aspects of linguistic behaviour.

including routines (e.g. modal particles, adverbials, formulas), and pragmatic interpretation (meta-pragmatic knowledge and assessment, e.g. in the form of ranking or rating).

In 2005, Müller provided one of the first comprehensive studies of discourse markers in learner English. While the use of discourse markers in native English has been studied extensively in pragmatics in the last decades, Müller concluded in her overview chapter on pragmatics in SLA that "there is little in the area of second language acquisition and applied linguistics which deals explicitly with discourse markers. The focus in this area is either on grammatical features or, as far as pragmatic competence goes, on speech acts" (2005: 23).

Callies (2009a) draws attention to the pragmalinguistic component of pragmatics and its interplay with grammar. He examined advanced L2 learners' comprehension and use of focus constructions, i.e. pragmatically-motivated variations of the basic word order. Outlining that knowledge of the principles of information organization in discourse, and the use of linguistic devices for information highlighting clearly relates to L2 pragmatic knowledge, Callies suggests that further research into L2 learners' abilities at the syntax-pragmatics interface may also be a rewarding enterprise with respect to the interplay of grammatical and pragmalinguistic knowledge, an important yet unresolved issue in ILP.

Dippold (2009) notes that ILP not only prioritizes research on the expression of L2 politeness and the acquisition of politeness strategies, but that it also does so in a decontextualized manner that takes little account of the situatedness of linguistic discourse. She argues that ILP should move away from its focus on politeness in a limited set of speech acts and focus also on self-presentation.

In sum, this clearly suggests that the significance of L2 pragmatic knowledge beyond the domain of speech acts has been neglected in ILP research to date. However, the scope of pragmatics in the context of SLA does not necessarily have to be a narrow one. In many broad definitions such as the one given by Kasper (2010: 141) ("the study of nonnative speakers' comprehension, production, and acquisition of linguistic action in L2") the scope of research in ILP is not restricted to issues of politeness and the domain of speech acts. Kasper and Rose (2002) have proposed the concept of "pragmatics-as-perspective" which "has the advantage of being inclusive and open to study new research objects *as* pragmatics, without precluding them from being examined from a different angle as well" (2002: 5; emphasis in original). In fact, recent developments suggest that there is a growing awareness in the field that L2 pragmatics is more than speech acts and that the scope of inquiry needs to be adjusted accordingly. For example, LoCastro (2011: 333) observes "a movement away from an almost exclusive focus on speech acts, particularly apologies, requests, refusals, and compliments, and formulaic language to a much broader view of language in use", pointing to studies that have examined topic marking, negation strategies, referent introduction and maintenance, self-qualification, discourse markers, modal particles, definiteness, and text organization. LoCastro also notes that "many of these studies delve into complexities in signaling pragmatic meaning beyond the more commonplace comparisons of a speech act in learners' L2 production and the native speaker enactment of the same speech act" (2011: 333).

1.2 Modeling L2 Pragmatic Knowledge

In this section, I argue that pragmatic knowledge in an L2 clearly includes more than the sociopragmatic and pragmalinguistic abilities for understanding and performing speech acts and propose a more encompassing definition of L2 pragmatic knowledge. Standard descriptions of ILP frequently use notions like "linguistic action in L2" (Kasper 2010: 141) and "L2 pragmatic knowledge" (Kasper and Rose 1999: 81; Gass and Selinker 2008: 287) respectively to refer to the general domain of inquiry. But what exactly constitutes L2 pragmatic knowledge? Definitions of pragmatic knowledge or competence[2] range from rather broad and general ones, e.g. "the ability to use language appropriately in a social context" (Taguchi 2009: 1) to more detailed ones, e.g. "the knowledge of the linguistic resources available in a given language for realizing particular illocutions, knowledge of the sequential aspects of speech acts and finally, knowledge of the appropriate contextual use of the particular languages' linguistic resources" (Barron 2003: 10). While Barron's proposal draws a useful distinction between pragmalinguistic and sociopragmatic knowledge, it reflects the bias in mainstream ILP in that it centers around the concept of illocutionary acts, thus narrowing down the scope of pragmatic knowledge to sociopragmatics.

There are a number of models of language proficiency that aim to capture the ability of L2 learners to use language in social interaction, all of which acknowledge to some degree the importance to acquire pragmatic competence in L2 learning. The two most influential constructs, communicative competence and communicative language ability, will be discussed briefly in turn. In general terms, communicative competence can be defined as "the fundamental concept of a pragmalinguistic model of linguistic communication: it refers to the repertoire of know-how that individuals must develop if they are to be able to communicate with one another appropriately in the changing situations and conditions" (Bußmann 1996: 84). In reaction to Chomsky's dichotomy of competence and performance, in which the notion of linguistic competence only includes knowledge of abstract grammatical rules and sets aside contextual factors of language use, Hymes (1972) introduced the concept of communicative competence, containing both grammatical competence and knowledge of the sociocultural rules of language use. Canale (1983), building on Canale and Swain (1980), suggested a model of communicative competence that includes four major components:

- GRAMMATICAL COMPETENCE (knowledge of the language code: vocabulary, phonology, spelling, morphology, and syntax needed to produce and understand well-formed sentences);
- SOCIOLINGUISTIC COMPETENCE (knowledge of appropriate use and understanding of language in different sociolinguistic contexts, with emphasis on appropriateness of both meanings and forms);

[2]The two terms are frequently used interchangeably in the literature.

- DISCOURSE COMPETENCE (knowledge of how to combine and interpret grammatical forms and meanings to achieve unified texts in different modes by using cohesion devices and coherence rules);
- STRATEGIC COMPETENCE (knowledge of the verbal and non-verbal strategies used to compensate for breakdowns in communication and to enhance the rhetorical effect of utterances).

Although these four components are described separately in Canale's model, it should be made clear that they interact with each other and also partly overlap. Pragmatic competence is not recognized separately here, but implicitly included in the sociolinguistic component in a predominantly sociopragmatic, that is speech-act based sense. In addition, Canale sees discourse competence as bridging the gap between grammatical and sociolinguistic competence and includes it as a separate component, predominantly understood in a textlinguistic sense (hence the focus on coherence and cohesion).

Building on the work of Hymes and Canale, Bachman (1990) introduces the model of communicative language ability which is composed of three components:

- LANGUAGE COMPETENCE, "a set of specific knowledge components that are utilized in communication via language";
- STRATEGIC COMPETENCE, "the mental capacity for implementing the components of language competence in contextualized communicative language use", and
- PSYCHOPHYSIOLOGICAL MECHANISMS, "the neurological and physiological processes involved in the actual execution of language as a physical phenomenon" (1990: 84).

Particularly interesting is the component of language competence which is further subdivided into

- ORGANISATIONAL COMPETENCE, which contains the modules of GRAMMATICAL COMPETENCE (the knowledge of vocabulary, morphology, syntax, and phonology), and TEXTUAL COMPETENCE, which "includes the knowledge of the conventions for joining utterances together to form a text, which is essentially a unit of language – spoken or written – consisting of two or more utterances or sentences that are structured according to rules of cohesion and rhetorical organisation" (1990: 88), and
- PRAGMATIC COMPETENCE, which intends to capture the speaker's or writer's ability to achieve his or her communicative intentions through the use of language, subsuming ILLOCUTIONARY COMPETENCE (knowledge of expressing and interpreting language functions and speech acts) and SOCIOLINGUISTIC COMPETENCE, or "sensitivity to, or control of the conventions of language use that are determined by the features of the specific language use context" (1990: 94).

Bachman's construct thus explicitly includes pragmatic competence, which is, however, described primarily in a sociopragmatic sense.

A more detailed model of discourse competence building on Canale's construct of communicative competence has been proposed by Archibald (1994: 59f.). It includes four components:

- COHESION: knowledge of how the lexico-grammatical structures of language may be used to produce connectedness in text;
- COHERENCE: knowledge of the principles of relevance and cooperation and the illocutionary functions of language;
- SITUATIONALITY: knowledge of how a text is related to discourse context, and the role of background knowledge;
- INFORMATION STRUCTURE: knowledge of thematic structure, the ordering of given and new information.

In sum, an integration of Canale's and Archibald's modules of discourse competence, largely covering the pragma- and textlinguistic component of pragmatics, and Bachman's definition of pragmatic competence, reflecting the sociopragmatic component, seems to account best for the complex nature of L2 pragmatic competence. I thus propose the following definition of pragmatic knowledge: L2 pragmatic knowledge is the knowledge of the (pragma-) linguistic resources available in a particular language for realizing communicative intentions, and the knowledge of the appropriate socio-contextual use of these resources. Pragmalinguistic knowledge is a component of L2 pragmatic knowledge which relates to learners' knowledge of the structural linguistic resources available in a given language for realizing particular communicative effects, and knowledge of the appropriate contextual use of these resources.

2 Going Beyond Speech Acts: The Role of Learner Corpora

Research in ILP has largely relied on elicited assessment and production data, most typically in the form of pseudo-oral discourse completion or production tasks. According to Bardovi-Harlig's meta-analysis, only 27 % of the studies she surveyed collected and analyzed authentic language samples (2010: 241). Despite the firm belief that the most authentic data in pragmatic research is provided by spontaneous speech gathered through observation, the discourse completion task (DCT) has become almost the standard technique due to the manifold administrative advantages of using written questionnaires.[3] The DCT is a data collection technique widely used to elicit production data about sociopragmatic behaviour in a specific communicative context. DCTs are usually administered in the form of written questionnaires that contain several contextualized descriptions designed to create communicative situations. Informants are then asked to provide direct speech in a written response to a stimulus, e.g. a first turn provided to them. DCTs come in

[3] LoCastro (2011: 331) sees this as another reason for the dominance of speech act research in ILP.

various formats. The classic format, in which informants have to fill in only one turn at talk, consists of an open turn for the required response (sometimes prefaced by an initiation of a fictitious interlocutor), and a rejoinder to the turn to be provided by the informant. The free DCT, also called dialogue construction task, has an open response format. It can be introduced by a first pair part, but includes no rejoinder to the required response. The response can be verbal, non-verbal, or the informant is given the possibility to opt out, i.e. to provide no response at all. Another type is the discourse production task in which participants are only provided with a contextualized situational description and have to construct a short dialogue sequence involving two or more participants.

The benefits and disadvantages of using elicitation data are widely recognized and discussed in the field, and there is by now a considerable amount of literature on various issues of research methodology in ILP.[4] Obviously, DCTs make it possible to collect large amounts of data in relatively short time and with comparatively little effort. Moreover, the context and situational descriptions can be manipulated to constrain the response so that the required, often highly specific linguistic structures can successfully be elicited. Also, social variables can be controlled much more systematically than in naturally-occurring situations. But there are also several disadvantages. The DCT is a pseudo-oral format, because despite its oral setting, it is more likely to elicit written than spoken language. Apparently, informants do not write as spontaneously as they would speak, and do not necessarily write down what they would say, but rather what they imagine is expected or should be said. Thus, data elicited in such a way are more likely to reflect interactive norms and underlying social and cultural values acquired in communication or learnt in the process of socialization. While the recording of naturally occurring talk enables the researcher to study the organization and realization of talk-in-interaction in natural settings, elicited data from DCTs indirectly reflect prior experience with language. Several studies have compared various formats of DCTs with other common data collection methods to investigate the effects of the instrument on the results (e.g. Sasaki 1998; Yuan 2001; Golato 2003). While oral formats, e.g. role plays, due to their interactive nature, induce longer responses and a larger number and greater variety of strategies/formulas than questionnaires, written formats produce more direct responses.

The compilation and accessibility of computer corpora and software tools for corpus analysis has revolutionized (applied) linguistics in the last two decades. Corpus linguistics and pragmatics can be considered related, but historically distinct disciplines in that the latter is a subfield of linguistics while the former is often considered a methodological approach to carrying out linguistic research (Andersen 2011: 588). Nevertheless, corpus linguistics and pragmatics can be said to form a "mutualistic entente" (Romero-Trillo 2008) in that they are joint forces in the common cause to work with real usage data, thus more convincingly addressing some specifics of language usage by combining the methodologies

[4] See e.g. the overviews by Kasper (2008) and Ellis (2008: 163–169). Callies (2012b) summarizes the advantages and disadvantages of the DCT.

that underlie both disciplines.[5] In fact, the marriage of corpus linguistics and pragmatics has more recently given rise to a new hybrid subfield referred to as "corpus pragmatics".[6]

In ILP, learner corpora – due to their very nature of being large systematic collections of authentic, continuous and contextualized language use (spoken or written) by L2 learners stored in electronic format – can help overcome several problems and limitations posed by the dominance of data elicitation techniques to date. Not only do learner corpora enable researchers to study a much broader range of different phenomena, but they can also provide results that may be viewed as more reliable, valid, and generalizable across populations without the lack of authenticity and replicability that often arises from the use of other types of data. Learner corpora also make it possible to abstract away from individual learners and identify a corpus-based, supra-individual description of a specific learner group while at the same time providing insights into intra-group variability. Such variability and individual differences have important implications for learner corpus analysis and compilation that will be addressed in detail in the case studies in Sect. 3. Additionally, learner corpora can be the basis for quantitatively oriented studies that are subjected to statistical analyses and create an opportunity for between-methods triangulation and alternative views to qualitative, ethnographic studies that have been common in pragmatics in general.

In particular, the availability of spoken learner corpora such as the *Louvain International Database of Spoken English Interlanguage* (LINDSEI, Gilquin et al. 2010) has enabled researchers to study a wider range of pragmatic features of learner language in the spoken mode.[7] The LINDSEI was compiled by an international research team and consists of spoken data, i.e. transcripts of interviews between learners of English as a foreign language (EFL) and English native-speaker or non-native-speaker interviewers. The learners are university undergraduates in their twenties whose proficiency level ranges from higher intermediate to advanced (being assessed on external criteria, most importantly their institutional status, e.g. the time they spent learning English at school and university and the fact that they are university undergraduates in English). The LINDSEI includes subcorpora of learners from 11 mother tongue backgrounds (e.g. German, French, Italian, Japanese, Polish, and Spanish) with 50 interview transcripts per subcorpus, i.e. a total of about 100,000 words per component. Each interview lasts approximately 15 min and involves three tasks: (1) a warm-up sequence in which interviewer and interviewee talk about a set topic, (2) a free discussion, and (3) a picture description.

[5] See Andersen (2011) and Rühlemann (2011) for recent overviews of the interrelation of the two fields.

[6] See e.g. the titles of the recent/upcoming publications by Felder et al. (2011) and Aijmer and Rühlemann (forthcoming).

[7] See e.g. the papers in Romero-Trillo (2008) and the studies on the list of publications based on the LINDSEI provided by the Centre for English Corpus Linguistics in Louvain-al-Neuve, Belgium, at http://www.uclouvain.be/en-cecl-lindsei-biblio.html.

Using data from corpora of spoken interlanguage, it is now possible to systematically examine lexico-grammatical patterns and syntactic structures that are part of the grammar of conversation on a broad empirical basis (see e.g. Mukherjee 2009 for a study along these lines). Recent studies have investigated individual pragmalinguistic units, e.g. discourse markers (e.g. Müller 2004, 2005; Aijmer 2004, 2009, 2011), modal particles (e.g. Belz and Vyatkina 2005) and tag questions (Ramirez and Romero-Trillo 2005), as well as other features of turn- and discourse structure, e.g. performance phenomena like hesitations, repetitions and disfluencies (Götz 2007; Gilquin 2008) or filled and unfilled pauses (see e.g. Brand and Götz 2011 and Götz 2013 for studies that examine and operationalize these features as measures of fluency). The present chapter makes a contribution to research on the grammar of conversation in learner English and focuses on the pragmalinguistic component of L2 pragmatic knowledge, in particular as it relates to information highlighting in discourse.

3 Case Studies

An area where pragmalinguistic devices abound and are of crucial importance is discourse pragmatics, the "general domain of inquiry into the relationship between grammar and discourse" (Lambrecht 1994: 2). More specifically, I will be concerned with lexico-grammatical and syntactic means of information highlighting located at the interface of lexico-grammar, syntax and pragmatics. This interface is often referred to as information structure or information packaging, viz. the structuring of sentences by syntactic, prosodic, or morphological means that arises from the need to meet certain communicative demands, e.g. emphasizing a certain point, correcting a misunderstanding, or repairing a communicative breakdown.[8] Information highlighting is clearly pragmatically motivated because, more generally speaking, it serves to express certain pragmatic functions in discourse, e.g. intensification or contrast. Compared to their frequency of occurrence and difficulty of acquisition there are still remarkably few (corpus-based) studies that have examined the linguistic means of information highlighting in learner language from a pragmalinguistic perspective (see e.g. Boström Aronsson 2003; Herriman and Boström Aronsson 2009; Callies 2008a, b, 2009a, b). L2 learners' knowledge (that includes awareness, comprehension, and production) of discourse organization and the (contextual) use of linguistic means of information highlighting is thus still an underexplored area in SLA research, as is the interplay of pragmalinguistic knowledge and discourse organization in general. Interface relations, opaque form-meaning mappings, optionality and discourse-motivated preferences are assumed to be the main areas of difficulty in advanced SLA (DeKeyser 2005). Recent findings

[8] Deppermann (2011) provides a recent overview of the role and relevance of pragmatics for grammar, in particular as to the structuring and packaging of information and the framing of discursive action by means of grammatical constructions such as clefts.

Table 1 Learner corpora used in the case studies

Name	Writers' L1	Professional status	No. of interviews	No. of turns (only interviewees)
LINDSEI-F	French	University students	50	5,504
LINDSEI-G	German	University students	50	6,051
LOCNEC	British English	University students	50	8,436

In view of the manifold problems to operationalize the concept of sentence in transcribed spoken language and thus, to count the amount of sentences in the corpora, I chose to apply the number of speech turns as a basis of comparison

suggest that information structure management is problematic even for advanced L2 learners and that such learners have only a limited awareness of the appropriate use of lexical and syntactic focusing devices in formal and informal registers (Callies 2009a).

The following sections report on two learner-corpus studies that investigate L2 learners' use of specific lexico-grammatical means of information highlighting in English: emphatic *do* and a special type of cleft construction introduced by the deictic demonstratives *that* or *this* (demonstrative clefts). Three research questions will be examined:

1. Are there differences in the frequencies of use of emphatic *do* and demonstrative clefts in the speech of native speakers of English and learners of English as a foreign language?
2. Are there differences in how native speakers and learners use these devices contextually, i.e. as to their discourse functions and characteristic lexical co-occurrence patterns?
3. Are there differences between learners from different L1 backgrounds, and if so, how can these be explained?

3.1 Data and Methodology

Both case studies are contrastive interlanguage analyses (CIA) based on corpora of spoken interlanguage. In a CIA, two types of comparisons are combined. First, the interlanguage of a certain learner group, e.g. German learners of English, is compared with the language of English native speakers in order to pinpoint possible differences between the two groups. This comparison is then subsequently combined with a corresponding analysis of the interlanguage produced by a second group of learners, e.g. French learners of English. For the present case studies, the learner data are drawn from the German and French components of the LINDSEI (Gilquin et al. 2010). For comparable native speaker data the *Louvain Corpus of Native English Conversations* (LOCNEC) was used. The LOCNEC contains transcribed interviews with native speakers of British English (university students at Lancaster university in the UK) aged between 18 and 30 years. The interviews involved the same tasks, topics and stimuli that were used for the interviews in the LINDSEI. Table 1 provides an overview of the corpora.

The target structures were extracted semi-automatically[9] using *WordSmith Tools 5* (Scott 2008), followed by manual inspection and filtering of false positives. The analysis of the data consisted in a quantitative analysis of frequencies of occurrence and a qualitative study of lexical co-occurrence patterns (e.g. verbs, connectives, pragmatic markers, intensifying adverbs) and discourse functions.

3.2 Emphatic *Do*

Emphatic *do* is a lexico-grammatical means of information highlighting that commonly serves to emphasize the meaning of a following predicate (underlined in example 1).

(1) <A> So you want to become a teacher now. <\A>
<B> I **do** want to become a teacher yeah I always thought I wanted to teach English. But now I want to teach French. <\B> (LOCNEC)[10]

Emphatic *do* is discussed only briefly in the standard reference grammars of English (Quirk et al. 1985; Biber et al. 1999; Huddleston 2002) and there are only very few corpus-based studies that have examined this feature in detail (Nevalainen and Rissanen 1986; Luzón Marco 1998/99). Emphatic *do* usually carries nuclear stress and is one of the few options to explicitly highlight a predicate. Syntactic options like predicate fronting or *wh*-clefting are available to highlight a verb phrase, but are contextually much more restricted.

Table 2 shows that the frequential distribution of emphatic *do* varies across spoken and written registers. Emphatic *do* is clearly most frequently used in spoken language. In addition, a breakdown of the individual genre sections for the spoken register in the BNC shows that it is particularly frequent in highly argumentative contexts such as (parliamentary) debates, meetings, lectures, interviews, and discussions, where its frequency even rises to more than a thousand occurrences per million words.

There are two views as to whether emphatic *do* expresses both contrastive and non-contrastive emphasis or whether it exclusively has a contrastive function. Quirk et al. (1985) argue that it focuses on the operator [i.e. the predicate, MC] either for contrastive or emotive emphasis. Huddleston (2002: 97f.) states that it expresses emphatic polarity, emphasizing the positive or negative polarity of a clause. As an

[9]To retrieve instances of emphatic *do* I ran a search for the forms *do*, *does* and *did* followed by an infinitive, excluding instances of grammatically conditioned inversion after negatives as in *Not only did they…*, *Even slower did …*, and elliptical sentence forms, e.g. *Yes we do* or *They never did so*. For demonstrative clefts the search involved all instances of *that* and *this* followed by a form of *be* (‘*s, is, was*) and a *wh*-word (*what, when, why, where, how*).

[10]In the LOCNEC and the LINDSEI, turns marked with <A> </A> indicate the interviewers’ turns, while turns marked with <B> </B> mark the interviewees’ turns. The transcription guidelines for the LINDSEI can be retrieved from the following webpage: http://www.uclouvain.be/en-307849.html. Unfortunately, some of the transcription conventions used for the LOCNEC have not been updated to follow those of the LINDSEI. For example, overlapping speech in the LOCNEC is still indicated by means of square brackets instead of the explicit tag <overlap />.

Table 2 Frequencies of occurrence of emphatic *do* across registers in four corpora (per million words)

Register corpus	Speaking	Fiction	News	Academic writing
Longman Spoken and Written English (LSWE) *Corpus* (Biber et al. 1999: 433)	400	300	150	150
Bank of English (Luzon-Marco 1998/99: 91)	~545	~218	~125	–
Corpus of Contemporary American English (COCA, Davies 2008)	576	212	172	169
British National Corpus (BYU-BNC, Davies 2004)	734	320	173	223

Note that the frequency counts for these registers are not completely comparable across the four corpora. The count for the spoken register on the basis of the LSWE corpus is given for "conversation", and the count for fiction provided by Luzon-Marco on the basis of the *Bank of English* corpus is given for "books". The counts for the *Bank of English* corpus are approximations, thus marked by a tilde

emphatic positive it contrasts a positive with a corresponding negative proposition that has been expressed or implicated in the preceding discourse. As an emphatic positive it may also occur to indicate the strength of one's beliefs or feelings. Lambrecht (1994) analyses emphatic *do* as a conventionalized, grammaticalized way of expressing emphasis that involved a gradual loss of the presupposition in three steps: (1) the construction originally required the presupposition that the truth of a proposition was questioned in the immediately preceding discourse (fully contrastive contradiction), (2) the presupposition weakened so that a contradiction was merely suggested and left implicit (implicit contradiction), and finally, (3) the presupposition disappeared completely with *do* functioning as an intensifier like *really* (non-contrastive emphasis). Nevalainen and Rissanen's (1986) analysis compared 358 instances of emphatic *do* in the London-Lund Corpus (spoken British English) and the Lancaster-Oslo-Bergen Corpus (written British English). Their findings lend support to the view that emphatic *do* can indeed express non-contrastive emphasis. While 63 (18 %) and 101 instances (28 %) in the two corpora signaled either explicit opposition or implicit contrast respectively, a majority of 194 instances (54 %) expressed neither opposition nor contrast.

Biber et al. (1999: 433) note that "emphatic *do* usually marks a state of affairs in contrast to some other expected state of affairs which is by implication denied". This contrast can then be explicitly marked by contrastive connectives such as *but*, *however*, *nevertheless* or *(al)though*. Similarly, Luzón Marco (1998/99) argues that contrastive and emotive emphasis are not two different functions of emphatic *do*. She suggests that it always implies contrast, concession or correction with regard to something that has been previously said or is supposed to be known, expected or assumed. Moreover, it expresses simultaneously contrastive emphasis and involvement (i.e. carries an emotive effect).

Emphatic *do* is also characterized by distinct lexical co-occurrence patterns that partially reflect its discourse functions. Contrastive uses are often explicitly marked by contrastive connectives (*but*, *however*, *nevertheless, [al]though*) as in example (2)

and can also occur in conditional sentences introduced by *(even) if*. Contrastive and non-contrastive instances frequently co-occur with intensifying adverbs (*really, certainly, indeed*) and pragmatic markers (*well*, *yes/yeah*, *actually*, *you know*, *I mean*) as in (3). The types of predicates that are highlighted often include cognition verbs (e.g. *think, know, believe*) and emotive verbs (e.g. *like, hope, feel, need, want*).

(2) <B> er … you know I I'm I'm not a real big fan of the cinema **but I do think it's a good night out** and I'd much prefer to go to the cinema than to watch er a video <\B> (LOCNEC)

(3) <A> must be quite hard after you you've played something [to to to find yourself back <\A>
<B> [oh … it d= **well yeah it it definitely does take a while to come back down** <\B> (LOCNEC)

In the present chapter, the manual qualitative analysis of the discourse functions of emphatic *do* is based on its contextual use and distinguishes between three functions: (1) an intensifying, non-contrastive use (e.g. to indicate the strength of one's beliefs or feelings), and two types of contrastive uses, i.e. (2) explicit contrast/opposition (both referents are explicitly mentioned and contrasted) and (3) implicit contrast (the contrasted referent is not explicitly mentioned but contextually implied, i.e. presupposed, expected or assumed). These three functions are illustrated in example (4).

(4) <A> I mean you're independent here you can do whatever you want to and then [you go back home. <\A>
<B> [Yes … mhm. <\B>
<A> How do you feel about that. is it sometimes difficult I mean. you have to to I guess to tell your parents where you're going to if you leave and that kind of thing.<\A>
<B> Erm … yeah it it is it is quite. difficult to I suppose it's something I've got used to a lot more **I do I do like going home** it has it has advan= some advantages over being here and being here <\B>
<A> You don't have to cook <laughs> <\A>
<B> <begin_laughter> **Well I do have to do some cooking** <end_laughter> but <\B>
<A> Yeah I mean but <\A>
<B> Yeah not so much yeah [so <\B>
<A> [not so much <\A>
<B> Er … yeah I I like going home <X> **I do get on with my parents** and they're not they're not very . strict but erm **Yes I d= I do . feel yeah I do have to . tell them . where I'm going** and <\B> (LOCNEC)

The first and the third instance can be classified as cases of implicit contrast. The interviewer (A) does not explicitly deny that the interviewee (B) does not like going home to his/her parents place or does not get on well with them, but this is implicitly questioned ("How do you feel about that. is it sometimes difficult") and subsequently

Table 3 Frequencies of occurrence of emphatic *do* in the three corpora

Corpus	Absolute frequency	Normalized frequency per thousand turns
LINDSEI-F	8	1.45
LINDSEI-G	22	3.64
LOCNEC	99	11.74

clarified by B ("I do like going home", "I do get on with my parents"). The second instance is a case of explicit contrast. A mistakenly presupposes that B does not have to do any cooking when spending time with his/her parents ("You don't have to cook") which B explicitly corrects ("Well I do have to do some cooking"). Finally, the fourth instance exemplifies the intensifying, non-contrastive use. B responds to A's earlier turn ("you have to to I guess to tell your parents where you're going to if you leave and that kind of thing") and emphasizes the truth of this statement by confirming it ("I do . feel yeah I do have to . tell them . where I'm going").

They only previous corpus study of emphatic *do* in learner language (Callies 2009a), was based on a subset of the German component of the *International Corpus of Learner English* (ICLE, Granger et al. 2009), a corpus of L2 learners' argumentative writing. This study found a significant underrepresentation of emphatic *do* when compared to similar NS writing, differences in contextual use and lexical co-occurrence patterns and several apparently unmotivated uses. The much higher frequency of occurrence in speaking and the strong intonational component of emphatic *do* makes it necessary to replicate this study on the basis of spoken learner data. On account of the previous research findings and the fact that French and German lack a clear one-to-one equivalent that expresses the functions of emphatic *do* in English, emphatic *do* is hypothesized to be underrepresented in both spoken learner corpora when compared to native speaker usage. In French and German the functions of emphatic *do* are often fulfilled by modal particles like *doch* or *schon* (in German) and *si* (in French) (König et al. 1990; Lambrecht 1994: 72), both of which can be translated as 'but'.

The quantitative analysis of the frequency of occurrence of emphatic *do* in the three corpora (Table 3) confirms the hypothesis and shows that *do* as a marker of emphasis is significantly underrepresented in the two learner corpora when compared to the native speaker corpus (LOCNEC vs. LINDSEI-F: Log Likelihood (LL)= −57.4***; LOCNEC vs. LINDSEI-G: LL= −30.7***). In particular, with only eight occurrences in total, it is largely absent in the LINDSEI-F.

When analyzing the use of emphatic *do* by individual learners (Figs. 1 and 2) it is striking that it is only very few learners who use it. In particular, in the LINDSEI-G there is a fairly uneven distribution with two learners (ge024 and ge034) producing 40 % of all instances (9 out of 22) whereas the majority of learners do not use emphatic *do* at all.

The comparative analysis of the discourse functions of emphatic *do* does not reveal any major differences between the corpora: it is mostly used to express contrast by all three groups. Native speakers and German learners show a fairly balanced distribution of the three functions (see Fig. 3). More interesting, however, is the qualitative analysis of the most frequent collocates and verbs that co-occur with

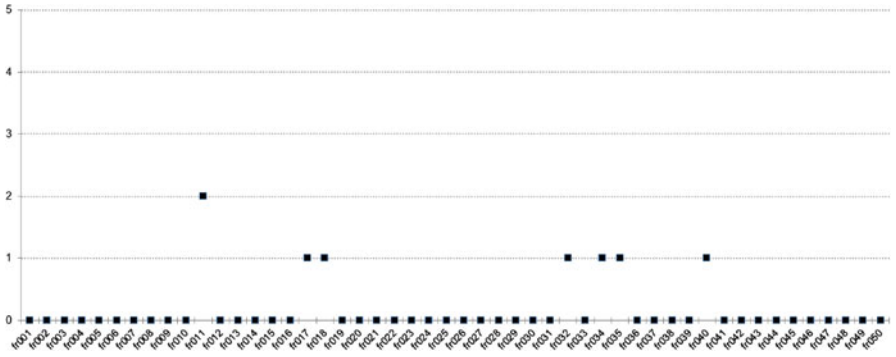


Fig. 1 Distribution of emphatic *do* in the LINDSEI-F

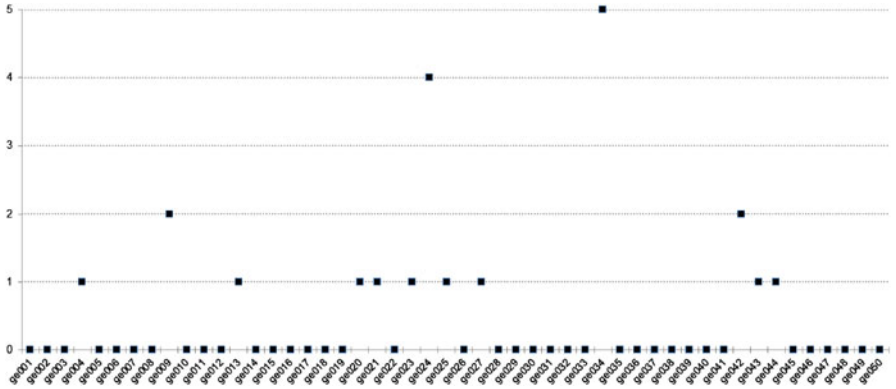


Fig. 2 Distribution of emphatic *do* in the LINDSEI-G

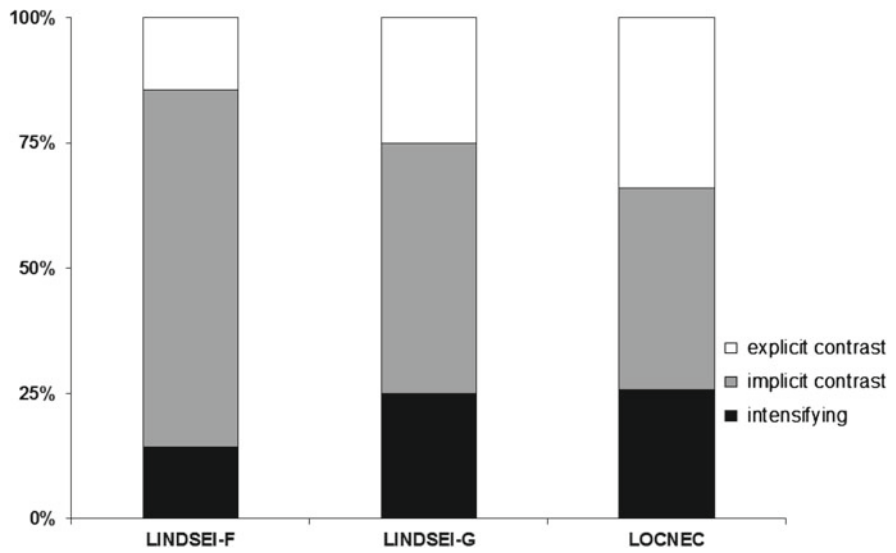


Fig. 3 Discourse functions of emphatic *do* in the three corpora

Table 4 Most frequent collocates and verbs occurring with emphatic *do* in the three corpora

Corpus	Collocate	N	All verbs (tokens)	All verbs (types)	TTR	Most freq. verbs (N≥3)	N
LINDSEI-F	*but*	4	8	6	0.75	–	–
LINDSEI-G	*but*	6	22	16	0.72	*have*	5
	yes, yeah	4				*like*	3
LOCNEC	*but*	24	99	48	0.48	*have (to)*	13
	yes, yeah	19				*like*	11
	I mean	8				*look*	8
	so	8				*get*	5
	actually	5				*think (about), work*	4 each
	well	4				*feel, go, know, miss*	3 each
	if	4					

emphatic *do*. It is striking that emphatic *do* is not only significantly underrepresented in the two learner corpora, but also that the few instances that can be found do not occur in their typical lexical co-occurrence patterns (contrastive connectives, intensifying adverbs, pragmatic markers, cognition verbs and emotive verbs, see Table 4).

How can the differences between native speakers and learners, and the differences between the two learner groups be explained? Considering recent findings that even advanced L2 learners have only a limited awareness of the appropriate use of lexical and syntactic focusing devices in formal and informal registers (Callies 2009a), the results are not surprising. Moreover, linguistic structures that are optional and subject to discourse-motivated preferences are assumed to be among the most difficult to acquire in advanced SLA (DeKeyser 2005). One explanation to account for the differences between the German and the French EFL learners could be that the German learners are benefitting from positive L1-transfer. In Standard German, the insertion of the semantically empty verb *tun* ('do') is obligatory in contexts where a lexical verb is topicalized and no other verb (auxiliary or modal) is present (Duden 1997: 726), see example (5a).

(5a) Tanzen **tut** Katja immer noch häufig.
Dance does Katja always still often.
'Katja does still dance often.'

Do-insertion is also frequently used in colloquial German and some German dialects to mark progressive aspect, see example (5b).

(5b) Sie **tut** gerade schreiben.
She does just now write
'She is writing just now.'

While another reason for why the Germans differ from the French learners may simply be differences in their general level of proficiency (see Sect. 3.3 for more explanation), further evidence for the influence of the learners' native language, possibly even in terms of a typological parameter, is suggested by the results of

preliminary analyses of other LINDSEI subcorpora: learners whose L1 is a (Germanic) language that has *do*-support seem to use emphatic *do* more often than learners from other L1 backgrounds (Callies in preparation).

The significantly lower frequency counts in the learner data may, however, also be an effect of the task and/or the interlocutor. It is a well-known fact that interlanguage variation is influenced by a number of external sociolinguistic factors that have to do with the situational context of language use, e.g. task, topic and interlocutor (see e.g. Ellis 2008: 141ff.). It is thus possible that L2 learners may be less inclined to disagree or object (hence experience much less need to make use of the linguistic means that convey contrastive emphasis) when they are interviewed by a native speaker who is of the opposite sex and not familiar to them rather than when interviewed by a same-sex non-native speaker who they know. Although variables such as the interviewer's mother tongue, gender and distance/closeness to the interviewee have been recorded in the LINDSEI, their influence cannot (yet) be assessed on a broad basis because of the small corpus size: strict control of all the relevant variables results in a very small database of sometimes only a handful of interviews.

3.3 Demonstrative Clefts

Cleft sentences are information packaging constructions that involve the splitting of a sentence into two clauses. They are pragmatically motivated and differ from their basic counterparts in that they serve to highlight a certain phrase or clause, the cleft constituent. The most common types are *it*-clefts and *wh*-clefts (also known as pseudo-clefts). There are also other types of cleft constructions one of which is the reverse *wh*-cleft, in which the order of *wh*- and cleft-clause is inverted. The vast majority of reverse *wh*-clefts feature the non-contrastive, non-focal deictic demonstratives *that* or *this* as the cleft constituent, see examples (6) and (7),[11] and therefore this type is also referred to as demonstrative cleft in the literature (Biber et al. 1999: 961; Calude 2008, 2009).

(6) <A> so you you did English and ling= and linguistics to: <\A>
<B> I did English and linguistics just because **that was what I was interested in** the the interest in going into film industry has only developed since I've been at university <\B> (LOCNEC)

(7) <A> so you had to cope with those kids <\A>
<B> I had to cope with those kids completely on my own with no back-up she said you know she w= she thought it was great having someone to help she said right you're gonna take half the kids …the worst half and you're going to teach them the same lesson as I'm teaching them here's the book **this is what I want you to teach them** go off and do it for a year <\B> (LOCNEC)

[11] Demonstrative clefts are given in bold print.

When compared to other types of cleft constructions, demonstrative clefts only rarely occur in written language but are clearly the most frequent variant in the spoken mode (Collins 1991: 178ff.; Oberlander and Delin 1996: 186; Weinert and Miller 1996: 176), occurring especially often in spontaneous spoken language, i.e. conversation (Biber et al. 1999: 961; Calude 2008: 86). Of the two demonstratives, *that* is much more frequent than *this* (Oberlander and Delin 1996: 189; Weinert and Miller 1996: 188; Biber et al. 1999: 962; Calude 2008: 79). Therefore, the majority of demonstrative clefts convey anaphoric deixis as in example (8),[12] but they can also express cataphoric deixis as in (9), function anaphorically and cataphorically simultaneously as in (10), or carry exophoric deixis, i.e. non-textual, extra-linguistic reference either in the form of shared world knowledge or physical/visual presence at the time of utterance, see example (11) (Calude 2008: 87ff.).

(8) <A> so what are you doing now as a major is it linguistics or is it <\A>
<B> <X> … I I thought I'd been accepted for Chinese and linguistics combined <\B>
<A> [mm <\A>
<B> [and **that's what they told me when I first . came here** but now they seem to think it's only linguistics <\B> (LOCNEC)

(9) <B> that we're living I mean I had my had my own flat and it's very difficult to: go from having your own flat and [<X> privacy to <\B>
<A> [and share a kitchen <\A>
<B> living in somewhere much smaller <\B>
<A> mhm <\A>
<B> but erm <\B>
<A> but I mean Graduate College is quite okay <\A>
<B> yeah I know **that's why I decided to pay a bit more** cos I thought sharing a kitchen and a bathroom with ten people <\B>
<A> yeah <\A>
<B> [I just couldn't <\B>
<A> [especially the bathroom <\A>
<B> yeah no I I really couldn't have faced that <\B> (LOCNEC)

(10) <A> and you don't live there and you you've never seen something like that before … but you you live in Sheffield <\A>
<B> yeah <\B>
<A> it's quite a big city isn't it <\A>
<B> it is quite big yeah **that's why I came here** cos I wanted to come to somewhere smaller <\B> (LOCNEC)

(11) <B> and she doesn't . it's not really a glamorous picture <\B>
<A> mhm <\A>
<B> or anything like that … erm the third one it looks like he's painted it again … erm … new hairstyle … smiling sat up … it makes her look more beautiful than she is <\B>

[12]The discourse segment(s) that the demonstrative *that* refers to are underlined.

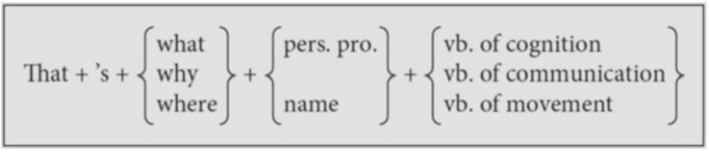


Fig. 4 The formulaic nature of demonstrative clefts (Reproduced from Calude 2009: 69)

<A> mhm <\A>
<B> <laughs> and in the fourth one she's telling all her friends of that's me **that's how I look** … things like that <\B> (LOCNEC)

In view of their relatively fixed structure, Calude (2009) argues that demonstrative clefts show characteristics of formulaic expressions, allowing only a narrow range of elements to occur in its structural "slots" (see Fig. 4). Prototypically, the demonstrative *that* occurs as the initial element. The copula *be* only occurs in simple present and simple past tense and is most commonly used in its contracted form'*s*. The copula is then most frequently followed by *what*, less frequently by *why*, *where, when* and *how* as *wh*-words in the cleft clause (Collins 1991: 28; Oberlander and Delin 1996: 187; Weinert and Miller 1996: 188). Moreover, demonstrative clefts have a distinct function in discourse as organizational and discourse-managing markers, and are typical of a specific register, i.e. conversation.[13]

Demonstrative clefts have multiple functions as to discourse organization and management. In particular, what sets them apart from other cleft types is their pointing function by means of the initial demonstrative pronoun (Weinert and Miller 1996: 188; Oberlander and Delin 1996: 189). They typically have extended text reference that spans over three or more turns prior to the cleft (Calude 2008: 79f.). With *that* as the initial element, demonstrative clefts have a strong anaphoric and attention-marking function (Weinert and Miller 1996: 192f.) and are typically used to underline or sum up previous discourse or to make reference to what has been said before (Collins 1991: 145f.; Weinert and Miller 1996: 192f.; Biber et al. 1999: 961ff.), while those introduced by *this* have a forward-pointing function and are also used as an attention marker (Weinert 1995).

Calude (2008: 99ff.; 108) suggests four discourse functions of demonstrative clefts. For the qualitative analysis of the discourse functions in the present case study, her taxonomy was adopted with slight modifications and two more functions (summarizing and projecting) were added. The six functions are exemplified in turn in (12)–(17).

(12) **quoting**: signaling direct speech, indirect speech or self-reported thought
<B> erm and I I wanted to come to university and do literature <XXX> interested<?> in that … and it was only really when I was looking through the prospectus sort of thinking well I don't just want to do literature what can I put [with it <\B>
<A> [mhm mhm <\A>

[13] One may add here that another feature that adds to their formulaicity is that in contrast to other types of clefts, demonstrative clefts are not reversible (Biber et al. 1999: 961).

<B> I sort of discovered the linguistics department and thought … ah yeah **that's what I've always wanted to do** <\B> (LOCNEC)

(13) **explaining**: giving a reason for a point previously made; explaining how two prior utterances relate to each other (linking function)
<B> yeah I think geography is interesting **that's why I study it** <laughs> </B> (LINDSEI-G)

(14) **evaluating**: giving opinions, evaluations or assessments; expressing agreement, disagreement or a neutral opinion with a previous comment
<B> yeah it wasn't much of a holiday really <\B>
<A> oh no <laughs> <\A>
<B> <laughs> <\B>
<A> it was just a a working holiday <X> <\A>
<B> a working holiday yeah <\B>
<A> just work <\A>
<B> well that's that's <X> **that's exactly what what our bosses were saying** exactly the same phrase said er you're here for no holiday you work you're here to work <\B> (LOCNEC)

(15) **highlighting**: singling out a preceding discourse element, thereby foregrounding it and giving it special prominence
<A> since you like the cinema so much <\A>
<B> [mhm <\B>
<A> [would you like to: to do: … later to work . in relation . to <\A>
<B> <X> what I'd like to do well I mean my degree is a primary school teaching degree **that's what I'm aiming to do at the[i:] end** <\B> (LOCNEC)

(16) **summarizing**: summing up a longer stretch of previous discourse
<B> he's changed the picture so that she's erm she looks considerably younger … erm obviously the hair's changed the face has changed <\B>
<A> [mhm <\A>
<B> [she's she's got a slight smile erm … and then now she's sort of erm just telling all her all of her friends sort of oh this is a picture of me isn't it lovely and doesn't it look so much like me but er \B>
<A> <laughs> <\A>
<B> **that's that's how I would say the story is going** she's er … she's she's eh this woman is actually quite vain <\B> (LOCNEC)

(17) **projecting**: drawing attention to a following stretch of discourse (only with cataphoric deixis)[14]

[14]This function is in line with Weinert's (1995) analysis of demonstrative clefts introduced by *this* as forward-pointing and attention marking devices. It is usually demonstrative clefts with cataphoric deixis that can be said to have a projecting function. In general, the development of cleft constructions in spoken English is strongly related to their discourse-pragmatic functions (see e.g. Callies 2012a for a study of the pragmaticalization of *wh*-clefts). For example, *wh*-clefts have been analysed as projector constructions that foreshadow upcoming discourse (e.g. Hopper and Thompson 2008) in which the *wh*-clause opens a projection span that draws the recipient's attention to the following highlighted constituent.

Table 5 Frequencies of occurrence of demonstrative clefts in the three corpora

Corpus	Absolute frequency	Normalized frequency per thousand turns
LINDSEI-F	27	4.72
LINDSEI-G	57	9.42
LOCNEC	73	8.65

> <B> so . it was a really nice (erm) .experience . I had and . what I found most (erm) impressive and I think **that's what everybody says when . he has seen Australia** is that . (erm) the distances are so huge . it's (er) that's really amazing so one day we drove for twelve hours and there was nothing . li<?> (eh) it's only dust . around us and so . but . it was really . yes impressive <laughs> </B> (LINDSEI-G)

Previous corpus-based studies of reversed *wh*-clefts in learner language are based on subsets of the ICLE. While Herriman and Boström Aronsson (2009) found an overrepresentation of reversed *wh*-clefts in the writing of Swedish EFL learners when compared to native speaker writing (93 vs. 62 instances), Callies (2009a) noted that native speakers used demonstrative clefts slightly more often when compared to the writing of German EFL learners (27 vs. 19 instances, but no statistically significant difference). Moreover, Callies observed that the learners showed little variation in how they used this construction: *what* was by far the most commonly used *wh*-word in reversed *wh*-clefts by both groups of writers, but the native speakers employed a broader range of *wh*-elements, while *how*, *where*, and *when* were completely absent from the learner data. They also strongly preferred *that* as a deictic marker and used the copula almost exclusively in its contracted form '*s*, which may indicate that the learners saw this as a formulaic expression. Non-deictic elements in reversed *wh*-clefts (e.g. Music *is what I like most*) were exclusively used by native speakers.

In view of these previous research findings and a contrastive analysis of such cleft types in French, German and English (see further below), the following two working hypotheses can be put forward for the case study: (1) demonstrative clefts are underrepresented in both learner corpora when compared to native speaker usage, and (2) advanced learner language is characterized by a narrower range of the formal and functional uses of this construction.

In fact, the quantitative analysis of the frequency of occurrence of demonstrative clefts in the three corpora (Table 5) shows that demonstrative clefts are significantly underrepresented in the LINDSEI-F when compared to the LOCNEC (LL= −7.7**), but that there is no statistically significant difference between the LINDSEI-G and the LOCNEC (LL= +0.23). Similar to emphatic *do*, the distribution of demonstrative clefts in the two learner corpora shows a high degree of inter-learner variability. In both corpora, it is merely a handful of learners who provide for almost 50 % of all tokens whereas half (or more) of the learners do not use this construction at all (see Figs. 5 and 6).

It is interesting to compare the two learner groups and the native speakers as to the relatively fixed structure of demonstrative clefts. Similar to the findings reported in the research literature, the deictic *that* and the *wh*-words *what* and *why* are the

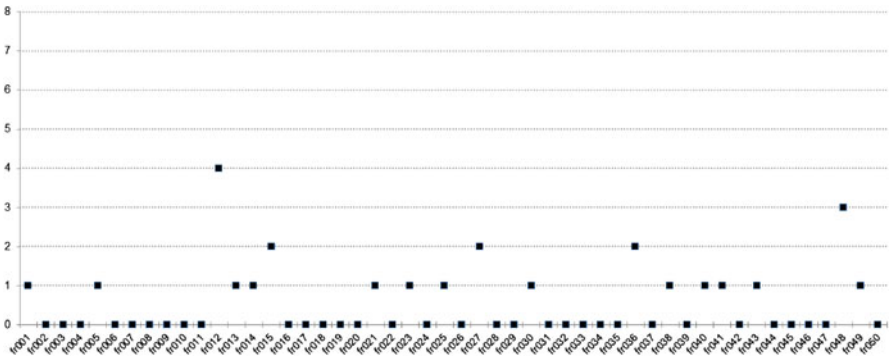


Fig. 5 Distribution of demonstrative clefts in the LINDSEI-F

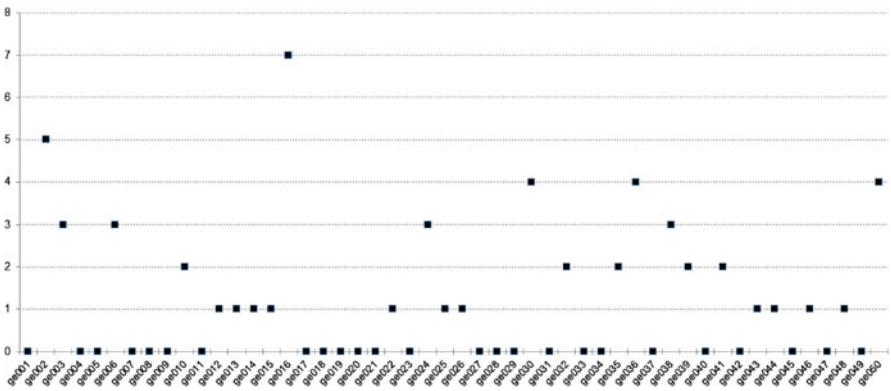


Fig. 6 Distribution of demonstrative clefts in the LINDSEI-G

most frequently occurring elements (Table 6). Demonstrative clefts primarily convey anaphoric deixis in all three corpora. While it is not surprising that the native speakers employ the full range of options that this construction allows in terms of the use of initial demonstratives, *wh*-words and deictic reference, it is indeed striking to see major differences between the two learner groups. The way in which the German learners use this construction very much resembles native speaker usage in terms of structural variation. By contrast, demonstrative clefts are not only significantly underrepresented in the spoken language of French learners, but the degree of formulaicity (or invariability) is also highest in the LINDSEI-F. A similar picture emerges when analyzing the discourse functions: the native speakers and the German learners use all six functions, but only four different ones occur in the LINDSEI-F (Fig. 7).

In this case, it is unlikely that the observed differences between native speakers and learners as well as the differences between the two learner groups are due to

Table 6 Use of demonstratives, *wh*-words and deictic reference in the three corpora

	LINDSEI-F	LINDSEI-G	LOCNEC
demonstrative			
that	26 (96 %)	44 (77 %)	67 (92 %)
this	1 (4 %)	13 (23 %)	6 (8 %)
***wh*-word**			
what	12 (44 %)	27 (47 %)	30 (41 %)
why	14 (52 %)	17 (30 %)	15 (21 %)
where	0	1 (2 %)	11 (15 %)
when	0	4 (7 %)	6 (8 %)
how	1 (4 %)	8 (14 %)	11 (15 %)
deixis			
anaphoric	26 (96 %)	42 (74 %)	57 (78)
cataphoric	0	5 (9 %)	4 (5 %)
both	1 (4 %)	4 (7 %)	6 (8 %)
exophoric	0	6 (11 %)	6 (8 %)

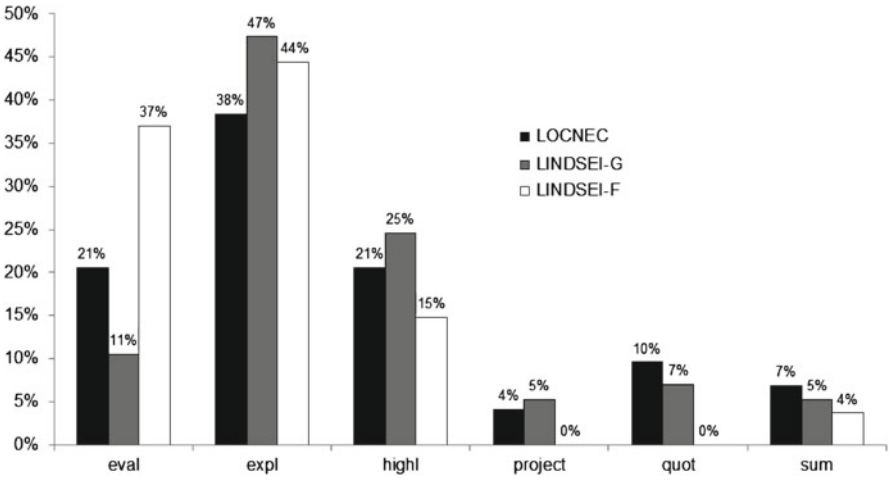


Fig. 7 Functions of demonstrative clefts in the three corpora

cross-linguistic influence, at least as far as the German learners are concerned. Although German does have cleft constructions, they are dispreferred options to convey focus and have only peripheral status because of the less restricted use of topicalization (see e.g. Weinert 1995 and Callies 2009a for discussion). Weinert (1995) compared *wh*- and reversed *wh*-clefts in English and German, contrasting their discourse functions with those of preposing/topicalization based on corpora of structured dialogue and conversation. Her findings showed that in contrast to speakers of English, Germans used only very few reversed *wh*-clefts because reversed clefts are extremely rare in German, structurally and functionally more restricted, and

often combine with focus or modal particles to supplement their focus, and thus create an even stronger focus than their English counterparts (Weinert 1995: 355). Moreover, topicalization in German is less restricted and not as strongly associated with contrastiveness as preposing in English. On account of this, demonstrative clefts should be expected to be underrepresented in LINDSEI-G, but this is clearly not the case.

Transfer in the form of underproduction may be an explanatory factor in the case of the French learners. French does have two types of clefts, the *c'est*-cleft, which often carries a contrastive and even exclusive value, and the *il y a*-cleft, which has presentational character, but in contrast to German and English, French does not have reversed *wh*-clefts because it does not allow pre-verbal focus (Lambrecht 2001: 492; Miller 2006: 185). The absence of this cleft type in the L1 may thus at least partially explain the observed underrepresentation.

It seems more likely that differences in general language proficiency may help explain the differences between the two learner groups. The assessment of language proficiency is a notoriously difficult (and also frequently neglected and underestimated) challenge in SLA and Learner Corpus Research (LCR).[15] In LCR, learners' proficiency level has been a fuzzy variable in that it has often been assessed globally by means of external criteria, most typically learner-centered criteria (e.g. Carlsen 2012). There are several problems connected with this practice (Thomas 1994, 2006). As a consequence, in some corpora learners' proficiency level varies considerably, both across and within subcorpora. This is also true for the LINDSEI, in the compilation of which proficiency was assessed globally on account of institutional status with learners being described as "university undergraduates in English (usually in their third or fourth year)" (Gilquin et al. 2010: 10). The proficiency level of learners who are represented in the LINDSEI in fact ranges from higher intermediate to advanced. While some LINDSEI subcorpora predominantly seem to include learners from either the C1 or C2 proficiency levels of the *Common European Framework of Reference for Languages*, e.g. Dutch, Swedish or German learners, others rather seem to include learners from higher intermediate (or lower) proficiency levels, e.g. those whose L1 is Italian, Spanish or French (Gilquin et al. 2010: 10f.). The LINDSEI handbook also provides information about two variables that have often been used to help operationalize proficiency: the amount of formal classroom instruction in the foreign language and time spent in a country where the target language is spoken. Comparing these two variables, it turns out that the number of years spent learning English in school and university is 4.6 and 3.8 on average in LINDSEI-F, while the German learners spent 8.6 and 3.6 years learning English. Thus, the Germans spent significantly more time learning English in school (they are also on average 2 years older than the French: 24.6 vs. 22.1 years). More important, though, is the difference in the time spent abroad: on average, speakers in LINDSEI-F spent only 1.9 months in an English-speaking country, while those in LINDSEI-G spent 9.3 months abroad (Gilquin et al. 2010: 40f.).

[15] It is not possible to go into detail here, but see Callies, Zaytseva & Present-Thomas (2013) for further discussion as to the operationalization and assessment of (advanced) proficiency in LCR.

4 Conclusion

This chapter has provided a critical assessment of research on pragmatics in the context of SLA showing that in mainstream ILP, the significance of L2 pragmatic knowledge beyond the domain of speech acts has been neglected to date. I have argued that the field of inquiry in ILP needs to be extended because pragmatic knowledge in an L2 includes more than sociopragmatic and pragmalinguistic abilities for understanding and performing speech acts. I have proposed a wider definition of L2 pragmatic knowledge and have highlighted the crucial role of learner corpora in the expansion of the narrow research agenda of ILP. Two case studies of EFL learners' use of emphatic *do* and demonstrative clefts have exemplified how spoken learner corpora enable researchers to study a much broader range of different pragmatic phenomena and can help overcome several problems and limitations posed by the dominance of data elicitation techniques in ILP to date.

The case studies have demonstrated the usefulness of corpora to abstract away from individual learners to identify a corpus-based description of a specific learner group while also providing insights into inter-learner variability. The individual differences found for both the French and the German EFL learners have important implications for learner corpus analysis and compilation in that they confirm that global proficiency measures based on external criteria alone are not reliable indicators of proficiency. However, in a substantial part of LCR to date individual differences often go unnoticed or tend to be disregarded and are thus not reported in favour of (possibly skewed) average frequency counts. Mukherjee (2009) is one study where the issue of inter-learner variability is explicitly addressed. Observing an extremely uneven distribution of the pragmatic marker *you know* in the LINDSEI-G, Mukherjee concludes that "the fiction of homogeneity that is often associated with the compilation of a learner corpus according to well-defined standards and design criteria may run counter to the wide range of differing individual levels of competence in the corpus" (2009: 216).

References

Aijmer, Karin. 2004. Pragmatic markers in spoken interlanguage. *Nordic Journal of English Studies* 3(1): 173–190.

Aijmer, Karin. 2009. 'So er I just sort I dunno I think it's just because…': A corpus study of *I don't know* and *dunno* in learners' spoken English. In *Corpora: Pragmatics and discourse*, ed. Andreas H. Jucker, Daniel Schreier, and Marrianne Hundt, 151–168. Amsterdam: Rodopi.

Aijmer, Karin. 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.

Aijmer, Karin, and Christoph Rühlemann. forthcoming. *Corpus pragmatics. Exploring speaker meaning in computerized corpora*. Cambridge: Cambridge University Press.

Andersen, Gisle. 2011. Corpus-based pragmatics I: Qualitative studies. In *Foundations of pragmatics*, Handbooks of pragmatics, vol. 1, ed. Wolfram Bublitz and Neil R. Norrick, 587–627. Berlin: DeGruyter Mouton.

Archibald, Alisdair N. 1994. *The acquisition of discourse proficiency*. Frankfurt/Main: Peter Lang.

Bachman, Lyle. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bardovi-Harlig, Kathleen. 2010. Exploring the pragmatics of interlanguage pragmatics: Definition by design. In *Pragmatics across languages and cultures*, Handbooks of pragmatics, vol. 7, ed. Anna Trosborg, 219–259. Berlin: DeGruyter Mouton.

Barron, Anne. 2003. *Acquisition in interlanguage pragmatics*. Amsterdam: John Benjamins.

Belz, Julie A., and Nina Vyatkina. 2005. Learner corpus analysis and the development of L2 pragmatic competence in networked intercultural language study: The case of German modal particles. *The Canadian Modern Language Review/La Revue canadienne des langues vivantes* 62(1): 17–48.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Boström Aronsson, Mia. 2003. On clefts and information structure in Swedish EFL writing. In *Extending the scope of corpus-based research. New applications, new challenges*, ed. Granger Sylviane and Petch-Tyson Stephanie, 197–210. Amsterdam: Rodopi.

Brand, Christiane, and Sandra Götz. 2011. Fluency versus accuracy in advanced spoken learner language: A multi-method approach. *International Journal of Corpus Linguistics* 16(2): 255–275.

Bußmann, Hadumod. 1996. *Routledge dictionary of language and linguistics*. London: Routledge.

Callies, Marcus. 2008a. Argument realization and information packaging in tough-movement constructions – A learner-corpus-based investigation. In *Morphosyntactic issues in second language acquisition studies*, ed. Danuta Gabrys-Barker, 29–46. Clevedon: Multilingual Matters.

Callies, Marcus. 2008b. Easy to understand but difficult to use? Raising constructions and information packaging in the advanced learner variety. In *Linking contrastive and learner corpus research*, ed. Gaëtanelle Gilquin, Maria Belen Diez Bedmar, and Szilvia Papp, 201–226. Amsterdam: Rodopi.

Callies, Marcus. 2009a. *Information highlighting in advanced learner English. The syntax-pragmatics interface in second language acquisition*. Amsterdam: John Benjamins.

Callies, Marcus. 2009b. 'What is even more alarming is…' – A contrastive learner-corpus study of *what*-clefts in advanced German and Polish L2 writing. In *On language structure, acquisition and teaching. Studies in honour of Janusz Arabski on the occasion of his 70th Birthday*, ed. Maria Wysocka, 283–292. Katowice: Wydawnictwo Uniwersyetu Slaskiego.

Callies, Marcus. 2012a. The grammaticalization and pragmaticalization of cleft constructions in present-day English. In *English corpus linguistics: Looking back, moving forward*, ed. Sebastian Hoffmann, Paul Rayson, and Geoffrey Leech, 5–21. Amsterdam: Rodopi.

Callies, Marcus. 2012b. Discourse completion task. In *Theories and methods in linguistics*, Wörterbücher zur Sprach- und Kommunikationswissenschaft [WSK] online, vol. 11, ed. Bernd Kortmann. Berlin: De Gruyter Mouton. doi: 10.1515/wsk.35.0.discoursecompletiontask

Callies, Marcus. in preparation. Emphatic *do* in advanced learner English. A contrastive interlanguage analysis of spoken and written corpora.

Callies, Marcus, Ekaterina Zaytseva, and Rebecca Present-Thomas. 2013. Writing assessment in higher education: Making the framework work. *Dutch Journal of Applied Linguistics* 2(1): 1–15.

Calude, Andreea. 2008. Demonstrative clefts and double cleft constructions in spoken English. *Studia Linguistica* 62(1): 78–118.

Calude, Andreea. 2009. Formulaic tendencies of demonstrative clefts in spoken English. In *Formulaic language, Vol. 1: Distribution and historical change*, ed. Roberta Corrigan, Edith A. Moravcsik, Hamid Quali, and Kathleen M. Wheatley, 55–76. Amsterdam: John Benjamins.

Canale, Michael. 1983. From communicative competence to communicative language pedagogy. In *Language and communication*, ed. Jack C. Richards and Richard W. Schmidt, 2–27. London/New York: Longman.

Canale, Michael, and Merril Swain. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1(1): 1–47.

Carlsen, Cecilie. 2012. Proficiency level – A fuzzy variable in computer learner corpora. *Applied Linguistics* 33(2): 161–183.

Collins, Peter C. 1991. *Cleft and pseudo-cleft constructions in English*. London: Routledge.

Crystal, David. 2003. *A dictionary of linguistics and phonetics*, 5th ed. Malden: Blackwell.

Davies, Mark. 2004. BYU-BNC (based on the British National Corpus from Oxford University Press). http://corpus.byu.edu/bnc/. Accessed 18 Dec 2012.
Davies, Mark. 2008. The corpus of contemporary American English: 450 million words, 1990-present. http://corpus.byu.edu/coca/. Accessed 18 Dec 2012.
DeKeyser, Robert M. 2005. What makes learning second language grammar difficult? A review of issues. *Language Learning* 55: 1–25.
Deppermann, Arnulf. 2011. Pragmatics and grammar. In *Foundations of pragmatics*, Handbooks of pragmatics, vol. 1, ed. Wolfram Bublitz and Neil R. Norrick, 425–460. Berlin: DeGruyter Mouton.
Dippold, Doris. 2009. Face and self-presentation in spoken L2 discourse: Renewing the research agenda in interlanguage pragmatics. *Intercultural Pragmatics* 6(1): 1–28.
Duden. 1997. Richtiges und gutes Deutsch: Wörterbuch der sprachlichen Zweifelsfälle, (Duden vol. 9). Mannheim: Dudenverlag.
Ellis, Rod. 2008. *The study of second language acquisition*, 2nd ed. Oxford: Oxford University Press.
Felder, Ekkehard, Marcus Müller, and Friedemann Vogel (eds.). 2011. *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen*. Berlin: DeGruyter.
Gass, Susan M., and Larry Selinker. 2008. *Second language acquisition. An introductory course*, 2nd ed. Mahwah: Erlbaum.
Gilquin, Gaëtanelle. 2008. Hesitation markers among EFL learners: Pragmatic deficiency or difference? In *Pragmatics and corpus linguistics. A mutualistic entente*, ed. Jesus Romero Trillo, 119–149. Berlin: Mouton de Gruyter.
Gilquin, Gaëtanelle, De.Cock. Sylvie, and Granger Sylviane. 2010. *The Louvain international database of spoken English interlanguage. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.
Golato, Andrea. 2003. Studying compliment responses: A comparison of DCTs and recordings of naturally occurring talk. *Applied Linguistics* 24(1): 90–121.
Götz, Sandra. 2007. Performanzphänomene in gesprochenem Lernerenglisch: eine korpusbasierte Pilotstudie. *Zeitschrift für Fremdsprachenforschung* 18(1): 67–84.
Götz, Sandra. 2013. *Fluency in native and nonnative English speech*. Amsterdam: John Benjamins.
Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. The international corpus of learner English. Version 2. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.
Herriman, Jennifer, and Mia Boström Aronsson. 2009. Themes in Swedish advanced learners' writing in English. In *Corpora and language teaching*, ed. Karin Aijmer, 101–120. Amsterdam: John Benjamins.
Hopper, Paul, and Sandra Thompson. 2008. Projectability and clause combining in interaction. In *Crosslinguistic studies of clause combining. The multifunctionality of conjunctions*, ed. Ritva Laury, 99–123. Amsterdam: John Benjamins.
Huddleston, Rodney. 2002. The verb. In *The Cambridge grammar of the English language*, ed. Rodney Huddleston and Geoffrey K. Pullum, 71–212. Cambridge: Cambridge University Press.
Hymes, Dell H. 1972. On communicative competence. In *Sociolinguistics. Selected readings*, ed. John B. Pride and Janet Holmes, 269–293. Harmondsworth: Penguin.
Kasper, Gabriele. 2008. Data collection in pragmatics research. In *Culture, communication and politeness theory*, 2nd ed, ed. Helen Spencer-Oatey, 279–303. London/New York: Continuum.
Kasper, Gabriele. 2010. Interlanguage pragmatics. In *Variation and change. Pragmatic perspectives*, Handbook of pragmatics highlights, vol. 6, ed. Mirjam Fried, Jan-Ola Östman, and Jef Verschueren, 141–154. Amsterdam: John Benjamins.
Kasper, Gabriele, and Merete Dahl. 1991. Research methods in interlanguage pragmatics. *Studies in Second Language Acquisition* 13: 215–247.
Kasper, Gabriele, and Kenneth R. Rose. 1999. Pragmatics and SLA. *Annual Review of Applied Linguistics* 19: 81–104.
Kasper, Gabriele, and Kenneth R. Rose. 2002. *Pragmatic development in a second language*. Oxford: Blackwell.

König, Ekkehard, Detlef Stark, and Susanne Requardt. 1990. *Adverbien und Partikeln: ein deutsch-englisches Wörterbuch*. Heidelberg: Groos.
Lambrecht, Knud. 1994. *Information structure and sentence form*. Cambridge: Cambridge University Press.
Lambrecht, Knud. 2001. A framework for the analysis of cleft constructions. *Linguistics* 39(3): 463–516.
Leech, Geoffrey. 1983. *Principles of pragmatics*. London: Longman.
LoCastro, Virginia. 2011. Second language pragmatics. In *Handbook of research in second language teaching and learning*, vol. 2, ed. Eli Hinkel, 319–344. New York: Routledge.
Luzón Marco, Maria Jose. 1998/99. The discursive function of *do*-support in positive clauses. *Revista Española de Lingüística Aplicada* 13: 87–102.
Miller, Jim. 2006. Focus in the languages of Europe. In *Pragmatic organization of discourse in the languages of Europe*, ed. Bernini Giuliano and Marcia L. Schwartz, 121–214. Berlin/New York: de Gruyter.
Mukherjee, Joybrato. 2009. The grammar of conversation in advanced spoken learner English: Learner corpus data and language-pedagogical implications. In *Corpora and language teaching*, ed. Karin Aijmer, 203–230. Amsterdam: John Benjamins.
Müller, Simone. 2004. 'Well you know that type of person': Functions of *well* in the speech of American and German students. *Journal of Pragmatics* 36(6): 1157–1182.
Müller, Simone. 2005. *Discourse markers in native and non-native English discourse*. Amsterdam: John Benjamins.
Nevalainen, Terttu, and Matti Rissanen. 1986. Do you support the *do*-support? Emphatic and non-emphatic *do* in affirmative statements in present-day spoken English. In *Papers from the third Scandinavian symposium on syntactic variation*, ed. Sven Jacobson, 35–50. Stockholm: Almquist and Wiksell.
Oberlander, Jon, and Judy Delin. 1996. The function and interpretation of reverse *wh*-clefts in spoken discourse. *Language and Speech* 39(2–3): 185–227.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
Ramírez Verdugo, Dolores, and Jesus Romero Trillo. 2005. The pragmatic function of intonation in L2 discourse: English tag questions used by Spanish speakers. *Intercultural Pragmatics* 2(2): 151–168.
Romero Trillo, Jesus (ed.). 2008. *Pragmatics and corpus linguistics. A mutualistic entente*. Berlin: Mouton de Gruyter.
Rühlemann, Christoph. 2011. Corpus-based pragmatics II: Quantitative studies. In *Foundations of pragmatics*, Handbooks of pragmatics, vol. 1, ed. Wolfram Bublitz and Neil R. Norrick, 629–656. Berlin: DeGruyter Mouton.
Sasaki, Miyuki. 1998. Investigating EFL students' production of speech acts: A comparison of production questionnaires and role plays. *Journal of Pragmatics* 30(4): 457–484.
Scott, Mike. 2008. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.
Taguchi, Naoko. 2009. Pragmatic competence in Japanese as a second language: An introduction. In *Pragmatic competence*, ed. Naoko Taguchi, 1–18. Berlin: Mouton de Gruyter.
Thomas, Margaret. 1994. Assessment of L2 proficiency in second language acquisition research. *Language Learning* 44(2): 307–336.
Thomas, Margaret. 2006. Research synthesis and historiography: The case of assessment of second language proficiency. In *Synthesizing research on language learning and teaching*, ed. John M. Norris and Lourdes Ortega, 279–298. Amsterdam: John Benjamins.
Weinert, Regina. 1995. Focusing constructions in spoken language. Clefts, Y-movement, thematization and deixis in English and German. *Linguistische Berichte* 159: 341–369.
Weinert, Regina, and Jim Miller. 1996. Cleft constructions in spoken language. *Journal of Pragmatics* 25: 173–206.
Yuan, Yi. 2001. An inquiry into empirical pragmatics data-gathering methods: Written DCTs, oral DCTs, field notes, and natural conversations. *Journal of Pragmatics* 33(2): 271–292.

Corpus Linguistics and Conversation Analysis at the Interface: Theoretical Perspectives, Practical Outcomes

Steve Walsh

1 Introduction

In this chapter, I offer a perspective on the merits and possible drawbacks of a combined corpus linguistics (CL) and conversation analysis (CA) methodology. The first part of the chapter provides a theoretical perspective of each methodology, considering their respective epistemological and ontological origins and traditions, before moving on to discuss how they might – in spite of their very different research positions – be used together, in combination. The broad argument for combining CL with CA is that CL is unable to account for some of the features of spoken interaction which occur at the levels of utterance and turn and largely ignores context, while CA is unable to identify linguistic patterns across larger corpora, limiting itself instead to detailed descriptions of small quantities of data. Each methodology, then, has its strengths and weaknesses – in combination, they have the potential to offer enhanced descriptions of spoken interaction. Using a combined CL and CA approach (henceforth, CLCA), cumulatively gives a more 'up-close' description of spoken interactions than that offered by using either one on its own. A CLCA analysis provides powerful insights into the ways in which interactants establish understandings and observe how words, utterances and text combine in the co-construction of meaning.

S. Walsh (✉)
School of Education, Newcastle University, King George VI Building,
Newcastle Upon Tyne NE17RU, UK
e-mail: steve.walsh@ncl.ac.uk

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_3,

2 Corpus Linguistics: Epistemology and Ontology

One of the key methodological underpinnings common to both CL and CA is that they make use of corpora; their point of departure is always the building of a corpus. A corpus is a collection of texts that is stored on a computer; texts may be spoken or written, but for the purposes of this chapter, we are concerned only with spoken texts. Texts are examples of spoken discourse which have been recorded and transcribed and which include conversations, phone calls, university seminars, debates, etc. Essentially, any spoken discourse, produced in context and for a genuine purpose, can be regarded as a text. A corpus therefore is a collection of real language that people use in all types of situations.

The emergence of corpus linguistics goes back to the 1970s and 1980s when computers were being developed that were powerful enough to store and search large databases of stored texts. At this time, the main use of corpora was in the production of dictionaries – today, all major publishers producing dictionaries use corpora. The main advantage is clear: rather than relying on intuition, lexicographers were able to search very large databases to find examples of real language in use. The use of invented – or idealised – examples became a thing of the past. Today, computers can be used to search up to a billion words at any one time to identify examples and see how language is really used. Perhaps the most revolutionary work in the area of dictionary production at this time was the *Collins Birmingham University International Language Database* (COBUILD) project. This was set up at the University of Birmingham in 1980 under the direction of John Sinclair. From this database, 16 dictionaries have been produced to date, most notably the *Collins COBUILD English Language Dictionary* (1987, 2nd edition 1995, 3rd edition 2001, 4th edition 2003) and the *Collins Cobuild Grammar Patterns* series (1996; 1998).

While the main focus of the early CL work was lexicography, these studies also led to a focus of attention on grammar, and, in particular, heightened understandings of the relationship between words and grammar: *lexico-grammatical* features of language. What this focus of attention did was to direct attention towards the importance of words and chunks of words in grammatical relationships, rather than regarding grammar as the most important language system. Vocabulary suddenly became at least as important as grammar in our emerging understandings of language systems. Many grammatical relationships could also be linked much more to words.

Today, most grammar books of English are corpus-informed, a process which has many advantages. First, like lexicographers, grammarians no longer have to rely on their intuitions – examples can be derived from a corpus; more importantly, the 'rules' of grammar can also be derived from the corpus since patterns can be more easily established by looking at numerous examples. A second advantage is that it is now much easier to identify relationships across different text types and study how, for example, spoken grammars are different to written ones (Carter and McCarthy 2006), or how certain language structures are more common than others in some text-types (e.g. newspaper articles). Related to this, is the point that

corpus-based grammars can now make clearer claims about regional varieties such as differences between American and British or Irish English. Corpora also allow comparisons to be made over time, allowing us to comment on how certain grammatical features are more or less widespread; for example, *can I* is more common today in most contexts than *may I*.

If one of the main concerns of linguistics and, to some extent applied linguistics, is the study of patterns of use in language, CL has made the whole process much easier and faster. It is now possible to compare huge databases and make reliable claims about how language is *actually* used in context, rather than prescribing how it *should* be used. From a pedagogic perspective, the advantages of this are obvious and too numerous to mention here. CL, then, gives us, at a glance, an overview of how a particular word or grammatical structure is used across a range of contexts and text-types.

When CL was in its infancy and being used mainly in the production of dictionaries, the main focus was on building large corpora: the bigger the better. The reason for this is both to ensure that as many examples as possible are available, and also to ensure that rarer words, or words which are less commonly used, could also be studied with the same degree of reliability. Essentially, the larger the sample, the more accurate are the claims which can be made about a particular feature. The trend of aiming for large or very large corpora has, to a large degree, come to an end. There has been a shift in recent years towards using smaller, more context-specific and locally derived corpora in order to highlight specific examples of language use in spheres such as business, medicine, science, classrooms or everyday conversations. These more specific corpora may be used, for example, by translators and materials designers. For a translator working on a medical document, a small corpus (for example five lakh to one million words) of medical articles, is more useful than a general corpus of ten million words. Equally, an author of business text books could find out a lot more from one million words of business language than from a much larger general corpus. Smaller still are corpora used for research; it would, for example, be quite feasible to conduct a small-scale research project using a corpus of 100,000 words, providing that it was designed with a specific context in mind.

In this chapter, then, CL is presented as a *methodological tool* which can be used to investigate, for example, small group interactions recorded in higher education. Using CL as a tool allows us to automatically search a large dataset, something which would have been impractical manually. However, while CL allows us to count frequencies and find key words in micro-seconds, thus revealing patterns that we could not otherwise find, it does not allow us to explain the dynamics of these interactions. One of the main reasons for using a combined CLCA methodology is that CA does allow us to reveal in some detail which is actually 'happening' in interactions. We return to this below.

To return to the arguments made above about the importance of CL in the study of language *use*, it is probably fair to say that CL is being increasingly applied to contexts and domains outside of the study of language itself where the focus is on the *use* of language in a given context. Such contexts include courtrooms and forensic linguistics (Cotterill 2010), the workplace, educational contexts (O'Keeffe and Farr 2003; Walsh and O'Keeffe 2007), political discourse (Ädel 2010), the media (O'Keeffe 2006), among other areas. In all of these cases, CL is used as a tool and

another approach, such as CA, discourse analysis or pragmatics, is drawn on as a framework. Under this 'applied' view of CL, language in use is the prime focus and the research endeavour is to uncover, using a complementary methodology, the broader interactional context in order to gain understandings of 'what is really happening'. The interest lies less in the linguistic features *per se* and more in what is being accomplished though their use. So, for example, we might be interested in studying the ways in which discourse markers are used in an educational context (cf. Yang 2013), or the use of modal verbs in transactional encounters. In both instances, the corpus and its description is not an end in itself, but a means to finding out more about a broader research question.

One of the consequences of the recent shift towards smaller corpora (O'Keeffe et al. 2007) is that there has been a corresponding movement towards combining CL with other methodologies, particularly when the focus is on spoken discourse. As McCarthy and O'Keeffe (2010) point out, in the early days of CL, the aim was to have very large written corpora to serve the needs of lexicographers, whose focus was obviously on semantic and lexical patterning rather than on discourse context. As a result, large corpora were lexically rich but contextually poor. That is, when a researcher looks at a lexical item in a mostly written corpus of 100 million words or more, it is detached from its context. However, when the researcher records, transcribes, annotates and builds a small contextualised spoken corpus, a different landscape of possibilities opens up in areas beyond lexis to areas of use (especially issues of pragmatics, interaction and discourse). We can say, then, that there has been some 'meeting of the ways' between CL and CA approaches: both CL and CA highlight the importance of context, albeit in different ways, and CL has recently started to recognise the value of smaller, context-specific corpora.

Before considering in more detail the relative merits (and shortcomings) of a combined CLCA methodology, I offer an overview of the origins and research traditions of CA.

3 Conversation Analysis: Epistemology and Ontology

The origins of conversation analysis (CA) lie in sociology, not linguistics or applied linguistics. The original interest arose out of a perceived need to study ordinary conversation as social action; CA's underlying philosophy is that social contexts are not static but are constantly being formed by participants through their use of language and the ways in which turn-taking, openings and closures, sequencing of acts, and so on are locally managed (Sacks et al. 1974). Interaction is examined in relation to meaning and context; the way in which actions are sequenced is central to the process. In the words of Heritage (1997, p. 162):

> In fact, CA embodies a theory which argues that sequences of actions are a major part of what we mean by context, that the meaning of an action is heavily shaped by the sequence of previous actions from which it emerges, and that social context is a dynamically created thing that is expressed in and through the sequential organisation of interaction.

According to this view, interaction is *context-shaped* and *context-renewing*; that is, one contribution, or 'turn-at-talk' is dependent on a previous one and subsequent contributions create a new context for later actions. Context is "both a project and a product of the participants' actions" (Heritage 1997, p. 163). According to Sidnell (2010, p. 1), CA aims to "describe, analyse, and understand talk as a basic and constitutive feature of human social life". In its early days, CA focused on describing conversations between friends; only later did it look at institutional settings (see below).

According to Seedhouse (2005, pp. 166–67), the basic principles which CA adopts are:

> There is order at all points in any interaction: talk- in-interaction is systematically organised, deeply ordered and methodic.
>
> Contributions to interaction are context-shaped and context-renewing (see above).
>
> No order of detail can be dismissed as disorderly, accidental, or irrelevant (cf. Heritage 1984): CA has a detailed transcription system, and a highly empirical orientation.
>
> The analysis is bottom-up and data driven: researchers should approach the data without prejudice or bias and adopt CA's principle of 'unmotivated looking'.

One of the main concerns of CA is turn-taking in talk-in-interaction (Hutchby and Wooffitt 2008). Adjacency pairs, repair, and preference are the other main foci of attention. In CA, the basic unit of analysis is a Turn Constructional Unit (TCU), approximately the same as a single utterance which carries meaning. A single turn may comprise several TCUs and any single TCU may indicate the end of a turn, marked by a transition relevance place (TRP), at which point any other speaker may take the floor, or the original speaker may retain his or her turn. This basic turn-taking mechanism underpins all CA research, which adopts the 'next turn proof procedure' (REF) as an indicator of the robustness of the method. Essentially, any one turn-at-talk can be related to any other turn in a logical and systematic way so that analysts view the interaction in the same way as participants.

Apart from turn-taking, another area of interest for CA is adjacency pairs, based on the premise that much human communication proceeds through paired utterances; greeting/greeting, question/response, invitation/acceptance, etc. An understanding of adjacency pairs entails a realisation that there are preferred and dispreferred second pair-parts. So, for example, the preferred second-pair part of invitation is acceptance. Space precludes a fuller treatment of adjacency pairs and preference structure, but see, for example Schegloff (2007) and Hutchby and Wooffitt (2008).

The final system which is of concern to CA is repair, defined as "the treatment of trouble occurring in interactive language use" (Seedhouse 2004, p. 34). Repair is essential for intersubjectivity, or mutual meaning-making, and interactants constantly make use of a range of repair strategies in order to understand and be understood. There is no limit to what can be repaired in spoken interaction, making it a key method for interactants to achieve mutual understanding.

Although the original focus of CA was naturally occurring conversation, it is perhaps in specific institutional settings, where the goals and actions of participants are clearly determined, that the value of CA approaches can be most vividly realised.

The discussion turns briefly to an institutional discourse perspective before looking specifically at CA in the L2 classroom.

An institutional discourse CA methodology takes as its starting-point the centrality of talk to many work tasks: quite simply, the majority of work-related tasks are completed through what is essentially conversation, or "talk-in-interaction" (Drew and Heritage 1992, p. 3); many interactions (for example, doctor-patient interviews, court-room examinations of a witness, classrooms) are completed through the exchange of talk between specialist and non-specialists (*ibid.*):

> Talk-in-interaction is the principal means through which lay persons pursue various practical goals and the central medium through which the daily lives of many professionals and organizational representatives are conducted.

The purpose of a CA methodology in an institutional setting is to account for the ways in which context is created for and by the participants in relation to the goal-oriented activity in which they are engaged (Heritage 1997, p. 163). All institutions have an over-riding goal or purpose which constrains both the actions and interactional contributions of the participants according to the business in hand, giving each institution a unique interactional "fingerprint" (Heritage and Greatbatch 1991, pp. 95–6). Thus, the interactional patterning (or "fingerprint") which is typical of, for example, a travel agent will be different from that of a classroom and different again from that of a doctor's surgery. In each context, there are well-defined roles and expectations which, to some extent, determine what is said.

By examining specific features in the institutional interaction, an understanding can be gained of the ways in which context is both constructed and sustained; features which can be usefully examined include turn-taking organisation, turn design, sequence organisation, lexical choice and asymmetry of roles (Heritage 1997). The second language classroom is, of course, a clear example of an institutional setting with asymmetrical roles, goal-oriented activities and a context which is constantly being created for and by participants through the classroom interaction. While the discourse of L2 classrooms does not and should not be interpreted as having any resemblance to conversation, there are nonetheless good reasons for using a CA methodology (Edwards and Westgate 1994, p. 116):

> The point is not that classroom talk 'should' resemble conversation, since most of the time for practical purposes it cannot, but that institutionalised talk […] shows a heightened use of procedures which have their 'base' in ordinary conversation and are more clearly understood through comparison with it.

The relevance of a CA approach to the L2 classroom context is not difficult to perceive. CA attempts to account for the practices at work which enable participants in a conversation to make sense of the interaction and contribute to it. There are clear parallels: classroom talk is made up of many participants; it involves turn-taking, -ceding, -holding and -gaining; there have to be smooth transitions and clearly defined expectations if meanings are to be made explicit. Possibly the most significant role of CA is to **interpret** from the data rather than **impose** pre-determined categories.

One of the biggest influences on CA-led classroom-based research was the call of Firth and Wagner (1997) for greater sensitivity towards contextual and interactional aspects of language use by focusing more on the participants in SLA research and less on cognitive processes. Since the late 1990s, these studies have highlighted the ways in which learning and interactional competence can be approached and described through a micro-analytic mode of inquiry (see, for example, Hellermann 2008; Markee 2008). From this body of research has emerged the field now known as CA-SLA or CA-for-SLA: Conversation Analysis for Second Language Acquisition. By focusing on micro-details of video- or audio-recorded interaction, CA-for-SLA aims to document micro-moments of learning and understanding by drawing upon participants' own understanding of the ongoing interaction, from an emic perspective. This perspective is revealed through a detailed analysis of vocal (words and grammar, suprasegmentals, pace of talk, etc.) and non-vocal (silence, body language, embodiment of surrounding artefacts, etc.) resources within the sequential development of talk. CA-for-SLA studies have succeeded in demonstrating 'good' examples of 'interactional competence' and/or understanding of certain information by students by using interactionally and pedagogically fruitful instances of talk; for instance through the use of repair sequences (e.g. Hellermann 2009, 2011).

To summarise this necessarily brief overview of the use of CA for the study of classroom discourse, we can make a number of claims concerning its appropriateness. Firstly, under CA, there is no preconceived set of descriptive categories at the outset. The aim of CA is to account for the structural organisation of the interaction as determined by the participants. That is, there should be no attempt to 'fit' the data to preconceived categories; evidence that such categories exist and are utilized by the participants must be demonstrated by reference to and examples from the data. Thus, the approach is strictly empirical. Secondly, there is a recognition that the context is not static and fixed, but dynamic and variable. A dynamic perspective on context allows for variability; contexts are not fixed entities which operate across a lesson, but dynamic and changing processes which vary from one stage of a lesson to another (Cullen 1998). A CA methodology is better-equipped to take variations in linguistic and pedagogic purpose into account since one contribution is dependent on another. Third, the approach recognises that all spoken interactions are goal-oriented. Under institutional discourse, the behaviour and discourse of the participants are goal-oriented in that they are striving towards some overall objective related to the institution. In a language classroom, for example, the discourse is influenced by the fact that all participants are focusing on some pre-determined aim, learning a second language. Different participants, depending on their own agenda may have different individual objectives; nonetheless, the discourse which is jointly constructed is dependent on both the goals and the related expectations of the participants. Finally, CA offers a multi-layered perspective on classroom discourse. Because no one utterance is categorised in isolation and because contributions are examined in sequence, a CA methodology is much better-equipped to interpret and account for the multi-layered structure of classroom interaction.

4 A CLCA Methodology

In light of the different research traditions of CL and CA outlined in the preceding sections, the reader might be forgiven for coming to the conclusion that the two methodologies are incompatible and that there is little point in pursuing the enterprise of CLCA. In this section, therefore, I present a practical example to demonstrate how this methodology was utilised in a recent study (see Walsh et al. 2011). The study reported here took place in a higher education, small group teaching (henceforth SGT) context, where seminars and tutorials are used to support larger lectures. These sessions are important in that they are designed to allow tutors and students to engage in debate and discussion. They account for up to 40 % of the time of undergraduate students and up to as much as 75 % of the time of postgraduate students (Bennett et al. 2002). The 2010 study used a corpus of 500,000 words taken from two universities in Ireland, one in the north, the other in the south.

Previous CL studies on spoken interaction in higher education have arisen principally from the Michigan Corpus of Academic Spoken English or MICASE (Simpson et al. 2002). This corpus comprises data from across a range of speech events in higher education. It includes contexts relevant to the study reported here, such as classroom discussions, seminars, lab work and advising sessions. Studies based on the MICASE corpus have explored a wide range of phenomena in academic spoken interaction, such as metadiscourse in lectures (Lorés 2006), the use of conditionals (Louwerse et al. 2008), and, of more direct relevance to this study, the effect of class size on lecture discourse (Lee 2009).

From a CA perspective, recent research on talk-in-interaction in SGT in higher education has uncovered important aspects of the processes or 'machinery' by which seminars and tutorials 'get done'. Such work has focused on cues and signals used to manage interaction and participant roles (Viechnicki 1997), sequential organisation and negotiation of meaning (Basturkmen 2002), the issue of 'topicality' in small group discussion (Stokoe 2000; Gibson et al. 2006), and the formulation and uptake of tasks and resistance to 'academic' identities (Benwell and Stokoe 2002). In most of these studies, SGT sessions are seen as locally produced accomplishments in which participants take actions to further their own goals and agendas and display their orientations to others' actions and make relevant certain identities. In SGT contexts, tutors will demonstrably orient to the accomplishment of pedagogical goals and tasks, and students may accept or resist these actions (Benwell and Stokoe 2002). At all times during interaction in these SGT contexts, as in other educational contexts, there is a complex relationship between pedagogic goals and the talk used to realise them. By looking closely at the interactions taking place in SGT settings, the aim of Walsh et al's 2010 study was to demonstrate how tutors and students engage in tightly organised and intricate negotiations of a set of pedagogic agendas, using both interactional and linguistic resources to achieve their goals.

A CLCA methodology essentially entails looking at the same data-set through two different lenses: one CL, the other CA. Thus, the same text is subjected to two

treatments, each offering a unique but complementary perspective on the data. A useful starting point is to use CL in the first layer of analysis as a means of scoping out and quantifying recurring linguistic features. This analysis enables the identification of recurring patterns, each specific to the context. The second layer of analysis (using CA) draws upon these contextual patterns in the quantitative analysis and investigates them more closely. For example, in the 2010 study, there were interesting findings around the frequency and use of certain discourse markers, which clustered around specific contexts. This led to a closer CA led investigation which, in turn, produced interesting findings above the level of turn and in relation to specific interactional features. The process adopted an iterative approach to analysis, from CL to CA, back to CL and so on. Key to this is the interdependence between the two modes of analysis, which was non-linear in that, for example, CL tools were sometimes used within the CA layer of analysis to quantify CA insights.

Using *WordSmith Tools* (Scott 2008) key words and word frequencies were identified for both single words and multi-word units (henceforth, MWU), units of two or more words sometimes referred to as *lexical bundles*, *lexical phrases*, *clusters*, *chunks*, though with slightly varying definitions (see Greaves and Warren 2010). Further analysis into the context using concordance lines revealed differences in the functioning of these key words. For example, *if* when used in 'first conditional' type structures had three main functions:

- pedagogic illustration of 'general truths/facts' *if John Kerry takes Texas*, … *he takes every vote*…;
- projecting, meaning 'when you find yourself in this situation' *if you are on TP and you have a class that*…;
- demonstrating, *if you click the mouse and then click*…

Other features which were identified through concordance line analysis include the prevalence of the interrogative pronoun *what* (e.g. *What do you think of it*?), discourse markers *so*, *okay*, *alright*, deictic *next* (as in *next week*, *next semester*, *next lecture*). Concordancing also showed that the relatively high frequency of *need* is related to the speech act of giving instructions (*what I need you to do*, *you need to* etc.).

At this lexical level therefore, the corpus data pointed towards certain contexts such as eliciting information, signposting the discourse, locating learning and teaching in time and giving instructions to learners to perform certain actions and carry out tasks. However, these are just pointers that are emerging as hypotheses as a result of key words, frequency counts, concordance searches. When the analysis was extended to patterns (2–6-word MWUs), concordance searches produced a total of 128 items which were salient to the SGT context. These items were then categorised according to their approximate functions in the discourse. The analysis, at this stage, was moving towards looking at longer stretches of discourse at the level of turn and longer sequences. At this point, the main focus switched to CA.

An initial CA analysis showed that the 128 items identified in the corpus as being salient played an important role as resources for participants' courses of action or 'interactional projects'. Schegloff (2007) describes interactional projects as a form

of interactional organisation in which a course of conduct "is developed over a span of time (not necessarily in consecutive sequences) to which co-participants may become sensitive, which may begin to inform their inspection of any next sequence start to see whether or how it relates to the suspected project, theme, stance, etc." (p. 244). These interactional projects are less tightly bound than the kinds of sequences or 'sequences of sequences' built up out of adjacency pairs, although they can themselves include such sequences, but they do set up specific types of identifiable speech exchange systems within SGT sessions.

In producing these speech exchange systems participants use the different 'organizations of practice' (Schegloff 2007, p. xiv) such as turn design, turn-taking, orientation to actions such as requesting and telling, building coherent sequences through adjacency pairs, repairing trouble, word selection and overall structuring of the interaction, in specific ways. In SGT interaction, in common with other types of pedagogical interaction, it is the tutor's interactional project to pursue pedagogical goals, and this leads to a reflexive relationship between such goals and the 'shape' of the interaction (Seedhouse 2004). In the dataset, four such speech exchange systems were identified, each with distinguishing interactional features and clear pedagogic goals (cf. Walsh 2006):

(a) Procedural talk, with a focus on organising learning and comprising long tutor turns and correspondingly little participation by students. Specific MWUs such as 'what I want you to do is' were also found in high frequency.
(b) Didactic talk, with a focus on eliciting information or giving feedback. The MWU *tell me* is prominent in this micro-context, while turn-taking is controlled tightly by the tutor. Display questions prevail and the three-part exchange structure IRF dominates. (Tutor Initiates, student Responds, tutor gives Feedback)
(c) Empathic talk. Here, students have more space and manage the floor, producing 'tellings' or accounts of personal experiences. There is more equality in turn-taking and roles are more symmetrical. Discourse markers play a key part in this micro-context, especially *you know* and *you see* which function to create 'shared space' for learning.
(d) Argumentational talk. This micro-context was found to occur when there was shared space, but the discussion was more combative, with a focus on agreeing and disagreeing. Words like *but* and *maybe* were used frequently to show disagreement or indicate stance.

5 Discussion

This aim of this chapter was to demonstrate the appropriateness of CL and CA in providing enhanced descriptions of spoken interactions in higher education small group settings. Four speech exchange systems (micro-contexts) were identified in the data, each with distinctive interactional, linguistic and pedagogic features or 'fingerprints' (Drew and Heritage 1992, p. 26). The four speech exchange systems are robust throughout the data. That is, at any point one or other will be

operating, whether for long spates of interaction or for shorter bursts. Using a CLCA methodology, I suggest, allows useful comparisons to be made both across and within these micro-contexts. For example, a comparison of didactic and empathic talk reveals very different profiles or 'fingerprints'. The former is characterised by short learner turns, tightly controlled turn-taking, evidence of IRF exchange structures, extensive use of the MWUs *tell me* and *can you tell me* and the main pedagogic function of eliciting. The main focus of empathic talk, on the other hand, is 'show and tell': the tutor's pedagogic goal is to promote debate and discussion and create a safe environment for that to take place.

When the CL analysis is related more closely to the CA findings, the single words and MWUs identified as being salient are found across all micro-contexts; more importantly, they are found to do different interactional work in relation to the particular agenda of the moment. Indeed, it is striking that the participants in this study used single words and MWUs to carry out specific actions that move forward their interactional projects. Thus they are helpful both to participants and analysts in solving what Schegloff (2007) describes as the 'action-formation' problem: that is, how language formations are designed to be recognizable by interlocutors as particular actions, such as requesting, telling, eliciting etc. Not only are these units used by participants to carry out specific acts, but they function as indices, both for participants and for analysts, of the current speech exchange system one is in. For this reason, they are bound up with the interactional competence displayed by participants in SGT sessions as they move forward their particular agendas and respond appropriately at any moment in the interaction.

It seems evident from the study presented here that there is much to be gained from using a combined CLCA methodology. First, the methodology allows two (at least) perspectives on the same dataset: one (using CL) offering an overview of the data and a profile of the most important recurring linguistic features in specific contexts of use; the other (using CA) offering a fine-grained, up-close view of the same data and highlighting the ways in which meanings are co-constructed. This dual perspective on the same dataset, I would suggest, facilitates a closer understanding of what linguistic and interactional resources are used to create meaning. Specifically, there is an opportunity for the analyst to examine in some detail the ways in which linguistic, interactional and textual features combine in any communicative encounter. Second, the methodology allows enhanced understandings of specific features of spoken discourse in a particular context. Arguably, it allows the analyst to focus more on language *use* (what we do with language) and less on language *usage* (what language is); the issue of what language does rather than what language is has been taxing applied linguists for many years (ref). Third, this methodology goes some way at least in compensating for the deficiencies of each method when used alone: CA, which is unable to extend its findings beyond the relatively small sample of data it typically utilises; CL which is only able to make general observations on the data, without offering the kind of interactional detail which CA provides. A CLCA methodology compensates for all these deficiencies and allows analysts to provide both greater depth and coverage in their findings.

There are, naturally, also some shortcomings to this methodology. The first is that there is a presupposition in the arguments put forward here that researchers are able to use both CL and CA. That is rarely the case since the two research traditions are, by definition, mutually exclusive. It would be unusual, but not unheard of, for a conversation analyst to use a CL methodology and the same is true in reverse. One way round this is for conversation analysts to work with corpus linguists in a spirit of shared expertise (cf. Walsh et al. 2010). A second shortcoming is that the methodology, while following an iterative process, is somewhat imprecise in terms of which steps should be taken and when. Should one, for example, commence with CL and then do CA, or vice versa? What precise steps should be taken once the first analysis has been completed and in what sequence? There are no exact answers to these issues; I would only say that with a little trial and error, it is possible to make effective use of the two methodologies.

6 Conclusion

This chapter set out with the proposition that CL and CA can be usefully combined in the analysis of spoken data. I have suggested how, in spite of their ontological and epistemological differences, these two research methodologies can be combined and offer a surprisingly rich and comprehensive perspective on a corpus. This combined CLCA approach has the potential to provide far more detailed analysis than that offered when each is used in isolation. In the study reported here, for example, detailed descriptions of the same corpus of academic spoken English were given from at least three perspectives: linguistic (portraying the use of high frequency items, key words, MWUs, discourse markers, question forms and so on), interactional (focusing on turn-taking and turn design, sequential organization) and pedagogic (looking at specific pedagogic functions at a given moment to include eliciting, explaining, instructing and so on). Arguably, a CLCA approach allows for a much more detailed description of a particular context (for example, small group teaching in higher education), offering insights into the ways in which language is used to mean, convey information and establish joint understandings. The approach, above all, underlines the centrality of joint enterprise in any spoken encounter: people establish understandings together and share equal responsibility for that goal in most cases.

While each methodology has its own merits, it also has significant shortcomings as outlined above. CL on its own, for example, may provide interesting lists of high frequency items which can then be explained functionally, but its perspective is a surface level one; a CA perspective, on the other hand, enables us to identify particular exchanges and sequence organisations, but misses the fact that particular linguistic features may occur in each exchange structure. Essentially, there is much to commend this combined methodology and the future is likely to

show further evidence of the power and potential of the two methodologies. Future research is likely to result in a narrowing of the perceived gap which currently exists between each approach: for example, there have already been moves to look more quantitatively at turn openings and closings using CL (refs), while there has been a corresponding prediction that CA will become more quantitative in the future (ref). By looking more at specific interactional features (such as discourse markers), it is not inconceivable that CL will begin to offer turn-level analyses which have relevance for CA. In short, we can predict that a combined CLCA methodology is here to stay and that we'll be witnessing a growth in its adoption in coming years.

References

Ädel, A. 2010. How to use corpus linguistics in the study of political discourse. In *The Routledge handbook of corpus linguistics*, ed. A. O'Keeffe and M.J. McCarthy, 591–604. London: Routledge.

Basturkmen, H. 2002. Negotiating meaning in seminar-type discussions and EAP. *English for Specific Purposes* 21(1): 233–242.

Bennett, C., C. Howe, and E. Truswell. 2002. *Small group teaching and learning in psychology*. York: LTSN Psychology University of York.

Benwell, B.M., and E.H. Stokoe. 2002. Constructing discussion tasks in university tutorials: Shifting dynamics and identities. *Discourse Studies* 4(4): 429–453.

Carter, R., and M.J. McCarthy. 2006. *Cambridge grammar of English. A comprehensive guide to spoken and written grammar and usage*. Cambridge: Cambridge University Press.

Cotterill, J. 2010. How to use corpus linguistics in forensic linguistics. In *The Routledge handbook of corpus linguistics*, ed. A. O'Keeffe and M.J. McCarthy, 578–590. London: Routledge.

Cullen, R. 1998. Teacher talk and the classroom context. *English Language Teaching Journal* 52(3): 179–187.

Drew, P., and J. Heritage. 1992. Analyzing talk at work: An introduction. In *Talk at work: Interaction in institutional settings*, ed. P. Drew and J. Heritage, 3–65. Cambridge: Cambridge University Press.

Edwards, A., and D. Westgate. 1994. *Investigating classroom talk*. London: Falmer.

Farr, F., B. Murphy, and A. O'Keeffe. 2004. The Limerick corpus of Irish English: Design, description and application. *Teanga* 21: 5–29.

Firth, A., and J. Wagner. 1997. On discourse, communication, and (some) fundamental concepts in SLA research. *The Modern Language Journal* 81: 285–300.

Gibson, W., A. Hall, and P. Callery. 2006. Topicality and the structure of interactive talk in face-to-face seminar discussions: Implications for research in distributed learning media. *British Educational Research Journal* 32(1): 77–94.

Greaves, C., and M. Warren. 2010. What can a corpus tell us about multi-word units? In *The Routledge handbook of corpus linguistic*, ed. A. O'Keeffe and M.J. McCarthy, 212–226. London: Routledge.

Hellermann, J. 2008. *Social actions for classroom language learning*. Clevedon: Multilingual Matters.

Hellermann, J. 2009. Looking for evidence of language learning in practices for repair: A case study of self-initiated self-repair by an adult learner of English. *Scandinavian Journal of Educational Research* 53(2): 113–132.

Hellermann, J. 2011. 'Members' methods, members' competencies: Looking for evidence of language learning in longitudinal investigations of other-initiated repair'. In *L2 Interactional competence and development*, ed. J.K. Hall, J. Hellermann, and S. Pekarek Doehler, 147–172. Bristol: Multilingual Matters.

Heritage, J. 1984. A change-of-state token and aspects of its sequential placement. In *Structures of social action*, ed. J.M. Atkinson and J. Heritage, 299–345. Cambridge: Cambridge University Press.

Heritage, J. 1997. Conversational analysis and institutional talk: Analysing data. In *Qualitative research: Theory, method and practice*, ed. D. Silverman. London: Sage Publications.

Heritage, J., and D. Greatbatch. 1991. On the institutional character of institutional talk: The case of news interviews. In *Talk and social structure: Studies in ethnomethodology and conversation analysis*, ed. D. Boden and D.H. Zimmerman. Berkeley: University of California Press.

Hutchby, I., and R. Wooffitt. 2008. *Conversation analysis*, 2nd ed. Cambridge: Polity Press.

Lee, J. 2009. Size matters: An exploratory comparison of small- and large-class university lecture introductions. *English for Specific Purposes* 28(1): 42–57.

Lorés, R. 2006. The referential function of metadiscourse: thing(s) and idea(s) in academic lectures. In *Corpus linguistics: Applications for the study of English*, ed. A. Hornero, M. Luzón, and S. Murillo, 315–334. Bern: Peter Lang.

Louwerse, M., S. Crossley, and P. Jeuniauxa. 2008. What if? Conditionals in educational registers. *Linguistics and Education* 19(1): 56–69.

Markee, N. 2008. Toward a learning behavior tracking methodology for CA-for-SLA. *Applied Linguistics* 29: 404–427.

McCarthy, M., and A. O'Keeffe. 2010. Historical perspective: What are corpora and how have they evolved? In *The Routledge handbook of corpus linguistics*, ed. A. O'Keeffe and M.J. McCarthy, 3–13. London: Routledge.

O'Keeffe, A. 2006. *Investigating media discourse*. London: Routledge.

O'Keeffe, A., and F. Farr. 2003. Using language corpora in language teacher education: Pedagogic, linguistic and cultural insights. *TESOL Quarterly* 37(3): 389–418.

O'Keeffe, A., M. McCarthy, and R. Carter. 2007. *From Corpus to classroom*. Cambridge: Cambridge University Press.

Sacks, H., E.A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4): 696–735.

Schegloff, E.A. 2007. *Sequence organization in interaction: A primer in conversation analysis*, vol. 1. Cambridge: Cambridge University Press.

Scott, M. 2008. *WordSmith tools (version 5)*. Liverpool: Lexical Analysis Software.

Seedhouse, P. 2004. *The interactional architecture of the language classroom: A conversation analysis perspective*. Oxford: Blackwell.

Seedhouse, P. 2005. Conversation analysis and language learning. *Language Teaching* 38(4): 165–187.

Sidnell, J. 2010. *Conversation analysis- an introduction*. West Sussex: Wiley-Blackwell.

Simpson, R.C., S.L. Briggs, J. Ovens, and J.M. Swales. 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor: The Regents of the University of Michigan.

Stokoe, E.H. 2000. Constructing topicality in university students' small-group discussion: A conversation analytic approach. *Language and Education* 14(3): 184–203.

Viechnicki, G.B. 1997. An empirical analysis of participant intentions: Discourse in a graduate seminar. *Language and Communication* 17(2): 103–131.

Walsh, S. 2006. *Investigating classroom discourse*. London: Routledge.

Walsh, S., and A. O'Keeffe. 2007. Applying CA to a modes analysis of third-level spoken academic discourse. In *Conversation analysis and languages for specific purposes*, ed. H. Bowles and P. Seedhouse. Bern: Peter Lang.

Walsh, S., T. Morton, and A. O'Keeffe. 2011. Space for learning: Language use, interaction and orientation to knowledge in small group teaching in higher education. *International Journal of Corpus Linguistics.*

Yang, S. 2013. Unpublished Ph.D. thesis, University of Newcastle.

Small Corpora and Pragmatics

Elaine Vaughan and Brian Clancy

1 Introduction

In this chapter we argue for the benefits of using small, domain-specific corpora in pragmatic research, and this position presupposes a number of questions. The first of these questions relates to the establishment of what we mean by 'small' corpora, in what context this characterisation developed, and how this is relevant to the type of studies we review and present. The second regards to what extent corpus methodology can assist research on pragmatic phenomena, and what type of insights this empirical orientation can generate. Below we attempt to answer these questions and frame them in general as well as in relation to two studies which use small corpora to investigate the pragmatics of how identities are indexed in two different speech contexts. With regard to our first question, any discussion of small corpora raises the question 'what do we mean by 'small'?', and this is worth pondering for two reasons: firstly, and instrumentally, answering this question will define our parameters in talking about 'small corpora and pragmatics' in general. More importantly, it raises some issues in connection with corpus linguistics as it has developed in the last few decades that prompts our position as to why 'small' corpora can be of benefit to pragmaticists.

The emergence of modern corpus linguistics is primarily associated with lexicography and the pioneering work of researchers such as John Sinclair. This work was predicated on creating the largest possible corpora in the 1960s and 1970s which, as Sinclair (2001: viii) points out, 'were simultaneously the largest and smallest of their type being the only ones'. Early COBUILD corpora contained tens

E. Vaughan (✉)
School of Languages, Literature, Culture and Communication,
University of Limerick, Limerick, Ireland
e-mail: Elaine.Vaughan@ul.ie

B. Clancy
Mary Immaculate College, University of Limerick, Limerick, Ireland
e-mail: Brian.Clancy@mic.ul.ie

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_4,

of millions of words and, as technology advanced, so too did the size of these corpora (Sinclair 2001; McCarthy and O'Keeffe 2010; Tognini-Bonelli 2010). The Collins corpus, which incorporates COBUILD, now contains 2.5 billion words, and the Oxford English Corpus approximately two billion words. The Cambridge English Corpus, which comprises samples of British, American and Learner English, consists of many billions of words. There does not appear to be any upper limit on language corpora; indeed, some discussions of 'corpus' and 'corpus linguistics' have explicitly (e.g. Biber et al. 1998) integrated 'large' as a defining feature, and the prevailing philosophy for corpora such as those mentioned above seems best summed up by the motto of the American-based Linguistic Data Consortium, there is 'no data like more data' (Sinclair 2001). As corpus linguistics has developed, it has come to be associated with many aspects of language study, such as language variation, historical linguistics or those studies with language pedagogy as their focus and it is now possible to access a range of large corpora designed for these purposes. Corpora such as the American National Corpus (ANC) and the British National Corpus (BNC) are designed to represent the language varieties of American and British English respectively and are also designed to be comparable across genres. The BNC contains 100 million words, of which ten million are spoken.[1] The International Corpus of English (ICE) brings together one-million-word samples from 18 countries which have English as their first or official language, with 60 % of each sample consisting of spoken texts, although some of these texts are scripted and/or monologues (see http://ice-corpora.net/ice/index.htm). The Corpus of Contemporary American English (COCA), the largest freely available corpus, is made up of over 450 million words in more than 175,000 texts, including 20 million words from each year from 1990 to 2011 (see corpus.byu.edu/coca/). The picture in terms of what are glossed as 'historical corpora' is no less impressive in terms of size. The Oxford Text Archive houses a number of these corpora (http://ota.ahds.ac.uk/). Table 1 below gives a brief overview of some of these large corpora.

With regard to corpora built to investigate aspects of language pedagogy, the Cambridge Learner Corpus (part of the Cambridge English Corpus mentioned above) contains 43 million words of written and spoken learner English across the proficiency levels and the International Corpus of Learner English is a 3.7 million word corpus of English as a Foreign Language writing from learners from 16 different mother tongue backgrounds (see http://www.uclouvain.be/en-cecl-icle.html).

1.1 Small Corpora in Corpus Linguistics

It is hard to imagine describing any of the corpora mentioned above as 'small', and, in fact, defining our terms here requires the caveat that 'small' is relative, related to modality (the term 'modality' is used here in its loosest sense as occupying some

[1] Almost 15 million words of the ANC are currently available. This is divided into approximately 11.5 million words of written language and 3.5 million words of spoken language (see www.anc.org).

Table 1 Examples of large corpora

Corpus	Number of words (approx.)	Overview of composition
Collins Corpus and the Bank of English™	2.5 billion	Written: e.g. websites, magazines, newspapers, books Spoken: e.g. radio, TV, everyday conversations
Oxford English Corpus	2 billion+	Mainly written material from World Wide Web, e.g. academic papers, technical manuals, corporate websites, personal websites, blogs
Cambridge English Corpus	2 billion+	Written and spoken English from a range of domains, e.g. books, newspapers, letters, e-mails, websites, conversations, meetings, radio
British National Corpus	100 million	Written (90 %): e.g. newspapers, books (fiction/non-fiction), letters, school/university essays Spoken (10 %): e.g. informal conversations, business and government meetings
International Corpus of English	1-million-word samples	Different varieties of English (e.g. British, Irish, Hong Kong, Singapore, East African English) Written: e.g. letters, academic writing, newspaper reports Spoken: e.g. conversations, meetings, radio
Corpus of Contemporary American English	450 million	Written: e.g. fiction, popular magazines, newspapers, academic texts Spoken: e.g. television, radio programmes

point on the speech to writing continuum, cf. Biber, e.g. 1988), and is, inevitably, 'frequently reinterpreted' (Sinclair 2001: xiii). Beside the behemoths of the major publishing houses, the corpora of national varieties mentioned above appear small. While it seems to be accepted that the upper limit of a small corpus is approximately 200,000–250,000 words (see Aston 1997; Flowerdew 2004), one- to five-million-word samples have also been described as 'small' (McCarthy 1998; Sinclair 2001). Aston (1997) notes that small corpora exist in the 20,000–200,000 word range, and are more specialized in terms of topic and/or genre than large corpora. In terms of modality, of relevance to corpus size is the type of corpus in question. Spoken corpora – the principle focus of this paper – are often, by necessity, smaller than written corpora. There are a great number of reasons for this, not least of which is the fact that spoken data still need to be manually transcribed to adequately represent the speech event, and even manual transcription does not completely represent the complexities of spoken interaction. Multi-modal corpora are still very much in the minority, although great strides have been made in this regard (see, for example, Knight et al. (2009)). A major factor behind the development of small corpora has not necessarily been the corpus linguistic research agenda *per se*, but something else entirely: the emergence of small corpora can be directly related to technological developments (Sinclair 2001). In the past, assembling a large amount of data was associated with high costs because of the difficulties involved in recording,

transcribing and coding the data. Data can now be easily collected, assembled, stored and analysed on a PC, arguably 'democratising' the notion of corpus building and corpus linguistics (cf. Rundell 2008: 26).

What we are implying is that it has not always been a given that corpora considered 'small' had full legitimacy in the field of corpus linguistics. A major reason for this reluctance to fully admit small corpora to the fold was rooted in, as previously mentioned, the predominant research agenda in corpus linguistics in its 'early modern' period, lexicography, and the remediation of concerns in relation to 'representativeness' and 'balance' in commercial corpus building. Corpora used for lexicographical research need to be as large as possible in order to generate sufficient occurrences which reflect how lexical items are used, and, as previously mentioned, these large corpora, such as the Bank of English, dominated research publications representing the 'output' of corpus linguistics. Representativeness, or 'or the extent to which a sample includes the full range of variability in a population' (Biber 1993: 243) has been a challenge in relation to language data, and, as Clear (1992: 21) points out, it is difficult to interpret the statistical notion of 'population' in relation to a phenomenon like language. One response to this difficulty has been to approach the sampling of language data in a different way. Biber (1993) proposes strata and sampling frames for representative corpus design based on 'register', or situationally defined text categories such as 'fiction', 'news article' etc., and linguistically defined text types, such as various written or spoken modes. In terms of the balance of a corpus, Sinclair (2005) refers to it as a rather vague notion but important nonetheless. Balance appears to rely heavily on intuition and best estimates (Atkins et al. 1992; Sinclair 2005; McEnery et al. 2006). In terms of a large corpus, the Longman Spoken and Written English Corpus (LSWE) is considered 'balanced'. According to Biber et al. (1999: 25), the registers contained within the corpus were selected on the basis of balance in that they 'include a manageable number of distinctions while covering much of the range of variation in English.' For example, conversation is the register most commonly encountered by native speakers whereas academic prose is a highly specialised register that native speakers encounter infrequently. Between these two extremes are the popular registers of newspapers and fiction. For a more specialised corpus, balance is reliant on the corpus containing a range of texts typical of what the corpus is designed to represent. In terms of small corpus compilation, a small corpus should be approached with as much caution as a large corpus, as issues of balance and representativeness are salient no matter the size of the corpus. A small corpus builder can address issues of representativeness by ensuring that the samples collected are typical of the speech domain represented by the corpus. For example, the corpus of family discourse discussed in Sect. 3 features members of that family talking while engaged in eating a meal, putting up the Christmas tree, talking about being a student in university and providing information about a city one of them is going to visit, interactions typical of many families and, therefore, considered 'representative' (Clancy 2010). McEnery et al. (2006: 5) maintain that if specialised corpora were discounted on the basis of sampling techniques used, then 'corpus linguistics would have contributed significantly less to language studies' and this is an enlightened and crucial point to keep in mind.

Sociolinguistic studies have shown that relatively small samples that could be considered technically unrepresentative are sufficient to account for language variation in large cities (see Sankoff 1988; Tagliamonte 2006). McEnery et al. (2006: 73) claim that although representativeness and balance are features that must be considered in relation to corpus design, they often depend on the ease with which the data can be collected (and, of course, the nature of the data itself) and, therefore, 'must be interpreted in relative terms i.e., a corpus should only be as representative as possible of the language variety under consideration.' They believe that corpus building is 'of necessity a marriage of perfection and pragmatism' (*ibid.*), echoing Stubbs' (2004) contention that corpus size tends to be 'a compromise between the desirable and the feasible' (p. 113). Flowerdew (2002: 96) maintains that 'the field [of corpus linguistics] has widened considerably to include the recognition of much smaller, specialised genre-based corpora'. Small corpora have been instrumental in pushing the boundaries of corpus linguistics as a field of enquiry, and have been similarly so in prompting a shift towards empiricism in the realm of pragmatics research (cf. Romero-Trillo 2008). The review below does not purport to, nor would it be possible to, represent the totality of the literature available on small corpora in relation to pragmatics; instead it is intended to be selective and illustrative of what working empirically with small corpora and a pragmatic agenda can uncover.

2 The Use of Small Corpora in Pragmatic Research: A Selective Review

The primary benefit of small corpora to the study of pragmatics is a fundamental one: they can enable the researcher to access authentic, naturally occurring language and to maintain a close connection between language and context. Indeed, in relation to contextual links and small corpora, Koester (2010) points out that small corpora have a clear advantage over larger ones. She maintains that large corpora are sampled from such a variety of different contexts that it is 'very difficult, if not impossible, to say anything about the original contexts of use of the utterances' (*ibid*: 66–67; see also Flowerdew 2004). While it is certainly possible to investigate phenomena such as hedging using large corpora, this can be a major challenge due to the variety of (para) linguistic selections available for use as hedges. Using a small, context-specific corpus offers significant advantages. These phenomena cannot only be investigated in their original context of use, it is also usually possible to investigate virtually all occurrences and essay a refined analysis which takes the polysemous nature of many pragmatic features into account. Therefore, we can move from quantitative observations regarding frequency of items with pragmatic potential, which only tell part of the story. The studies summarised below have turned up contextualised findings in relation to the pragmatic significance of linguistic and extra-linguistic strategies as diverse as question forms, modality, small talk, humour and (evaluative) speech acts.

In the public sphere of media discourse, O'Keeffe (2005) used a 55,000 word corpus from radio phone-in to focus on question forms as they are used in this context, which from other analytical perspectives – for example, conversation analysis – displays a fairly typical (and canonical) turn-taking structure with the presenter holding the discursive power. However, although many asymmetrical norms of institutional discourse do apply to this context, there is widespread downtoning of power at a lexico-grammatical level. In addition to using pragmatic markers to hedge, the presenter of the radio show employs a variety of features such as first name vocatives, latching and reflexive pronouns, as in *you've a daughter yourself?*, to create a 'pseudo-intimate' (p. 340) environment between speaker and caller. Also in the public sphere, but in a more difficult to access 'occluded genre' (Swales 1996; Loudermilk 2007), Koester (2006) created a 34,000 word corpus of American and British office talk and demonstrated the influence of local contexts on frequency and use of various phenomena, such as hedging and modality. She identifies a number of genres within the workplace discourse she investigates, and finds that modal verbs of obligation are more frequent in collaborative genres (for example, decision making or planning) than in unidirectional genres (for example, giving instructions). The boundaries between the genres she identifies are, however, fluid. She notes that there is no easy distinction between 'on-task' transactional talk and small or relational talk essential for building speaker relationships (p. 161) due to the complex nature of speakers' interactional goals. Vaughan (2007, 2008) employs a 40,000 word corpus of meetings of English language teachers (C-MELT, see Sect. 3 below) to explore particular linguistic features characteristic of this community of practice (Wenger 1998). Part of this study involved exploring how the community managed and maintained itself, and looked at how power and solidarity are negotiated, for example through humour. The size of C-MELT allowed specific instances of humour to be isolated in order that they might be assigned a function. Vaughan (2007: 186) found that teachers 'use [humour] to establish the social space they share, and implicitly define who they are, and what their attitude is to the work they do'; humour in this context is in fact a highly salient, 'powerful, polyvalent pragmatic resource' (Vaughan and Clancy 2011: 51). Finally, in a study that is also situated in the institutional domain, Farr (2007) demonstrates how in teacher education, 'a spoken language corpus can be a valuable instrument in the toolbox of professional development' (p. 254) and her 80,000 word professional talk corpus allowed the identification of areas for development and, equally, good professional practice. She explores the use of relational strategies present in the data to demonstrate how trainers work to lessen asymmetrical speech relationships and claims that small talk, in particular talk about health issues, is a typical way of establishing solidarity between speakers in this context (Farr 2005). Furthermore, she demonstrates how shared socio-cultural references such as *muinteóir*, the Gaelic word for *teacher*, are a method of diluting institutional power on the part of the teacher trainer in interaction with the trainee.

At this juncture, it is important to note that for all the studies mentioned here, the researchers were also the corpus compilers (and often participants also), and this close relationship between corpus and researcher further strengthens the advantage of small

corpus research for pragmatics. As Koester (2010: 67) points out a feature of small corpus research is that the researchers themselves often have a high degree of familiarity with the context and this ensures that the quantitative corpus results generated can be 'balanced and complemented with qualitative findings' such as information about setting, participants and purpose. Cutting (2001) investigated the evaluative speech acts of six students on a taught Master's course in Applied Linguistics as they became members of an academic discourse community. Cutting isolated and tagged each of these speech acts and found that positive acts increase as the course progresses and participants build solidarity. She also found that negative speech acts are most common in conversations about the course. Cutting explicitly states that she deliberately limited the corpus used to 26,000 words so that she 'could become familiar enough with each one's [participant's] linguistic idiosyncrasies, personalities and attitudes to interpret the findings' (pp. 1208–1209), an approach that would be very difficult with a larger corpus. This is not to say that a similar level of familiarity cannot not be attained by researchers who are not the corpus compilers in these cases. The cogent point is that the smaller sizes of the corpora facilitate ease of familiarisation.

A significant advantage of using small corpora for this type of research, as previously touched upon, is that frequency information, while interesting, is insufficient for pragmatic characterisation or categorisation. In relation to a study of the epistemic function of modal markers in English for Academic Purposes (EAP), Holmes (1988) notes that there was little corpus frequency information in relation to the occurrence of modal markers for this specific context. This, she claims, is unsurprising given that a million-word corpus, even if it contains data from EAP, when searched will provide the analyst with approximately 3,000–4,000 tokens of modal forms such as *would*, and each of these tokens requires detailed contextual analysis in order to assign function. She maintains that with a smaller domain-specific corpus, however, 'it is possible to establish both the range and the frequency of modal verbs expressing epistemic modality' (p. 28). Within this wider issue of a surfeit of data is a connected and rather human one: as Orpin (2005: 39) suggests, 'an attendant danger in using a large corpus is that the researcher may feel swamped by the huge amount of data s/he is faced with.' In order to overcome this analytical barrier of large frequency count results, researchers seek to 'manage' the data, primarily through the processes of sampling from it and normalising frequency count information to aid comparisons across corpora. For example, Torgersen et al. (2011) analysed the use of pragmatic markers in the Linguistic Innovators Corpus and the Corpus of London Teenage Language. The 2,000 instances of *yeah* they examined in the study comprise only 10.7 % of the total number of instances of the marker (18,693). Similarly, Clancy and Vaughan (2012), faced with 4,860 instances of the item *now* in the Limerick Corpus of Irish English (LCIE), provide a detailed analysis of 500 random instances (for both of these studies, it must therefore be acknowledged that the normalized frequency information presented in the discussion of the results is based on extrapolated figures).

Researchers using large corpora for pragmatic research have also used an iterative approach with smaller and larger corpora in order to fully interpret the initial frequency information the larger corpora generate. For example, O'Keeffe

and Adolphs (2008) investigate the form and function of response tokens in two one-million-word spoken corpora: the Limerick Corpus of Irish English and a one-million-word sample from the Cambridge and Nottingham Corpus of Discourse in English. To put the sort of data generated in context, response tokens tend to be very high frequency items in spoken corpora. They examined the form taken by response tokens, a largely quantitative enterprise, in Irish and British English using the one-million-word samples and found that British speakers in general use a broader range of single and two-word response token forms than their Irish counterparts. However, in order to investigate response token *function* across the two corpora, a more qualitative and detailed process, they constructed two parallel 20,000-word corpora taken from the private sphere. These corpora were comprised of the speech of Irish and British females, all around the age of 20. The female participants were students and close friends who, in most cases, shared accommodation. They found that, again, in these smaller corpora, response tokens are more frequent in British English speech. However, they found no real variation at the level of the response tokens' pragmatic functions. In other words, drilling down into quantitative results using qualitative methodologies uncovers the subtleties of the pragmatic profile of particular items which extends beyond the limited, albeit interesting, information frequency provides.

3 A Case Study: 'We' in Small Corpora

Many of the studies above have in common a methodological approach that moves from general frequency counts to investigating items in context. What they also have in common is a focus on particular locations of discourse, for example, classroom discourse, family discourse or workplace discourse, and the linguistic features that characterise them. In many cases, the research explicitly details the pragmatic norms of the contexts or communities they study. This idea of being able to use a small domain-specific corpus to characterise the discourse of a particular community is intriguing, and, with this in mind, we show below two approaches to identifying the pragmatic function of the personal pronoun *we*. What we are looking at in a broad sense is indexicality, a central notion in pragmatics. It is axiomatic that for the study of pragmatics language and context are inseparable, and it has been argued that the '*single* most obvious way in which the relationship between language and context is reflected in the structures of the languages themselves, is through the phenomenon of deixis' (Levinson 1983: 54, our italics). The phenomenon of deixis, therefore, serves as a constant reminder to us that language can only be interpreted within its context of use, moreover, as Hanks (1992: 48) observes, '… deixis links language to context in distinguishable ways, the better we understand it, the more we know about context'. A significant criticism of corpus linguistics in the past was its abstraction of language from its original context, and to an appreciable extent the fact that most small corpora contain samples of 'complete' texts mitigates this quite valid point: small

Table 2 Description of the two corpora

	C-MELT	Family corpus
Length of recording	3.5 h	1 h
Number of speakers	33	6
Number of words	39,975	12,531

corpus-based pragmatic research is often conducted iteratively, with quantitative observations investigated in qualitative detail to account for frequency/infrequency. For the case studies presented below, while corpus methodologies dictated the research agenda and highlighted pragmatic areas of focus, the fact they were based on small corpora allowed us to investigate the phenomena in context, and thus reanimate the disembodied data returned by corpus searches.

The principal purpose of these two case studies was to investigate how identity is expressed by two quite different communities in two quite different contexts. The first case study uses a small corpus of family discourse recorded in Limerick, a city in the south of the Republic of Ireland. The second study uses a corpus of the meetings of English language teachers (C-MELT) compiled by recording meetings in two different geographical locations, México and Ireland. Table 2 provides more detail on the two different small corpora consulted for the studies described.

A point of confluence for both studies was the contention that if pragmatics is about exploring how context and speaker relationship impact on language, then, as a corollary, control (or not) of pragmatic norms is also about demonstrating membership of a community.[2] Hence uncovering the pragmatic norms around identity work in these particular communities was the primary research focus for both studies. Identities are not monolithic however (De Fina et al. 2006), but mutable, dynamic and situated (Tracy 2002). We propose to look at how identities are expressed though a detailed examination of linguistic proxies for identity, personal pronouns, in order to get at the social relationships being indexed in talk, and the pragmatic management engendered in this process.

Rees (1983) posits pronominal use on a scale of 'distance from the self', where 'I' is closest to the self, and 'they' is the most distant:

0	1	2	3	4	5	6	7	8
I	we	you	one	you	it	she	he	they

The complexity of reference encoded in any one of these pronouns has been the subject of much linguistic, though not necessarily always pragmatic, research. If only because, as Mühlhäusler and Harré (1990) have argued, 'any pronoun can be used for any person'. Complexities in what aspect of identity or what speech act a

[2]There are various frameworks and conceptualisations of 'community', such as the 'speech community' (e.g. Patrick 2002), 'discourse community' (e.g. Swales 1990), or 'community of practice' (Lave and Wenger 1991; Wenger 1998). Both of the studies reported on in Sect. 3 operationalise the notion of community of practice.

speaker invokes with 'I' are not immediately obvious, and it may appear one of the 'least ambiguous' pronouns (Fasulo and Zucchermaglio 2002), though this is only at first sight. 'I' may not always index the speaker only, as in reporting direct speech, for example. In addition, say in the case of ventriloquising (Tannen 2007), 'I' may not refer to the 'animator' of the statement, but to a postulated 'author' (Goffman 1981), in the case of Tannen's research (2007) the family dog.[3] Fasulo and Zucchermaglio (2002) investigate the multiple identities speakers invoke with 'I' in Italian work-place meetings. They present how various role identities are enacted, and show how these identities are situated, highlighting how the meanings of pronouns (for this research 'I') are layered according to the context in which they are invoked. 'You', which has a singular and plural reference in English (plural 'you' is positioned in the middle of Rees' scale above), has an obvious addressee referent. However, it can also can be used in a generalised, 'generic' or 'impersonal' (e.g. Whitley 1978) way, for example, to create a sense of distance or objectivity, or, alternatively, to emphasise or recruit involvement (O'Connor 1994; Stirling and Manderson 2011). *I* and *you* are prominent features on most spoken corpus frequency lists reflecting the canonical conversational dyad. Their high frequency is also due to features of online speech production such as repetition and reduplication, as well as their frequency in fixed pragmatic clusters such as *I think*, *I mean*, *you know* and so on.

Íñigo-Mora (2004: 41) observes of the pronoun *we* that 'depending on the speaker's intention, "we" is the only personal pronoun that can (a) be inclusive and exclusive and (b) claim authority and communality at the same time' (see also Pennycook 1994). While we could argue that it is not the 'only' pronoun to display this type of complexity (see above), it does present an interesting case. As previously mentioned, Mühlhäusler and Harré (1990) have shown that *we* is sufficiently flexible and multifunctional to encode any of the six persons that are usually referred to in English. Biber et al. (1999: 329) assert that 'the meaning of the first person plural pronoun [*we*] is often vague: *we* usually refers to the speaker/writer and the addressee (inclusive *we*), or to the speaker/writer and some other person or persons associated with him/her (exclusive *we*). The intended reference can even vary in the same context.' Inclusive and exclusive *we* can be used to create a perspective of: *I* the speaker + *you* the addressee(s) in the immediate context ('inclusive *we*') and *I* the speaker + someone else not in the immediate context ('exclusive *we*'). An investigation of *we* allows us to examine how different speaker relationships *and* identities are negotiated locally and what this negotiation reveals and entails. In this sense, the pragmatics of personal pronoun usage and invocation of identity becomes critical to conceptions of community, with their natural and appropriate use about demonstrating membership of the community. Understanding speaker identity is crucial to understanding context and it has been shown in research on intercultural pragmatics that inability to inhabit appropriate identities

[3] In this research, Tannen examines how speakers in family discourse use the family pet to interact with one another, allowing them 'to distance themselves figuratively from their own utterances' (2007: 417), for example, to defuse a potential conflict.

Table 3 Top 25 word frequency counts for four spoken corpora (*we* is shaded)

	C-MELT	Family corpus	LCIE	BNC (Spoken)
1	the	the	the	the
2	to	you	I	I
3	I	it	and	you
4	and	I	you	and
5	yeah	to	to	it
6	that	a	it	that
7	of	and	a	a
8	you	of	that	's
9	a	that	of	to
10	it	in	yeah	of
11	**we**	is	in	n't
12	they	yeah	was	in
13	in	no	is	**we**
14	so	it's	like	is
15	is	on	know	do
16	but	what	he	they
17	have	do	on	er
18	do	**we**	they	was
19	think	now	have	yeah
20	be	was	there	have
21	know	have	no	what
22	if	there	but	he
23	just	like	for	to
24	what	all	be	but
25	for	not	what	for

in context can lead to pragmatic 'failure' (Thomas 1983). The first step in the analysis will be to see what looking at frequency information using corpus linguistic methodology can tell us about the pronoun *we*.

3.1 Frequency

Table 3 illustrates that *we* features prominently in the top 25 words of the C-MELT (position 11), family (position 18) and the British National Corpus (BNC) (position 13) corpora; however, interestingly, it does not feature in the top 25 words of the Limerick Corpus of Irish English (LCIE) corpus. *We* lies just outside the top 25 words in the LCIE, in position 28 (this is a potentially interesting anomaly which it is outside the scope of the current research to investigate). If there was no other agenda, the basis of frequency alone could make this item deserving of attention.

Obviously, C-MELT, the family corpus, LCIE and the BNC are different sizes and represent different types of talk. As already detailed in Table 2, C-MELT is comprised of c.40,000 words and the family corpus, c.12,500 words. As we know, the Limerick Corpus of Irish English is a one-million-word corpus,

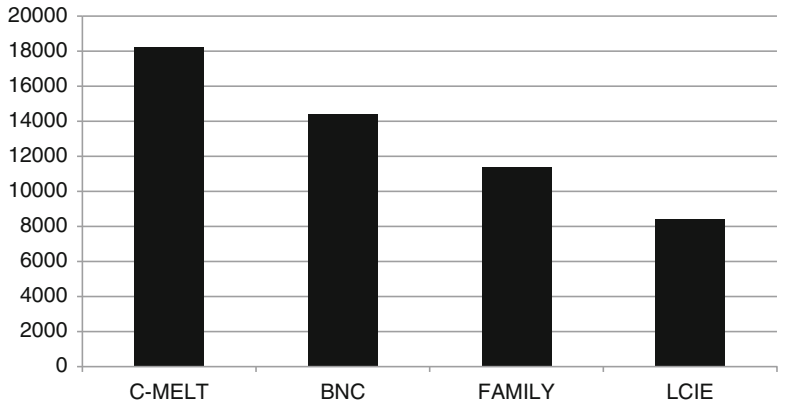


Fig. 1 Distribution of WE across the four corpora (normalised per million words)

whereas the spoken component of the British National Corpus is comprised of ten million words. Therefore, in order to properly compare the frequency of *we* across the four corpora it is necessary to normalise the frequency counts (in this case, *we* is normalised per million words). Additionally, in order to provide a more accurate picture of *we* across the four corpora, Fig. 1 presents the normalised frequency per million words for the lemmatised WE, where WE includes *we*, *we'd*, *we'll*, *we're*, *we've* and *us*.

Figure 1 demonstrates that the lemmatised WE is most frequent per million words in C-MELT, followed by the BNC, the family corpus and LCIE. Of note here is that WE in C-MELT is more than twice as frequent as in LCIE. The reason for this is more than likely related directly to context. As already detailed, C-MELT is a corpus of workplace meetings, a professional context-type. LCIE is predominantly a corpus of informal spoken Irish English, where the casual conversation component accounts for over 70 % of the corpus. This may indicate the importance of WE as a pragmatic item in professional discourse in comparison to informal discourse, such as that between family and friends. This appears to be supported by the fact that WE is 1.6 times more frequent in C-MELT than in the family corpus. In addition, C-MELT is composed of many speakers of different nationalities – Irish, British, American, Jamaican and Ugandan, for example, whereas in LCIE and the family corpus, the speakers are all Irish. Our speculations on why WE occurs with such frequency are inevitably linked to context, but also to intuitions regarding how and why certain identities are indexed in specific communities. Understanding how these identities are indexed is crucial to interpreting how these communities are supported, created and realised pragmatically. In the sections that follow, we will analyse the pragmatics of WE in the context of the family in relation to their inclusive and exclusive WE references, and for the workplace in relation to the indexical ground of WE (cf. Hanks 1992).

3.2 Family Discourse: Inclusive and Exclusive WE

As has been mentioned, person reference, manifest in personal pronouns, is concerned with the orientation to identity of participants in the communicative situation. In order to investigate the identity orientation of family members, it is interesting to examine WE. There were 143 occurrences of the WE lemma in the family corpus, and in order to categorise how WE was being used, each of the 143 occurrences were tagged pragmatically as either referring inclusively to the family itself (inclusive WE) or to some other community external to the family to which the family member speaking belonged (exclusive WE). Thus tagged, it was possible to generate quantitative information on this functional difference in the use of WE. Inclusive WE was found to be notably more frequent than exclusive WE, accounting for 88 % of the occurrences. This, it can be argued, indicates that this family primarily utilise WE to create a perspective of *I*, the speaker + *you*, the addressee(s) in the immediate context. This use of inclusive WE is evident from the following extract (1) from the family corpus. The siblings are in the living room discussing the origins of the name of their dog, Goldie:

(1)

<Son 1>	But Goldie's a girl's name like.
<Daughter 1>	Yeah b = **we** didn't give her the name.
<Son 1>	What?
<Daughter>	<$O> **We** didn't give her the name <\$O>.
<Son 2>	<$O> **We** didn't give her the name <\$O>. Although she was so young she wouldn't notice it.
<Son 1>	She wouldn't have a clue shur.
<Son 2>	**We** could've changed it. **We** could call her am Alex.
<Son 1>	Shit for brains.
<Daughter>	Alex.

Earlier in the conversation, son 1 has been complaining about the dog's name, and suggesting different names for her. The other siblings use *we* (marked in bold) in the repeated utterance *We didn't give her the name* as a form of 'safety in numbers' defence to deflect the criticism of the dog's name from themselves. Mühlhäusler and Harré (1990: 174) claim that in this integrative use of *we*, 'the social bonding aspect and the establishment of solidarity is of importance.' The siblings create an in-group, 'we the family', in opposition to the person who originally named the dog. Further to this, son 2 adds *We could've changed it. We could call her am Alex*, reaffirms this solidarity by invoking the power that the family had, and still have, to change the name of the dog should they choose to do so. In contrast, exclusive WE accounts for just 17 of 143 instances in the family corpus, or 12 % of instances. Exclusive WE in the family corpus refers to a range of outgroups (marked in bold and underlined to the left) and these are illustrated in extracts (2)–(5):

(2)

Friends	**<Son>**	Yeah but the=or they often say members and regulars. But a bouncer would just turn around to you if you said anything like that and go they're members.
	<Daughter>	Mm. Because one night **we** were goin right and **we** got stopped. Another two got in in front of us and **we** said what oh they're gold cards.

(3)

Workplace	**<Daughter>**	**We** have them outside too the 80 mini bulbs. Is that what they are? Eighty mini bulbs <$G3> yeah **we**'ve them too.

(4)

University	**<Son>**	Are you doin corpus stuff?
	<Daughter>	Ah **we** hit at it last semester like.

(5)

Limerick	**<Son>**	+aren't **we** already twinned with Quimper?
	<Daughter>	It's in France.
	<Son>	Yeah.

Exclusive WE demonstrates that the family, in addition to identifying themselves as members of their family community, also identify themselves as members of a wider Irish society. This finding may indicate the nature of the different identities around which members of the families can construct their reference system. The family in this study have several 'pivots', around which to organise reference such as other communities to which they belong, for example, family, friends, the workplace or education. By invoking inclusive WE, the family is simultaneously defining its identity. The fact that the members of the family are involved in 'we' identities external to the family indicates interaction with a broader society. In findings reported elsewhere (see Clancy 2011a, b), where family discourse representing a different ethnic and socio-economic grouping in the Republic of Ireland was compared with the family discourse described above, the use or non-use of pragmatic items has been shown to have implications in terms of access or non-access to the dominant culture in Irish society.

3.3 Workplace Discourse: The Indexical Ground of WE

Moving now to a very different speech context, the workplace. WE is again tagged pragmatically in terms of reference. This time, relying on a distinction such as 'inclusive' and 'exclusive' WE does not cover the plethora of referents within the discourse. While it is absolutely true to say that WE operates inclusively and exclusively, there are multiple inclusive and exclusive WE identities indexed, and therefore further classification and categorisation was necessary. This was done in order to trace the interactional 'footing' (Goffman 1979) displayed by participants, get at their various roles in the discourse (Wortham 1996), and thus delineate the 'participant framework' (Goffman 1979). As Wortham (1996: 332) points out, 'acculturated individuals' come to expect standard participation frameworks in given situations, and this, obviously, has resonance in terms of understanding how the individuals in the workplace operate as a community. Borthen (2010: 1809), in quite a different study admittedly, has noted that '… the capacity of human beings to pragmatically enrich utterances with a seemingly sparse semantics should not be underestimated' and this certainly holds true for this data set. Bargiela-Chiappini and Harris observe a very instrumental function of *we* in the institutional context which makes it an interesting item to study: 'in a professional business setting, negotiating between "I" as an individual and some form of collective identity "we" is an everyday matter involving tactical choices, whether conscious or unconscious' (1997: 175). As part of a more general exploration of pronominal reference, the referents contained in the lemma WE were investigated in context. While multiple referents were identified, it was possible to apply a generic taxonomy for quantitative purposes, and distinguish and tag the following referents in WE:

1. Professional ([PROF]): WE as professionals, for example, in the classroom with our students; this use of WE related specifically to language teaching and its practices;
2. Departmental/Subgroup ([DEPT]/[SUB]): A local, situated WE ([DEPT]) which referred to the group of teachers in the department as part of, or as distinct from, the university as an institution. This superordinate WE [DEPT] was subdivided ([SUB]) where teachers referred to themselves in relation to particular subgroups they were part of, such as subgroups teaching different proficiency levels, or working groups set up for other purposes;
3. Procedural ([MEET]): A procedural use emerging from the speech situation itself, the meeting, and referring to everyone in the room at that point in time as a participant in the meeting;
4. Other ([OTHER]): The 'other' category held occurrences such as fixed phrases, e.g., *a bit of both as we say in Ireland*, which in this case also indexes an exclusive use of WE which refers to a national grouping not shared by all of the speakers present.

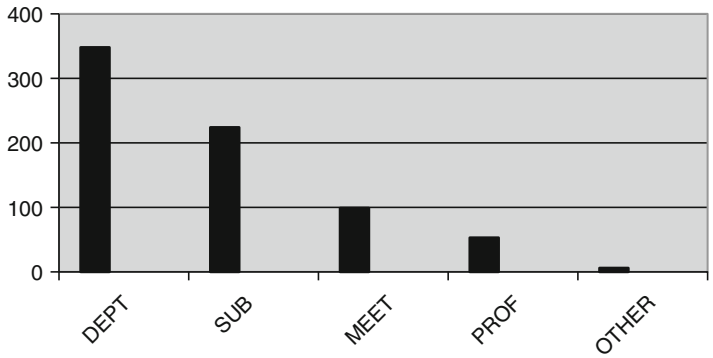


Fig. 2 WE by reference in context

By tagging the reference of WE, it was possible to generate some data about how frequently each identity was indexed (see Fig. 2), and, on the basis of these results, note patterns and problematise why these patterns might occur.

An interesting pattern here is for WE to refer primarily to the institutional context – to the fact that the teachers are members of a department or engaged in a meeting – rather than the broader professional context of the enterprise they are engaged in (being English language teachers). As other personal pronouns, including YOU, were also analysed, a potential explanation for this can also be offered. A similar process of reference retrieval was conducted for the lemma YOU (this lemma includes singular and plural reference, but for the purpose of the study reported here excluded fixed pragmatic clusters such as *you know* and *you see* which were analysed separately). Obviously, singular and plural entity reference was implicated, as well as specific addressee and generic reference. The generic use was interesting for its strong tendency to signal a generic, impersonal reference to the professional/teacher, e.g. in the classroom with students. In other words, in a way that mirrored the [PROF] category of WE, but is somehow qualitatively different in context. This extract (6) from the C-MELT corpus gives a brief view of the professional YOU in context:

(6)
[Kate is reporting on a pilot course she taught on the previous semester, specifically how she put together a syllabus for the class in the absence of a specific textbook]

Kate: But em in that in that kind of respect there was no focus. So the classes developed according to what the students wanted to do and what they needed to do and what as the classes went by what **you [PROF]** could perceive that they needed to do and what they asked for themselves. Basically so they they the course kind of grew as opposed to was there initially.

We can argue that by invoking the YOU [PROF] reference here Kate is inviting engagement and alignment with her process in teaching this class, and suggesting that

any professional would recognise this type of organic, responsive syllabus planning for a pilot course. In a broader sense, this YOU can also more safely stake out and reference professional common ground where WE might be slightly more face-threatening. The use of YOU facilitates both an invitation to render subjective judgements shared knowledge, but also, crucially in this case, a way of providing a sufficiently distant 'professional footing' (Vaughan 2009). Additionally, the speaker may be pre-emptively staving off any criticism which might be made of not starting the course with a pre-determined syllabus. The potential of WE and YOU (and, indeed, all the other personal pronouns) to do this sort of complex pragmatic work makes them a rich area for investigation.

We contend that small, domain-specific corpora provide a rich resource for investigating the pragmatic systems of different communities in detail, and here a corpus-based investigation revealed the high frequency of personal pronouns in general. Our broader focus on the idea of 'community' and 'identity', with the attendant questions about how these are manifested linguistically, led us to a focus on isolating and categorising instances of *we* as (arguably) a linguistic proxy for both. What is striking is how the complexity of reference in a potentially loaded item such as *we* can resolve itself when investigated in context. This reflects the canonical concerns of pragmatics as a discipline, and through the use of a small corpus that can be tagged, in this case in terms of sphere of reference, the pragmatic nuances of how an item can be explored quantitatively as well. The broad framework used – that of identifying and isolating an 'inclusive' and 'exclusive' WE – held across both corpora, though required more elaboration in relation to the workplace context, which raises the question of why this may be the case. We suggest that this is related to the nature of the communities: the family community's use of WE operates to define itself and its own identity, through explicitly identifying the out-groups that contrast to the core in-group. In the case of the workplace community, the pragmatic work that WE does becomes ever more complicated: in a context where the members of the community do not share the same closeness as the family, WE is required not only as an expression of the community's identity, defining its in-groups and out-groups (and hence the parameters of the community), but must also to perform more complex functions in relation to politeness. Clancy (2011b) has demonstrated how the family represents a kind of politeness 'ground zero' (after Levinson 2004). The findings for the case studies reported above in relation to WE would appear to bear this out, showing how the referential potential of a single item is complex within the family itself, and how this complexity multiplies in another, different, context, the workplace.

4 Summary and Conclusions

In the two case studies summarised above, we have focussed on WE in relation to its intriguing 'complexity with regard to personal, social and other deixis' (Mühlhäusler and Harré 1990: 47), and a number of points can be made now by way of summing up our observations, and underlining the case for using small corpora for investigating

pragmatic phenomena. Firstly, although there are various conceptions of what 'small' might mean, as we have shown, many small corpora successfully used in the analysis of pragmatic features appear to be in the 20,000–50,000 word range. Their stance on what constitutes 'balance' and/or 'representativeness' differ from how these were traditionally perceived; however, these corpora can be argued to be both balanced and representative in terms of the speech situations they are designed to characterise. As the literature reported above illustrates, small corpora are eminently suitable for investigating phenomena in context given the constant interpretative dialectic between features of texts and the contexts in which they are produced. Another benefit of using small corpora to do empirical pragmatic research is that the results produced are manageable. In the two case studies reported, it was feasible to isolate each instance of the feature under investigation and assign it a pragmatic tag, which was in turn used to generate quantitative results. This is possible because the small corpus researchers had access to comprehensive metadata and other background knowledge of the context.

That is not to say that corpus-based research in pragmatics is without its difficulties. It is relatively straightforward to search a corpus for an item with pragmatic potential if we can connect that potential with a linguistic form or forms (as in Sect. 3 we connect personal pronouns and the pragmatic management of identity within communities). Research has shown that investigating speech acts, such as apologies, thanks or requests, is a more difficult process. Archer et al. (2012) point out that a weakness associated with using corpora for speech act research lies in the difficulties in automatically retrieving all the linguistic manifestations of a particular speech act, or identifying an appropriate 'lexical hook', to use Rühlemann's phrase (2010: 290), for extracting quantitative information. As Jautz (2008: 147) observes in relation to the speech act of thanking 'it is difficult…to investigate phenomena above the level of the word or phrase in corpora…since corpora are not (yet) tagged for speech acts, it is not possible to search for all instances of gratitude in a speech act theoretical sense.' To an extent, these reported limitations can be mitigated by using a small corpus: a speech situation in which these acts are likely to occur can be identified, data collected, and a corpus it is possible to manually tag compiled. In fact, given that pragmatic phenomena are extremely context-sensitive and occasionally completely resistant to automatic retrieval, we should accept that larger corpora are simply not suitable for some of our purposes, despite the volume of potential data they contain. The middle ground lies in the design and exploitation of small corpora for pragmatic research.

References

Archer, D., K. Aijmer, and A. Wichmann. 2012. *Pragmatics: An advanced resource book for students*. London: Routledge.

Aston, G. 1997. Large and small corpora in language learning. In *PALC97: Practical applications in language corpora*, ed. B. Lewandowska-Tomaszczyk and P.J. Melia, 51–62. Łodz: Łodz University Press.

Atkins, S., J. Clear, and N. Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7(1): 1–16.

Bargiela-Chiappini, F., and S. Harris. 1997. *Managing language: The discourse of corporate meetings*. Amsterdam: John Benjamins.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4): 243–257.

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

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

Borthen, K. 2010. On how we interpret plural pronouns. *Journal of Pragmatics* 42(7): 1799–1815.

Clancy, B. 2010. Building a corpus to represent a variety of language. In *The Routledge handbook of corpus linguistics*, ed. A. O'Keeffe and M. McCarthy, 80–92. London: Routledge.

Clancy, B. 2011a. Complementary perspectives on hedging behaviour in family discourse: The analytical synergy of variational pragmatics and corpus linguistics. *International Journal of Corpus Linguistics* 16(3): 371–390.

Clancy, B. 2011b. *Do you want to do it yourself like?* Hedging in Irish traveller and settled family discourse. In *Situated politeness*, ed. B. Davies, M. Haugh, and A. Merrison, 129–146. London: Continuum.

Clancy, B., and E. Vaughan (2012). *It's lunacy now*: A corpus-based pragmatic analysis of the use of *now* in contemporary Irish English. In *New perspectives on Irish English*, ed. B. Migge and M. Ní Choisáin, 225–246. Amsterdam: John Benjamins.

Clear, J. 1992. Corpus sampling. In *New directions in english language corpus methodology*, ed. G. Leitner, 21–31. Berlin: Mouton de Gruyter.

Cutting, J. 2001. The speech acts of the in-group. *Journal of Pragmatics* 33(8): 1207–1233.

De Fina, A., D. Schiffrin, and M. Bamberg (eds.). 2006. *Discourse and identity*. Cambridge: Cambridge University Press.

Farr, F. 2005. Relational strategies in the discourse of professional performance review in an Irish academic environment: The case of language teacher education. In *The pragmatics of Irish English*, ed. A. Barron and K. Schneider, 203–234. Berlin: Mouton de Gruyter.

Farr, F. 2007. Spoken language analysis as an aid to reflective practice in language teacher education: Using a specialised corpus to establish a genetic fingerprint. In *Spoken corpora in applied linguistics*, ed. M.C. Campoy and M.J. Luzón, 235–258. Bern: Peter Lang.

Fasulo, A., and C. Zucchermaglio. 2002. My selves and I: Identity markers in work meeting talk. *Journal of Pragmatics* 34(9): 1119–1144.

Flowerdew, L. 2002. Corpus-based analyses in EAP. In *Academic discourse*, ed. J. Flowerdew, 95–114. London: Longman.

Flowerdew, L. 2004. The argument for using English specialised corpora to understand academic and professional settings. In *Discourse in the professions: Perspectives from corpus linguistics*, ed. U. Connor and T. Upton, 11–33. Amsterdam: John Benjamins.

Goffman, E. 1979. Footing. *Semiotica* 25: 1–29.

Goffman, E. 1981. *Forms of talk*. Oxford: Blackwell.

Hanks, W. 1992. The indexical ground of deictic reference. In *Rethinking context. Language as an interactive phenomenon*, ed. A. Duranti and C. Goodwin, 43–77. Cambridge: Cambridge University Press.

Holmes, J. 1988. Doubt and certainty in ESL textbooks. *Applied Linguistics* 9(1): 21–44.

Íñigo-Mora, I. 2004. On the use of the personal pronoun we in communities. *Journal of Language and Politics* 3(1): 27–52.

Jautz, S. 2008. Gratitude in British and New Zealand radio programmes: Nothing but gushing? In *Variational pragmatics: A focus on regional varieties in pluricentric languages*, ed. K. Schneider and A. Barron, 141–178. Amsterdam: John Benjamins.

Knight, D., D. Evans, R. Carter, and S. Adolphs. 2009. HeadTalk, HandTalk and the corpus: Towards a framework for multi-modal, multi-media corpus development. *Corpora* 4(1): 1–32.

Koester, A. 2006. *Investigating workplace discourse*. London: Routledge.

Koester, A. 2010. Building small specialised corpora. In *The Routledge handbook of corpus linguistics*, ed. A. O'Keeffe and M. McCarthy, 66–79. London: Routledge.

Lave, J., and E. Wenger. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.
Levinson, S. 1983. *Pragmatics*. Cambridge: Cambridge University Press.
Levinson, S. 2004. Deixis. In *The handbook of pragmatics*, ed. L. Horn and G. Ward, 97–121. Oxford: Blackwell.
Loudermilk, B.C. 2007. Occluded academic genres: An analysis of the MBA thought essay. *English for Academic Purposes* 6(3): 190–205.
McCarthy, M. 1998. *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.
McCarthy, M., and A. O'Keeffe. 2010. Historical perspective: What are corpora and how have they evolved? In *The Routledge handbook of corpus linguistics*, ed. A. O'Keeffe and M. McCarthy, 3–13. London: Routledge.
McEnery, T., R. Xiao, and Y. Tono. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.
Mühlhäusler, P., and R. Harré. 1990. *Pronouns and people: The linguistic construction of social and personal identity*. Oxford: Blackwell.
O'Connor, P. 1994. "You could feel it through the skin": Agency and positioning in prisoners' stabbing stories. *Text* 14(1): 45–75.
O'Keeffe, A. 2005. You've a daughter yourself? A corpus-based look at question forms in an Irish radio phone-in. In *The pragmatics of Irish English*, ed. A. Barron and K. Schneider, 339–366. Berlin: Mouton de Gruyter.
O'Keeffe, A., and S. Adolphs. 2008. Response tokens in British and Irish discourse: Corpus, context and variational pragmatics. In *Variational pragmatics: A focus on regional varieties in pluricentric languages*, ed. K. Schneider and A. Barron, 69–98. Amsterdam: John Benjamins.
Orpin, D. 2005. Corpus linguistics and critical discourse analysis: Examining the ideology of sleaze. *International Journal of Corpus Linguistics* 10(1): 37–61.
Patrick, P. 2002. The speech community. In *The handbook of language variation and change*, ed. J.K. Chambers, P. Trudgill, and N. Schilling-Estes, 573–597. Oxford: Blackwell.
Pennycook, A. 1994. The politics of pronouns. *ELT Journal* 48(2): 173–178.
Rees, A. 1983. *Pronouns of person and power: A study of personal pronouns in public discourse*. Unpublished MA dissertation, Sheffield University.
Romero-Trillo, J. (ed.). 2008. *Corpus linguistics and pragmatics: A mutualistic entente*. Berlin: Mouton de Gruyter.
Rühlemann, C. 2010. What can a corpus tell us about pragmatics? In *The Routledge handbook of corpus linguistics*, ed. A. O'Keeffe and M. McCarthy, 288–301. London: Routledge.
Rundell, M. 2008. The corpus revolution revisited. *English Today* 24(1): 23–27.
Sankoff, D. 1988. Problems of representativeness. In *Sociolinguistics: An international handbook of the science of language and society*, ed. U. Ammon, N. Dittmar, and K. Mattheier, 899–903. Berlin: Walter de Gruyter.
Sinclair, J.M. 2001. Preface. In *Small corpus studies and ELT: Theory and practice*, ed. M. Ghadessy, A. Henry, and R.L. Roseberry, vii–xv. Amsterdam: John Benjamins.
Sinclair, J. 2005. Corpus and text: Basic principles. In *Developing linguistic corpora: A guide to good practice*, ed. M. Wynne, 1–16. Oxford: Oxbow Books. Available online at http://ota.ahds.ac.uk/documents/creating/dlc/chapter1.htm. Date accessed 25 June 2012.
Stirling, L., and L. Manderson. 2011. About you: Empathy, objectivity and authority. *Journal of Pragmatics* 43(6): 1581–1602.
Stubbs, M. 2004. Language corpora. In *The handbook of applied linguistics*, ed. A. Davies and C. Elder, 106–132. Oxford: Blackwell.
Swales, J.M. 1990. *Genre analysis: English and academic research settings*. Cambridge: Cambridge University Press.
Swales, J. 1996. Occluded genres in the academy: The case of the submission letter. In *Academic writing: Intercultural and textual issues*, ed. E. Ventola and A. Mauranen, 45–58. Amsterdam: John Benjamins.

Tagliamonte, S. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.

Tannen, D. 2007. Talking the dog: Framing pets as interactional resources in family discourse. In *Family talk: Discourse and identity in four American families*, ed. D. Tannen, S. Kendall, and C. Gordon, 49–70. New York: Oxford University Press.

Thomas, J. 1983. Cross-cultural pragmatic failure. *Applied Linguistics* 4(2): 91–112.

Tognini-Bonelli, E. 2010. Theoretical overview of the evolution of corpus linguistics. In *The Routledge handbook of corpus linguistics*, ed. A. O'Keeffe and M. McCarthy, 14–27. London: Routledge.

Torgersen, E.N., C. Gabrielatos, S. Hoffmann, and S. Fox. 2011. A corpus-based study of pragmatic markers in London English. *Corpus Linguistics and Linguistic Theory* 7(1): 93–118.

Tracy, K. 2002. *Everyday talk: Building and reflecting identities*. New York: Guilford.

Vaughan, E. 2007. *I think we should just accept…our horrible lowly status*: Analysing teacher-teacher talk within the context of community of practice. *Language Awareness* 16(3): 173–189.

Vaughan, E. 2008. "Got a date or something?": An analysis of the role of humour and laughter in the workplace meetings of English language teachers. In *Corpora and discourse: The challenge of different settings*, ed. A. Ädel and R. Reppen, 95–115. Amsterdam: John Benjamins.

Vaughan, E. 2009. *Just say something and we can all argue then: Community and identity in the workplace talk of English language teachers*. Unpublished Ph.D. thesis, University of Limerick.

Vaughan, E., and B. Clancy. 2011. The pragmatics of Irish English. *English Today* 27(2): 47–52.

Wenger, E. 1998. *Communities of practice learning. Meaning and identity*. Cambridge: Cambridge University Press.

Whitley, M.S. 1978. Person and number in the use of WE, YOU, and THEY. *American Speech* 53(1): 18–39.

Wortham, S. 1996. Mapping participant deictics: A technique for discovering speakers' footing. *Journal of Pragmatics* 25(3): 331–348.

Part II
New Domains for Corpus Linguistics and Pragmatics

Multiword Structures in Different Materials, and with Different Goals and Methodologies

Britt Erman, Margareta Lewis, and Lars Fant

1 Introduction

Combinations of words that fulfill specific functions have come to be known by the term 'formulaic language'. Instantiations of formulaic language have been referred to by different names in the literature, such as formulaic sequences, multiword units, prefabricated patterns. We have opted for the term multiword structures because these instantiations typically have identifiable structural characteristics, e.g. phrases as in *do a jigsaw puzzle*, and full-length clauses as in *Could you give me a hand?*, or functional characteristics such as hesitation markers, for example *sort of, I guess,* and discourse markers as in *so you're saying (that)*.

Research on formulaic language has been performed under different conditions and with different goals. Some approaches focus on specific multiword structures, while others use holistic methods, scanning entire texts for multiword structures. Three current methods are presented and compared from qualitative aspects, such as size of material, amount of manual work involved, control of task, topic and discipline. This is followed by a presentation of a small-scale study using two of these methods, one automatic and one manual, applied to the same material of L1 and L2 English and Spanish. Both methods have been developed for the analysis of entire texts. In the review of literature we will refer to all the instantiations of formulaic language as multiword structures (henceforth MWSs), except where especially indicated.

B. Erman (✉) • M. Lewis
Department of English, Stockholm University, Stockholm, Sweden
e-mail: britt.erman@english.su.se

L. Fant
Department of Spanish and Portuguese, Stockholm University,
Stockholm 10691, Sweden

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_5,

Corpora have been collected for different purposes serving different functions. The review of literature below will in the main concern studies of native and non-native spoken and written *production* within SLA research. We discuss the studies according to: (1) the size of the corpus; (2) the selection of MWSs; (3) the methodology used.

Following a brief overview of the methodology of early frequency-based research on large corpora, studies representing three current methods within SLA are presented. The first of these involves studies of specifically selected MWSs or types of MWSs in speech and writing based on smaller specialized corpora of native and non-native speaker production using the phraseological method. We then present two methods which involve no pre-selection of items to be studied but are applied to entire texts, notably the lexical bundle method, and the comprehensive method. Finally, we describe an empirical small-scale study using these two methods applied to the same material, i.e. the spoken production of advanced non-native Swedish speakers of English and Spanish and English and Spanish native speakers. To our knowledge no previous study has used two different methods applied to the same data. All four groups perform the same task. This entails that topic is controlled for, which is a particular strength of the study, and crucial, in view of the theoretical orientation of the study.

The theoretical stance pervading the empirical study, apparent in choice of primary material as well as methodology, is that it is contexts and situations that trigger the use of MWSs, and ultimately the user's exposure to and experience of these. In other words, the present study belongs to a usage-based and cognitive theoretical framework. The two researchers that have been most explicit about the role of exposure and about restrictions imposed on language through situation and overall context are Nick Ellis (1996) and Michael Hoey (2005), but contributions by Fillmore, Kay and O'Conor through frames and construction grammar should also be acknowledged (1988). According to Ellis *frequency*, *recency* of use and *context* are essential features in all learning (2006: 105). A key concept in Ellis' work is 'chunking' brought about by automatic frequency counts (on the part of the learner) and associative learning, which underlie the acquisition of combinations of sounds, as well as combinations of morphemes and words (Ellis 1996). A key concept in Hoey's psycholinguistic approach is 'lexical priming' involving the interaction between the repeated experience of language in context and the mental lexicon (2005). That is to say, co-occurring words and contextual features become part of the meaning, function and collocational range of the word, all of which are stored in the mental lexicon. Hoey further observes that priming may result in 'nesting', that is "the product of priming becomes itself primed in ways that do not apply to the individual words making up the combination" (2005: 8), and the extended combination is in turn a carrier of the co-text and context where it occurred.

Most of the studies of MWSs of both written and spoken materials have been carried out within SLA, which is also the focus of the present article. This overview will mainly deal with research concerned with native (NS) and non-native (NNS) data, and with methodology rather than results.

2 Forerunners: Concordances, Collocational Frames and Collocation

The starting point for research of concordances, i.e. recurring patterns and co-occurring words in large written corpora, gave rise to the revolutionary insight that language use is largely formulaic. Sinclair (1991) was among the first to show that certain words frequently co-occur and that grammar and vocabulary interact. This insight led him to formulate his idiom principle (Sinclair 1991). Another approach to corpus-based investigations of the interaction between grammar and lexis is captured in the notion of collocational frameworks. Renouf and Sinclair discuss continuous sequences with one or more free slots (1991: 129), showing that words are not random in a given frame but belong to particular semantic classes associated with the framework in question. For instance, *too* + ? + *to* is predominantly filled with non-verbal adjectives such as *easy, good, late* and *young*. Hunston and Francis (2000) developed the notion 'pattern grammar', e.g. patterns emerging around specific sets of verbs, concluding that patterns carry meanings.

To researchers in the frequency-based tradition collocation is a statistical concept, quantifiable as a probability of co-occurrence. Although Firth (1957) was the first to make the term 'collocation' widely known, he did not propose a consistent definition of the term. One of the first attempts in the frequency-based tradition at defining collocation is assigned to Halliday as "the syntagmatic association of lexical items" (1961: 276). More recent definitions include, Stubbs who define collocation as "a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text" (2002: 24), and Lewis as "the way in which words co-occur in natural text in statistically significant ways" (2000: 132). Hoey defines collocation from a psycholinguistic and corpus linguistic perspective as follows: "a psychological association between words (rather than lemmas) up to four words apart and is evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution" (2005: 5).

Altenberg (1998) made a frequency-based study of recurrent combinations of words in a spoken corpus, the London-Lund Corpus of Spoken English. A recurrent word combination was defined as "any continuous string of words occurring more than once in identical form" (1998: 101). Altenberg found that a substantial part of the words in the corpus formed part of recurrent word combinations, such as connectors (*first of all*), sentence stems (*it seems to me that*) (1998: 102), and collocational frameworks (the + N + of, as in *the whole of*).

Granger compared specific word combinations in native and non-native writing using a frequency-based approach, notably adverb-adjective collocations (e.g. *perfectly natural*) and 'pragmatic phrases' (e.g. *it seems (to me) (that) X*) (1998: 147). Granger found that some collocations were more frequently used by the NNSs, especially those that had equivalents in the participants' L1, French (1998: 149). Granger uses the most widely accepted definition of a collocation as "the linguistic phenomenon whereby a given vocabulary item prefers the company of another item rather than its 'synonyms'" (1998: 145). In fact, this definition is at

the core of the phraseological approach, which means that Granger positions herself somewhere between the frequency-based and phraseological methods.

Next we will present research of formulaic language and MWSs still with a focus on native and non-native speaker performance, starting with a selection of studies using the phraseological method.

3 Three Methods Exploring MWSs in SLA

3.1 The Phraseological Method

The majority of phraseological studies involve comparison between native and non-native speakers' written production, using smaller specialized corpora. This account will start by reviewing a selection of studies based on written production, to be followed by a couple of studies on spoken production, most of which are aimed at NS and NNS differences.

Studies using the phraseological method tend to focus on specific MWSs frequently combined with computer-assisted methodology for extracting data. Basing their view of phraseology in the Russian tradition (Vinogradov 1947; Amosova 1963, in Cowie 1998), British phraseologists see combinations of words as ordered on a continuum from free to fixed word combinations (Cowie 1998: 217; Howarth 1998b: 164). As mentioned, the notion of 'substitutability', meaning that at least one of the members should be used in a specialized, restricted sense, precluding the substitutability of a synonymous word, is at the core of the phraseological method. Howarth orders the following five groups of word combinations from free combinations to idioms: 'free collocations' (*blow the trumpet*), 'restricted collocations' (*blow a fuse*), 'figurative idioms' (*blow your own trumpet*), and 'pure idioms' (*blow the gaff*) (Howarth 1998b: 164). The phraseological method is based on traditional grammar and part of speech, and the data are frequently extracted by drawing on part-of-speech-tagged corpora. Many studies using this method thus focus on the production of specific MWSs in the writings of native and non-native speakers, e.g. verb-noun collocations in L2 English (Howarth 1998a, b), and adverb-adjective collocations in L2 English (Granger 1998). Howarth found that the NNS types of collocations deviated from those of the NS group (1998b: 178). Nesselhauf (2003) in her study of non-native written material found that the most decisive factor in the learners' mistakes in the production of English collocations was a strong influence from their L1 (German).

Compared to written corpora spoken corpora are usually quite small. Studies of MWSs in spoken production are either directed towards specific patterns, McCarthy and Carter 2004 (e.g. 'this that and the other'), or on specific markers (Hancock 2000; Denke 2009).

Hancock (2000) studied NNS and NS usage of a number of French markers and connectors (e.g. *mais, donc, parce que*). Following Raupach (1984) Hancock also investigated some formulaic/'prefabricated' 'epistemic phrases' (e.g. *je crois/pense/*

trouve que) and found that these phrases were used by the non-native speakers in connection with hesitation as a production strategy – compensatory fluency – which suggested less familiarity with nativelike uses, such as for organising discourse.

Denke investigated 'pragmatic markers' with different subfunctions (e.g. social functions as repair markers, and textual functions as discourse markers) in L1 and L2 English (*well, I mean, y'know*) in native and non-native oral presentations within engineering and natural science programs at Swedish and British universities (Denke 2009). Discourse markers is thus a subcategory of pragmatic markers in Denke's work. In line with work by Schiffrin (1987), Redeker (1990), and Brinton (1996) Denke sees the discourse marking function as a means of signalling structural units in discourse as well as their boundaries. Denke's main finding was that the NSs use these markers as discourse markers, i.e. in organizing the text, whereas the NNS tend to use them in connection with stalling and repair. In these two studies markers are thus used as fluency devices, and discourse organizers.

3.2 The Lexical Bundle Method

Formulaic language in written academic discourse has been gaining interest in the last decade through the method commonly referred to as the lexical bundle method. This approach is frequently applied to large written corpora, involves no set conditions for selection other than restrictions on length, frequency and dispersion among texts, and is largely computer-driven (Biber et al. 2004; Cortes 2004; Hyland 2008; Römer 2009; Chen and Baker 2010; Ädel and Erman 2012). The only manual component is the assignment of functional categories, and some additional sorting of the data. Biber et al. (2004: 384) describe the three main categories in their functional classification of lexical bundles as follows: *referential bundles* make direct reference through a content word to physical or abstract entities, *stance bundles* express attitudes or assessments of certainty that frame some other proposition, and *discourse organizers* reflect relationships between prior and coming discourse.

Ädel and Erman (2012) studied four-word lexical bundles in two specialized corpora made up of English academic essays by a group of advanced Swedish and a group of English undergraduate students, written within the discipline of linguistics. They found that the native English speakers had a larger number of types of lexical bundles, which were also more varied than those of the Swedish group, such as lexical bundles headed by 'unattended' *this* (e.g. *this shows that the, this is supported by*), and *there*-headed bundles (*there has been a, there is evidence of*). Some, especially lexical bundles with a hedging function, were missing in the NNS material (e.g. *there appears to be, this may be because*). Preposition-headed bundles were used differently in the two groups. The natives used *in*-bundles with abstract nouns (*in an attempt to*), whereas the non-natives' *in*-bundles usually involved concrete nouns (*in this essay we, in the table above*).

In sum, certain lexical bundles are underused or used differently by non-native writers, while others are overused, such as generalizing bundles as in *all over the world* reported by Chen and Baker (2010: 30) less appropriate for academic discourse. Groom investigated lexical bundles involving frequent prepositions for two groups of Swedish writers of L2 English, one group having spent less than 1 month and one having spent more than 12 months in the target language country. The results of his quantitative method showed that time spent in an L2 environment had an effect on the results.

Inspired by above all Altenberg's (1998) work on recurrent word combinations in spoken English in NS speech (in the London-Lund Corpus of Spoken English), De Cock (2004) made a large-scale corpus-driven, frequency-based study exploring recurrent two-, three-, four-, five-, and six-word sequences occurring at least 12, 6, 4, 3, and 3 times, respectively, in NS and NNS speech. De Cock's study showed, among other things, that there was a lack of certain prefabricated sequences in speech, 'vagueness tags' (*and things, and stuff (like that), (and) that sort of thing)* in the learners' production. De Cock (2004) interprets this lack as a failure on the part of the learner to master the informal register.

3.3 The Comprehensive Method

In the comprehensive method MWSs are identified according to expectations raised in the context and situation at hand, inspired by Mel'cuk (1998) and Hoey (2005). The comprehensive method is holistic and involves the scanning of entire texts for detection of MWSs. In this method the MWSs are extracted manually, and as a first step checked in reference corpora, and dictionaries. However, it is the overall context and situation that will ultimately decide whether they should be included. Furthermore, the comprehensive method takes a maximalist perspective, that is two (or more) contingent MWSs adequate and expected in the context are marked as one unit (*further down + the production line* in the context of the film 'Modern Times'), and modifications are considered part of the MWS if they are expected (*get into a (terrible) mess, make (all kinds of) gestures*). Forsberg (2008) made a comprehensive study of quantity and distribution of MWSs over categories in NS and NNS spoken French, which she combined with a phraseological methodology using substitutability tests for inclusion. Following Forsberg's methodology in the 2008 paper, Forsberg and Fant (2010) investigated L2 French and L2 Spanish and native controls. Wiktorsson (2003) and Lewis (2009) used the comprehensive method to establish the quantity and distribution of MWSs over categories in NS and NNS written material. All four studies referred to above found that the MWSs were used more frequently and with more variation by NSs than the NNSs and that the high-proficient NNSs used larger quantities than the low-proficient NNSs.

4 Comparison Between the Phraseological, Lexical Bundles and Comprehensive Methods: Time-Economy and Quality

We will start this section by pointing to some general differences between the three methods. The main difference on a general level between the comprehensive and lexical bundle methods on the one hand, and the phraseological method on the other, is that the former two are applied to entire texts, whereas the latter is frequently applied to pre-selected types of MWSs. Furthermore, the comprehensive method, as the name suggests, entails a holistic approach by including various subtypes and sub-categories of MWSs, and is maximalist by including modifications and extensions appropriate in the context.

4.1 Time-Economy and Quality

Time-economy and quality are naturally important in all research. In view of this, any method will have advantages as well as disadvantages. This should not be seen in a binary fashion, and it will be shown that methods are complementary in some respects, but not in others. This part will address a selection of the works cited in the previous sections.

Of the three methods the lexical bundle and comprehensive methods are at the extreme end points on a time-economy scale with the phraseological method ranging in between these two. Extraction of lexical bundles is automatic and therefore quick, the only manual component being assignment of function, and various sorting processes, some of which can be done using appropriate software. The phraseological method is frequently applied to tagged corpora, which allows computerized extraction. The manual component involves ascertaining phraseological status of the items extracted (Howarth 1998a, b; Granger 1998; Nesselhauf 2003). In the comprehensive method, too, assigning multi-word status to candidate MWSs is done manually, whereas, like in the lexical bundle method, compilation, sorting and comparison of data once retrieved and categorized may be supported by software, such as the Microsoft software Excel, which facilitates all further analyses. All in all, of the three methods, the comprehensive method involves the most manual work and is thus the most time-consuming method.

The qualitative aspects of the research are considered in terms of the aims of the study, the size of the corpus or corpora, the number and nationality of participants, the medium (spoken or written), and the extent to which variables such as task and topic have been controlled for. The next section will discuss these questions, winding up by pointing to advantages and disadvantages of the three methods under scrutiny.

4.2 Qualitative Aspects: The Phraseological Method

Howarth (1998a, b) used two different written corpora, one native and one non-native, collected in different ways. The NS corpus comprised 238,000 words from two corpora, the Lancaster-Oslo-Bergen (LOB) Corpus (58,000 words, 29 texts), and a collection of texts donated by staff of Leeds University (180,000 words; number of texts unknown) from which all verb-noun combinations were extracted. There were high frequency restrictions on the verbs (at least ten instances) for inclusion. The NNS corpus comprised 25,000 words over ten texts from which verb-noun collocations were extracted manually. Howarth's NS corpus is thus of a considerable size, whereas the NNS corpus is of moderate size. The NNS group is very heterogeneous consisting of a variety of different nationalities, and the texts largely within the disciplines of linguistics and language teaching. The two NS groups are homogenous in that they have the same L1, whereas the texts are more varied from different disciplines within the social sciences and law. The main objective of Howarth's study was to conduct "an empirical study of non-native academic writing aimed at identifying and analyzing non-standard phraseology" (1998b: 163), because "[non-standard forms] /…/ can /…/ help in understanding what a non-native has done on a particular occasion" (Howarth 1998a: 39).

Granger (1998) used two large written corpora for her study: the International Corpus of Learner English (ICLE) totaling about 250,000 words for the NNS part of the study (French-speaking learners of English), consisting of literature exam papers and argumentative essays on a variety of topics. The NS part of the study included the Louvain essay corpus, the student essay component of the International Corpus of English (ICE) and the Belles Lettres category of the LOB corpus, totaling about 235,000 words, consisting of argumentative and general essays on different topics, and literature exam papers. There is no indication of the number of texts used. We can conclude that the two corpora in Granger's study are of considerable size, and that there is some control of genre but little or no control of topic. Granger's hypothesis regarding the adverb-adjective collocations was that "learners would make less use of /…/ conventionalized language in their writing than their native counterparts" (1998: 146). She also predicted that their L1 (French) would have an impact on the results. This last point is apparent also in Nesselhauf's study (2003) of German learners of English. Nesselhauf restricts her investigation to non-native writers, basing her study on verb-noun collocations extracted manually from a corpus of 32 argumentative essays from the German sub-corpus of ICLE of an average length of 500 words, totaling 16,000 words on a variety of topics. All three studies above are directed towards finding errors and non-standard phraseological usage in learner writing.

4.3 Qualitative Aspects: The Lexical Bundle Method

As mentioned, what distinguishes this method from the phraseological and comprehensive methods is that it is computer-driven. This means that there is theoretically no upper limit as to the size of the corpus or corpora used. Nevertheless, most

studies of lexical bundles within SLA have used specialized corpora, which can still be sizeable. Chen and Baker's study (2010) of English language four-word lexical bundles involved three groups of writers, one English native expert group drawn from Freiburg-Lancaster-Oslo/Bergen corpus, FLOB-J (academic prose including articles and book chapters) involving 80 texts, one English native peer group extracted from the British Academic Written English (BAWE) corpus, BAWE-EN (an English sub-corpus) consisting of 60 essays, and one non-native group from BAWE-CH (a Chinese sub-corpus), 53 essays. The size of each finalized corpus was around 150,000 words, totaling 450,000 words, and covering a wide range of disciplines including arts and humanities, life sciences, physical sciences and social sciences. Chen and Baker thus used sizeable corpora, and had some control of genre, but no control of discipline or topic. A great deal of lexical bundle research has been directed towards academic writing. Cortes (2004) compared novice student writing to published academic writing and found that there were differences in the manner lexical bundles were used. Following Cortes (2004) and Hyland (2008) Chen and Baker's study aims to establish lexical bundle usage at different levels of writing proficiency, published academic prose compared to student writing. Chen and Baker had the additional aim of comparing learner and native undergraduate writing. Published academic writing turned out to have the widest range of lexical bundles while L2 student writing (the Chinese sub-corpus) showed the smallest range.

Considerably larger corpora were used in Ädel and Erman's study (2012) of English language four-word lexical bundles in advanced learner writing by speakers of L1 Swedish from an English department at a Swedish university and in comparable native-speaker writing from a linguistics department at a British university. The Swedish corpus consisted of 836,200 words over 243 texts, and the English corpus was about a quarter of that size, 248,000 words over 82 texts, totaling over 1,000,000 words. All the texts were thus produced by undergraduate university students in the discipline of linguistics. This is an advantage in view of the fact that there is great variation between disciplines (Hyland 2008). The aim of Ädel and Erman's study was to compare the usage of English lexical bundles in academic writing in two student groups, one non-native advanced group and one native group. Unlike several earlier studies in lexical bundle research both genre and discipline were controlled for.

4.4 Qualitative Aspects: The Comprehensive Method

Wiktorsson (2003) made a study of the written production of three groups of informants, two Swedish groups, and one native English group. The material consisted of 30 essays by students at Upper Secondary level, 19 essays from the Swedish – corpus of the ICLE (SWICLE), and 16 essays from the native English sub-corpus of the ICLE (LOCNESS), totaling around 32,000 words, a good 10,000 words per group. Wiktorsson's study is comparative and cross-sectional, and does not control for topic. Her main focus was to compare the numbers and

types of MWSs ('prefabs') in her data, first of all between native and non-native speakers, and secondly, at two different levels of advancement. Lewis (2009) studied MWSs in compositions by a group of students at Upper Secondary level, 19 students, around 5,000 words, and a group of native speakers, 13, around 5,000 words, totaling a good 10,000 words. Topic was controlled for, and there were three L1s involved, although the majority were Swedish speakers (11 Swedish, 4 Farsi and 4 Spanish speakers). Lewis' aim was to explore differences between the two groups over numbers and types of MWSs, and within the NNS group to determine correlation with grade.

Forsberg (2008) investigated the spoken production in semi-directed interviews extracted from four corpora of Swedish speakers of L2 French at different acquisitional levels, involving a group of early Swedish learners, a group of students at Upper Secondary level studying French, third term university students of L2 French, and one group living and working in France (average 6.3 years). These are compared with two groups of French native speakers, one group of exchange students studying at a Swedish University within the Erasmus program, and one group living in France (Paris). Each group consisted of 6 participants, totaling 36 participants and between 10,000 and 20,000 words per corpus, yielding a total of around 90,000 words. Like Wiktorsson (2003) Forsberg's study is comparative aimed at comparing the numbers and types of MWSs ('séquences préfabriquées') in the speech of native and non-native speakers, but its main objective is to examine how the use of MWSs develops over time in L2 French, i.e. comparing the number and types of MWSs at different acquisitional levels.

In conclusion, even a brief glance at the literature on MWSs has revealed that there is a bias with regard to studies of MWSs towards written material, MWSs in spoken production thus being clearly underrepresented. This is presumably explained by the collection of speech data being time-consuming and costly, involving recorded material, which has to be transcribed to be turned into searchable texts.

4.5 Main Points of Comparison Between the Three Methods

We will focus on four questions that will get different answers depending on the method used. We wind up this section by mentioning advantages and disadvantages of each of the three methods.

What Criteria of MWS Identification are Applied? The phraseological method is focused on the system by using strict semantic criteria for assigning phraseological status to combinations of words. As a consequence, many adequate word combinations are filtered away. In the comprehensive method the context and co-text of a word combination will ultimately decide its multi-word status. One of the core ideas in the comprehensive method is bringing in the notions of expectation and overall situation, which makes this method the most usage-based of the three. The comprehensive and lexical bundle methods are more exploratory than the phraseological method by including any word

combination that satisfies the criteria applied, but with theoretically no upper limit with regard to length. As a consequence these methods can also display patterns of different sizes and functions, some of which may be quite informative. However, both these methods could be said to include at the same time too much and too little although in different ways. The comprehensive method, having a subjective component, cannot claim to have included all, and only, MWSs. The lexical bundle method includes too much because some lexical bundles, although meeting the criteria for frequency and dispersion, have neither structure nor function, (e.g. *again and they, oh what is*), and too little because the method does not capture syntagmatic relations that are variable in terms of modification (*to a (very) large extent*).

How does the Question of Time and Economy Relate to these Methods? The comprehensive method is the most time-consuming of the three methods, foremost because it applies to entire texts, does not involve pre-selection of targeted MWSs, and because a large part is done manually. The lexical bundle framework is appealing in that it is essentially computer-driven and that the retrieval can be fully automatized. This method involves very little manual work, once the data have been prepared for computerized searches. Scanning entire texts as within the lexical bundle and comprehensive frameworks, involves for spoken language turning recorded material into a searchable format, which is a time-consuming and costly procedure. This explains why spoken corpora are small compared to written corpora. The phraseological method involving manual as well as automatic procedures and often applied to written material ranges somewhere between those extreme points on a time scale.

What are the Main Aims and Objectives of Studies Using the Different Methods? The phraseological method has been used to detect idiosyncrasies and non-standard word combinations in L2 usage mainly in written material. The comprehensive method is targeted towards adequate word combinations given the context, topic, co-text and overall frame in L2 usage at different acquisitional levels. The lexical bundle method is used to detect frequent building blocks in the construction of discourse. It has been applied to a variety of material, foremost NS and NNS but also expert and non-expert written production of academic texts. As mentioned, this makes the lexical bundle and comprehensive methods more exploratory and more usage-based compared to the phraseological method. All three methods use native data for control.

What Variables have been Controlled in Studies Using the Three Methods? We have seen above that some studies using the phraseological and lexical bundle methods within SLA have little control of topic (Granger 1998; Howarth 1998b; Nesselhauf 2003; Wiktorsson 2003; Chen and Baker 2010) and background languages of the L2 participants (Howarth 1998b), which might skew the results. This contrasts with several studies using the comprehensive method (Forsberg 2008; Lewis 2009; Erman 2009).

What Advantages and Disadvantages can be Discerned in each of the three Methods? All methods will have advantages and disadvantages. We have pointed to a couple of these, which will be repeated here. One advantage of the phraseological

method is that the MWSs investigated are established, conventionalized word combinations. However, the criteria for inclusion are quite strict and many appropriate word combinations are filtered away. Appealing characteristics of the lexical bundle method are that it is automatic, objective and effective. However, the method is by definition too inclusive, and, although some bundles can be very informative and suggestive, quite a few will be left without further analysis. An advantage of the comprehensive method is that it captures a language user's repertoire as well as choice of expression triggered by the situation, context and topic. Both the lexical bundle and the comprehensive methods allow multiword patterns to be displayed, which is a particular strength considering the exploratory character of these two methods. A disadvantage of the comprehensive method is that it is time-consuming, something that has to be taken into account in the planning of larger projects. The important thing is to choose a method that can help the researcher not only to get answers to the questions addressed, but also to open up for new questions and perspectives. We believe that the lexical bundles and comprehensive methods do this but in different ways.

Finally, regardless of whether MWSs are selected manually, or automatically, and regardless of method used, the results frequently converge so that native speakers use larger quantities of MWS types than non-native speakers, expert writers use larger quantities of MWS types than non-expert writers, and high-level proficient L2 users use larger quantities of MWS types than low-level proficient L2 users.

In the next section we will account for an empirical study comparing the comprehensive and the lexical bundle methods applied to the same material, i.e. transcriptions of recorded on-line retelling of a film clip involving very advanced Swedish long-residency L2 English and L2 Spanish users and matched native English and Spanish speakers.

5 An Empirical Study: Two Methods Illustrated on the Basis of the Same Material

The two methods selected for the empirical study are the lexical bundle method and the comprehensive method. The methods are alike in that they lend themselves to the scanning of entire texts and to detection of patterns, which means that there will be some overlap of results; they are different in that one is purely based on frequency and dispersion, the lexical bundle method, and one is based on the identification of MWSs expected given the topic and the specific situation investigated. The two methods could also be said to be complementary, in that, as we will see, some lexical bundles may point to MWSs.

This section will describe and compare the comprehensive and the lexical bundle methods applied to the same material, notably the spoken production of advanced non-native Swedish speakers of English and Spanish and English and Spanish native speakers. Finally, suggestions for future research using these methods will be given.

5.1 Material

The participants in the present study include long-residency Swedish L2 users of English, and Spanish along with native speakers of these two languages used as control groups, totaling 40 participants, 20 NNSs and 20 NSs. The groups are matched regarding education (all four groups are educated up to university level) and age, the English NS and NNS groups, 32.1 and 33.3 years, respectively, and the Spanish NS and NNS groups 38.8 and 39.8 years, respectively. Mean length of stay for the English NNS group is 7.3 years, and for the Spanish NNS group 9.9 years in London, UK, and Santiago, Chile, respectively. The English sub-corpora consist of NS 17,693 words and NNS 16,236 words, totaling around 34,000 words. The corresponding number of words for the two Spanish sub-corpora is NS 16,109 and NNS 14,168, totaling around 30,000 words. The total number of words for all four sub-corpora amounts to over 64,000 words.

5.2 Task

The task is what might be called a classic narrative SLA task. The participant is asked to narrate simultaneously a video clip consisting of the first 14½ min from Charlie Chaplin's film *Modern Times.* In this task, the participants were not allowed any planning time and were only told to imagine that they were describing what they saw on the screen to someone who could not see it.

This task puts high demands on the participants, because encoding new visual information into linguistic form under time pressure is cognitively complex, and because the task involves infrequent lexis as well as production on-line. In the following two sections we describe the two methods applied to the above material, the comprehensive and the lexical bundle methods.

5.3 The Comprehensive Method: Categories and Inclusion

All the manually extracted MWSs have been checked using dictionaries, and reference corpora. Two main categories of MWSs are distinguished: lexical MWSs and procedural MWSs. The former includes structures that have at least one element with a *lexical* (denotative or referential) meaning, the latter includes structures that have primarily been selected on the basis of their function as operators on, or qualifiers of, other structures. Lexical MWSs are divided into two subcategories: Phrases and Clauses (or clause-shaped utterances). Examples of phrasal MWSs for English include: *Swat the wasp, do a jigsaw puzzle, right now, all over, close-up shot, light a cigarette, have a bad day, in quick succession, a production line, cross the road, main course, turn the wheel, up and down,* and for Spanish (an English 'functional' translation within brackets is provided

throughout) *marcar la tarjeta* ('clocking in'), *la semana pasada* ('last week'), *limarse las uñas* ('do the nails'), *llegar a la hora justa* ('arrive at the right time'), *apretar tuercas* ('tighten nuts'), *presidente de la compañía* ('director of the company'), or *volverse loco* ('go crazy').

In contrast to phrasal MWSs, clausal MWSs require little or no morpho-syntactic manipulation as these examples show: (English) *no problem, do you want to have a think about it, how are you doing, how lovely, take care, I need to ask you a big favour, I do realize that, there's an argument, time to have a break*, and (Spanish) *hagamos una cosa* ('I have a suggestion'), *esa es una posibilidad* ('that's a possibility'), *me parece que no* ('I don't think so'), *yo muy bien* ('I'm fine'), *muchas gracias* ('thanks a lot'), *nos estamos viendo* ('we'll keep in touch'), *queda la escoba* (Chilean Sp. 'you got the old maid').

Procedural MWSs function as signposts or markers guiding the addressee in her/his interpretation of the message, thus 'qualifying' the message in various ways. Procedural MWSs are typically fixed. Depending on whether they have words/phrases or clauses within their scope (i.e. operate at phrase or sentence level) or wider stretches of discourse (at text level), they are subdivided into 'grammatical' and 'discursive', respectively. There is no sharp dividing line between grammatical and discursive MWSs. Conjunctions, for instance, can be conceived of as operating at both sentence and text level; in this study they are referred to the grammatical category.

Grammatical MWSs have grammatical functions as quantifiers, auxiliaries, catenatives, prepositions, conjunctions, etc. They operate at phrasal and clausal levels. Among the subcategory quantifiers, we find examples such as English *loads of*, Spanish *un montón de*. Examples of catenative MWSs include English *try to*, Spanish *tratar de*. In this sub-category we also find conjunctions like English *even though*, and Spanish *para que ('in order for … to')*.

Discursive MWSs can be defined as 'multiword pragmatic markers'. Some verbalize the illocutionary meaning of the embedded clause, frequently involving a mitigating value, such as *I was wondering [if/whether + Clause]*. Others carry an evidential meaning, such as Spanish *me parece [(que) + Clause],* English *I think [(that) + Clause]*. Yet others have a text-organising function, such as *I mean [+Clause)],* or Spanish *o sea [+Clause]* 'that is' [+Clause]. A modalising interactional function is to be found in expressions such as Spanish *¿te parece?* ('don't you think?' or 'do you/we agree?'), and English *you see what I mean.*

The acquisition of formulaic language is presumably favored by immersion in the target language country (cf. Forsberg 2008; Groom 2009) The main aim of using the comprehensive method in the present study is to establish the possible effects on the acquisition of MWSs as a result of daily exposure to and use of the L2, which is the situation of the L2 participants of this study. The categorisation of multiword structures within the comprehensive framework applied in the works cited above as well as in the present study is a modified version of the model presented in Erman and Warren (2000).

5.4 The Lexical Bundle Method: Length of Bundles

It should be borne in mind that lexical bundle studies in general are applied to large written corpora, so the present study involving transcribed speech, which puts restrictions on size, should be seen as an illustration of and a complement to the comprehensive method. As mentioned, lexical bundles are included on the basis of frequency and dispersion. The recommendation for frequency is between 25 and 40 times in one million words and dispersion at between 3 and 5 texts (Biber and Barbieri 2007). In view of the fact that the two corpora are small (each sub-corpus is 14,000–17,000 words), the cut-off point for frequency was set at two, and dispersion in at least two texts, i.e. involving at least two speakers. The length, again in view of the small corpora, was set at three-word lexical bundles to be comparable to the average length of MWSs in the comprehensive framework, which is about three words on average. In order to retrieve the three-word lexical bundles, the WordList cluster function of the software WordSmith Tools (Scott 2007) was used.

6 Comparison of a Selection of Results from the Empirical Study

The following account will include a comparison of the two methods used in the empirical study: the comprehensive method and the lexical bundle method; the term MWSs is reserved for the comprehensive method and lexical bundles (LBs) for the lexical bundle method to keep the two methods apart. The following aspects in the four sub-corpora will be compared and discussed: Total number of words, total number of types of MWSs and LBs, and T/T ratios (Sect. 6.1); the 20 most frequent MWSs and LBs (Sect. 6.2); specific patterns captured through both methods (Sect. 6.3); a selection of patterns captured through one but not the other method (Sect. 6.4). In the final Sect. (7), we wind up by commenting on the main findings of this empirical study comparing two different methods, and by giving some suggestions for future research.

6.1 Numbers of MWS and LB Types in the Four Sub-corpora

We remind the reader that MWSs and LBs have been included on the basis of different criteria. In other words, there is no restriction on frequency or on the number of speakers for an MWS to be categorized as such, since other criteria have been used (see Sect. 3.3). In contrast, LBs are defined on the basis of frequency and dispersion; in the present study the cut-off point has been set at ≥2 speakers. Table 1 shows the results for the total number of words, total number MWS types + Type/Token (T/T) ratios, total number of LB types + Type/Token (T/T) ratios over the four groups.

Table 1 Total no. of words, total number of types, and T/T ratios for MWSs and LBs over English and Spanish NS and NNS groups

NS/NNS / MWS/LB	English NS	English NNS	Spanish NS	Spanish NNS
Tot no. of words	17,693	16,236	16,109	14,168
MWS types	1,555	1,245	972	754
MWS T/T	0.65	0.66	0.56	0.54
LB types; ≥2 texts	1,248	1,157	863	852
LB T/T; ≥2 texts	0.29	0.29	0.26	0.27

Both methods yield lower numbers of MWS and LB types for the NNSs compared to the NSs, the difference between NNS and NS being more noticeable for the number of MWS types than LB types. It is worth noting in this context that MWSs, in contrast to LBs, by definition are conventionalized semantic or pragmatic units, which have to be learnt. LBs, on the other hand, through their very nature seldom occur as units but rather as building blocks for the construction of discourse, whose size is determined by the researcher (e.g. three- and four-word bundles). It is interesting to note that both NNS groups come close to their native counterparts not only in numbers of types of LBs, but also in the Type/Token measurements of both MWSs and LBs.

There is a considerable difference between the Type/Token ratios of MWSs and LBs for all four sub-corpora, the LB ratios being considerably lower than the MWS ratios. This can be explained by LBs containing many high-frequency words. The Type/Token results for MWSs indicate that both Spanish groups recycle their MWSs to a higher degree compared to the English groups.

Other results worth observing concern differences between the two languages. The Spanish sub-groups exhibit considerably lower numbers of both LBs and MWSs, as well as lower numbers for total number of words, compared to the English sub-groups. Tentative explanations for these differences are that Spanish features, such as Pro-Drop (involving zero subject pronouns), and preposition and article contractions have no correspondences in English. Furthermore, Spanish has no operator corresponding to English *do,* and also makes less use of auxiliaries. Another contributing factor is that all contractions in the English material (which are abundant in the spoken register) have been removed (e.g. *he's* has been changed into *he is*). These differences between the English and Spanish systems are reflected in the lower numbers for the LBs and MWSs in the Spanish material, which frequently involve high-frequency words such as pronouns, and auxiliaries.

6.2 The Most Frequent MWSs and LBs

The top 20 LBs and MWSs have been ordered in terms of frequency (see Appendices A and B, respectively, for the English sub-corpora, and D and E for the Spanish sub-corpora). Those bundles that appear in only one list are in bold. This does not mean

that for instance *the production line* (Appendix A) or *limar las uñas* ('do the nails') for Spanish (Appendix D) are not used at all by the NNSs, only that they were not used frequently enough to appear in the top 20 list.

What a top 20 list thus can show are tendencies, which then have to be further analyzed. An interesting tendency can, for instance, be seen in the English subcorpora with regard to *and*-headed bundles in the LB lists: these are more than twice the number for the NNSs compared to the NSs. The same tendency, albeit not as marked, can be found in the Spanish LB lists, where the NNSs produce 20 % more *y*-headed bundles (*y*='and') than the NSs. Furthermore, it is clear that certain vocabulary tied to the overall topic (Charlie Chaplin), such as *the conveyor belt* or *presidente de la /compañía/* ('the director of /the company/'; see Appendix D: the Spanish LB *presidente de la* is followed by the LB *de la compañia*) in the LB lists, and *tighten the nuts* or *marcar la tarjeta* ('clocking in') in the MWS lists, does not make it to the top 20 list in the NNS groups.

6.3 Types Captured by Both Methods

There is only around 20 % direct overlap in the data between LBs and MWSs, which is a result of using different criteria for inclusion. In order to find directly overlapping structures we made separate lists of MWSs and LBs ordered alphabetically (see Appendix C for *a*-headed LBs in English, e.g. *a bow tie*). A comparison between the alphabetical lists of MWSs and LBs revealed that the two methods for some categories yield similar results. Phrases, which are one of the basic categories in the comprehensive framework, are not normally captured by the lexical bundle approach, since lexical bundles are defined on the basis of form and frequency. It should be pointed out that in the comprehensive framework all MWSs are registered in their basic form in the lists, thus resembling dictionary entries (*do a jigsaw puzzle, lunch break, nuts and bolts,* etc.). This means that whenever a noun phrase or a verb phrase in the LB list is used in the basic form frequently enough to meet the criteria for inclusion there will be overlaps with MWSs. As a consequence, there are fewer overlaps involving verb phrases than noun phrases, since the verb phrases in both languages appear more often as inflected forms than in the infinitive. Notwithstanding, the two methods sometimes yield similar results, which can be seen in the alphabetical lists, exemplified in Appendix C. Below are examples of phrases captured by both methods.

- **Noun phrases:** e.g. (English) *a good idea, a thermos flask, nuts and bolts*, *a railway station, a soup bowl*; (Spanish) *compañero de trabajo* 'work mate', *hora de almuerzo* 'lunch time', *vaso de agua* 'glass of water', *plato de sopa* 'soup bowl'
- **Quantifiers/Qualifiers:** (English) *a kind of, a herd of, a couple of, a lot of,* (Spanish) *una especie de* 'a kind of', *una tropa de* 'a herd of', *un montón de* 'a lot of'
- **Preposition phrases:** *at the moment, around the corner, down the road*; (Spanish) *fuera de control* 'out of control', *por el estilo* 'of the kind', *en la tarde* (Chilean Sp.) 'in the afternoon'.

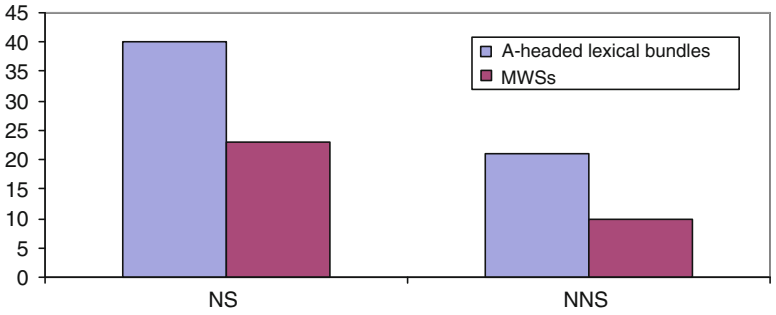


Fig. 1 Numbers of occurrence of *A*-headed lexical bundles and MWSs in the English NS and NNS sub-corpora

A somewhat different picture emerges when we take into account not only *total* overlap between an LB and a MWS, but also those cases in which a given LB *includes* a smaller-sized MWS, and even those cases in which a given LB *strongly suggests* the existence of a partly overlapping MWS. Doing this would, in fact, be a way of manually detecting true MWSs in the LB lists. In our top 20 LB lists, all of these instances have been marked with an asterisk (*). In the English data, 10/20 items are categorized in this way for the NSs, compared to 8/20 for the NNSs. This difference between the NS and NNS groups might suggest a higher degree of consensus among the natives. The same tendency is apparent in the Spanish data: 8/20 for the NSs and 5/20 for the NNSs.

6.4 Patterns Captured by One Method Only

The two methods yield different patterns, so that phrasal patterns are more frequently captured in the MWS lists for reasons accounted for above. For instance, verbal patterns such as the verb 'turn' taking different complements (e.g. *turn a dial, turn a handle, turn a knob, turn a switch*) are only captured in the MWS lists. In fact, phrasal patterns tied to specific lexemes (e.g. *turn,* and *down*) show interesting differences between the NNS and NS speakers, among other things, that the NS group has a richer repertoire in terms of types. This tendency is supported by the results apparent in the *a*-headed LB lists (Appendix C), where the NSs show more variation throughout (see Fig. 1). In fact, both methods show the same tendency concerning number of types as can be seen in numbers of *a*-headed lexical bundles from the English sub-corpus in Fig. 1 below. Although there are fewer MWSs than lexical bundles for both groups the proportions between NSs and NNSs are similar.

On the other hand, structures which do not coincide with syntactic categories are normally captured only in the LB lists. This, as stated above, is the case of *y/and-headed* LBs, which are far more frequent among NNSs, in particular in the English data. In fact, most *y/and*-bundles in the non-native production appear in contexts

where the conjunction is used as a discourse marker to establish links between the events accounted for in the retelling task (e.g. *and it looks…, and he falls…; y se sienta…* 'and he sits down', *y parece que…* 'and it seems that').

Other conjunction-headed bundles, for instance, the *so*-headed bundles of the English data, show the opposite tendency, being nearly twice as frequent in the native as in the non-native data (NS 22 vs. NNS 12). This gives us interesting hints about the order of accessibility for second language acquisition as regards conjunctions, and possibly also about preferences regarding discourse phenomena such as the expression of argumentative structure.

A similar reasoning could be applied to preposition-headed bundles. In the Spanish data, *a* and *de*-headed bundles are far more frequent among the natives (NS 67 vs. NNS 41 for *a* and NS 72 vs. NNS 59 for *de*). By contrast, bundles headed by the preposition *en* 'in, on' appear in the same proportion among NSs (33) and NNSs (32). Now, the latter preposition is not only semantically more 'concrete' but also corresponds closely to the use of the Swedish prepositions *i* 'in' and *på* 'on', whereas the former two have a more abstract character and hardly show any 1-to-1 correspondence with Swedish forms.

The similarity with L1 patterns could not, however, account for the more frequent use of *with*-headed bundles in the English native data than among the non-natives (NS 28 vs. NNS 12), considering the fact that the usage of the Swedish preposition *med* frequently corresponds to that of English *with*.

The results concerning prepositions and conjunctions suggest that features such as complexity and discourse structure may be different in native and non-native data. More research is clearly required to shed light on these differences between NS and NNS usage.

An interesting case is bundles headed by the indefinite or the definite article, which appear more frequently among the NSs than the NNSs in both data sets. As stated above, *a*-headed bundles in the English data (see Appendix C) are produced twice as frequently by natives (51) as by non-natives (25). In the Spanish data, the same tendency although less marked, was observed (NS 36 vs. NNS 31). As for bundles headed by the definite articles *el* and *la*, the difference is considerably stronger: the frequency of these bundles is about 50 % higher in the native productions (NS 159 vs. NNS 108; there is no appendix showing Spanish alphabetical lists). Altogether, this clearly suggests that noun phrases are more frequent in the NS material, which in turn may indicate a more complex language.

Both methods thus yield interesting patterns worth further investigation. Furthermore the majority of patterns (even non-sensical ones, as is the case of several LBs) point to interesting differences between NS and NNS usage.

7 Conclusions

On the basis of an overview of three current methods of analyzing recurrent word combinations, two were selected for a small-scale empirical study, the lexical bundle method and the comprehensive method. Both methods are applied to entire

texts, one being automatic, the lexical bundle method, and the other largely manual, the comprehensive method. In the present study they were applied to the same material, the spoken retelling of a film clip (from *Modern Times*) by two groups of advanced, long-residency Swedish users of L2 English and L2 Spanish and native English and Spanish controls. The results of this small-scale empirical study show that both NNS groups approach nativelike levels in quantity of LBs, and Type/Token ratios of both MWSs and LBs. The two NNS groups are similar in that they have somewhat lower quantities of MWSs compared to the NS groups. Furthermore, conjunction-, preposition-, and *a*-headed LBs showed patterns in the NNS data, which deviated from those of the NS groups.

There are also general differences between the English and the Spanish sub-corpora worth pointing out. The figures for the Spanish material are lower for the total number of words, as well as for the total number of LBs and MWSs. One reasonable explanation for the difference between the Spanish and English sub-corpora is to be found on the system level, for instance the Spanish feature Pro-Drop (involving subject pronouns), and preposition and article contractions, which have no correspondences in English.

In conclusion, this small-scale study has shown that two different methods, yielding different results, can fruitfully be used together, not only to inform and complement one another, but to broaden our knowledge of what being nativelike involves, not to mention the vast implications such increased knowledge may have for teaching and learning an L2. Finally, regarding differences between long-residency L2 users and native speakers, it needs to be pointed out that, in spite of having lived and worked for a long time in the target language country, the NNSs, naturally, do not even come close to the life-long input that the NSs have received. This is the most plausible explanation for the divergences found in this as well as in other studies focusing on differences between NSs and high-proficient NNSs regarding recurrent word combinations.

Acknowledgments Thanks are due to generous funding by *The Bank of Sweden Tercentenary Foundation*, Sweden. *We also wish to thank Professor Annelie Ädel for providing the WordSmith files.*

Appendices

Appendix A. Lexical Bundles – English

NS Frequency list				NNS Frequency list			
N	*NS LB*	*Freq.*	*Texts*	*N*	*NNS LB*	*Freq.*	*Texts*
1	and he is	62	8	1	and he is	101	10
2	***THE PRODUCTION LINE**	45	6	2	*there is a	41	8
3	*there is a	43	8	3	**AND THEY ARE**	36	10
4	now he is	40	8	4	*is trying to	33	9

(continued)

(continued)

NS Frequency list				NNS Frequency list			
N	*NS LB*	*Freq.*	*Texts*	*N*	*NNS LB*	*Freq.*	*Texts*
5	***THE CONVEYOR BELT**	36	6	5	now he is	22	9
6	*one of the	26	8	6	**THE MACHINE AND**	22	7
7	*seems to be	26	5	7	and it is	21	9
8	(*)and there is	25	6	8	**IN THE FACTORY**	21	8
9	**(*)ON THE PRODUCTION**	22	6	9	(*)and there is	19	7
10	**(*)ON THE CONVEYOR**	21	6	10	*corn on the	19	8
11	and now he	20	7	11	**(*)HE IS TRYING**	19	7
12	*corn on the	20	8	12	and now he	18	8
13	***SOME KIND OF**	20	5	13	***THE FACTORY FLOOR**	18	7
14	**THE MACHINE IS**	19	9	14	***AND NOW THE**	17	6
15	and it is	18	9	15	**INTO THE FACTORY**	17	9
16	***BACK TO THE**	14	7	16	**HE IS NOW**	16	8
17	*is trying to	13	8	17	**INTO THE MACHINE**	16	8
18	***OUT OF THE**	13	6	18	***IS GOING TO**	16	7
19	**HE CAN NOT**	12	6	19	*one of the	16	7
20	**THE PRESIDENT IS**	12	5	20	*seems to be	16	6

UPPER CASE = unique for each top 20 list (i.e. not shared between the NS or NNS top 20 list): 10/20 items

* = constitutes, includes or suggests a MWS. NS: 10/20 items. NNS: 8/20 items

(*) = constitutes, includes or suggests a MWS already present in the top 20 list

Appendix B. MWS – English

NS Frequency list					NNS Frequency list				
Order of freq	MWS	Category	Freq	N of texts	Order of freq	MWS	Category	Freq	N of texts
1	there is	Gram	60	7	1	try to	Gram	86	9
2	**A PRODUCTION LINE**	Phrase	48	6	2	and now	Disc	60	10
3	**A CONVEYOR BELT**	Phrase	45	6	3	there is	Gram	39	7
4	try to	Gram	37	10	4	**SEEM TO**	Gram.	27	7
5	and now	Disc	32	8	5	**I THINK**	Disc	26	6
6	**KIND OF**	Disc	29	6	6	and then	Disc	22	8
7	one of	Gram	26	8	7	corn on the cob	Phrase	20	8
8	**PULL A LEVER**	Phrase	23	7	8	**RUN AFTER**	Phrase	18	7
9	some kind of	Gram	21	5	9	**COME UP**	Phrase	17	5
10	come in	Phrase	20	8	10	come in	Phrase	16	7
11	a corn on the cob	Phrase	18	8	11	**A FACTORY FLOOR**	Phrase	16	7
12	lots of	Gram	17	8	12	one of	Gram	16	7
13	**THERE IS A**	Gram	16	6	13	**READ A PAPER**	Phrase	13	5
14	and then	Disc.	16	6	14	**HAVE TO**	Gram	13	6

(continued)

(continued)

NS Frequency list					NNS Frequency list				
Order of freq	MWS	Category	Freq	N of texts	Order of freq	MWS	Category	Freq	N of texts
15	**TIGHTEN THE NUTS**	Phrase	14	6	15	lots of	Gram	13	7
16	**LOOK AT**	Phrase	13	6	16	some kind of	Gram	12	5
17	**LOOK LIKE**	Phrase	13	6	17	**TALK TO**	Phrase	11	5
18	**RUN AWAY**	Phrase	11	7	18	**WANT TO**	Gram	11	5
19	**BE ABOUT TO**	Gram	10	5	19	**OR SOMETHING**	Disc	11	6
20	**CLOCK BACK IN**	Phrase	9	5	20	**CLOCK IN**	Phrase	10	5

UPPER CASE = unique for each top 20 list (i.e. not shared between the NS or NNS top 20 list): 11/20 items

Distribution of MWS categories:

NS: 10 phrasal, 7 grammatical, and 3 discursive

NNS: 8 phrasal, 8 grammatical, and 4 discursive

Appendix C. NSs and NNSs: Alphabetical Lists of Bundles

Example: *A*-Headed Bundles in the English Material

NS: Alphabetical List of *A*-Headed Bundles

Alphabetical list *A*-headed	Freq.	No. Texts
A A RECORD	2	2
A BIG MACHINE	3	3
A BIT BORED	2	2
A BIT CONFUSED	2	2
A BIT OF	6	4
A BIT TOO	3	3
A BOW TIE	2	2
A CIGARETTE AND	2	2
A CONVEYOR BELT	4	4
A CORN ON	5	5
A COUPLE OF	5	4
A FIRE HYDRANT	4	4
A FLASK INTO	2	2
A GLASS OF	7	7
A GOOD IDEA	2	2
A GROUP OF	4	3
A HERD OF	6	4
A IN A	2	2
A JIGSAW PUZZLE	3	3
A KIND OF	9	5

(continued)

(continued)

Alphabetical list *A*-headed	Freq.	No. Texts
A LADDER AND	2	2
A LADY WALKING	2	2
A LARGE LADY	3	3
A LITTLE BIT	5	2
A LOT OF	6	4
A LUNCH BREAK	2	2
A MAN COMES	3	2
A MAN IN	2	2
A MAN WALKING	2	2
A PIECE OF	2	2
A POLICEMAN AND	2	2
A POLICEMAN WHO	2	2
A RECORD PLAYER	3	2
A ROTATING CORN	2	2
A SORT OF	2	2
A SUIT AND	3	3
A TABLET AND	2	2
A THERMOS FLASK	2	2
A WOMAN COMES	2	2
A WORD WITH	2	2
NP = 10/40		
DET = 6/40		
HEDGE = 7/40		
23/40 = MWSs		

NNS: Alphabetical List of *A*-Headed Bundles

Alphabetical list *A*-headed	Freq.	No. texts
A BEE IN	2	2
A BIG SCREEN	3	3
A BIT OF	9	4
A BOWL OF	2	2
A CIGARETTE AND	2	2
A CORN ON	5	5
A FEEDING MACHINE	2	2
A GLASS OF	7	7
A IN A	4	3
A JIGSAW PUZZLE	3	3
A LITTLE BIT	3	3
A LOT OF	8	6
A MACHINE THAT	3	2
A MAN AND	2	2
A MAN WHO	3	2
A PIECE OF	2	2
A PILL AND	2	2
A PLATE AND	3	3

(continued)

(continued)

Alphabetical list *A*-headed	Freq.	No. texts
A POLICE OFFICER	2	2
A WOMAN COMES	2	2
NP = 4/21		
DET = 4/21		
Hedge = 2/21		
MWSs = 10/21		

Appendix D. Lexical Bundles – Spanish

NS frequency list				NNS frequency list			
N	*NS LB*	*Freq.*	*Texts*	*N*	*NNS LB*	*Freq.*	*Texts*
1	**DE LA MÁQUINA**	32	11	1	la boca y	20	8
2	**LA MÁQUINA Y**	24	9	2	***Y AHORA VIENE**	18	8
3	*en la cara	22	7	3	***CÓMO SE LLAMA**	17	8
4	***PRESIDENTE DE LA**	20	7	4	***QUE TIENE QUE**	14	6
5	**DE LA COMPAÑÍA**	17	–5	5	**A LA BOCA**	13	5
6	**LA MÁQUINA LE**	16	6	6	**A LA FÁBRICA**	13	8
7	***EN LA BOCA**	15	7	7	**(*)TIENE QUE HACER**	13	7
8	en el plato	14	8	8	**(*)AHORA VIENE EL**	12	7
9	**DE LA EMPRESA**	13	6	9	***DENTRO DE LA**	12	6
10	***Y EMPIEZA A**	13	5	10	en el plato	12	8
11	**(*)EL PRESIDENTE DE**	12	6	11	*en la cara	12	5
12	***LA LÍNEA DE**	12	4	12	**(*)Y TIENE QUE**	12	6
13	**A LA SECRETARIA**	11	6	13	**CON LA SOPA**	11	8
14	***LIMPIA LA BOCA**	11	7	14	**EN LA FÁBRICA**	11	5
15	**DE LOS TRABAJADORES**	10	5	15	**(*)Y AHORA LE**	11	7
16	**EL RITMO DE**	10	5	16	**DE LA FÁBRICA**	10	7
17	la boca y	10	6	17	**(*)Y AHORA SE**	10	6
18	**LA CARA Y**	10	6	18	**NO SÉ LO**	10	4
19	***LE LIMPIA LA**	10	6	19	**LA FÁBRICA Y**	9	6
20	***UNA ESPECIE DE**	10	6	20	***QUE VA A**	9	6

UPPER CASE = unique for each top 20 list (i.e. not shared between the NS or NNS top 20 lists): 17 items

* = constitutes, includes or suggests a MWS. NS: 8/20 items. NNS: 5/20 items

(*) = constitutes, includes or suggests a MWS already present in the top 20 list

Appendix E. MWSs – Spanish

NS frequency list					NNS frequency list				
Order of freq	MWS	Category	Freq	N of texts	Order of freq	MWS	Category	Freq	N of texts
1	de nuevo	gram	43	8	1	de nuevo	gram	63	10
2	**VOLVER A**	gram	41	10	2	tener que	gram	47	9
3	empezar a	gram	41	5	3	parece que	disc	32	8
4	**MARCAR LA TARJETA**	phrase	24	10	4	empezar a	gram	23	8
5	tratar de	gram	23	8	5	limpiar la boca	phrase	21	7
6	limpiar la boca	phrase	20	7	6	tratar de	gram	19	6
7	tener que	gram	19	9	7	**POR FAVOR**	phrase	19	4
8	para que	gram	18	9	8	(a)dentro de	gram	12	7
9	**DAR VUELTA**	phrase	18	7	9	**EL POBRE**	phrase	14	6
10	**APRETAR TUERCAS**	phrase	17	5	10	para que	gram	11	5
10	parece que	disc	17	5	11	**ME PARECE**	disc	9	4
12	volverse loco	phrase	13	5	12	CÓMO SE LLAMA	disc	9	4
13	**DEJAR DE**	gram	11	8	13	**UN MONTÓN DE**	gram	8	5
14	**UNA ESPECIE DE**	gram	11	6	14	ir al baño	phrase	7	6
15	adentro de	gram	10	6	15	**HACER SU TRABAJO**	phrase	7	4
16	**PRESIDENTE DE LA COMPANIA**	phrase	9	6	16	**HASTA QUE**	gram	7	4
17	**LIMAR LAS UÑAS**	phrase	8	7	17	**EN LA CALLE**	phrase	6	5
18	**DETRÁS DE**	gram	7	5	18	**YO CREO**	disc	5	4
19	ir al baño	phrase	6	6	19	volverse loco	phrase	5	4
20	**EN VEZ DE**	gram	5	5	20	**INTENTAR DE**	gram	4	4

UPPER CASE = unique for each top 20 list (i.e. not shared between the NS or NNS top 20 lists): 10/20 items

Distribution of MWS categories:

NS: 8 phrasal, 11 grammatical, and 1 discursive

NNS: 7 phrasal, 9 grammatical, and 4 discursive

References

Ädel, Annelie, and Britt Erman. 2012. Recurrent word combinations in academic writing by native and non-native students: A lexical bundles approach. *English for Specific Purposes* P33: 81–92.

Altenberg, Bengt. 1998. On phraseology of spoken English: The evidence of recurrent word combinations. In *Phraseology: Theory, analysis and applications*, ed. A.P. Cowie, 101–122. Oxford: Oxford University Press.

Biber, D., and F. Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26: 263–286.

Biber, D., S. Conrad, and V. Cortes. 2004. 'If you look at…': Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25: 371–405.

Brinton, L.J. 1996. *Pragmatic markers in English: Grammaticalization and discourse functions*. Berlin: Mouton de Gruyter.

Chen, Y.-H., and P. Baker. 2010. Lexical bundles in L1 and L2 academic writing. *Language, Learning and Technology* 14(2): 30–49.

Cortes, V. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23(4): 397–423.

Cowie, A.P. 1998. Phraseological dictionaries: Some east–west comparisons. In *Phraseology: Theory, analysis and applications*, ed. A.P. Cowie, 209–228. Oxford: Oxford University Press.

De Cock, S. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL)*, New Series 2: 225–246.

Denke, A. 2009. *Nativelike performance: A corpus study of pragmatic markers, repairs and repetition in native and non-native English speech*. Saarbrücken: VdmVerlag.

Ellis, N.C. 1996. Sequencing in SLA, phonological memory, chunking, and points of order. *Studies in Second Language Acquisition* 18(1): 91–126.

Ellis, N.C. 2006. Cognitive perspectives on SLA: The associative-cognitive CREED. *Aila Review* 19: 100–122.

Erman, B. 2009. Formulaic language from a learner perspective: What the learner needs to know. In *Formulaic language, Vol. 2: Acquisition, loss, psychological reality and functional explanations*, ed. R. Corrigan, E.A. Moravcsik, H. Quali, and K.M. Wheatly, 323–346. Amsterdam/Philadelphia: John Benjamins.

Erman, B., and B. Warren. 2000. The idiom principle and the open choice principle. *Text* 20(1): 29–62.

Fillmore, C.J., P. Kay, and M.C. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64(3): 501–538.

Firth, J.R. 1957. A synopsis of linguistic theory. 1930–1955. In *Selected papers of J.R. Firth*, ed. F.R. Palmer, 168–205. London: Longman.

Forsberg, F. 2008. *Le Langage Préfabriqué: Formes, fonctions et frequences en français parlé L2 et L1*. Bern: Peter Lang.

Forsberg, F., and L. Fant. 2010. Idiomatically speaking: Effects of task variation on formulaic language in highly proficient users of L2 French and Spanish. In *Perspectives on formulaic language: Acquisition and communication*, ed. Wood David, 47–70. London/New York: Continuum.

Granger, S. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In *Phraseology: Theory, analysis and applications*, ed. A.P. Cowie, 145–160. Oxford: Oxford University Press.

Groom, N. 2009. Effects of second language immersion on second language collocational development. In *Researching collocations in another language: Multiple interpretations*, ed. A. Barfield and H. Gyllstad, 21–33. Basingstoke: Palgrave Macmillan.

Halliday, M.A.K. 1961. Categories of the theory of grammar. *Word* 17: 241–292.

Hancock, V. 2000. *Quelques connecteurs et modalisateurs dans le français parlé d'apprenants universitaires*. Cahiers de la recherche 16. Doctoral dissertation, Department of French and Italian, Stockholm University.

Hoey, M. 2005. *Lexical priming: A new theory of words and language*. London/New York: Routledge.

Howarth, P. 1998a. Phraseology and second language proficiency. *Applied Linguistics* 19(1): 24–44.

Howarth, P. 1998b. The phraseology of learners' academic writing. In *Phraseology: Theory, analysis and applications*, ed. A.P. Cowie, 161–186. Oxford: Oxford University Press.

Hunston, S., and G. Francis. 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hyland, K. 2008. Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics* 18(1): 41–62.

Lewis, Michael. 2000. *Teaching collocation: Further developments in the lexical approach*. Hove: Language Teaching Publications.

Lewis, Margareta. 2009. *The idiom principle in L2 English: Assessing elusive formulaic sequences as indicators of idiomaticity, fluency, and proficiency*. Saarbrücken: VDM Verlag.

McCarthy, M., and R. Carter. 2004. This that and the other: Multi-word clusters in spoken English as visible patterns of interaction. *Teanga: The Irish Yearbook of Applied Linguistics* 21: 30–53.

Mel'čuk, I. 1998. Collocations and lexical functions. In *Phraseology: Theory, analysis, and applications*, ed. A.P. Cowie, 23–53. Oxford: Clarendon.

Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24(2): 223–242.

Raupach, M. 1984. Formulae in second language speech production. In *Second language production*, ed. H.W. Dechert, D. Möhle, and M. Raupach, 114–137. Tübingen: Günter Narr Verlag.

Redeker, G. 1990. Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics* 14: 367–381.

Renouf, Antoinette, and John Sinclair. 1991. Collocational frameworks in English. In *English corpus linguistics: Studies in honour of Jan Svartvik*, ed. Karin Aijmer and Bengt Altenberg, 128–144. London/New York: Longman.

Römer, U. 2009. English in academia: Does nativeness matter? *Anglistik: International Journal of English Studies* 20(2): 89–100.

Schiffrin, D. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

Scott, M. 2007. *WordSmith Tools (version 4.0)* (Computer software). Oxford: Oxford University Press.

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stubbs, Michael. 2002. *Words and phrases*. Oxford: Blackwell Publishing.

Wiktorsson, M. 2003. *Learning idiomaticity: A corpus-based study of idiomatic expressions in learners' written production*, Lund studies in English 105. Lund: Lund University.

Discourse Functions of Recurrent Multi-word Sequences in Online and Spoken Intercultural Communication

Yen-Liang Lin

1 Introduction

Kjellmer (1994) suggests that "[t]here is no doubt that natural language has a certain block-like character. Words tend to occur in the same clusters again and again" (p. ix). Previous research has indeed highlighted the fact that both written and spoken discourse contains a large proportion of highly recurrent sequences of words, reflecting the phrasal nature of the English language (Adolphs 2006; Biber et al. 1999; Greaves and Warren 2010; Nation and Webb 2011; Schmitt 2010; Wood 2010; Wray 2002). Biber et al. (1999), for example, illustrate that two-word (e.g., *I think*), three-word (e.g. *a lot of*) and four-word (e.g. *what do you think*) recurrent sequences made up nearly 45 % of the spoken conversation and approximately 21 % of the academic written discourse they studied (the cut-off was set at a frequency of 20 occurrences per million words). Erman and Warren (2000) also calculated that recurrent multi-word units constituted 58.6 % of the spoken corpus and 52.3 % of the written discourse analysed in their study. In addition, Foster (2001) analysed the transcripts of unplanned speech of English native speakers and found that 32.3 % consisted of recurrent multi-word sequences, while in Hill's (2001) study up to 70 % of language (spoken and written discourse) comprised fixed expressions. Despite the variation in the reported percentage of multi-word sequences encountered in language in these studies, they all indicate an observable tendency for particular items to co-occur in the written and spoken discourse of both native and non-native speakers of English, and for these co-occurrences to make up an appreciable proportion of authentic language use.

There has been a burgeoning field of research looking at multi-word sequences in different registers and settings, identifying different kinds of sequences and

Y.-L. Lin (✉)
School of English, University of Nottingham, University Park NG7 2RD, Nottingham, UK
e-mail: Yenliang.lin@nottingham.ac.uk

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_6,

describing how they are employed in a particular context, such as academic writing (Chen and Baker 2010; Simpson-Vlach and Ellis 2010), university classroom teaching (Biber et al. 2004), small group teaching contexts such as tutorials and seminars (Walsh et al. 2011), textbook discourse (Chen 2010; Wood 2010), and spoken interview discourse (Adolphs and Durow 2004). Although multi-word sequences used in various contexts have been extensively studied, relatively few have focused on the sequences in an intercultural setting and further compared their use in two important registers, namely computer-mediated communication (CMC) and spoken interaction. The research of intercultural discourse increasingly represents a particularly important endeavour as it offers insights into language variety which reflects social and cultural differences of the writers and speakers (Hanna and de Nooy 2003; Liaw and Master 2010). The present study aims to offer a comparative investigation of the most frequent multi-word sequences in the two different communicative modes and identify their primary discourse functions in an intercultural setting. The differences in how two groups of participants – British and Taiwanese teenagers – use multi-word sequences for different discourse functions and in different registers are also examined.

2 What Are Multi-word Sequences?

An enormous technical vocabulary is used to describe the phrasal nature of language, including *prefabricated patterns* (Hakuta 1974), *routine formulae* (Coulmas 1979), *lexical phrases* (Nattinger and DeCarrico 1992), *lexical clusters* (Wood 2010), *chunks* (De Cock 2004; O'Keeffe et al. 2007), *clusters* (Scott 2010), *multi-word units* (Greaves and Warren 2010; Nation and Webb 2011), *formulaic sequences* (Schmitt 2004, 2010; Wray 2002) and *lexical bundles* (Biber et al. 2004; Biber 2009). This notwithstanding, the various terms all seem to share the consensus that many words have an observable tendency to "occur in multiple word phraseological units" (Schmitt 2010, p. 117) and such a phenomenon is one of the most significant features of language use.

Wray (2002) defines formulaic sequence as follows:

> A sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar (p. 9).

This definition shows that in many cases sequences of words are processed mentally as if single words and stored in mind as a chunk that can be retrieved holistically at the time of use. For example, in the case of a self-introduction, learners just need to retrieve the chunk *my name is …*, *I am from …* etc. instead of building the sentence word by word. This concept of holistic storage and retrieval of formulaic language has been demonstrated in a range of studies (e.g., Conklin and Schmitt 2008; Tremblay and Baayen 2010). Nevertheless, whether or not the recurrent elements are prefabricated in speakers' or writers' mind is still debatable. In this chapter, I use *multi-word sequence* as an umbrella term to cover all types of recurrent sequences of words, that is, "frequently occurring contiguous words that constitute

a phrase or a pattern of use" (Greaves and Warren 2010, p. 213). In this regard, a multi-word sequence does not necessarily have to be complete grammatical structure or idiom. This definition is similar with Biber et al.'s (1999) definition of lexical bundles, which are defined as "recurrent expressions, regardless of the idiomaticity, and regardless of their structural status. That is, lexical bundles are simply sequences of word forms that commonly go together in discourse" (p. 990).

3 Multi-word Sequences and Functional Language Use

Recurrent multi-word sequences are generally seen as serving various types of discourse functions in language use as demonstrated in a range of studies (see Biber et al. 2004; Biber 2009; Nattinger and DeCarrico 1992; Wood 2010; Wray and Perkins 2000; Wray 2002) that look at the interrelationship between language and context. Discourse functions in this chapter refer to the function that each sequence serves in the context of online and spoken intercultural communication. Nattinger and DeCarrico (1992) develop a taxonomy that captures three central functions served by what they called *lexical phrases*: (1) social interaction, (2) necessary topics and (3) discourse devices. In their framework, social interaction sequences are associated with social relationships, consisting of conversational maintenance (e.g., *excuse me; how are you?*), and functional meaning relating to conversational purpose, such as expressing politeness (e.g., *thanks very much*), questioning (e.g., *do you like X?*), requesting (e.g., *may I X?*), offering (e.g., *would you like X?*), complying (e.g., *of course*), responding (e.g., *oh, I see*) and asserting (e.g., *I think that X; there is/are/was/were X*). Necessary topics are phrases marking domain-specific topics that often feature in daily conversation. For example, in autobiography, formulaic expressions such as *my name is*, *I am from*, and *I'm X years old* would be quite helpful. With regard to shopping, expressions such as *how much is X?*, *I want to buy X*, *too expensive*, *cost X dollars* may be highly recurrent sequences for use in daily conversation. Discourse devices are lexical phrases that connect the meaning and structure of the discourse, such as logical connectors (e.g., *as a result*), temporal connectors (e.g., *and then*), fluency devices (e.g., *you know; it seems to me that*), exemplifier (e.g., *for example; it's like*), evaluators (e.g., *as far as I know*) and so on. Each of these three main categories has a number of sub-categories associated with more specific functions and meanings (Nattinger and DeCarrico 1992).

Biber et al. (2004, p. 384) identify three primary functions for multi-word sequences in English (they use the term *lexical bundles*): (1) stance expressions, (2) discourse organizers and (3) referential expressions. According to them, stance bundles "express attitudes or assessments of certainty that frame some other proposition", such as *I want you to* and *I don't think so*. These are usually used to convey personal attitudes, intention, prediction and so on. Discourse organizers "reflect relationships between prior and coming discourse," serving two major functions: topic introduction/focus (e.g., *what I want to do is*; *if you look at*) and topic elaboration/clarification (e.g., *on the other hand; I mean you know*). Referential bundles "make direct reference to physical or abstract entities, or to the textual context

itself". Examples of this include identification bundles (e.g., *those of you who*), imprecision bundles (e.g., *and things like that*), bundles specifying attribute (e.g., *have a lot of*) and time/place/text-deixis bundles (e.g., *in the United States; the end of the*).

In addition, Carter and McCarthy (2006) illustrate the functions of recurrent sequences (they use the term *clusters*): relations of time and space (e.g., *in the; on the; the bottom of the*), other prepositional relations (e.g., *with a; for the*), interpersonal functions (e.g., *I don't know what; you know what I mean*), vague language (e.g., *sort of; and stuff; something like that*), linking functions (e.g., *and it was; but I mean*) and turn-taking (e.g., *what do you; do you think*) (pp. 834–837). These studies all demonstrate that multi-word expressions in English have systematic discourse functions although most of them are not semantically or grammatically complete patterns. As claimed by Biber (2009),

> although they are neither idiomatic nor structurally complete, lexical bundles are important building blocks in discourse. Lexical bundles provide a kind of discourse 'head' for larger phrases and clauses, where they function as discourse frames for the expression of new information. (pp. 284–285)

4 Data

The data that forms the basis of the British and Taiwanese Intercultural Communication Corpus (hereafter BATICC) was collected from messages posted to an electronic discussion board (BATICC-O) and recorded face-to-face interaction (BATICC-F) of 60 teenagers between 13 and 14 years of age from Taiwan and the UK. The BATICC-O amounts to a total of 1,035 messages, comprising 31,910 words, while the BATICC-F includes approximately 20,099 words, transcribed in accordance with standard orthographic practices in order to facilitate analysis by currently available corpus tools.

Although this collection of data is relatively small compared to existing corpora, many corpus linguists remark that the size of corpus needed depends upon the purposes of the research study and the language features to be analysed (Sinclair 2001; Hunston 2002; Adolphs 2006; McEnery et al. 2006; Handford 2010; Koester 2010). In this study, focusing on different patterns of language use in different communication modes by different groups of people, the comparative nature of study may make this size of corpus valuable. As Sinclair claims, "comparison uncovers differences almost regardless of size" (2001, p. xii). That is, small specialised corpora may well also contain sufficient examples of frequent linguistic features for illuminating the comparison between different types of different interaction. In addition, the size and composition of a specialized corpus makes it more manageable for qualitative studies as it is feasible to examine most of the concordance lines (not just a random sample) of particular linguistic features in contexts, which can provide a rich source of data to complement the more quantitative-based studies (Hunston 2002; Flowerdew 2004; Handford 2010; Koester 2010).

5 Method

The present study investigates the discourse functions of recurrent multi-word sequences by a group of British and Taiwanese adolescents over a one-year period of intercultural CMC, followed by FTF spoken interaction. A frequency-driven approach is mainly employed for the identification of recurrent sequences as it is a good starting point for subsequent analysis of this study. Research has shown that frequency facilitates enquiry across different corpora, different language varieties and different contexts of use (Baker 2006; Leech et al. 2001; O'Keeffe et al. 2007, 2011), and it is also more systematic and less subjective (Adolphs and Durow 2004; Biber, et al. 2004).

This chapter focuses particularly on the recurrent three-word sequences derived from BATICC and reference corpora CANELC and CANCODE using the programme *Wordsmith Tools 5.0* (Scott 2008). The unit of three words per sequence (e.g., *I don't know, I would like*) was chosen as three-word sequences include sufficient contextual information for the assessment of their discourse functions. They are also analytically more manageable since analysing two-word sequences (e.g., *of the, to be*) would include too many phrasal verbs and grammatical colligations, while considering a larger unit, four or more words (e.g., *at the end of the*) in the sequence would reveal too few examples.

In order to obtain a deeper insight into the use of three-word sequences over time in CMC and spoken interaction, the electronic messages are divided into three data subsets according to the time of posting, resulting in three four-month phases. The highly recurrent three-word units retrieved from different phases of the program are then examined. I further compare the three-word sequences in BATICC-O and BATICC-F and, as a reference for comparison, also list the high-frequency sequences in a large corpus of online discourse (CANELC) and spoken discourse (CANCODE). The online and informal spoken nature of these respective corpora resemble the computer-mediated and face-to-face interaction in this project and thus makes them suitable resources as reference corpora. The 50 most common three-word sequences retrieved from the four datasets are then inductively grouped into three central categories with regard to the discourse function that they serve in the data. In addition, BATICC is also divided according to users of different countries, so more information about cultural differences as well as differences in the use of multi-word sequences by British and Taiwanese participants can be revealed.

To explore the functions of multi-word sequences, they were examined in their extended discourse context to identify their primary functions. The framework for this analysis was principally drawn from Nattinger and DeCarrico (1992), and partly adapted from taxonomy work carried out by Biber et al. (2004) and Carter and McCarthy (2006). Nattinger and DeCarrico's (1992) function-based description of multi-word expressions is based on casual conversation and particularly developed for learners of English as a second language, thus making their taxonomy useful for the present research. While Biber et al.'s (2004) framework is comprehensive, it is based on classroom teaching and university-level textbooks, which are

not the main concern of this study. Additionally, although classroom teaching is a spoken register, it typically focuses on specific topics and much of the content may be pre-planned by the instructors, factors which problematise comparison with the informal and unplanned interactions in the BATICC. Although this study mainly utilises Nattinger and DeCarrico's (1992) framework in the analysis of multi-word sequences, their work includes only structurally and semantically complete sequences (e.g., *as a result*; *and then; I think*), while structurally incomplete but high frequency sequences, such as *and I think that, and it was, but I mean*, are ignored. Therefore, Biber et al.'s (2004) and Carter and McCarthy's (2006) taxonomies can be useful supplements.

It should be emphasised that it is sometimes difficult to assign a sequence to a particular category since; in some cases, a sequence serves multiple functions and is functionally ambiguous. For example, the three-word unit *would you like* in an interrogative clause might function as an offer, an invitation, a request or simply a question. It sometimes can be used to perform two or more speech acts at the same time. As Tsui (1994) argues, the source of multiple functions often lies in the sequential environment of the conversation in which the utterance occurs (p. 45). As such, the use of each sequence in its discourse context is examined to identify the primary function of each sequence in different datasets.

6 Results and Discussion

6.1 Most Frequent Sequences from CMC to FTF

The ten most-frequent three-word sequences are derived from the participants' discourse as displayed in Table 1, showing their use over time from a three-phase CMC to FTF communication by the Taiwanese and British participants. In Phase 1, it can be seen that the majority of the most-frequently used sequences are in relation to personal introductions regarding names (e.g., *my name is*), ages (e.g., *am # years*), and birthdays (e.g., *my birthday is*). This may be due to the nature of the online community, starting from a personal introduction. From the table, it can also be noted that there are two sequences that are frequently used by both groups of young people, namely *my favourite food* and *favourite food is*, which are likely to both be part of an extended sequence such as *my favourite food is*. The sequences surrounding the frequently-used lexical item *food* indicate that food culture is commonly discussed in Phases 1 and 2 of CMC.

When entering Phases 2 and 3, on the other hand, the participants used more expressions concerned with the elicitation of opinions and knowledge such as *do you like*, *do you have* and *what kind of*. That is, the types of discourse tend to shift from basically transactional, transmitting factual information, to increasingly interactional, used for maintaining social relationships (Nattinger and DeCarrico 1992). Such sequences can be found with an extremely high frequency in FTF interaction.

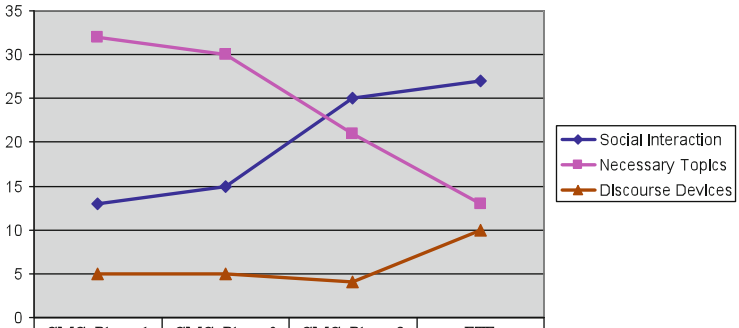
Table 1 The ten most frequent three-word sequences over time

	CMC_Phase 1		CMC_Phase 2		CMC_Phase 3		FTF interaction	
	Sequence	Freq.	Sequence	Freq.	Sequence	Freq.	Sequence	Freq.
1	my name is	64	I like to	43	I like to	34	do you like	41
2	I am #	48	a lot of	30	I am very	27	do you have	32
3	am # years	39	my name is	28	my name is	27	I don't know	24
4	I like to	38	my school is	26	I can't wait	23	do you want	20
5	my favourite food	30	Chinese New Year	26	it is very	22	what do you	19
6	favourite food is	28	I have a	25	do you have	21	you want to	18
7	junior high school	28	do you have	24	a lot of	21	I want to	18
8	and I am	25	do you like	23	with my friends	20	it was very	18
9	I have a	25	we have a	18	but I am	17	fish and chips	16
10	my birthday is	24	my favourite food	18	what kind of	15	go to school	16

For example, the top five high-frequency sequences (e.g., *do you have, I don't know*) marks the highly interactional nature of FTF communication.

From Table 1, we can also see the different use of personal pronouns. Phase 2, in particular, the sequence *we have a*, with 18 occurrences is ranked the ninth. The use of *we* in this case indicates the level of focus on involvement with others, shifting from somewhat more self-identity to group identity (Liaw and Master 2010). Other sequences in Phases 2 and 3 involving the use of *we* and occurring in more than five occurrences include *we had a*, *when we are* and *we sometimes do*. The frequency information also reveals that the numbers of *we* in three stages of CMC are 91, 141 and 138 respectively. Such increasing tendency of the use of *we* from Phase 1 to the second and third phases is probably a natural process of relationship building in that young people share their experiences and identify themselves as the group member of online community.

Furthermore, the top 50 highly recurrent three-word sequences retrieved from different phases are inductively grouped to their functional categories based on their usage in the texts. Figure 1 below illustrates the distribution of functional use across the four phases of the intercultural exchange programme. With regard to the sequences for social interaction, a subtle growth can be seen from Phase 1 (26 %) to Phase 2 (30 %), and the rate of increase is much more extreme from Phase 2 (30 %) to Phase 3 (50 %). On the other hand, there is a clear decreasing trend in the use of sequences for necessary topics in the percentage of top 50 sequences over time, which began at 64 % in Phase 1, followed by a slight decline to 60 % in Phase 2 and then a considerable drop to 42 % in Phase 3. It seems to show that the participants frequently talked about specific topics on the discussion board in the first two phases of CMC, and as their relationship was gradually built over time, they used increasingly large numbers of three-word sequences for social interaction. Particularly in



	CMC_Phase 1	CMC_Phase 2	CMC_Phase 3	FTF
Social Interaction	13 26%	15 30%	25 50%	27 54%
Necessary Topics	32 64%	30 60%	21 42%	13 26%
Discourse Devices	5 10%	5 10%	4 8%	10 20%

Fig. 1 Functional use of three-word sequences across four phases of intercultural exchange

Phase 3, more sequences of personal assertion are found, such as *I think I, I can't wait, looking forward to*, which occurred in very few instances in Phases 1 and 2. Moreover, when entering the spoken mode of communication, even more three-word sequences were used for social interaction, whereas the amount of use of necessary topics sequences dropped appreciably. With reference to the use of sequences as discourse devices, it is apparent from Fig. 1 that no improvement can be seen in terms of the amount of use in the online exchange over a year, and the sequences of this type used in different phases are also quite similar. For example, the sequences serving linking functions such as *and I am, and I love* and *and we have* occur in the top 50 highly recurrent sequences in the three phases of CMC.

6.2 *Functional Categories of Three-Word Sequences*

The previous section has demonstrated the use of three-word sequences over time in the one-year programme. This section takes a closer look at the functions that the high-frequency three-word units serve in the intercultural context, and also compares their use across two modes of communication. Table 2 presents the 50 most common three-word units retrieved from BATICC-O and BATICC-F, as well as large general corpora CANELC and CANCODE for further comparison. It is, however, worth noticing that a number of top 50 sequences do not have a clearly recognizable function, such as *to be a, to go to, to be the, to do it* and *to have a*, which are mainly constituted of high-frequency lexical words. These lexical sequences appear widely across online and spoken data, and this may be simply due to the highly recurrent nature of these grammatical fragments (Biber 2009; Ellis et al. 2008). As such, this current analysis excludes the five three-word sequences mentioned above. The following subsections will focus on the three different functional categories: social interactions, necessary topics and discourse devices.

Table 2 Functions of 50 most common three-word sequences across corpora

BATICC-O	CANELC	BATICC-F	CANCODE
SOCIAL INTERACTION:			
Summoning and greeting			
nice to meet	how are you	nice to meet	
to meet you		to meet you	
		how are you	
Questioning			
do you have	do you think	do you like	do you think
do you like	what do you	do you have	do you want
what do you		what do you	do you know
		how do you	you want to
		have you ever	what do you
		how is the	have you got
		did you see	
		do you do	
		do you go	
		how about you	
		do you think	
Offering			
	you want to	do you want	
	if you want	you want to	
	would you like		
Inviting			
come to Taiwan		come to Taiwan	
Expressing politeness			
	thanks for the		
Responding to requests			
	it would be		it would be
	would be a		
Asserting (personal)			
I have a	looking forward to	I don't know	I don't know
I go to	I have a	I want to	I don't think
I want to	I don't think	we have a	you have to
we have a	be able to	we went to	I think it's
I can play	I don't know	I have a	I think it
I can't wait	I want to	I think it's	I think I
we go to	I'm going to	we have to	you've got to
would like to	you have to	I think I	
I would like	I have to		
I would love	I think I		
	I am not		
	I think it		
Asserting (impersonal)			
It is very	going to be	it was very	it was a

(continued)

Table 2 (continued)

BATICC-O	CANELC	BATICC-F	CANCODE
there are many	It was a	it's very nice	there was a
	this is a		
	there is a		
	It is a		
	this is the		
Complying			
			yeah you know
			yeah I mean
			yeah I think
Responding			
			yeah yeah yeah
			no no no
NECESSARY TOPICS:			
Autobiography			
my name is		my name is	
I am #			
am # years			
My birthday is			
I live in			
birthday is on			
I am not			
I am very			
I come from			
Time/location			
in your country	in the UK	in the UK	at the moment
	of the year	in your country	the end of
	at the moment		all the time
	the end of		at the end
	of the day		
	the first time		
	at the end		
	out of the		
Quantity			
a lot of (54)	a lot of	a lot of	a lot of
	a couple of	a lot more	one of the
	a bit of		a bit of
	one of the		a couple of
	some of the		a little bit
	part of the		
	one of my		
	the rest of		
	one of those		
	most of the		

(continued)

Table 2 (continued)

BATICC-O	CANELC	BATICC-F	CANCODE
Likes			
I like to		fish and chips	
my favourite food		what's your favourite	
favourite food is		I don't like	
like to play		I like it	
I also like			
I love my			
I love to			
pearl milk tea			
Schools			
junior high school		go to school	
my school is		in your school	
I study in			
go to school			
Other topics			
with my friends	Happy New Year	Dragon Boat Festival	
Chinese New Year			
my friend and			
go to bed			
DISCOURSE DEVICES:			
Linking functions			
and I am	as well as	so do you	but I mean
but I am	but I think	and it was	and it was
and I love	and I have	and we have	and you know
but I don't		and we were	and I think
			and I was
Fluency devices/Elaboration			
	the fact that	so er erm	I mean I
		sort of like	you know I
		you know I	you know and
		I was like	you know what
		it was like	you know the
			you know yeah
			you know you
			I mean it's
			you know it's
			what I mean
			I mean you
			mm you know
			that you know
			know what I
Shifting topics			
by the way			
Exemplifiers			
		sort of thing	sort of thing
Evaluators			
		to be honest	

6.2.1 Social Interaction

One common function for which sequences are often employed is the maintenance of social interaction (see Nattinger and DeCarrico 1992; Schmitt and Carter 2004; Wray and Perkins 2000). In this category, a large amount of conventionalized language is typically associated with different speech acts in social interaction, such as *thanks for the* to express politeness, *it would be* to comply with a request, *I can't wait* to express personal intention, and *would you like* to express an offer. However, a single unit might sometimes serve multiple functions. In this case, for example, the sequence *do you want* derived from BATICC-F is used by participants for different speech acts. One such is an offer, which is a speech act in which "the speakers volunteer to do something beneficial for the listener (or a third party) or give something to the listener (or a third party)" (Carter and McCarthy 2006, p. 699). In the following two examples, although the surface form is a question, it is apparent to show the offers made by both British (i.e. <BT06>) in (1) and Taiwanese (i.e. <TW10>) students in (2).

(1) <BT06[1]>: Someone gave it to me. [laughter] **Do you want** it?
<TW05>: Yeah. Thanks. (passing the item)
(2) <TW10>: Look at that … **Do you want** to write England in Chinese?
<BT13>: Yeah. Have a go.
(<TW10> is writing on <BT13>'s workbook.)

In (1), according to Levinson (1983), <BT06>'s utterance is both an offer and a question in that *yeah* is used to respond to the question, while *thanks* responds to the offer. In this particular situation, BT06 passes the item to TW05 following TW05's response. It is obvious that *do you want* in (1) is used as an offer of the physical thing since if <BT06>'s utterance is simply a question, BT06 would have no obligation to do anything. Similarly, in (2), TW10 offers to do the Chinese writing for BT13 although no *thank you* is included in the response.

In other cases, *do you want* is used in the form of questions as requests or polite directives, which have the purpose of eliciting information (Tsui 1994), as in (3) and (4). In such cases, the speaker TW15 wishes the interlocutor to write his/her birthday on a card, and TW07 makes a request for everyone's mobile phone number. *Do you want* is also used simply as questioning in an interrogative form in that speakers ask specific information about a particular issue, event or other related topics, as in (5) and (6):

(3) <TW15>: Erm … .**do you want** to write your birthday on a card? I have a pen.
(<BT17> is writing on the card.)
<BT17>: We'll have to send a birthday card.
(4) <TW07>: **Do you want** to ask anyone's cell phone number or

[1]The speaker codes (e.g., <TW01>, <TW02>, <BT01>, <BT02>,…) represent participants from two different countries, namely TW and BT referring to Taiwanese and British learners respectively.

(passing a sheet) (Everyone is writing on the sheet.)
<BT09>: I'd just like to say thank you very much.

(5) <BT13>: What **do you want** <TW26>?
<TW26>: I want a paper and pen.

(6) <TW10>: So **do you want** to see anything in the Temple like …erm … what er

The differences between "questions as requests" and "questions as questions" can be seen from the preceding excerpts. According to Tsui (1994), questions simply elicit an obligatory verbal response so that "the interaction between the speaker and the addressee is completed entirely at the verbal level" (p. 80), as in (5) and (6). Requests, however, elicit "an obligatory non-verbal response with perhaps an accompanying verbal response, and the interaction is completed at the non-verbal level" (ibid.), as in (3) and (4). This can be also applied to the distinction between the "questions as offers" and "questions as questions", which were shown in the excerpts (1) and (2). Since more instances of the use of *do you want* serve the function of offers in BATICC-F, the sequence is categorized to this sub-group.

In addition, with respect to sequences expressing speech acts such as complying, offering, responding to requests and making personal assertions, it can be seen that a number of sequences include the modal verb *would*, such as *would you like, I would love, it would be, would be a*. A further look at the users of these sequences indicates that Taiwanese students use relatively few *would* expressions, particularly in CMC. Examples of such use in online correspondences, retrieved from the BATICC-O, are shown in (7).

(7) – Nice to see you. I am <TW28>. **Can we** make friends? (from <TW28>)
– yes sure. **I would love to** be friends. **It would be** really nice. (from <BT30>)
– **I would very much like** to be friends with you and I too am hoping to make lots of friends through the connecting classrooms experience! (from <BT32>)

As (7) shows, the Taiwanese learner TW28 expresses his/her desire to maintain the friendship by asking *Can we make friends*?. The British participant BT30 then responds to the request with the lexical phrases *I would love* and *it would be* to demonstrate willingness; similarly, the other British learner BT32 uses *I would very much like to be friends*. Many similar instances can be found in BATICC-O, indicating the high frequency with which modal *would* is used by British students to respond to requests in online communication. The calculation of the word frequencies of modal verbs indicates that young Taiwanese learners significantly underuse *could* and *would* and overuse *can* and *will* when compared to British English speakers. According to Carter and McCarthy (2006), *could* and *would* are generally perceived as more polite and less forceful than *can* and *will*.

As can be seen in Table 3, the majority of three-word sequences used in social interaction are for the expression of assertions, similar to what Biber et al. (2004) call *attitudinal/modality stance bundles*, which express "attitudes toward the actions or event described in the following proposition" (p. 390). Asserting sequences are divided into personal and impersonal; in interaction, most are overtly personal and express desire (e.g., *I want to*; *I can't wait*), personal opinions (e.g., *I think it*; *I think I*),

Table 3 Distribution of common three-word sequences across corpora

	BATICC-O		BATIC-F		CANELC		CANCODE	
Social interaction	18	36 %	27	54 %	27	54 %	23	46 %
Necessary topics	27	54 %	12	24 %	19	38 %	9	18 %
Discourse devices	5	10 %	11	22 %	4	8 %	18	36 %

intention/prediction (e.g., *I'm going to*; *I hope you*), ability (e.g., *I can play*; *be able to*), or obligation (e.g., *you have to*). These sequences are directly attributed to the speaker or writer. However, some sequences of asserting do not explicitly mention the speaker or writer, such as in descriptions of existence (e.g., *there are many*; *there was a*), evaluations of specific things or events (e.g., *it is very*; *it's very nice*), narratives of past events (e.g., *it was a, it was very*), or expressing predictions of future events (e.g., *going to be*).

A comparison of different columns in Table 3 illustrates the different uses of three-word sequences in different communication modes. In particular, it can be seen that sequences for asserting are more commonly used in CMC (i.e., BATICC-O and CANELC), while questioning, complying, and responding are considerably more frequent in FTF interaction (i.e., BATTIC-F and CANCODE). These differences can be explained in part by the highly interactive nature of FTF conversation, in which people are consistently asking each other questions, clarifying questions, and responding to questions. This, in turn, may facilitate personal relationship building (Belz 2007). In particular, the number of different questioning sequences in BATICC-F is extremely high. This suggests that the participants in intercultural exchange project also demonstrate *skills of discovery and interaction*, namely "the ability to employ a variety of questioning techniques in order to elicit from members of the foreign culture" (Byram 1997, p. 61), which are some of the most important skills that constitute intercultural competence. A closer examination of the sequences such as *do you have, how do you, do you do, do you think*, reveals that young learners in this project demonstrate a willingness to engage with others and a curiosity in discovering different perspectives regarding their own and other cultures. This is the prerequisite attitude of being an intercultural speaker, as has been discussed by a number of scholars (see Byram 1997; Belz 2007; Fantini 2012).

6.2.2 Necessary Topics

Another specific function of the use of sequences is that of introducing or progressing necessary topics (the second section of Table 3), that is, topics about which questions are often asked or which are necessary in daily conversation (Nattinger and DeCarrico 1992). These sequences provide overt signals on specific themes, such as autobiography (e.g., *my name is*), food (e.g., *fish and chips*), time/location (e.g., *in the UK*), school life (e.g., *in your school*), likes (e.g., *what's your favourite*), quantity (e.g., *a lot of*), and some culturally specific topics (*dragon boat festival*). However, very few high-frequency sequences found in the CANELC and CANCODE

reference corpora are grouped in the domains of autobiography, likes/food, or school life. For example, the autobiography sequence *my name is* is frequently used by both Taiwanese and British pupils, but only four occurrences are identified in the CANELC.

The sequences included in the *necessary topics* category are similar to what Biber (2009) calls *referential bundles*, which "identify an entity or single out some particular attribute of an entity as especially important" (p. 285). For example, a number of sequences refer to particular places or locations (e.g., *in the UK*; *in your country*) in BATICC-O and BATICC-F, while these are not found with a high frequency in CANELC and CANCODE. This is probably not surprising, as this project involved learners from different countries and they frequently talked about their own and other cultures during the exchange programme.

With regard to the domain of quantity, most of the sequences, such as *a couple of, a lot of, a bit of*, describe amounts or quantities of the subsequent head noun, as in (8)–(11).

(8) <TW07>: Really?
<BT07>: Not all the time – for **a couple of** days – and then there's **a couple of** months and it's quite warm. (BATICC-F)
(9) We have **a lot of** snow at the moment and I love it! (BATICC-O)
(10) my worst school memory was when i didnt get a part in the school play in primary school, **a couple of** years ago now. (BATICC-O)
(11) So it was **a bit of** a shame that the Waterside had only ordered two casks and they ran out about 20 min. (CANELC)

From the extracts above, we see how the participants use vagueness and approximations in both FTF and CMC. This shows that the young learners prefer to describe quantities with vague language to avoid being too precise and pedantic in intercultural exchange. As was claimed by O'Keeffe et al. (2007), especially "in such domains as references to number and quantity, where approximation rather than precision is the norm in conversation" (p. 74). This is sometimes described as "vague additives" (Channell 1994) or "vague approximators" (Koester 2007).

In addition, some sequences in this category may serve as focus markers, such that new information or the focus of the utterance often follows. Biber et al. (2004) label these *identification/focus bundles*, "focusing on the noun phrase following the bundle as especially important" (p. 394). This can be seen in (12), with *one of my*, and (13), with *one of those*, which identify the food and platforms respectively and are the focus of the utterances.

(12) **One of my** favourite food is Tomatoes on sticks. It is very sweet! (BATICC-O)
(13) If you have a blog at **one of those** platforms, follow us there. If not, just choose one (CANELC)

It is also worth noting that the time/location and quantity sequences found in CANELC outnumber those in the other three datasets. Some of the sequences, such as *one of my*, *some of the*, *part of the*, and *one of those*, do not commonly occur in FTF spoken discourse. Most of these sequences incorporate noun phrase and

prepositional phrase fragments that have been shown to be one of the typical features of written discourse (Biber 2009; Carter and McCarthy 2006; O'Keeffe et al. 2007). The use of three-word sequences in asynchronous CMC therefore demonstrate a closer approximation to those sequences used in ordinary written discourse as displayed in corpora of native speakers of English.

6.2.3 Discourse Devices

The third category of multi-word sequences is discourse devices, which refers to "lexical phrases that connect the meaning and the structure of the discourse" (Nattinger and DeCarrico 1992, p. 64). As such, they serve an organising function for the flow of information being transmitted, and further improve the fluency of utterances. From Table 3, the sequences which serve linking functions, such as *and I am*, *but I am, and I love, but I don't, and it was, and we were* are common in both online and FTF interaction. Examples of these sequences can be seen in the following excerpts:

(14) Yes there is a lot of snow in England **but I am** OK there is not too much around my area which is good I don't like too much snow **but I like** some (BATICC-O)
(15) My brother plays the guitar and is teaching me. he is nine **and I am** 13 **but he is** so much better than me!! (BATICC-O)
(16) <BT17>:… I walked around with Aiden and Katie **and it was** very fun.
(17) <BT09>: Yeah, yeah it was straight and like others we walked together.
<BT07>: **And we were** listening to music and … (BATICC-F)

It is evident that coordinating conjunctions *and* and *but* are frequently used by the participants to express a variety of logical relations between phrases and sentences in both online and spoken datasets. As shown in Table 3, somewhat equal numbers of three-word sequences which serve linking functions can be found across the four datasets. This result is slightly different from Crossley and Louwerse's (2007) study, which examines two-word sequences (bigrams) and found that the use of coordinating conjunctions collocating with first person pronouns, such as *and I, so I, but I,* and *and we*, is an important feature distinguishing natural dialogues from written discourse. However, their finding can be generated when comparing written discourse and unplanned real-time communication, while in this present study, the online discourse also exhibits this feature of unplanned speech.

Nevertheless, the sequences including coordinating conjunctions found in spoken discourse have a slightly different function to those in the CMC corpora. Many of them are used as a turn-initial resource for speakers, as in (25), and such use is not common in CMC. Evison (2008) defines such units *flexible instalment openers* because their "lack of specificity means that they can begin an instalment of talk without having to commit to a more complex relationship between upcoming and prior talk from the outset of the turn" (p. 223). As such, the turn is still occupied, and the processing load can further be eased.

The differing use of fluency devices in spoken discourse is also notable between the target and reference corpora. From the data in Table 2, it is apparent that larger numbers of sequences are found in CANCODE as compared with BATICC, and in particular, most of the sequences are centred by the two-word sequence *you know* or *I mean*, such as *you know I*, *you know it's*, *I mean I*. Many of the instances of *you know* sequences in BATICC-F function as interpersonal discourse markers, marking statements as assumed shared knowledge or experience between speakers and hearers (Östman 1981; Schiffrin 1987; Jucker and Smith 1998; Fox Tree and Schrock 2002; Fung and Carter 2007; Hellermann and Vergun 2007; O'Keeffe et al. 2007; House 2009). As Östman (1981) proposes, the highly frequent use of *you know* is to show that "[t]he speaker strives towards getting the addressee to cooperate and/or to accept the propositional content of his utterance as mutual background knowledge" (p. 17). For example:

(18) <BT13>: Hey Aiden – **you know** last night at the meeting thing –
<BT14>: Yeah.
<BT13>: … did you see that cat man who was there?

(19) <TW11>: In in typhoon, it's very … very bad, **you know**, it's it's wet =
<BT15>: Yeah.
<TW11>: =because it's raining and it's cold.
<BT15>: Windy as well.

In both cases, *you know* is used by speakers to invite addressee inferences based on their shared experience or knowledge. In (18), both BT13 and BT14 might be familiar with what BT13 said *last night at the meeting thing*; in (19), TW11 is talking about typhoon and is appealing to BT15's shared understanding of it. Moreover, in the conversations, *yeah* used by BT14 and BT15 serves as an acknowledgement, and this is expected since in inviting inferences, participants in conversation normally back-channel to show their understanding. Therefore the speakers may not only want to appeal to shared knowledge but also desire the interlocutors to participate and share more about their own ideas. As Jucker and Smith (1998) argue, *you know* does not just simply indicate that the recipient knows the information, but it often serve as "a device to aid in the joint construction of the representation of the event being described … *you know* invites the addressee to recognize both the relevance and the implications of the utterance" (p. 194). In this case, both Taiwanese and British participants utilize *you know* to seek speaker involvement in spoken interaction although it is used less frequently in Taiwanese discourse.

In the domain of fluency devices, it is also considering the sequence *so er erm*, which marks the speakers' hesitation in their utterance. Marking hesitation feature frequently in spontaneous conversations and fulfils an important pragmatic function, and it is pervasive in that *er* and *erm* are ranked 16th and 20th in the most frequent words in BATICC-F. Other three-word sequences that serve a similar function and occur at least three times include *I er I, er er I, I I I, er I er, I like erm* and *er I think*, which all contain hesitation items and/or repeats. What needs to be emphasized is that these sequences are mainly found in Taiwanese learners' speech; although *er* and *erm* are used slightly more frequently by the

British participants, there are very few three-word sequences containing these two items in their top 50 sequences and that the first sequence of such type is *erm I think* (rank 77). This also accords with De Cock's (2004) findings, which indicate that EFL learners use significantly more sequences which contain repeats and/or hesitation items than native speakers of English. In her analysis, 12 out of the top 20 high-frequency sequences are of this type, and the total numbers of hesitation or repeat sequences in her learner corpus are approximately three to four times larger than those found in native speakers' discourse.

With regard to the exemplifiers in discourse devices, the sequence *sort of thing* appears in the top 50 items in both BATICC-F and CANCODE. Such expression is often referred to as vague language in previous studies (e.g., Carter and McCarthy 2006). One of the primary functions of being vague is to "indicate assumed or shared knowledge and mark in-group membership" (O'Keeffe et al. 2007, p. 177). In this way, it is not necessary for speakers/writers to convey precise and concrete information, and the hearers/readers in most instances know what a vague expression refers to. As apparent in the following extract, BT13 and BT14 are discussing gift ideas for their fathers. BT14 uses the vague expression *that sort of thing* twice, and BT13 knows what he/she means.

(20) <BT14>: Right, because all of them … all the presents I've made … you know what I mean, like I made all the key rings they're more for Mum then you know … my dad doesn't like **that sort of thing**.
<BT13>: Yeah, I bought a load of rope bracelets for my dad.
<BT14>: My Dad's not into **that sort of thing**. I was going to get him like a model **or something** … If I do, I'll get him some alcohol from duty free …

The first use of *that sort of thing* refers to the presents that the speakers have made during the programme, and this reference appears to be a marker of shared knowledge and experience on which they can draw. The second use of *that sort of thing* refers to the gift that BT13 bought for his/her father so that the speaker BT14 does not necessarily need to repeat the noun phrase *a load of rope bracelets*, which are part of a larger category that is implicitly understood by speaker and hearer. Moreover, the use of *or something* basically indicates an alternative category of gifts, and such usage simply "keeps options open" (Carter and McCarthy 2006, p. 202). Such use of vague language when describing categories of items is sometimes referred to as a "vague category identifier" (Channell 1994) which is made up of an exemplar (i.e. *a model*) plus a vague tag (i.e. *or something*), where the exemplar directs listeners to identify the category referred to. In addition, the online discourse also presents the use of vague language, for example:

(21) i love summer but then i dont think it get to hot and all the bee **and stuff** are horrible x In the summer i go camping with my dad. (BATICC-O)
(22) cool, I want to be a doctor because I have always liked **that sort of** job. (BATICC-O)
(23) it would be enough wouldn't it… to write **something like that**. Even just once… (CANELC)

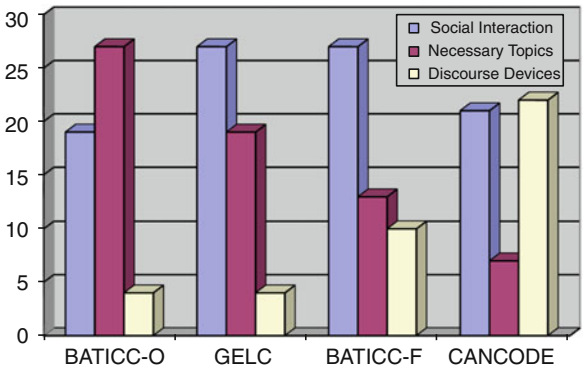


Fig. 2 Distribution of top 50 three-word sequences across functional types

(24) a Big Mac or a bucket of KFC?! It's usually the **sort of thing** we eat from the bag rather than taking it home and serving it up on a plate. (CANELC)

In these examples, the use of *and stuff, that sort of, something like that* and *sort of thing* indicate an assumption of shared personal experience with the audience on the part of the writer. As a result, these expressions are often used as "a marker of intersubjectivity" with the implicit meaning that "there is more to say on this, but I don't have to because you know what I mean" (Overstreet and Yule 2002, p. 787). Such forms of vague exemplifier have found to be particularly distinctive features in adolescent speech and informal online messages in terms of use and frequency as compared with adult talk (see Martínez 2011; Tagliamonte and Denis 2010).

6.2.4 Distribution of Common Three-Word Sequences Across Functional Types

The previous sections have demonstrated that the use of three-word sequences is often tied to particular conditions of use, and can be identified according to Nattinger and DeCarrico's (1992) three functional categories: social interaction, necessary topics and discourse devices. Nevertheless, it can be clearly seen that the use of sequences in different communication modes differs in relation to the functional types. Table 3 and Fig. 2 present the distribution of functions served by three-word sequences across corpora.

The sequences for social interaction are extremely common across each dataset, ranging from 38 to 54 % of the top 50 high-frequency sequences in each corpus. On the other hand, a more noticeable distribution difference of sequences across corpora can be seen in the percentage of necessary topics and discourse devices, which range from 18 to 54 % and from 8 to 36 % respectively. In necessary topics, overall they are particularly common in CMC, presented in BATICC-O (54 %) and CANELC (38 %), which are strikingly higher than the percentage figures of the spoken data presented in BATICC-F (24 %) and CANCODE (18 %). The use of multi-word sequences as discourse devices demonstrates the opposite pattern in that

Table 4 Accumulative frequencies of sequences and the statistical test of significance

	BATICC-O	BATICC-F	CANELC	CANCODE	Significance
Social interaction	361	278			$p<.05$
			1237	31227	$p<.001$
Necessary topics	838	118			$p<.001$
			1178	10782	$p<.01$
Discourse devices	92	78			
			284	18274	$p<.001$

a largely higher rate can be found in spoken discourse than CMC. As shown in Table 3, the sequences of discourse devices can be found only in four instances (8 %) out of the first 50 high-frequency sequences in both BATICC-O and CANELC, while the percentage figures of BATICC-F and CANCODE reached 22 and 42 % respectively.

Table 4 shows statistically significant differences in the use of three-word sequences with different functions among the corpora using log-likelihood (LL) ratio (Rayson 2008) based on the accumulative frequencies of sequences. The table indicates significant differences between CANELC and CANCODE in three functional categories: social interaction (LL = −1374.98; $p<.001$), necessary topics (LL = 8.13; $p<.01$) and discourse devices (LL = −1907.71; $p<.001$). The negative values indicate a significant higher rate of sequences as social interaction and discourse devices in CANCODE. The difference in distribution of functional categories between BATICC-O and BATICC-F is significant in social interaction (LL = −6.28; $p<.05$) and necessary topics (LL = 328.57; $p<.001$). However, the distribution difference in the use of discourse devices between BATICC-O and BATICC-F does not reach significant level. This notwithstanding, the three-word sequences commonly used in the two datasets are largely different. For example, the sequences as discourse devices in BATICC-O mainly serve a linking function, while the ones in BATICC-F include four different functional types (see Table 2).

The highly frequent use of three-word sequences for social interaction in BATICC-F is likely due to the phatic nature of FTF communication in that young learners focused more on social interaction instead of specific information when they meet FTF. It may be also because of the fact that the multi-word expressions in CMC are less interactional in nature. Concerning necessary topics, the significantly higher rate in BATICC-O might be due to the fact that the patterns of language use on electronic discussion boards reflect the particular topics that the participants are interested in, while in FTF interaction, topics are more easily shifted in the immediate environment. The discourse is also more likely to be oriented to topics, which can be referred to pronominally (e.g., *it, this, that*) (higher frequencies of these proximal deictic forms can be found in BATICC-F), rather than written forms, which tend to require full lexicalization, and hence more topic-instantiating sequences. In addition, the notably greater discourse devices used in BATICC-F, compared with BATICC-O can possibly be attributed to differences in the nature of spoken and written modes. Nattinger and DeCarrico explain that

> writers are removed from their audience in a way that speakers are not from theirs. Speakers and hearers work jointly, in a rather spontaneous, unplanned manner, to establish meaning inside the immediate context in which the interaction takes place. They can thus rely on shared signals … to regulate the speed and content of the message (1992, p. 83).

It emerges that participants in online interaction based on written form may not have such proximate relationships with each other since the discourse is more explicit, with less recurrent discourse devices in the mediation of online discussion. Fewer high-frequency discourse marking sequences may be due to the fact that the foregoing discourse is preserved in online communication, rather than real-time speech, which needs to be more explicitly organized.

7 Conclusion

The present study has explored the discourse functions of recurrent three-word sequences in intercultural CMC and FTF communication (BATICC-O and BATICC-F), as well as two large reference corpora (CANELC and CANCODE). It is evident that the high-frequency three-word sequences in the four datasets serve three central functions: social interaction, necessary topics and discourse devices. These findings add to a growing body of literature on the functional use of multi-word expressions, which has been shown in a range of previous studies (e.g., Biber et al. 2004; Nattinger and DeCarrico 1992; Schmitt and Carter 2004; Wray and Perkins 2000). It further appears that three-word sequences often perform systematic discourse functions, even though they do not usually constitute complete grammatical or idiomatic structures. They function as "important building blocks in discourse" (Biber 2009, p. 284), and accord with interlocutors' expectations and preferences, which may facilitate efficient and effective communication for different communicative purposes (Schmitt and Carter 2004; Wood 2010).

It is also apparent that three-word sequences employed in CMC and spoken conversation are significantly different. In the category of social interaction, questioning, complying and responding are generally more frequently used in spoken communication, while sequences used for making assertions in both personal and impersonal contexts are found in more instances in CMC. In addition, the sequences employed in the area of necessary topics are particularly common in CMC, reflecting topics such as autobiography, time/location, likes and interests, quantity, and schools, while these sequences are not found with a high frequency in FTF talk. With regard to the sequences functioning as discourse devices, a large number of highly recurrent sequences in BATIC-F and CANCODE are not commonly used in online discussion. Some examples include fluency devices (e.g., *you know I, I mean I, it was like*), exemplifiers (e.g., *sort of thing*) and evaluators (e.g., *to be honest*). Nevertheless, the sequences that serve linking functions (e.g., *and I am, and it was, but I don't, and we have*) are very common in both CMC and spoken interaction.

The analysis of the use of three-word sequences by different groups of participants also reveals a number of differences between British and Taiwanese

participants discourse. For example, some sequences that frequently used by British participants can only be found in a very few instances in Taiwanese learners' discourse, such as sequences serving linking functions (e.g. *and I love, but I think*), expressions with *would* for responding to requests (e.g. *it would be, would love to*), vague exemplifiers (e.g. *sort of like, sort of thing, things like that*), vague quantifiers (e.g. *a couple of*) and hedges for downtoning their utterances (e.g. *a bit of, a little bit*). These findings highlight the need for teachers and material developers to incorporate multi-word formulaic expressions commonly used by native English speakers in learning materials and EFL instruction. Learners will thus be exposed to appropriate expressions in different communicative situations. As Schmitt and Carter (2004) claim, "formulaic sequences are not only helpful for efficient language usage; they are essential for appropriate language use" (p. 10).

In conclusion, this study sheds light on how multi-word sequences are used in different modes of intercultural communication by adolescents and in general corpora of online and spoken discourse. However, there remain a number of caveats to be noted regarding the present study, most notably that, due to the small size of the BATICC datasets, the present results are not necessarily generalizable to other data. This notwithstanding, the size and composition of the specialized corpus makes it more manageable for qualitative studies and permits a closer link between the corpus and the contexts of its data in order to understand the functional features of three-word sequences in CMC and FTF interaction. Secondly, participants in the present study were teenagers from Taiwan and England and consequently the findings may not be transferable to the use of multi-word sequences by native and non-native speakers of English generally. Moreover, I have only focused on the 50 most common three-word sequences retrieved from four corpora. A number of sequences that may be unique to this particular intercultural setting but which have a lower frequency are therefore neglected. Future research may also analyse these context-specific, lower-frequency sequences or those used by other second language learners/speakers. While some possible limitations are recognised, this study nevertheless illuminates important functional aspects of recurrent multi-word sequences in an intercultural setting, and has further pedagogical implications in relation to EFL course design for online and FTF intercultural communication.

References

Adolphs, S. 2006. *Introducing electronic text analysis. A practical guide for language and literary studies*. New York: Routledge.

Adolphs, S., and V. Durow. 2004. Social-cultural integration and the development of formulaic sequences. In *Formulaic sequences: Acquisition, processing and use*, ed. N. Schmitt, 107–126. Amsterdam: John Benjamins.

Baker, P. 2006. *Using corpora in discourse analysis*. London: Continuum.

Belz, J.A. 2007. The development of intercultural competence in online interaction. In *Online intercultural exchange: An introduction for foreign language teachers*, ed. R. O'Dowd, 127–166. Clevedon: Multilingual Matters.

Biber, D. 2009. A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics* 14(3): 275–311.

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

Biber, D., S. Conrad, and V. Cortes. 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371–405.

Byram, M. 1997. *Teaching and assessing intercultural communicative competence*. Clevedon: Multilingual Matters.

Carter, R., and M. McCarthy. 2006. *Cambridge grammar of English: A comprehensive guide*. Cambridge: Cambridge University Press.

Channell, J. 1994. *Vague language*. Oxford: Oxford University Press.

Chen, L. 2010. An investigation of lexical bundles in ESP textbooks and electrical engineering introductory textbooks. In *Perspectives on formulaic language: Acquisition and communication*, ed. D. Wood, 107–128. London/New York: Continuum.

Chen, Y.H., and P. Baker. 2010. Lexical bundles in L1 and L2 academic writing. *Language, Learning and Technology* 14(2): 30–49.

Conklin, K., and N. Schmitt. 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics* 29(1): 72–89.

Coulmas, F. 1979. On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics* 3: 239–266.

Crossley, S., and M. Louwerse. 2007. Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics* 12(4): 453–478.

De Cock, S. 2004. Preferred sequences of words in NS and NNS speech. *Belgium Journal of English and Literatures (BELL)*, New Series 2: 225–246.

Ellis, N., R. Simpson-Vlach, and C. Maynard. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 42(3): 375–396.

Erman, B., and B. Warren. 2000. The idiom principle and the open-choice principle. *Text* 20: 29–62.

Evison, J. 2008. *Turn-openers in academic talk: An exploration of discourse responsibility*. Unpublished doctoral thesis, University of Nottingham, UK.

Fantini, A.E. 2012. Language: an essential component of intercultural communicative competence. In *The Routledge handbook of language and intercultural communication*, ed. J. Jackson, 263–278. New York: Routledge.

Flowerdew, L. 2004. The argument for using English specialized corpora to understand academic and professional language. In *Discourse in the professions: Perspectives from corpus linguistics*, ed. U. Connor and T.A. Upton. Amsterdam: John Benjamins Publishing Company.

Foster, P. 2001. Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In *Language tasks: Teaching, learning and testing*, ed. M. Bygate, P. Skehan, and M. Swain, 74–93. Harlow: Longman.

Fox Tree, J.E., and J.C. Schrock. 2002. Basic meanings of *you know* and *I mean*. *Journal of Pragmatics* 34: 727–747.

Fung, L., and R. Carter. 2007. Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics* 28(3): 410–439.

Greaves, C., and M. Warren. 2010. What can a corpus tell us about multi-word units? In *The Routledge handbook of corpus linguistics*, ed. M. McCarthy and A. O'Keeffe, 212–226. London: Routledge.

Hakuta, K. 1974. Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning* 24: 287–298.

Handford, M. 2010. What can a corpus tell us about specialist genres? In *The Routledge handbook of corpus linguistics*, ed. M. McCarthy and A. O'Keeffe, 255–269. London: Routledge.

Hanna, B., and J. de Nooy. 2003. A funny thing happened on the way to the forum: Electronic discussion and foreign language learning. *Language, Learning and Technology* 7(1): 71–85.

Hellermann, J., and A. Vergun. 2007. Language which is not taught: The discourse marker use of beginning adult learners of English. *Journal of Pragmatics* 39: 157–179.

Hill, J. 2001. Revising priorities: From grammatical failure to collocational success. In *Teaching collocation: Further development in the lexical approach*, ed. M. Lewis, 47–69. Hove: LTP.

House, J. 2009. Subjectivity in English as Lingua Franca discourse: The case of *you know*. *Intercultural Pragmatics* 2: 171–193.

Hunston, S. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Jucker, A.H., and S.W. Smith. 1998. And people just you know like 'wow': Discourse markers as negotiating strategies. In *Discourse markers. Descriptions and theory*, ed. A.H. Jucker and Y. Ziv, 171–201. Amsterdam: John Benjamins.

Kjellmer, G. 1994. *A dictionary of English collocations*, 3 Vols. Oxford: Clarendon Press.

Koester, A. 2007. "About twelve thousand or so": Vagueness in north American and UK offices. In *Vague language explored*, ed. J. Cutting, 40–61. Basingstoke: Palgrave Macmillan.

Koester, A. 2010. Building small specialised corpora. In *The Routledge handbook of corpus linguistics*, ed. M. McCarthy and A. O'Keeffe, 66–79. London: Routledge.

Leech, G., P. Rayson, and A. Wilson. 2001. *Word frequencies in written and spoken English*. Harlow: Longman.

Levinson, S. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Liaw, M.L., and S.B. Master. 2010. Understanding telecollaboration through an analysis of intercultural discourse. *Computer Assisted Language Learning* 23(1): 21–40.

Martínez, I.M.P. 2011. I might, I might go I mean it depends on money things and stuff. A preliminary analysis of general extenders in British teenagers' discourse. *Journal of Pragmatics* 43: 2452–2470.

McEnery, T., R. Xiao, and Y. Tono. 2006. *Corpus-based language studies*. London: Routledge.

Nation, I.S.P., and S. Webb. 2011. *Researching and analyzing vocabulary*. Boston: Heinle.

Nattinger, J.R., and J.S. DeCarrico. 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

O'Keeffe, A., M. McCarthy, and R. Carter. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

O'Keeffe, A., B. Clancy, and S. Adolphs. 2011. *Introducing pragmatics in use*. New York/ Abingdon: Routledge.

Östman, J.O. 1981. *You know: A discourse functional approach, pragmatics and beyond II: 7*. Amsterdam: John Benjamins.

Overstreet, M., and G. Yule. 2002. The metapragmatics of *and everything*. *Journal of Pragmatics* 34(6): 785–794.

Rayson, P. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics* 13(4): 519–549.

Schiffrin, D. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

Schmitt, N. 2010. *Researching vocabulary: A vocabulary manual*. London: Palgrave Macmillan.

Schmitt, N., and R. Carter. 2004. Formulaic sequences in action. In *Formulaic sequences: Acquisition, processing and use*, ed. N. Schmitt, 1–22. Amsterdam: John Benjamins Publishing Company.

Schmitt, N. (ed.). 2004. *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins Publishing Company.

Scott, M.R. 2008. *WordSmith Tools version 5.0*. Liverpool: Lexical Analysis Software.

Scott, M.R. 2010. *WordSmith Tools help manual*. Version 5.0. Liverpool: Lexical Analysis Software.

Simpson-Vlach, R., and N.C. Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31(4): 487–512.

Sinclair, J. 2001. Preface to small corpus studies and ELT. In *Small corpus studies and ELT*, ed. M. Ghadessy, A. Henry, and R. Roseberry, vii–xv. Amsterdam: John Benjamins.

Tagliamonte, S., and D. Denis. 2010. The 'stuff' of change: general extenders in Toronto, Canada. *Journal of English Linguistics* 38: 335–368.

Tremblay, A., and H. Baayen. 2010. Holistic processing of regular four-word sequences: A behavioural and ERP study of the effects of structure, frequency, and probability on immediate free recall. In *Perspectives on formulaic language: Acquisition and communication*, ed. D. Wood, 151–173. London/New York: Continuum.

Tsui, A.B.M. 1994. *English conversation*. Oxford: Oxford University Press.

Walsh, S., T. Morton, and A. O'Keeffe. 2011. Analysing university spoken interaction. *International Journal of Corpus Linguistics* 16(3): 325–344.

Wood, D. 2010. Lexical clusters in an EAP textbook corpus. In *Perspectives on formulaic language: Acquisition and communication*, ed. D. Wood, 88–106. London/New York: Continuum.

Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A., and M. Perkins. 2000. The functions of formulaic language: An integrated model. *Language and Communication* 20(1): 1–28.

Formality in Digital Discourse: A Study of Hedging in CANELC

Dawn Knight, Svenja Adolphs, and Ronald Carter

1 Introduction

Technology has transformed the way we communicate in the modern digital age. No longer do we simply rely on speech and writing but also on a range of different forms of 'e-language'. E-language is defined here as any communicative, interactive and/or linguistic stimulus that is digitally based and 'incorporates multiple forms of media bridging the physical and digital' (Boyd and Heer 2006: 1): from e-mails to discussion board threads, SMS messages and so on ('e-language' is also known as Computer Mediated Communication, CMC: see Walther 1996; Garcia and Jacobs 1999; Herring 1999 and Thurlow et al. 2004, and 'netspeak', Crystal 2003: 17). As a relatively new 'genre' of communication (Herring 2002), the definition and description of the features of e-language and how it compares and contrasts with spoken and written genres of communication is an on-going concern in studies of CMC, Applied Linguistics, Corpus Linguistics and beyond. This is something that will be examined in more detail in the current chapter.

Based on Crystal (2003: 17), there is a suggestion that spoken and written language effectively exist on a 'continuum' of formality (also see Condon and Cech 1996; Ko 1996; Herring 2007 for further discussions on the differences between spoken and written discourse). The 'more' formal language structures exist on the

D. Knight (✉)
School of Education, Communication and Language Sciences,
Newcastle University, Newcastle NE1 7RU, UK
e-mail: Dawn.Knight@ncl.ac.uk

S. Adolphs • R. Carter
School of English Studies, The University of Nottingham,
University Park, Nottingham NG7 2RD, UK

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_7,

left of the continuum, where written language is conventionally positioned, and the least formal exists towards the right end of the continuum, where spoken language is conventionally perceived to be positioned (although obviously their positioning is somewhat fluid as no absolute positioning in this abstract notion can ever exist – it is a theoretical continuum not a static classification system).

Considered as a distinct genre of communication, Crystal suggests that 'netspeak' is perhaps somewhere in the middle, between spoken and written language (2003: 17). He suggests that there is essentially a blurring of traditional characteristics of spoken and written language, in digital communication, making it a combination of both of the more 'traditional' genres (also Biber 1993; Collot and Belmore 1996; Yates 1996; Crystal 2001 for further discussion). Others have added to this notion, instead suggesting that each e-language 'mode' (Murray 1988) is structurally, semantically and pragmatically different from one another as well as spoken and written language types, making their relative positioning along this continuum of formality highly variable (see Murray 1988; Baym 1995; Cherny 1999; Herring 1996).

Levels of formality in *specific* modes of e-language have already received attention from researchers (see works by Sutherland 2002; Hard af Segersteg 2002; Shortis 2007; Crystal 2008 for further details). For example, Tagg (2009) and Ling (2003) both report on the tendency for SMS messages to be immediate and personal, written in the first person and directed to specific recipients. Tagg adds to this, underlining that 'the informal and intimate nature of texting encourages the use of speech-like language' in this e-language mode (2009: 17, also see Crystal 2003; Oksman and Turtianen 2004). Similarly, Baron highlights that although email, as with texting and other common forms of e-language, is typed or 'written' rather than spoken, 'participants exploit it for typically spoken purposes' (1998: 36), and it therefore shares more similarities with communication situated at the spoken rather than written end of the continuum.

Levels of formality across e-language as a specific *genre* and the relationships that exist between individual *modes*, however, is something that remains under-explored in corpus-based analyses of real-life data. Initial developments in this area of research have been made by Knight et al. (forthcoming, 2012) who provided some preliminary observations about the frequency of pronouns and deictic markers in e-language, compared to written and spoken excerpts from the BNC.[1] This study is extended in the present chapter but with a focus, instead, on the use of forms of hedging in e-language. The corpus used in this chapter is CANELC, the Cambridge and Nottingham e-language Corpus, a one-million-word corpus of digital discourse taken from British contributors or those posting to British websites in 2010–2011. It includes data from discussion boards, blogs, tweets, emails and SMS messages, distributed according to Fig. 1 (word counts for each mode are included in this figure).

[1]The British National Corpus, BNC, is a 100 million word corpus of written and spoken discourse in English. For more information see: http://www.natcorp.ox.ac.uk/

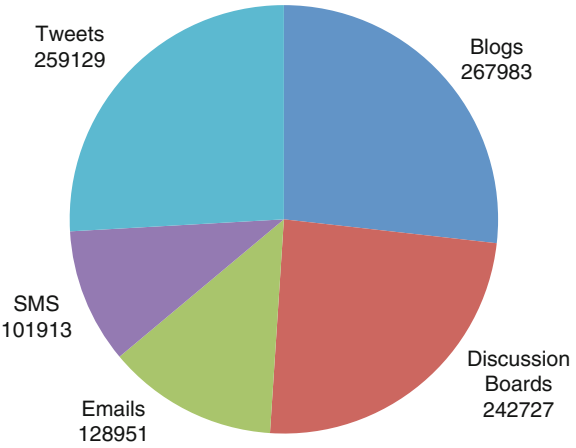


Fig. 1 The contents of the CANELC corpus

CANELC was built to allow for the querying of data at the general level of the *genre* of interaction as well as at the level of individual the communicative *mode*. So, using results from corpus-pragmatic based enquiries of CANELC, we will aim to create a deeper understanding of how different modes of e-language relate to Crystal's notion of the 'continuum' of formality.

2 Corpus Pragmatics

2.1 Overview

The study of the pragmatics of language use has traditionally concentrated on spoken registers rather than written language because the latter tends to be 'referentially explicit' (McEnery et al. 2006: 104) while the former allows for a more 'extensive reference to the physical and temporal situation of discourse' (Biber 1988: 144) in the construction of meaning. Spoken interaction is, in other words, highly context specific, and meaning is not only determined by the specific spoken or written 'sign' (Morris 1946: 287) used, but by a range of other 'extrinsic'; 'social, cultural and interactive' factors, and 'intrinsic', 'cognitive, affective and conative' factors that exist (Kopytko 2003: 45; also see Labov 1972; van Dijk 1977; Duranti and Goodwin 1992; Eckert and Rickford 2001; Fetzer 2004, for further discussion on language and context).

There is no one-to-one relationship between language form and function as the interpretation of a given message is highly dependent on the communicative function of a word or utterance, in a specific discursive context (for discussions of language and context see Labov 1972; Bates 1976; Nelson et al. 1985; Brown 1989; Halliday and Hasan 1989; Duranti and Goodwin 1992; Widdowson 1998; Green 2002; Scollon and Scollon 2003). In spoken communication, much of the discursive context is 'shared' (McEnery et al. 2006: 105) between a speaker and an interlocutor.

This affects the type of language used as there is a temporal and/or physical closeness in spoken discourse between the individuals as well as a shared knowledge about the immediate communicative context. This provides a 'clear advantage in using contextual expressions such as *I*, *there*, or *now*, [for example,] which are shorter and more direct' (Heylighen and Dewaele 2002: 301). Depending on the relationship and social distance between the speaker and interlocutor, speakers can thus use less formal expressions and a larger number of pronouns and deictic markers in this shared communicative space (see Fowler and Kress 1979; Chafe and Danielewicz 1987; Biber 1992; Biber et al. 1999; Leech 2000; Carter and McCarthy 2006; Atkins 2011). There is more of a gulf in spatial distance and time between writers and readers of written texts as there is no guarantee of when a text may be read or by whom. Written texts are not as contextually bound and thus often lack the shared knowledge and understanding between writer and reader, which often correlates with a decrease in the use of contextual (deictic) expressions in these texts.

While not necessarily true of all forms of e-language (instant messaging, IM, for example), the different modes of data included in CANELC are somewhat similar to one another in the fact that they do not 'require that users be logged on at the same time in order to send and receive messages' (Herring 2007: 13). The content sent via these different modes are 'stored at the addressee's site until they can be read' by the recipient (Herring 2007: 13). They are not forms of communication which necessarily require an instant response as, again, IMs do and face-to-face (spoken) interaction does. They are, therefore, asynchronous (for more detailed discussion of synchronicity see Condon and Cech 1996; Ko 1996; Herring 2007).

This asynchronicity means that the data in CANELC is arguably structurally organised in a way that is more consistent with written than spoken language (which is also asynchronous). It is interesting, then, to note that it is actually often the case that only a few seconds or minutes passes between the time when a message is sent and attended to across different e-language modes, despite this asynchronicity. There may in fact only be a short delay between the time a message is composed and read/responded to (although there is likely to be some inconsistency in the average time taken across the different modes of e-language). This is likely to reduce the temporal and social distance between sender and receiver as highly context-specific information about the message (related to time) is more likely to be shared and understood.

As a consequence of this, as outlined in Knight et al. (forthcoming, 2012), there is often a frequent use of 'temporal referents….deictic marking (as with the prolific use of personal pronouns)' in e-language. These discursive features again hint at forms of communication that are potentially allowing for an immediate or near-immediate information exchange, a forum for communicating reports of events and incidents in near real-time, as the understanding of the temporal referent is shared'. There is a shared digital space rather than physical space, within which 'the social, physical and temporal context is frequently changeable' (Knight et al. forthcoming, 2012). This is contrary to what is expected from asynchronous

Contribution name: F&D.19 (British Female Student, aged 20-24, sent 20:12:00 on 26/04/11)

hmm ***kind of*** *mixed opinions here* ***I figure*** *its not a* ***particularly*** *busy pub, especially on a Wednesday evening and* ***maybe*** *they'll be nice seeing as its my birthday *wishful thinking**

Fig. 2 An example of hedging, taken from the discussion board data in CANELC

communicating, aligning e-language more closely to more informal, spoken discourse, despite the fact it is not synchronous and is typed/written rather than spoken.

2.2 Hedging

In addition to pronouns and deictic markers, another pervasive feature that relates to levels of formality in discourse is the use of hedging (first coined by Lakoff 1972: 195). In pragmatics, hedges are 'expression[s] of tentativeness and possibility' (Hyland 1996: 433) which operate to 'mitigate the directness of what we say and so operate as face-saving devices' (O'Keeffe et al. 2007: 174 – for more information on politeness theory and the notion of 'face', see Brown and Levinson 1978, 1987). They are 'pragmatic markers' (Carter and McCarthy 2006: 223) which can be used 'to downtone…..the force of an utterance for various reasons e.g. politeness, indirectness, vagueness and understatement' (Farr et al. 2004: 13). The specific form, frequency and functions that hedges adopt also 'vary relative to context' (O'Keeffe et al. 2007: 174). Examples of hedging are seen in Fig. 2:

We see the use of four hedges (in bold) in this discussion board thread. The contributor is making plans for her birthday evening, discussing the possibility of inviting a party of friends to a local pub to celebrate. *Kind of* operates as an inexact stance adverb, softening the content of the thread. As with *maybe*, *kind of* acts almost as a 'downtoner', as instead of saying '*it would be nice to go the pub, especially since it is my birthday*', the use of this hedge provides an approximate reflection of what the contributor really means (Hübler 1983: 68). *I figure* also functions in a similar way, acting as a verb with a modal meaning, used to soften the meaning of the assumption about the pub, in order to mitigate against a potential face threat for the sender or receiver of the message, while *particularly* also has a similar effect as an omission of the adverb in this context would result in the utterance seeming blunt.

As face-saving devices, 'softeners' (Nikula 1997: 188), the frequent use of hedges is often linked to formal rather than informal contexts of communication (this is true of both spoken and written discourse, but given the tendency for written to be 'more' formal, the level of hedging is generally higher for written discourse vs. spoken discourse). Farr and O'Keeffe's (2002) study of hedging in the spoken

LCIE corpus (Limerick Corpus of Irish English[2]) best illustrates this pattern (2002). In this study, hedges were found to be most frequently used in institutional settings including teacher training contexts and radio discourse, with their use reducing in conversations between family and friends (see Farr et al. 2004) where there 'fixed relationships' (Clancy 2002), a closeness between speakers and listeners (creating less of need for participants to save face). The context where the fewest hedges were used in the corpus was in shop encounters. This is 'perhaps explained by the lesser need to protect face in service encounters, where a customer and a server do not know each other, and where they are interacting within transactional roles' (O'Keeffe et al. 2007: 176). The potential face threat is lower so the use of the mitigating hedging devices is not as essential in such discursive contexts.

Having said this, other studies have suggested that since it is performed in 'real-time' (Leech 2000), spoken 'conversation is [often] more vague than written genres' (McEnery et al. 2006: 105), so an increase in the frequency of certain forms of hedging functioning as vague language markers is often seen. For example, based on queries of the World Edition of the BNC (British National Corpus), Gries and David (2007) discovered that *kind of* and *sort of* were both forms of hedges functioning as vague stance adverbs that are frequently used in spoken discourse, in comparison to written discourse. Although, of these two clusters, *sort of* was significantly more common in written mode than *kind of*, while the reverse was found to be true of the spoken mode. Of written communication specifically, Biber et al. reported that the clusters *kind of* and *sort of* are both used more frequently in formal, academic prose than in other written registers (based on a study of the Longman Spoken and Written English Corpus, 1999: 560–561, other studies of these clusters have been carried out by Crystal and Davy 1975 and Quirk et al. 1985 – comparing their frequency of use between British and American English).

This pattern is inversely true of more private and personal forms of communication as opposed to more public forms (Carter and McCarthy 2006: 9–16). So written interaction, for example, that is most public (professional) and formal in nature (a government policy document for example), will likely see an increase in the number of vague stance adverbs used, when compared to a more personal expression of feelings, for example as this 'softening' function is unlikely to be required with close or intimate relationships.

Numerous other studies have been carried out on hedging in written discourse (Dubois 1987; Channell 1990; Drave 1995; Allison 1995), spoken interaction (see Crystal and Davy 1975; Brown and Yule 1983; McCarthy 1991; Cheng and Warren 1999; Jucker et al. 2003 for examples) and individual modes of e-language including SMS messages (Crystal 2001; Tagg 2009), Blogs (Myers 2010), Instant Messaging (IMs – Brennan and O'Haeri 1999), Discussion Boards (Atkins 2011) and Twitter (Benjamin 2011). More large scale corpus-based, studies have also examined vague language (arguably a sub-set of hedging) in both written and written discourse

[2]The Limerick Corpus of Irish English, LCIE, is a one million word corpus of spoken interaction from a range of different speech genres in Irish.

(Channell 1985, 1994; Kennedy 1987). To date, however, no studies offer an insight into hedging use across these different communicative genres. The current study aims to fill this research 'gap'.

3 Analysis

3.1 Study Questions

To build on the foundations of what was previously discovered about levels of formality in e-language (using CANELC – Knight et al. forthcoming, 2012), the following sections focus on the use of hedges in more detail. The analyses address the following research questions:

- Is there a significant difference in the frequency of hedging used:
  - Between all modes of e-language in CANELC, compared with data from the spoken and written BNC?
  - Between the different topic categories of data included in CANELC?
- What do the frequency and use of this phenomenon reveal about the levels of formality within and across the different modes of e-language in CANELC?

To answer these questions, the following sections present results from an analysis the use of hedges in e-language compared to one-million-word samples from the written and spoken BNC samples (which contain 968,267 and 982,712 words respectively). Given that the size of the corpora used are slightly inconsistent, the results are normalised using statistical measures so accurate comparisons can be made. The analyses are conducted out using Rayson's WMatrix software (2003) which includes utilities for carrying out word, cluster and parts of speech queries (centring around the production of key word lists and key-word-in-context, KWIC, outputs), and allows researchers to explore the patterned use of these features in a corpus. With the use of the WMatrix semantic tagger, common themes and semantic associations connected with corpora can also be queried using the software.

In addition to the 'data' taken from communication performed across the different e-language modes, CANELC also contains detailed metadata records: data about the data. Metadata is critical to a corpus as without it 'the investigator has nothing but disconnected words of unknowable provenance or authenticity' (Burnard 2005) to examine. As outlined by Knight (2011: 31, based on Burnard 2005) 'the inclusion of this information assists in identifying the name of the corpus (administrative metadata), who constructed it, and where and when this was completed (editorial metadata), together with details of how components of the corpus have been tagged, classified (descriptive metadata), encoded and analysed (analytic metadata)'. Collectively, this information allows us to reconstruct aspects of the reality of the discursive context in which specific e-language messages were sent, allowing us to

Topic / Genre Codes:		Topic / Genre Codes:	
	News, Media and Current Affairs	D	Music
A	Politics		Sports
	Business and Finance		
	Weather and the Environment		Celebrity news and gossip
		E	TV
	Culture, Literature and the Arts		Humour
B	Fashion		
	Teaching, Academia and Education		Health and Beauty
		F	Parenting and Family Life
	Technology, Computers and gaming		Personal and Daily Life
C	Hobbies and Pastimes		
	Travel		
	Cookery		

Fig. 3 Topics featured in CANELC

frame the language in a more contextually accurate way. The following metadata is included in CANELC:

- Author's (and receivers) name, age, gender, nationality
- Date and time composed
- Intended recipient
- Content
- General topic of content
- Follow up comments/responses
- 'Other' relevant information

Regarding 'general topic of content', it is viable to note that in addition to the metadata information, data in CANELC is also broadly categorised by topic. This is based on the schema presented in Fig. 3.

Topics in category 'A' are aligned with more public concerns such as news, politics and current affairs, while those in category 'F' are more aligned with personal issues such as personal and daily life (with B-E existing almost on a continuum between these poles). The distribution of the CANELC data, by number of words, across these different topic categories is represented in Fig. 4.

Figure 4 illustrates that across the entire corpus there is a dominance of contributions in categories 'F' and 'A'. The majority of data in category 'F' is included in the SMS messages and personal emails included in the corpus, which primarily contain language discussing topics concerning aspects of personal and daily life. More public, outward facing, topics such as business, finance and the news are frequently featured in the language of the blogs, tweets and discussion boards, although the tweet and blog sub-corpora have the most balanced distribution of contributions/word count across each of the thematic categories. Finally, CANELC also includes a number of business emails, which contribute to the high frequency of data type 'A'.

While the assignment of the content to these thematic groupings was fairly transparent in some cases, other messages were slightly more 'fuzzy' and flexible, insofar as they discussed multiple topics ranging across the different categories. In these

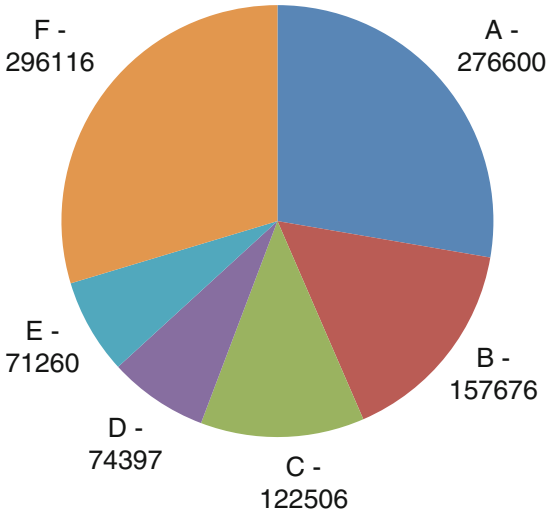


Fig. 4 Approximate distribution of words across the 6 topic categories of CANELC (refer to Fig. 3 for data key)

instances, when compiling CANELC, the data was given a range of category codes, so A/B/C rather than simply 'A'. For the purpose of Fig. 4 and the analysis seen in Sect. 3.3, individual contributions are counted once across these groupings, so they are classified according to, crudely, their 'best fit'. That is, even in instances where multiple categories were assigned, only one single category was counted. This was, subjectively, the category which is descriptively the 'most' appropriate for these contributions, that is, the one that is approximately the most representative/appropriate of that data. In other words if data was assigned the categories A/B/C, for example, and the content was described as being most dominantly 'business related' [i.e. category A], content was re-labelled as being category 'A' only.

The inclusion of this categorisation scheme provides a helpful way-in to querying levels of formality in CANELC as, in parallel with previous comments, the division of public vs. private can affect the levels of formality in a text. So comparisons of hedging within and across both the modes of data in CANELC and these different topics, can help us to assess how closely e-language compares with more formal (akin to the written end of the continuum) and informal discourse (positioned toward the spoken end of the continuum).

Given the level of contextual specificity, 'hedging can be achieved in indefinite numbers of surface forms' (Brown and Levinson 1987: 146), making it potentially difficult to draw up a 'list of hedges' (Clemen 1997: 236, 243; Nikula 1997: 190) to use as a basis of a study of this phenomenon. Despite this, across the literature there are specific words or expressions that are *often* used as hedges. For example, as outlined by Farr et al. (2004: 13–14) the most salient hedges are 'core modal verbs' and 'verbs with modal meaning' (O'Keeffe et al. 2007: 175 – e.g. *might*, *may*), 'clausal items' (e.g. *I think*, *you know*), 'noun based expressions' (e.g. *the thing is*), 'degree adverbs' (e.g. *really*, *necessarily*) and 'stance adverbs' (e.g. *of course*, *sort of*)

Actually	Generally	Likely	Only	Really	Surely
Apparently	Guess	Maybe	Partially	Relatively	Thing
Arguably	I think	Necessarily	Possibility	Roughly	Typically
Broadly	Just	Normally	Probably	Seemingly	Usually
Frequently	Kind of	Of course	Quite	Sort of	You know

Fig. 5 Some common hedges in spoken and written discourse

and so on. The hedges that the present study will focus on are some of the most common forms that have been examined in past studies of this topic (based on Biber et al. 1999; Carter and McCarthy 2006; O'Keeffe et al. 2007: 175), and are forms which are frequent in the CANCODE[3] (Cambridge and Nottingham Corpus of Discourse in English), BNC, CEC[4] (Cambridge English Corpus) and CANELC corpora. These are listed in Fig. 5. These terms were queried in the CANELC data.

Some of the adverbs listed here, such as *just*, have the softening hedging function, but are also often used with intensifying and specifying functions in discourse. *Just do it*; *it's just about five o'clock* and *we'll only be a couple of minutes late* are examples of this. Of course is another examples of this, this cluster can be used as a hedge when it has a pragmatic function but it can also be emphatically and directly; *Are you coming*? *Of course*. So although we can define some frequent forms of hedges, a more qualitative screen by screen study is needed if we are to drill down into specific functions. The current study undertakes a more quantitative approach, but a more qualitative assessment of the data would be welcomed in future studies of this nature and are, indeed, necessary.

3.2 Frequency of Hedges

The frequency of use of the terms in Fig. 5 were queried across the entire corpus as well as each mode is presented and compared, along with the frequency of use seen in the written and spoken BNC sub-corpora. Results are shown in Fig. 6. Log-likelihood scores are also presented in this figure. These provide a statistical measure of the relationship between the frequencies, indicating whether specific patterns of significant differences are likely to exist by chance or not. In this figure, a '+' log-likelihood score indicates that a particular rate of use is statistically higher in the CANELC corpus compared to the other parameter defined,

[3]CANCODE stands for *Cambridge and Nottingham Corpus of Discourse in English*. This corpus has been built as part of a collaborative project between The University of Nottingham and Cambridge University Press with whom sole copyright resides. CANCODE is comprised of five-million words of (mainly casual) conversation recorded in different contexts across the British Isles.

[4]CEC stands for Cambridge English Corpus, a corpus of over one billion written and spoken words in English. For more information visit: http://www.cambridge.org/

Word/ cluster	Freq. in CANELC	Spoken			Written			Blogs			Discussion Boards			Emails			SMS			Tweets		
		Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL
Actually	538	**1228**	-	**270.17**	**116**	+	**293.2**	151	-	0.66	161	-	5.89	49	+	5.37	57	-	0.09	120	+	3.99
Apparently	142	**79**	+	**18.84**	**90**	+	**11.5**	54	-	5.29	22	+	4.02	10	+	3.83	13	+	0.13	43	-	0.38
Arguably	5	1	+	2.95	3	+	0.5	3	-	1.19	2	-	0.35	0	+	1.19	0	+	0.97	0	+	2.4
Broadly	6	5	+	0.01	**26**	-	**13.58**	2	-	0.09	2	-	0.15	0	+	1.43	0	+	1.16	2	-	0.06
Frequently	27	12	+	6.07	**57**	-	**11.1**	9	-	0.04	13	-	3.89	1	+	2.18	0	+	5.23	4	+	1.49
Generally	74	50	+	4.91	*113*	-	*8.39*	30	-	3.91	29	-	4.67	11	-	0.23	**0**	+	**14.33**	**4**	+	**16.37**
Guess	140	**58**	+	**35.81**	**34**	+	**68.79**	*20*	+	*7.3*	32	+	0.06	9	+	4.77	**42**	-	**30.65**	37	+	0.02
I think	240	280	-	2.7	**17**	+	**229.98**	50	+	2.15	44	+	2.83	*49*	-	*8.32*	20	+	0.77	77	-	1.56
Just	3641	**4076**	-	**20.45**	**919**	+	**1724.96**	**677**	+	**69.72**	913	-	1.51	422	+	3.08	**669**	-	**172.6**	960	+	0.7
Kind of	35	39	-	0.18	**5**	+	**25.16**	11	-	0.29	6	+	0.62	3	+	0.47	2	+	0.75	13	-	0.88
Likely	173	**62**	+	**55.67**	**277**	-	**24.78**	*67*	-	*7.16*	58	-	4.63	20	+	0.15	8	+	6.08	**20**	+	**16.38**
Maybe	444	**320**	+	**21.45**	**106**	+	**221.61**	*82*	+	*8.81*	80	+	5.87	72	-	3.55	**103**	-	**47.65**	107	+	1.28
Necessarily	22	**51**	-	**11.56**	*44*	-	*7.59*	10	-	1.97	6	-	0.08	1	+	1.39	1	+	0.8	4	+	0.6
Normally	43	**141**	-	**54.04**	60	-	2.9	12	-	0.04	21	-	6.5	3	+	1.19	*0*	+	*8.33*	7	+	1.78
Of course	338	*414*	-	*6.97*	**235**	+	**18.11**	116	-	6.3	*110*	-	*7.39*	*23*	+	*10*	27	+	1.56	*62*	+	*8.85*
Only	1328	*1191*	+	*8.86*	1366	-	0.74	360	-	0.46	360	-	4.23	187	-	1.77	107	+	5.71	314	+	5.07
Partially	4	0	+	5.58	7	-	0.84	2	-	0.52	1	-	0	1	-	0.32	0	+	0.77	0	+	1.92
Possibility	28	40	-	2.01	**74**	-	**21.74**	7	+	0.01	3	+	2.18	*11*	-	*8.35*	2	+	0.26	5	+	0.82
Probably	376	**545**	-	**29.55**	**376**	+	**59.01**	107	-	0.65	97	-	0.42	*71*	-	*8.67*	41	-	0.18	**60**	+	**16.64**
Quite	529	**928**	-	**106.79**	**297**	+	**64.9**	162	-	3.18	149	-	2.96	48	+	5.39	54	-	0	116	+	4.58
Really	1434	**1747**	-	**27.85**	**296**	+	**809.25**	331	+	3.99	*404*	-	*8.04*	154	+	3.98	*183*	-	*7.92*	362	+	1.6
Relatively	32	23	+	1.57	**94**	-	**32.19**	17	-	5.17	11	-	1	2	+	1.16	1	+	2	*1*	+	*9.51*
Roughly	18	15	+	0.3	28	-	2.24	3	+	0.57	6	-	0.46	6	-	6.52	3	-	0.56	*0*	+	*8.65*
Seemingly	14	*3*	+	*7.83*	17	-	0.31	7	-	1.83	7	-	2.29	0	+	3.34	0	+	2.71	*0*	+	*6.73*
Sort of	56	**661**	-	**594.97**	*28*	+	*9.68*	26	-	5.49	23	-	4.36	4	+	1.45	1	+	5.54	**2**	+	**15.7**
Surely	87	*50*	+	*10.48*	67	+	2.51	25	-	0.19	*36*	-	*7*	4	+	5.42	*2*	+	*7.24*	20	+	0.47
Thing	527	**1090**	-	**194.7**	**212**	+	**137.1**	150	-	0.92	**176**	-	**13.75**	**34**	+	**17.8**	*32*	+	*9.35*	135	+	0.38
Typically	18	7	+	5.12	**7**	+	**43.96**	5	-	0.02	9	-	2.95	4	-	0.91	0	+	3.49	*0*	+	*8.65*
Usually	115	*171*	-	*10.49*	**202**	-	**24.62**	31	-	0.03	**51**	-	**12.26**	*5*	+	*7.72*	4	+	6.32	24	+	1.47
You know	211	**211**	-	**1625.6**	**82**	+	**58.15**	*32*	+	*9.24*	**26**	+	**12.14**	*43*	-	*7.25*	**41**	-	**12.43**	69	-	1.72
TOTAL	**10645**	**13287**			**5255**			**2559**			**2832**			**1247**			**1413**			**2568**		

Fig. 6 The frequency of common forms of hedges used in CANELC, compared to the spoken and written sub-corpora from the BNC

while a '–' log-likelihood indicates a statistically lower frequency of use in CANELC. Numbers in **bold** indicate that there is a statistical difference (measured using a log-likelihood score) in the frequency of usage across specific modes/ genres to a p value of <0.01 (with a critical value range of 6.63–10.82) while those in *italics* mark a significant to p value <0.001 (critical value of 10.83). So an '+' indicates an overuse in CANELC compared to the listed parameter and thus an underuse in the given category.

In Fig. 6 we see that, for the terms *actually*, *just*, *you know*, *probably*, *quite*, *really*, *thing*, there is a significant underuse in CANELC compared to the written BNC corpus, while there is a significant overuse compared to the spoken BNC sub-corpus (to $p<0.001$). *Probably* is significantly underused in the twitter data and overused in the email data (to $p<0.01$ and $p<0.001$) while *really* is overused in the discussion boards and SMS messages compared to rate of use across CANELC (to $p<0.001$). *Just* is significantly underused in the blog data and overused in the SMS data, while *you know* is underused in the blog and discussion board data but overused in the email and SMS data and *just* is underused in the email but overused in the discussion board data. Finally, there is no real significant difference in the rate of use of *quite* and *actually* across the different e-language modes.

The only item that is significantly overused, at $p<0.01$, in the spoken BNC **and** underused in the written compared to CANELC is *likely*. There are, however, some terms which are overused in CANELC, compared to both sub-corpora. These include *apparently*, *guess* and *maybe*. Of these terms, *apparently* is used at a near-consistent rate across all of the modes in CANELC, while *guess* is underused (to $p<0.001$) in the blogs and significantly overused in the SMS (to $p<0.01$) when compared to the other modes. *Maybe* and *likely*, on the other hand, are both underused in the blogs (to $p<0.001$ respectively) but the former is overused in the SMS messages and the latter in the tweets (both to $p<0.01$).

I think, *kind of*, *broadly*, *typically* and, to some extent *of course* are used at a significantly higher rate in CANELC than the written BNC (to $p<0.01$), but no significant difference exists between the rate that they are used in the spoken BNC (aside from *of course* where the difference is to ($p<0.001$)). Conversely, there is an underuse of the expression *normally* in CANELC compared to the spoken data (to $p<0.01$) while there is no significant difference between the use of this term when compared to the written corpus. *Kind of* is used at a consistent rate across all modes in the corpus, while *typically* and *normally* are used at consistent rates across all modes aside from tweets and SMS messages where a slight underuse occurs when compared to CANELC respectively (to $p<0.001$). Similarly *of course* is slightly underused in the SMS messages but slightly overused in the discussion board data (to $p<0.001$) and *I think* is slightly overused in the email data, but used consistently across the other modes in CANELC.

Figure 6 also indicates that there is a slight overuse of *only*, *seemingly* and *surely* compared to the spoken BNC (to $p<0.01$) while no difference exists between the rate of use of these words in CANELC versus the written BNC.

Frequently, *possibility*, *relatively* and, to some extent, *generally* are all underused in CANELC compared to the written BNC, while there is a near-consistent rate of use of these terms when compared to the spoken BNC data (to $p<0.01$ aside from *generally* which is to $p<0.001$). The rate at which *frequently* is used across each of the modes in CANELC is near-consistent while there is an overuse of *possibility* in the email data, an underuse of *relatively* in the tweets (both to $p<0.001$) and a significant underuse of *generally* in the SMS and tweet data (to $p<0.01$). Similarly, *only* is used at a near-consistent rate across the different modes while *seemingly* is slightly underused in the twitter data and *surely* is underused in the SMS data but overused in the discussion board data (to $p<0.001$).

Necessarily, *usually* and *sort of* are all underused in CANELC when compared to the spoken BNC (to $p<0.01$, $p<0.01$ and $p<0.01$ respectively) and, similarly, the first two of these terms are also underused compared to the written data (to $p<0.001$ and $p<0.01$ respectively) while *sort of* is slightly overused compared to the written BNC (to $p<0.001$). *Necessarily* and *sort of* are used at consistent rates across all modes aside from the tweets, where a significant underuse *of sort* of can be seen when compared to CANELC (to $p<0.01$). Comparatively, *usually* is significantly overused in the discussion board data and underused in the email data compared to the other modes included in CANELC (to $p<0.01$ and $p<0.001$ respectively).

Word/ cluster	Freq.	A			B			C			D			E			F		
	in CANELC	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL
Actually	538	*116*	+	*6.8*	98	-	0.21	72	-	3.99	45	+	0.18	32	+	0.89	175	-	2.34
Apparently	142	34	+	0.68	24	+	0.01	12	+	0.47	12	+	0.03	3	+	6.43	57	-	4.59
Arguably	5	2	-	0.17	1	-	0.02	0	+	0.98	2	-	2.48	0	+	0.68	0	+	2.5
Broadly	6	4	-	1.67	1	+	0	0	+	1.18	1	-	0.29	0	+	0.82	0	+	3
Frequently	27	14	-	3.26	3	+	0.6	6	-	2.44	2	+	0.07	1	+	0.49	*1*	+	*7.9*
Generally	74	22	-	0.06	19	-	2.17	13	-	2.75	8	-	0.25	4	+	0.28	*8*	+	*8.73*
Guess	140	24	+	5.49	21	+	0.39	17	-	0.38	10	+	0.5	6	+	1.65	*62*	-	*7.93*
I think	240	55	+	1.84	54	-	2.86	21	+	0.56	20	+	0.09	10	+	3.1	80	-	1.48
Just	3641	**778**	+	**49.04**	593	+	1.94	390	-	0.46	291	+	3.46	248	+	0.24	**1341**	-	**63.01**
Kind of	35	4	+	3.63	5	+	0.17	5	-	0.42	7	-	3.15	0	+	4.76	14	-	1.11
Likely	173	61	-	1.52	23	+	1.52	15	+	0.44	20	-	1.11	12	+	0	42	+	0.87
Maybe	444	**69**	+	**23.69**	67	+	1.14	47	-	0.03	33	+	1.1	19	+	5.25	**209**	-	**33.76**
Necessarily	22	6	+	0	10	-	5.54	1	+	0.84	1	+	0.54	0	+	2.99	4	+	0.75
Normally	43	14	-	0.24	*18*	-	*8.6*	4	+	0.04	2	+	1	2	+	0.37	3	+	8.39
Of course	338	85	+	0.78	69	-	1.5	36	-	0.03	29	+	0.04	16	+	2.7	103	-	0.38
Only	1328	382	-	0.24	220	+	0.37	112	+	4.5	137	-	2.47	120	-	6.45	357	+	0.89
Partially	4	2	-	0.42	2	-	1.29	0	+	0.79	0	+	0.68	0	+	0.54	0	+	2
Possibility	28	12	-	1.44	3	+	0.71	0	+	5.5	2	+	0.1	0	+	3.81	11	-	0.79
Probably	376	81	+	4.78	77	-	1.73	35	+	0.36	39	-	0.76	16	+	4.54	128	-	3.01
Quite	529	**97**	+	**16.04**	**137**	-	**16.26**	61	-	0.64	**26**	+	**10.66**	35	+	0.12	173	-	2.5
Really	1434	**242**	+	**59.32**	**362**	-	**38.02**	**106**	+	**12.18**	150	-	3.23	90	+	1.13	**484**	-	**10.35**
Relatively	32	10	-	0.09	3	+	1.22	6	-	1.56	4	-	0.37	4	-	1.01	5	+	1.79
Roughly	18	9	-	1.88	2	+	0.4	0	+	3.54	2	-	0.08	1	+	0.06	4	+	0.21
Seemingly	14	6	-	0.72	2	+	0.07	0	+	2.75	1	+	0.05	4	-	4.61	1	+	2.67
Sort of	56	8	+	3.72	11	-	0.14	5	+	0.1	11	-	4.77	10	-	5.92	11	+	2.04
Surely	87	31	-	1.29	10	+	1.68	5	+	1.94	**21**	-	**13.52**	8	-	0.49	12	+	6.6
Thing	527	113	+	6.39	100	-	0.69	73	-	5.13	*70*	-	*8.78*	50	-	3.77	121	+	4.69
Typically	18	*15*	-	*9.03*	2	+	0.4	0	+	3.54	0	+	3.08	0	+	2.45	1	+	4.19
Usually	115	31	+	0.03	32	-	5.13	22	-	6.09	11	-	0.05	5	+	1.29	**14**	+	**11.17**
You know	211	51	+	0.9	*21*	+	*6.82*	22	-	0	10	+	4.68	9	+	2.52	**97**	-	**14.37**
TOTAL	**10645**	**2378**			**1990**			**1086**			**967**			**705**			**3518**		

Fig. 7 The use of hedges in the topic categories in CANELC

Finally, we see no statistical difference in the use of *arguably* and *partially* when comparing CANELC to the spoken and written BNC, or across the individual modes of e-language.

3.3 Patterns of Use Across Topics

In addition to exploring the use of the hedges across the different modes in CANELC, we are able to look in more detail at differences in use across the topic categories detailed in Fig. 3. Figure 7 documents the frequency of word use across the different topic categories and provides a log-likelihood score of difference in use for each category compared to CANELC (note – a '+' indicates an overuse in CANELC compared to a category, thus an underuse in the given category), while Figs. 8 and 9 tabulate the frequency of use across these topics compared to the spoken and written BNC (note – a '+' indicates an overuse in the BNC compared to a category).

Word/ cluster	Freq. in spoken BNC	A			B			C			D			E			F		
		Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL
Actually	1228	**116**	+	**164.8**	**98**	+	**66.17**	**72**	+	**24.92**	**45**	+	**44.89**	**32**	+	**41.91**	**175**	+	**83.9**
Apparently	79	34	-	4.34	24	-	5.41	12	-	1.48	12	-	2.66	3	+	1.29	**57**	-	**27.01**
Arguably	1	2	-	2.78	1	-	1.39	0	+	0	**2**	-	**96.39**	0	+	0	0	+	0
Broadly	5	4	-	2.31	1	-	0.02	0	+	0	1	-	0.46	0	+	0	0	+	0
Frequently	12	**14**	-	**12.78**	3	-	0.32	*6*	-	*7.95*	2	-	0.59	1	-	0.03	1	+	1.93
Generally	50	22	-	3.07	*19*	-	*7.63*	*13*	-	*7.41*	8	-	2.1	4	-	0.07	8	+	2.52
Guess	58	24	-	2.58	*21*	-	*7.58*	**17**	-	**11.85**	10	-	3.24	6	-	0.77	**62**	-	**50.49**
I think	280	55	+	5.84	54	-	0.61	21	+	2.06	20	+	0.91	10	+	5.34	80	-	0.01
Just	4076	**778**	+	**98.1**	**593**	+	**14.55**	390	+	1.61	**291**	+	**13.31**	248	+	4.43	**1341**	-	**23.86**
Kind of	39	4	+	4.68	5	+	0.4	5	-	0.21	7	-	2.52	0	+	0	14	-	0.58
Likely	62	**61**	-	**46.3**	*23*	-	*8.77*	*15*	-	*7.46*	**20**	-	**19.8**	*12*	-	*9.31*	**42**	-	**17.76**
Maybe	320	69	+	3.74	67	-	2.12	47	-	4.9	33	-	0.67	19	+	0.48	**209**	-	**82.39**
Necessarily	51	6	+	4.91	10	-	0.15	1	+	4.8	1	+	3.78	0	+	0	*4*	+	*8.76*
Normally	141	**14**	+	**17.7**	18	+	1.5	*4*	+	*9.87*	**2**	+	**12.9**	*2*	+	*8.86*	**3**	+	**49.88**
Of course	414	*85*	+	*6.72*	69	+	0.05	36	+	0.91	29	+	1.58	16	+	6.44	103	+	1.29
Only	1191	382	-	6.11	220	-	1.01	112	+	0.74	*137*	-	*7.9*	**120**	-	**13.46**	357	-	1.07
Partially	0	2	-	6.11	*2*	-	*7.69*	0	+	0	0	+	0	0	+	0	0	+	0
Possibility	40	12	-	0.06	3	+	2.43	0	+	0	2	+	0.74	0	+	0	11	+	0
Probably	545	**81**	+	**31.73**	77	+	2.65	*35*	+	*8.18*	39	+	1.74	**16**	+	**15.43**	128	+	3.53
Quite	928	**97**	+	**108.2**	137	+	2.77	**61**	+	**12.76**	**26**	+	**49.24**	**35**	+	**15.43**	**173**	+	**27.17**
Really	1747	**242**	+	**121**	**362**	-	**10.32**	**106**	+	**31.89**	150	+	0.13	*90*	+	*8.52*	484	+	0.1
Relatively	23	10	-	1.33	3	+	0.21	6	-	3.44	4	-	1.33	4	-	2.3	5	+	0.29
Roughly	15	9	-	3.09	2	+	0.12	0	+	0	2	-	0.27	1	+	0	4	+	0.01
Seemingly	3	*6*	-	*8.35*	2	-	1.91	0	+	0	1	-	1.03	**4**	-	**12.7**	1	-	0.02
Sort of	661	**8**	+	**260.8**	**11**	+	**139.2**	**5**	+	**93.63**	**11**	+	**55**	**10**	+	**39.66**	**11**	+	**254.26**
Surely	50	**31**	-	**11.43**	10	-	0.19	5	+	0	**21**	-	**27.66**	8	-	3.9	12	+	0.25
Thing	1090	**113**	+	**128.7**	**100**	+	**42.66**	**73**	+	**13.91**	*70*	+	*7.43*	*50*	+	*9.48*	**121**	+	**120.6**
Typically	7	**15**	-	**21.76**	2	-	0.37	0	+	0	0	+	0	0	+	0	1	+	0.48
Usually	171	31	+	5.18	32	-	0.2	22	-	0.97	11	+	1.15	5	+	4.87	**14**	+	**28.08**
You know	211	**51**	+	**645**	**21**	+	**472.7**	**22**	+	**247.15**	**10**	+	**259.31**	**9**	+	**200**	**97**	+	**490.48**
TOTAL	**13498**	**2378**			**1990**			**1086**			**967**			**705**			**3518**		

Fig. 8 The rate of use of hedges in the topic categories in CANELC, compared to the spoken BNC

Six sub-corpora of the CANELC data were created (for A–F) to draw these comparisons in the data.

From Fig. 7 we can see that none of the hedging terms are overused in data classified under topic category 'A' compared to CANELC, although *just*, *maybe*, *quite* and *really* are all significantly underused (to $p<0.01$) and *actually* and *typically* are slightly underused (to $p<0.001$). Similarly, Fig. 7 shows an underuse of *a bit*, *like* and *stuff* in this category when compared to the corpus as a whole (to $p<0.01$). As documented in Figs. 8 and 9, *actually*, as used in category 'A' in CANELC occurs at a far less frequent rate than it does in the spoken and written BNC (both to $p<0.01$) and the converse is true for *relatively* (to $p<0.01$). While for *frequently*, *likely*, *seemingly* and *partially*, there is a higher rate of use in category 'A' than the spoken BNC, but a near consistent rate of use to the written corpus (to $p<0.01$, $p<0.01$ and $p<0.001$ respectively).

Surely and *typically* are used at a higher rate in the category 'A' data in the spoken BNC data, but while *surely* is used at a near consistent rate to the written BNC, *typically* is far less frequent in A. The converse of this is true for *typically*. While *arguably*, *possibility*, *roughly*, *only* and *generally*, when classified in category 'A'

Word/ cluster	Freq. in written BNC	A			B			C			D			E			F		
		Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL
Actually	116	**116**	+	**87.8**	**98**	+	**116.24**	**72**	+	**113.1**	**45**	+	**53.83**	**32**	+	**35.26**	**175**	+	**193.36**
Apparently	90	34	+	2.06	24	+	3.17	12	+	0.62	12	+	1.5	3	-	2.09	**57**	+	**20.39**
Arguably	3	2	+	0.82	1	+	0.28	0	-	0	2	+	3.77	0	-	0	0	-	0
Broadly	26	4	-	1.45	1	-	3.61	0	-	0	1	-	0.91	0	-	0	0	-	0
Frequently	57	14	-	0.21	3	-	5.92	6	+	0	2	-	2.32	1	-	3.12	**1**	-	**21.55**
Generally	113	22	-	2.69	19	-	0.02	13	+	0.12	8	-	0.45	4	-	2.31	**8**	-	**21.94**
Guess	34	**24**	+	**10.99**	**21**	+	**17.94**	**17**	+	**22.17**	*10*	+	*8.61*	6	+	3.44	**62**	+	**78.8**
I think	17	**55**	+	**96.6**	**54**	+	**133.47**	**21**	+	**50.39**	**20**	+	**51.72**	**10**	+	**21.09**	**80**	+	**159.15**
Just	919	**778**	+	**474.94**	**593**	+	**533.75**	**390**	+	**430.79**	**291**	+	**275.84**	**248**	+	**266.3**	**1341**	+	**1442.06**
Kind of	5	4	+	2.25	*5*	+	*6.83*	**5**	+	**10.76**	**7**	+	**19.51**	0	-	0	**14**	+	**22.73**
Likely	277	61	-	3.12	**23**	-	**14.36**	*15*	-	*7.36*	20	-	0.95	12	-	3.21	**42**	-	**17.05**
Maybe	106	**69**	+	**27.14**	**67**	+	**58.86**	**47**	+	**54.4**	**33**	+	**30.61**	**19**	+	**11.22**	**209**	+	**279.6**
Necessarily	44	6	-	3.3	10	+	0.55	1	-	3.83	1	-	2.97	0	-	0	4	-	6.61
Normally	60	14	-	0.41	18	+	3.71	4	-	0.84	2	-	2.63	2	-	1.2	**3**	-	**15.06**
Of course	235	85	+	3.82	**69**	+	**13.21**	36	+	4.3	29	+	2.4	16	-	0.02	**103**	+	**12.34**
Only	1366	382	-	0.01	220	-	1.16	112	-	6.16	137	+	1.48	120	+	4.91	357	-	2.29
Partially	7	2	+	0	2	+	0.35	0	-	0	0	-	0	0	-	0	0	-	0
Possibility	74	12	-	3.55	*3*	-	*9.85*	0	-	0	2	-	4.21	0	-	0	11	-	4.79
Probably	376	*81*	+	*8.66*	**77**	+	**33.11**	*35*	+	*8.09*	**39**	+	**17.77**	16	+	0.36	**128**	+	**50.3**
Quite	297	97	+	1.61	**137**	+	**77.12**	**61**	+	**20.21**	26	-	0.02	*35*	+	*7.16*	**173**	+	**51.24**
Really	296	**242**	+	**140.48**	**362**	+	**571.56**	**106**	+	**95.83**	**150**	+	**230.17**	**90**	+	**110.4**	**484**	+	**569.17**
Relatively	94	**10**	-	**11.05**	**3**	-	**14.85**	6	-	1.54	4	-	2.71	4	-	1.16	**5**	-	**22.69**
Roughly	28	9	+	0.12	2	-	1.92	0	-	0	2	-	0.1	1	-	0.56	4	-	1.99
Seemingly	17	6	+	0.22	2	-	0.3	0	-	0	1	-	0.19	4	+	3.62	1	-	3.83
Sort of	28	8	+	0	11	+	4.58	5	+	1.1	**11**	+	**13.33**	**10**	+	**14.38**	11	+	0.35
Surely	67	31	+	4.92	10	-	0.21	5	-	0.56	**21**	+	**19.65**	8	+	1.71	12	-	2.47
Thing	212	**113**	+	**27.95**	**100**	+	**58.42**	**73**	+	**62.7**	**70**	+	**69.87**	**50**	+	**45.33**	**121**	+	**34.09**
Typically	7	**15**	+	**21.46**	2	+	0.35	0	-	0	0	-	0	0	-	0	1	-	0.5
Usually	202	**31**	-	**11.36**	32	-	0.25	22	+	0.05	11	-	3.04	*5*	-	*7.66*	**14**	-	**39.96**
You know	82	**51**	+	**18.27**	21	+	2.31	**22**	+	**12.89**	10	+	0.77	9	+	1.4	**97**	+	**86.12**
TOTAL	5255	2378			1990			1086			967			705			3518		

Fig. 9 The rate of use of hedges in the topic categories in CANELC compared to the written BNC

occur at near-consistent rates to the spoken and written BNC data (as seen in Fig. 8) and *relatively*, although nearly-consistent to the spoken BNC, is used at a much higher rate in the topic 'A' data than the written BNC (to $p < 0.01$, as seen in Fig. 9).

For topic 'B', that is topics covering 'culture, literature and the arts', 'fashion' and 'teaching, academia and education', Fig. 7 indicates that the only significant differences seen are in the rate of use of *quite* and *really*, both of which are used at a rate higher than the average rate seen in CANELC.

Necessarily, *normally*, *broadly* and *usually* are terms that are most commonly classified under topic category 'B' in CANELC. The rate of use of these terms, in this category are shown to be nearly consistent to the rates of use in the spoken and written BNC, as no real significant differences are outlined in Figs. 8 and 9. There is, however, an underuse of *sort of*, in the category 'B' data compared to the spoken BNC (which is also most commonly classified under category 'B'), while near consistent rates to the written BNC are shown.

Figure 7 indicates that there are no significant differences in the use of the search terms for topic 'E'. There is, however, a significant underuse of *really* in CANELC compared to 'C', and an underuse of *quite* and an overuse of *surely* compared to 'D'.

These are the only real difference seen for these categories (to $p<0.01$). None of the hedges explored were more frequently used in the data classified under topic category 'E' or 'C' than the other topic categories. The only ones frequently used in 'D' were *arguably* and *sort of*. *Arguably* is overused in this category compared to the average use in the spoken BNC, but near-consistent with rates of use in the written BNC, while *sort of* is used at a significantly lower rate in the topic 'D' data than the spoken and written BNC (to $p<0.01$).

Finally, Fig. 7 highlights that *just*, *maybe* and *really* are all used at a significantly higher rate in the data for category 'F' than the CANELC average (all to $p<0.01$) and *usually* is used at a lower rate than the CANELC average (both to $p<0.01$). The first of these terms are also significantly overused compared to the spoken BNC, but significantly underused compared to the written BNC. It is the use of terms in this category that we see the most marked difference in frequency rates when compared to the written and spoken BNC data (Figs. 8 and 9).

Apparently, *guess*, *just*, *maybe*, *stuff*, *or so* and *a bit* are all used at a significantly higher rate in CANELC compared to both the spoken and written data (all to $p<0.01$ aside from *a bit* and *or so* which are to $p<0.001$ for the spoken and written data respectively) while *like*, *quite*, *you know* and *thing* are all underused in the category 'F' data compared to the spoken BNC but overused when compared to the written data (all to $p<0.01$). *Kind of*, *I think*, *probably* and *really* are all significantly overused in the category 'F' data when compared to the written BNC but are used at near consistent rates to the spoken excerpt (to $p<0.01$). Conversely, *sort of* is significantly underused in this data compared to the spoken BNC, but used at near-consistent compared to the written data and *of course* is used at near-consistent rates in the category 'F' data compared to both the written and spoken BNC.

4 Discussion

Of the hedges examined, the most commonly used forms featured in CANELC were:

From this we can surmise that:

1. Of the forms examined, the most frequent hedge used in CANELC is the adverb *just*, followed by *really* and *only*.

Seven of the top ten of these hedges featured in Fig. 10 were shown to be significantly underused in CANELC compared to the spoken BNC but overused compared to the written BNC. The first of these adverbs were also shown to be frequently used in the study of hedging in LCIE (Farr et al. 2004), but none of noted as common hedges in studies of written academic discourse (see Channell 1990; Clemen 1997; Gries and David 2007). As discussed by Atai and Sadr (2006) the use of full verbs, nouns and adjectives as hedges (in that order) are often the most commonly used forms in more formal, written contexts. Although hedges of these forms were common in the data, they were used far less frequently than the adverbial forms.

No	Form	Freq	No	Form	Freq	No	Form	Freq	No	Form	Freq
1	Just	3641	9	Of course	338	17	Generally	74	25	Roughly	18
2	Really	1434	10	I think	240	18	Sort of	56	26	Typically	18
3	Only	1328	11	You know	211	19	Normally	43	27	Seemingly	14
4	Actually	538	12	Likely	173	20	Kind of	35	28	Broadly	6
5	Quite	529	13	Apparently	142	21	Relatively	32	29	Arguably	5
6	Thing	527	14	Guess	140	22	Possibility	28	30	Partially	4
7	Maybe	444	15	Usually	115	23	Frequently	27			
8	Probably	376	16	Surely	87	24	Necessarily	22			

Fig. 10 Rank order of the 30 hedges in CANELC (by frequency of use)

This suggests that, by form alone, the use of hedging in e-language shows some clear similarities with those used in more informal, spoken discourse.

More generally, of the 30 hedges examined, 15 were found to be more frequent in the spoken than written BNC sample than in CANELC. Of these terms, 11 were significantly underused in CANELC compared to the BNC (10 to $p<0.01$ and 1 to $p<0.001$) while only 2 were overused in CANELC. Similarly, there was a higher rate of underuse of the 15 terms most frequently used in the written data, although this was only seen with 7 of the terms (with 2 of these 15 being overused in CANELC). Across all 30 terms, we saw that 12 of them were significantly underused and 7 overused in CANELC compared to the spoken data, while 15 were overused and 8 were underused in CANELC compared to the written data. This can be summarised as follows:

2. Hedges that were most frequently used in the spoken rather than written BNC sample (and vice versa) were used at a significantly lower rate in the e-language data.
3. Of the forms analysed, a higher proportion were significantly overused rather than underused in CANELC when compared to the written data (15 vs. 8).
4. Of the forms analysed, a higher proportion were significantly underused rather than overused in CANELC when compared to the spoken data (12 vs. 7).

These findings suggest that the rate of hedging use in the e-language data is inconsistent with typical rates in spoken and written discourse. While more hedges were used compared to the written data, far fewer were used than in the spoken data. This provides an argument for classifying e-language as its own distinct genre (as suggested in Sect. 2).

When comparing the patterns of use across the different modes of data we also see the following:

5. Emails and discussion boards contained fewer disparities in the rate of under/overuse of specific hedging forms than other modes of e-language (i.e. they were most 'similar').
6. The SMS, discussion board and twitter data contained the most disparities in the rate of under/overuse of specific hedging forms than other modes of e-language (i.e. they were the least 'similar' modes of e-language).

In terms of relative frequencies (calculated as the number of hedges used per word in each of the modes) we see that:

7. Hedges were used at a more frequent rate in the SMS and discussion board data than the other modes (1:72 words and 1:86 words), while they were used at a near consistent rate across the twitter, email and blog modes (1:101, 1:103 and 1:105 respectively).

Again, this is an interesting finding as it is in the 'most immediate' form of e-language, SMS messages (which, from show a shorter delay in the response times to messages in CANELC), there is a tendency for a higher number of hedges to be used. For the SMS messages, given that the relationship between the sender and sendee is often 'fixed', with messages being directed at individuals or groups of people known to the sender, and are often classified as being of the 'personal and daily life' topic, the need for hedging to mitigate against potential face threats is assumed to be reduced, so the reverse of this is interesting here. Similarly, while it is not necessarily the case that discussion board members 'know' each other personally, this mode of e-language often involves a fixed community of contributors who respond to each other regularly, creating a closeness between those involved.

The data also reveals that dramatic differences are seen in frequency rates across the different topic categories, compared to corpus as a whole. Of all the hedges analysed, the most common topic of the content was classified under category 'F'. When compared to the BNC, we saw that those terms in category 'F' were statistically overused in the 'F' data than in both the written and spoken BNC. This was true of 8 of the 17 terms featured under the category 'F' data in Fig. 8 (to $p<0.01$ or $p<0.001$). These patterns can be summarised as follows:

8. Based on frequency, content classified under the topics in categories 'A' and 'F' used more hedging than the other topic categories.
9. Of the hedges analysed, all were, on average, used at a less frequent rate in each of the topic sub-corpora when compared to the written BNC.
10. While all hedges were also used at a less frequent rate in the topic sub-corpora than in the spoken BNC, the difference in rate of use was less significant than when compared to the written BNC.
11. Hedges used in topic categories 'B', 'C' and 'D' were underused and overused a near-consistent rate when compared to the spoken BNC. Hedges used in the category 'A' data were most significantly underused in the data when compared to the spoken BNC.

As is perhaps to be expected, then, the more formal and the more 'spoken' topic categories (i.e. interpersonal contexts, category 'F') witnessed a higher rate of hedging use than was the case with the other topics. As we saw earlier, spoken discourse often utilises more hedges than written discourse, but more formal spoken and written contexts use more hedges than the informal ones. The content which concerns matters related to personal and daily life are more akin to spoken discourse (although at the more informal end) so the more extensive use of hedging in this category is as expected. Similarly, the topics in category 'A' are most akin to 'formal' discursive

contexts (both across written and spoken genres) so the frequent use of hedging also aligns with expectations.

If we look at some specific forms of hedging in more detail we see that *kind of* and *sort of* are two hedges which have previously been found to be particularly frequent in formal language contexts, specifically academic discourse (Biber et al. 1999: 560–56; Poos and Simpson 2002: 1). We would thus expect them to be more prevalent in the content classified under category B, in 'teaching, academia and education'. This pattern was not mirrored in the e-language content and, in fact, there was a general underuse of both of these terms across the topics, modes and corpus when compared to the spoken and written data.

5 Summary

This chapter has revealed that there is no clear-cut relationship between the use of hedging in e-language compared to written and spoken genres of discourse. The use of hedging across different communicative contexts (defined by topic categories) and across the different modes of e-language is fluid and not necessarily fixed, although when compared to standard (BNC) written and spoken modes of discourse the forms of hedging isolated for the purposes of this study appear to behave in a way that suggests greater internal similarity across the modes than similarity with the standard (BNC) written and spoken data. As initially suggested by Crystal (2003), there appears to be an argument to conceptualise e-language as its own distinct variety on the continuum of formality: between spoken and written discourse. The more immediate forms of e-language (e.g. SMS messages) are positioned closer to the 'spoken' end while the emails and blogs are better positioned towards the more formal, written end (based on what we have found here).

To build on what has been found here, a more qualitative, screen by screen study of the data would allow us to examine, more closely, specific functions of the common hedging forms analysed here. A closer observation of hedging use between specific contributors (according to gender and relationship, for example) may also help us to create a clearer profile of use across the different modes. Finally, a focus on a wider range of hedging forms and a clearer distinction between the individual functions of forms, in specific contexts, as well as extending the focus to synchronous forms of e-language (e.g. IMs) would add to the discussions. There is scope to carry out such investigations in future studies of this nature.

References

Allison, D. 1995. Assertions and alternatives: Helping ESL undergraduates extend their choice in academic writing. *Journal of Second Language Writing* 4: 1–16.

Atai, M., and L. Sadr. 2006. A cross-cultural genre study on hedging devices in discussion section of applied linguistics research articles. In Proceedings of the 11th conference of Pan-Pacific Association of Applied Linguistics, 42–57. Hong Kong: The Chinese University of Hong Kong.

Atkins, S. 2011. *A cognitive linguistic perspective on social space in online health communities*. Unpublished Ph.D. thesis. Nottingham: The University of Nottingham.

Baron, N. 1998. Writing in the age of email: The impact of ideology versus technology. *Visible Language* 32(1): 35–53.

Bates, E. 1976. *Language and context*. New York: Academic.

Baym, N. 1995. The emergence of community in computer-mediated communication. In *Cybersociety: Computer-mediated communication and community*, ed. S.G. Jones, 138–163. Thousand Oaks: Sage.

Benjamin, J. 2011. Tweets, Blogs, Facebook and the Ethics of 21st- Century Communication Technology. In *Social media: Usage and impact*, ed. Noor Al-Deen, H.S. and J.A. Hendricks, 271–288. Maryland: Lexington Books.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. 1992. On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes* 15: 133–163.

Biber, D. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4): 243–257.

Biber, D., S. Conrad, G. Leech, J. Svartvik, and E. Finegan. 1999. *The Longman grammar of spoken and written English*. Harlow: Longman.

Boyd, D., and J. Heer. 2006. Profiles as conversation: Networked identity performance on Friendster. In Proceedings of the Hawaii International Conference on System Sciences (HICSS-39), Persistent Conversation Track, 4–7 Jan 2006. Kauai: IEEE Computer Society.

Brennan, S.E., and J.O. Ohaeri. 1999. Why do electronic conversations seem less polite? The costs and benefits of hedging. In Proceedings, International Joint Conference on Work Activities, Coordination, and Collaboration (WACC '99), 227–235, San Francisco.

Brown, G. 1989. Making sense: The interaction of linguistic expression and contextual information. *Applied Linguistics* 10(1): 97–108.

Brown, P., and S.C. Levinson. 1978. Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*, ed. E. Goody, 56–311. Cambridge: Cambridge University Press.

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

Brown, G., and G. Yule. 1983. *Discourse analysis*. Cambridge: Cambridge University Press.

Burnard, L. 2005. Developing linguistic corpora: Metadata for corpus work. In *Developing linguistic corpora: A guide to good practice*, ed. M. Wynne, 30–46. Oxford: Oxbow Books.

Carter, R.A., and M.J. McCarthy. 2006. *Cambridge grammar of English*. Cambridge: Cambridge University Press.

Chafe, W.L., and J. Danielewicz. 1987. Properties of spoken and written language. In *Comprehending oral and written language*, ed. R. Horowitz and S.J. Samuels, 83–113. New York: Academic.

Channell, J. 1985. Vagueness as a conversational strategy. *Nottingham Linguistic Circular* 14: 3–24.

Channell, J. 1990. Precise and vague quantities in academic writing. In *The writing scholar: Studies in the language and conventions of academic discourse*, ed. W. Nash, 95–117. Newbury Park: Sage Publications.

Channell, J. 1994. *Vague language*. Oxford: Oxford University Press.

Cheng, W., and M. Warren. 1999. Inexplicitness: What is it and should we be teaching it? *Applied Linguistics* 20(3): 293–315.

Cherny, L. 1999. *Conversation and community: Chat in a virtual world*. Stanford: Center for the Study of Language and Information.

Clancy, B. 2002. The exchange in family discourse. *Teanga* 21: 134–150.

Clemen, G. 1997. The concept of hedging: Origins, approaches and definitions. In *Hedging and discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts*, ed. R. Markkanen and H. Schröder, 235–248. Berlin: Walter de Gruyter.

Collot, M., and N. Belmore. 1996. Electronic language: A new variety of English. In *Computer mediated communication: Linguistic, social and cross-cultural perspectives*, ed. S.C. Herring, 13–28. Amsterdam: John Benjamins.

Condon, S., and C. Cech. 1996. Functional comparison of face-to-face and computer-mediated decision-making interactions. In *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*, ed. S. Herring, 65–80. Philadelphia: John Benjamins.
Crystal, D. 2001. *Language and the internet*. Cambridge: Cambridge University Press.
Crystal, D. 2003. The joy of text. *Spotlight magazine,* 16–17.
Crystal, D. 2008. *Txtng: The Gr8 Db8*. Oxford: Oxford University Press.
Crystal, D., and D. Davy. 1975. *Advanced conversational English*. London: Longman.
Drave, N. 1995. *The pragmatics of vague language*: *A corpus-based study of vagueness in national vocational qualifications*. Unpublished Master's dissertation, University of Birmingham.
Dubois, B.L. 1987. 'Something on the order of around forty to forty-four': Imprecise numerical expressions in biomedical slide talks. *Language in Society* 16: 527–541.
Duranti, A., and C. Goodwin (eds.). 1992. *Rethinking context: Language as an interactive phenomenon*. Cambridge: Cambridge University Press.
Eckert, P., and J.R. Rickford (eds.). 2001. *Style and sociolinguistic variation*. Cambridge: Cambridge University Press.
Farr, F., and A. O'Keeffe. 2002. Would as a hedging device in an Irish context: An intra-varietal comparison of institutionalised spoken interaction. In *Using corpora to explore linguistic variation*, ed. R. Reppen, S. Fitzmaurice, and D. Biber, 25–48. Amsterdam: John Benjamins.
Farr, F., B. Murphy, and A. O'Keeffe. 2004. The limerick corpus of Irish English: Design, description and application. *Teanga* 21: 5–29.
Fetzer, A. 2004. *Recontextualizing context: Grammaticality meets appropriateness*. London: John Benjamins Publishing Company.
Fowler, R., and G. Kress. 1979. Critical linguistics. In *Language and control*, ed. R. Fowler, B. Hodge, G. Kress, and T. Trew, 185–213. London: Routledge Kegan Paul.
Garcia, A.C., and J.B. Jacobs. 1999. The eyes of the beholder: Understanding the turn-taking system in quasi-synchronous computer-mediated communication. *Research on Language and Social Interaction* 32: 337–367.
Green, L.J. 2002. *African American English: A linguistic introduction*. Cambridge: Cambridge University Press.
Gries, S.Th., and C.V. David. 2007. This is kind of/sort of interesting: Variation in hedging in English. In *Studies in variation*, *contacts and change in English 2*: *Towards multimedia in corpus studies*, Vol. 2. Helsinki: Varieng.
Halliday, M.A.K., and R. Hasan. 1989. *Language, context and text: Aspects of language in a social semiotic perspective*. Oxford: OUP.
Hard af Segerstag, Y. 2002. *Use and adaptation of the written language to the conditions of computer-mediated communication*. Unpublished Ph.D. thesis, University of Goteborg.
Herring, S.C. (ed.). 1996. *Computer-mediated communication: Linguistic, social and crosscultural perspectives*. Amsterdam: John Benjamins.
Herring, S. 1999. Interactional coherence in CMC. *Journal of Computer-Mediated Communication* 4(4): 1–13.
Herring, S. 2002. Computer-mediated communication on the Internet. *Annual Review of Information Science and Technology* 36: 109–168.
Herring, S. 2007. A faceted classification scheme for computer-mediated discourse. *Language@Internet 4*(1): 1–37.
Heylighen, F., and J.-M. Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science* 6: 293–340.
Hübler, A. 1983. *Understatements and hedges in English*. Amsterdam: John Benjamins.
Hyland, K. 1996. Writing without conviction? Hedging in scientific research articles. *Applied Linguistics* 17(4): 433–454.
Jucker, A.H., S.W. Smith, and T. Ludge. 2003. Interactive aspects of vagueness in conversation. *Journal of Pragmatics* 35: 1737–1769.
Kennedy, G. 1987. Quantification and the use of English: A case study of one aspect of the learner's task. *Applied Linguistics* 8(3): 264–286.
Knight, D. 2011. *Multimodality and active listenership*. London: Continuum.

Knight, D., S. Adolphs, and R. Carter. 2014. CANELC – Constructing an e-language corpus. *Corpora Journal*. 9(1).
Ko, K. 1996. Structural characteristics of computer-mediated language: A comparative analysis of InterChange discourse. *Electronic Journal of Communication* 6(3).
Kopytko, R. 2003. What is wrong with modern accounts of context in linguistics? *Vienna English Working Papers* 12: 45–60.
Labov, W. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Lakoff, R. 1972. Language in context. *Language* 48(4): 907–927.
Leech, G. 2000. Grammar of spoken English: New outcomes of corpus-oriented research. *Language Learning* 50(4): 675–724.
Ling, R. 2003. The socio-linguistic of SMS: An analysis of SMS use by random sample of Norwegians. In *Mobile communications: Renegotiation of the social sphere*, ed. R. Ling and P. Pedersen, 335–349. London: Springer.
McCarthy, M. 1991. *Discourse analysis for language teachers*. Cambridge: Cambridge University Press.
McEnery, T., R. Xiao, and Y. Tono. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.
Morris, C. 1946. *Signs, language and behaviour*. Englewood-Cliffs: Prentice Hall.
Murray, D.E. 1988. The context of oral and written language: A framework for mode and medium switching. *Language in Society* 17: 351–373.
Myers, G. 2010. *The discourse of blogs and wikis*. London: Continuum.
Nelson, K., S. Engel, and A. Kyratzis. 1985. The evolution of meaning in context. *Journal of Pragmatics* 9: 453–474.
Nikula, T. 1997. Interlanguage view on hedging. In *Hedging and discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts*, ed. R. Markkanen and H. Schröder, 188–207. Berlin: Walter de Gruyter.
O'Keeffe, A., M. McCarthy, and R. Carter. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.
Oksman, V., and J. Turtianen. 2004. Mobile communication as a social stage: Meanings of mobile communication in everyday life among teenagers in Finland. *New Media and Society* 6(3): 319–339.
Poos, D., and R.C. Simpson. 2002. Cross-disciplinary comparisons of hedging: Some findings from the Michigan Corpus of Academic Spoken English. In *Using corpora to explore linguistic variation*, ed. R. Reppen, S.M. Fitzmaurice, and D. Biber, 3–23. Amsterdam: John Benjamins.
Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
Rayson, P. 2003. *Matrix*: *A statistical method and software tool for linguistic analysis through corpus comparison*. Unpublished Ph.D. thesis, Lancaster University.
Scollon, R., and S. Scollon. 2003. *Discourse in place: Language in the material world*. London: Routledge.
Shortis, T. 2007. Gr8 Txtpectations: The creativity of text spelling. *English Drama Media Journal* 8: 21–26.
Sutherland, J. 2002. Cn u txt? *The Guardian*, 11 Nov 2002. http://www.guardian.co.uk/technology/2002/nov/11/mobilephones2. Accessed 22 April 2013.
Tagg, C. 2009. *A corpus linguistics study of SMS text messaging*. Unpublished Ph.D. thesis. Birmingham: The University of Birmingham.
Thurlow, C., L. Lengel, and A. Tomic. 2004. *Computer mediated communication: Social interaction and the internet*. London: Sage.
van Dijk, T.A. (ed.). 1977. *Text and context: Explorations in the semantics and pragmatics of discourse*. London: Longman.
Walther, J.B. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research* 23: 3–43.
Widdowson, H.G. 1998. Communication and community: The pragmatics of ESP. *English for Specific Purposes* 17(1): 3–14.
Yates, S. 1996. Oral and written linguistic aspects of computer conferencing: A corpus based study. In *Computer mediated communication: Linguistic, social and cross-cultural perspectives*, ed. S.C. Herring, 29–56. Amsterdam: John Benjamins.

A Corpus-Based Classification of Commitments in Business English

Rachele De Felice

1 Introduction

This chapter presents a corpus-based, data-driven study of the speech act of commitments in Business English (henceforth BE) emails. Based on a sample of 1,200 instances, a detailed analysis of this speech act from a variety of linguistic perspectives is given: lexical, grammatical, and syntactic. We assess how well 'real-world' commitments compare to the traditional theoretical definitions given in the literature, whether they are in fact a homogeneous category, and how we 'do politeness' when we perform a commitment. By looking at the commitments' main lexico-grammatical characteristics, including frequent phrases and patterns, we propose a revised description of this speech act, which expands the definitions given by Austin (1962) and Searle (1969, 1976). This work is a case study showing how the tools of corpus pragmatics can lead to new developments not just in the description of language, but also in pragmatic theories. As O'Keeffe et al. (2011) note, "There can be tensions between speech act classifications and taxonomies which were developed on the basis of invented examples, and the analysis of speech acts in corpus data" (O'Keeffe et al. 2011:97). In using corpus analysis to further our understanding of pragmatics, and speech acts in particular, this research aims to resolve and clarify some of these tensions.

In particular, the questions posed by this paper are as follows:

1. What do commitments look like?
2. Can we update or expand their definition?
3. What is their range of functions?

R. De Felice (✉)
Department of English, University College London, Gower Street, WC1E 6BT London, UK
e-mail: r.defelice@ucl.ac.uk

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_8,

2 Related Work

There is a growing body of work on the use of corpora in pragmatics, exemplifying what Romero-Trillo has described as the "mutualistic entente" between the two disciplines of corpus linguistics and pragmatics (Romero-Trillo 2008:3). This involves both the use of corpus-based studies of pragmatics topics, and the use of "pragmatics as a model for the interpretation of [corpus] data" (ib.). Rühlemann (2010) provides an overview of some of the advantages and disadvantages of corpus pragmatics research, as well as discussing some of the key studies in this field; lengthier discussions of the topic are found in Romero-Trillo (2008), Jucker et al. (2009), and O'Keeffe et al. (2011).

There are also several monographs exploring particular pragmatic phenomena from a corpus-based perspective, such as Aijmer (1996, 2002) and Brinton (1996) on discourse markers, and Adolphs (2008), which relies on corpus analysis and real data, as in this case, to determine the "functional profiles" of the phrases used in the speech act of suggestions. The research described in the present paper has some similarities with Adolphs (2008) in that both make use of real world data, and, as we will see below, both posit a key role for collocations and phrases in determining the intended function of the speech act. However, there are two significant differences: the work described in this article looks at email communication rather than spoken language, and it aims to be an unbiased analysis of the characteristics of speech acts. While Adolphs (2008) began her analysis by searching for specific strings, in this research we begin by looking at the entire speech act category. The corpus used is annotated for speech acts (cf. De Felice et al. 2013), making it possible to analyse an entire category for any features, and minimising the risk of missing potentially informative phrases or other items. Furthermore, to the best of our knowledge, the present research is one of very few to focus on the speech act of commitments rather than the more-often discussed requests, suggestions, and other directives.

Regarding BE specifically, there is a vast amount of work in this field, some of it corpus-based, like this study (e.g. Holmes and Stubbe 2003; Koester 2006, 2010; Handford 2010; Holmes et al. 2011); however, the focus of those works is usually on spoken rather than written communication, and, while many features are analysed, including effective pragmatic use, they are not systematic accounts of particular speech acts such as is done here.

3 Background to the Research

This research is part of the larger 2-year project PROBE (PRagmatics of Business English), which uses corpus and computational linguistics to create a description of the pragmatic features of email BE, in particular speech acts. One of the central questions of the project is: "what do speech acts look like?". This takes into account two lines of inquiry, their identification and interpretation by computers

Table 1 Speech act categories used

Speech act	Tag	Example
Direct request	DR	Please send me the files.
Question-request	QR	Could you send me the files?
Open question	QQ	What time is the meeting?
First person commitment	FPC	I will attend the meeting.
First person expression of feeling	FPF	I am uncertain about the agenda.
First person other	FPO	I am an employee of this company.
Other statements	OT	The meeting is at 8 tomorrow. You always work so hard.

(e.g. for automated tagging) and by humans (e.g. how, in everyday interaction, we correctly interpret the intention of the speaker,[1] and communicate our own intentions). The research presented in this paper focuses on the latter aspect, speech act creation and interpretation by humans. Central to the work is the creation of a high quality manually speech act annotated corpus of real email data. This resource is described in detail in De Felice et al. (2013). In brief, it consists of approximately 20,700 utterances (263,100 words) from the EnronSent email corpus (Styler 2011)[2]; in the context of the PROBE project, each utterance has been manually annotated with one of seven speech act tags, as shown in Table 1. This is, to the best of our knowledge, one of very few such annotated resources. Though other pragmatically annotated corpora exist (e.g. Archer 2005; Maynard and Leicher 2006; Kallen and Kirk 2012), these are very domain-specific and identify pragmatic categories typical of the particular domain under investigation, such as judicial, spoken,[3] or academic language. In this project, on the other hand, there is a broader focus on general pragmatic features, to enable reuse within the academic community.[4]

The goal of this paper is to use corpus data to propose revised, better-fitting definitions of this speech act category, expanding the traditional definitions originally given by Austin and Searle. Austin's traditional definition of commissives

[1] Although the research deals with written language, for convenience and uniformity with spoken language research, I refer to speaker and hearer instead of writer and reader, respectively.

[2] The Enron email corpus consists of the unedited, unmodified collection of Enron employees' mailboxes; this data was made publicly available following legal proceedings against the corporation. It is the largest publicly available collection of real-world BE email data.

[3] I am grateful to the anonymous reviewer who drew my attention to the pragmatically annotated spoken language corpus by Kallen and Kirk (2012); unfortunately I have not yet been able to review it so I cannot assess its suitability for the present research.

[4] There is also limited overlap between the research presented here and the well-established field of dialogue act classification (for example Core and Allen 1997; Georgila et al. 2009; Stolcke et al. 2000). The focus there is on synchronous communication, with few complete sentences, and categories which do not reflect well the types of utterances found in written language.

(essentially the same as what we are here referring to as commitments) is that they "commit the speaker to a certain course of action" (Austin 1962:156). They are "an obligation or declaring of an intention" (ib.). Typical examples of commissives include promising, intending, and betting. Searle (1969, 1976) uses a very similar definition, saying that commissives "commit the speaker to a future course of action" (Searle 1976:11), and includes among his examples promising, offering, and vowing.

4 A Revised Classification of Commitments

The proposed classification of commitment functions follows from a detailed corpus analysis of data. As Adolphs (2001) notes, "lexicogrammar has not traditionally been part of the description of speech acts" (Adolphs 2001:63), with the focus having been instead on felicity conditions or philosophical parameters. This research, which does instead put lexicogrammar at the forefront, contributes to our understanding of corpus linguistics, pragmatics, and business communication, showing the potential of combining insights from different disciplines to advance our knowledge of language use. In particular, the contributions of this work are:

1. to further our understanding of the pragmatics of email communication;
2. to offer an expanded view of commitments in speech act theory;
3. to provide new corpus-based insights into business email communication.

This work can in turn inform further lines of inquiry. For example, the study of the language used in business communication is relevant to those learning it as non-native speakers, as it is helpful to know how to communicate one's intentions clearly in a professional context, to avoid costly misunderstandings (this is a rich area of research, cf. for example Koester 2002, 2010; Gimenez 2006; Bargiela-Chiappini et al. 2007; Bargiela-Chiappini 2009; Handford 2010; Newton and Kusmierczyk 2011). Furthermore, from a technical perspective, a thorough understanding of the lexical and grammatical properties of commitments and all other speech acts allows us to develop more refined corpus analysis and computational linguistic tools, such as automated speech act taggers, offering new opportunities for corpus linguistics (cf. e.g. De Felice 2012; De Felice and Deane 2012).

In the dataset, currently, commitments make up around 10.5 % of the annotated speech acts. To allow a more in-depth analysis, a set of 1,200 examples was extracted for the purpose of this work. This amounts to 14,663 words, giving an average length per instance of 12.2 words.

Typical examples from the commitments data are shown in the following examples:

(a) I will send out the report tomorrow.
(b) I am going to work on this today.
(c) I will be out of my office on Monday.

Table 2 A 3×2 matrix for the classification of commitments

	SOFT	HARD
ACTION	√	√
INFORMATION	√	√
SOCIAL	√	X

(d) I can be reached on my mobile phone.
(e) I'll let you know when I figure it out.
(f) I'll keep you posted.

These examples share some surface similarities: first person subjects, modal verbs with a main verb or main verbs followed by a complement clause, and often an expression of time at the end. However, a closer look at their content reveals important functional differences related to the use of different verbs and auxiliaries, the presence of expression of time, and other surface features.

We propose that these definitions are not sufficiently fine-grained to appropriately capture the main functions performed by commitments as revealed by corpus data, at least in the domain of business communication. A more informative representation of this speech act category can instead consist of a 3×2 matrix highlighting the three main functions assigned to commitments, and the two ways in which their content is constrained by a particular time frame. The basic matrix is shown in Table 2; the final, populated version is shown in Table 5. We will briefly introduce this descriptive framework before providing the corpus analysis data supporting this categorisation.

The three main functional classes of commitments are:

- action
- availability information
- social function

The action class more or less corresponds to our general understanding of prototypical commitments, where there is an offer or promise of a specific action given by the speaker, often with a tangible outcome. Typical examples include (a) and (b). This class is related to the notion of transactional or task goals discussed, among many others, by Brown and Yule (1983:1–3), Clark (1996) and, in the context of business communication, by Koester (2004a, b); its function is to "get things done" (Koester 2004b:1406).

The availability information class includes commitments such as (c) and (d). Its function is to convey information about the speaker's availability and plans rather than committing to any particular concrete action. As the examples above suggest, this group of commitments is used by the speaker to communicate his or her absence or presence at work, and other ways in which they can be available for communication. This function is particularly salient in business communication since, where people work in teams, the absence or lack of availability of a colleague can have consequences for his or her co-workers, too. It might be argued that the lack of concrete action in this group of speech acts invalidates their inclusion in the

commitments category. However, they do contain an undertaking on the part of the speaker to do something (e.g. travel, or be in the office). Furthermore, as we have seen above, the traditional definition of this category includes "declaring of an intention, committing to a future course of action" (Austin 1962; Searle 1976), which corresponds to these examples: the speaker is stating his or her intention to be away or in the office (or other relevant location), which represents a future course of action to be undertaken.

Finally, the social function class, exemplified by (e) and (f) above, refers to those commitments which contain a promise of future contact, of keeping the hearer up-to-date about something. They are generally somewhat vague, and their main function is related to what have been variously referred to as interactional goals (Brown and Yule 1983:1–3), interpersonal goals (Clark 1996), and relational goals (Koester 2004a, b) among other terms. They respond to the need for social harmony in the workplace. This group of commitments is used to keep channels of communication open and maintain good working relations by reassuring the hearer that matters that concern them are not being neglected. As Clark (1996) notes, the function of interpersonal goals is "maintaining contact with the other participants, impressing them, being polite, maintaining self-respect" (Clark 1996:34). This function and the availability information one are closely related since, on the one hand, giving information about one's availability can also contribute to good working relations and, on the other, stating that further communication will follow is also a form of information-giving. However, the stronger focus on the hearer's positive face (cf. e.g. Brown and Levinson 1987) in the case of the social function supports the decision to maintain the distinction between these two functional categories.

4.1 The Hard-Soft Time Constraint

As introduced in the previous section, the other main dimension along which we propose to classify commitments regards whether any temporal constraints are associated with the commitment made. This is a transversal classification that applies to all three functional categories. A commitment is defined as 'hard' if there is a specific time or date given as the deadline for the commitment's fulfilment (cf. *tomorrow*, *today*, *on Monday* in examples (a), (b), (c) respectively). These expressions of time add accountability; for ease of reference, they are referred to as 'timestamps'. A commitment is 'soft' if it either has no timestamp at all (cf. examples (d) and (f)), or only a vague one, referring to an unquantified amount of time (e.g. *in a few weeks*) or to a vague event without a specific point in time (cf. example (e), *when I figure it out*).

Overall, there are 529 expressions of dates and times in the data, according to the named entity recogniser (cf. Sect. 5). Dates, which include days, months, and specific dates, make up the majority of these, with a total frequency of 389. Many (n = 85) are days of the week, with *Monday* and *Friday* the most frequent. These are also often further defined by the use of next, such as *next week* or *next [number] days*. Specific calendar dates (e.g. *Jan 11th*) also occur relatively often (n = 44),

while months are mentioned much less frequently; taken together, these features contribute to the impression that (self-imposed?) hard deadlines are frequent in commitments. Times, which refer to specific times of the day (e.g. *3 o'clock* or *10:30*) are less frequent, with a total of 140 occurrences. The focus of commitments on immediate time spans and quick deadlines might have led us to expect this category of times to be larger, as times of the day are generally representative of short-term commitments. We suggest that their relative infrequency is due to the fact that there is a limit to how 'hard' people are willing to commit: while a generic part of the day is acceptable (e.g. morning or afternoon), specific times are avoided as such a precise deadline might be harder to meet.

As we will see in the analysis of the corpus data, action and availability commitments can be both soft and hard, while social function commitments are soft only. Action commitments are both hard and soft as they encompass a wide range of workplace actions, not all of which require a clear deadline. As we will see below, there are interesting differences in the phraseology of the two groups. The availability commitments are often hard, as is to be expected given their function: for the information to be useful, it needs to include details about the time frame it refers to. They do also occur as soft ones, which appear to be vaguer (cf. (d) above), but we can imagine that they gain clearer significance in the wider context of the email. The social functional category is only of the soft type because the main role of these commitments is to facilitate agreeable working relations, and foster friendly interaction: these sorts of actions naturally do not require a time frame.

4.2 Requirements for Classification

The two dimensions of classification – the functional and the temporal – correspond to what we maintain are the two key requirements for something to be classified as a commitment. These are, in a way, data-driven versions of felicity conditions (Austin 1962):

1. intentionality: the speaker has to declare an intention to do something;
2. fulfillability: the statement has to be something that can be fulfilled, or whose fulfilment can be checked, rather than just a state of being; it refers to something that ought to happen. Hence *I am away this week* counts as a commitment, because once the week is over, or during the week, one can check if I have been away; but *I am Italian* does not, because it just describes an ongoing state of being rather than something I have declared an intention of doing.

5 Tools and Methodology

The corpus analysis underlying the development of the functional categorisation encompasses the use of many tools. For reasons of space, the full details of the analysis cannot be provided here; only the main findings will be discussed.

To obtain a detailed and multi-level description of commitments, several tools are used in the analysis of the data. In the PROBE project, the multi-level analysis considers all levels of language, namely:

- Vocabulary and phrases/n-grams
- Use of proper nouns
- Grammar and part of speech behaviour
- Syntactic structures
- Discourse sequences

This information is collected using a variety of tools from computational and corpus linguistics, mainly the C&C toolkit (Curran et al. 2007), which performs part of speech tagging, parsing (for syntactic information, cf. Clark and Curran 2007), and named entity recognition (NER; i.e. the automatic classification of proper nouns into categories such as dates, times, names, organisations, places, cf. Curran and Clark 2003), and WordSmith 5 (Scott 2010). Further analysis is also provided by the speech act analyser tool developed by the author and described e.g. in De Felice (2012) and De Felice et al. (2013). The analysis presented here focuses mainly on the lexical data extracted and analysed using WordSmith. The lexicon of business commitments will be briefly introduced, followed by a detailed description of their relationship to the different functional categories previously established.

6 The Vocabulary of Business Commitments

The set of commitments analysed, which consists of 14,663 words, does not have a very varied vocabulary: there are only 2,171 types, suggesting that there is a core vocabulary of phrases and chunks used in formulating this speech act. Table 3 lists the 30 most frequent words in the corpus. *Will* and *'ll* have been collapsed together, as have *am* and *'m*. The symbol # replaces any numbers. Table 4 lists the 30 most frequent bigrams (two-word sequences) in the data; as we can see, there is a very high degree of overlap between the two tables. Overwhelmingly, the vocabulary is dominated by *I*, *will*, and their combination *I will*. This is a straightforward, canonical way to create a commitment, so its high frequency is at first glance not remarkable. However, frequency lists are not very meaningful without the more in-depth view of the data provided by close reading of concordance lines, which, as we will see in the following sections, reveal how the same words and bigrams can actually fulfil very different functions.

6.1 Action Function

The function of this subgroup of commitments is for the speaker to offer or promise to do some future action. As mentioned in the previous section, the bigram *I will*,

Table 3 Top 30 words in commitments corpus

Word	Freq.	% of total
I	910	6.21
WILL	734	5
TO	655	4.47
THE	616	4.2
YOU	325	2.22
AND	305	2.08
WE	280	1.91
#	263	1.79
BE	244	1.66
A	237	1.62
IN	228	1.55
ON	225	1.53
OF	183	1.25
AM	174	1.18
FOR	148	1.01
THAT	141	0.96
CAN	136	0.93
IT	133	0.91
WITH	125	0.85
IF	121	0.83
THIS	118	0.8
AS	117	0.8
AT	115	0.78
HAVE	96	0.65
GET	95	0.65
TOMORROW	89	0.61
CALL	88	0.6
S	79	0.54
OUT	72	0.49
UP	70	0.48

and the less frequent *(am) going to*, are standard ways of indicating future intention and could be said to often function as the Illocutionary Force Indication Devices (IFID, i.e. as a lexico-grammatical structure used to indicate the illocutionary function of the utterance, in this case future action or intention; cf. Levinson 1983:238) for this function. These bigrams are complemented by a wider range of verbs denoting actions typical of a workplace, such as:

(g) I will send out a real legislative report tomorrow.
(h) We'll get the first one out Monday.
(i) I am going to work on this today and hope to e-mail it out by 9 a.m. tomorrow.
(j) I'll pick up my ticket on Monday.
(k) I am seeing Tom today.

A distinctive characteristic of this group of commitments is that they almost always occur with a specific expression of time, that is, following the distinction introduced above, they fall into the category of hard commitments. The use of the

Table 4 Top 30 bigrams in commitments corpus

Word	Freq.	% of total
I WILL	490	3.34
WILL BE	148	1
I AM	142	0.96
WE WILL	122	0.84
I CAN	83	0.57
IN THE	69	0.47
ON THE	61	0.42
TO YOU	53	0.36
OF THE	48	0.33
TO THE	45	0.31
GOING TO	43	0.29
AND I	41	0.28
IF YOU	36	0.25
AT #	34	0.23
WE CAN	34	0.23
TRY TO	33	0.23
WE ARE	33	0.23
WILL HAVE	31	0.21
AM GOING	30	0.2
SOON AS	30	0.2
BE IN	29	0.2
HAPPY TO	29	0.2
WILL TRY	28	0.19
LET YOU	27	0.18
YOU A	27	0.18
YOU KNOW	27	0.18
WILL FORWARD	26	0.18
WILL GET	26	0.18
WILL LET	26	0.18
SEE YOU	26	0.18

timestamp, as mentioned, gives a fixed point in time at which the speaker can be held accountable for fulfilling the action committed to.

Another bigram which is usually considered a typical IFID is *I can*. Indeed, it, too, indicates the speaker's orientation towards doing something, and it is often followed by verbs denoting specific actions, similarly to the previous examples:

(l) I can [call and get the address, draft up something, follow up with details, give you a ride].

However, the use of a different modal verb weakens the commitment: the speaker is showing willingness but not necessarily entering into an obligation unless required or encouraged by the hearer. This interpretation is supported by the fact that commitments with *I can* tend to be of the soft type, that is, without a timestamp (as in the examples given in (l)). This further weakens the commitment by removing clear accountability and presenting it as a suggestion or offer rather than a clear undertaking to do something. Of course, knowledge of the relevant contextual variables, such as speaker status, could help determine the true extent of the strength of these

commitments – a large difference in power between hearer and speaker, for example, would yield quite a different interpretation.[5]

Closely related to this bigram is the use of the adjective *able (to)*. There are 15 occurrences of it in this data, which all pattern with modal verbs, as follows: *should be able to (5), will be able to (3), will not be able to (2), may not be able to (2), other (3)*; there are no bald statements of the form *I am able to*. These constructions are even more tentative and hedging than *I can*, adding a further layer of possibility and distance from the action discussed; similarly, they are always soft commitments. Examples of the use of these phrases are:

(m) Once you let me know how this looks to you, I will be able to move forward with the creation of the products.
(n) We then plan to request the Board's approval for this plan on March 15th in order to be able to implement the plan by summer.

The main difference between *can* and *able to* – which are, on the surface, nearly synonymous – is that the latter phrase is used in making commitments where the speaker's action depends not on the speaker him/herself, despite their willingness, but on some external event. In examples (m) and (n), we can see that a chain of events is indicated, allowing the speaker to show goodwill and proactivity without having to necessarily act on their commitment immediately. This can be seen as a face-preserving strategy for the speaker: he or she is displaying willingness, while at the same time deflecting potential criticisms of the timescale of activities by shifting the burden of responsibility.

The corpus analysis also reveals a further pattern for soft commitments of this kind, which uses if-clauses or the subordinating conjunctions *until* and *unless* together with, often, negation of the main action:

(o) If they don't change the date, I won't go to D.C.
(p) If I can't get the materials off Ian, I'll let you know.
(q) Until we finalize those business decisions, we cannot make that determination.
(r) Unless you instruct otherwise, I won't drop off the check in person.
(s) I won't contact others until we discuss Enron's position on co-sponsorship.

Again, in these cases the speaker is making it clear that their good intentions are limited by external events. Although often the main action proposed is negated, this is only a 'pending' negation which holds only as long as the external event is also still pending. The speaker's intention to carry out an action is still evident, but the introduction of other actions outside their control is a useful form of hedging and avoiding too much overt responsibility.

Interestingly, the bigram *I promise*, one of the prototypical examples of performative verbs in the literature, never occurs in this data.

[5]For a related discussion, cf. Koester (2004a:62), where the author notes how the use of modals such as *will* and *be going to* indicate confidence and assertiveness on the part of the speaker, in contrast to the more tentative modals *could* or *might*.

To summarise this section, there are clear and well-established lexical patterns in the corpus data which support the creation of the offer-of-action function of commitments. These include both hard ones, which use *I will/am going to* together with a timestamp, and soft ones, which use a range of hedging lexical and syntactic devices such as *I can*, *able to*, if-clauses, and subordinating conjunctions to mitigate the strength of the action promised.

6.2 Availability Information Function

The function of this subgroup of commitments, as explained above, is to convey information about the speaker's intentions regarding his or her travel and absence from the office or other locations. As we have seen in the previous section, *I will* is a typical IFID for the action function of commitments; however, the analysis of concordance lines reveals that it also fulfils a major role in the function of giving information about availability. In particular, unlike the previous function, this function makes frequent use of the bigram (*I*) *will be* (the second most frequent bigram according to the table). In this pattern, *be* is the main, stative, verb and is followed by one of the four prepositions: *at, in, on*, or *out*, together with expressions of time or place. For example:

(t) I will be at the meeting and will report back.
(u) I will be in at 7 a.m. tomorrow.
(v) I will be on vacation on November 26.
(w) I will be out of the office next week.

In these commitments, the intention of the speaker refers not to a tangible action, but to their being (or not being) somewhere. We noted above that availability commitments are mainly hard, as is to be expected given their function; indeed, examples with *I will be* always follow this pattern. The other phrase that is used regularly for this function is *I am*, as in the following examples:

(x) I am out of the office/on vacation.
(y) I am available on [Monday/Wednesday/next week/etc.].

These can be either soft or hard, as the examples show. As we suggested above, the soft ones can still be informative for the hearer, particularly if he or she has access to the wider discourse context which will assist in the interpretation of the statement's relevance. This also applies to the third bigram used for this function, *I can*, which is usually part of soft commitments:

(z) I can [be available, be reached at this number, do any of these dates, make time for the meeting].

Although the function of this sub-group of commitments is limited and highly specialised, I believe it is justified to consider it as a stand-alone category because of the high frequency of the patterns identified for it, and its strong homogeneity.

6.3 Social Function

Finally, this section describes the corpus data supporting the creation of the third function of commitments, the social or phatic communion function. As previously discussed, these commitments, which are always only soft, have the role of maintaining good interpersonal relations in the workplace by showing consideration of the hearer's needs and positive face.

Many of the bigrams in the bottom half of Table 3, which are combinations of *will* together with another main verb (*try/forward/get/let*), relate to this function. Examples are:

(aa) I will get back to you as soon as possible.
(bb) I will let you know [as soon as I have heard back from them/when I figure it out/if there are any omissions].

These soft commitments belong to interpersonal rather than transactional discourse, as they fulfil the social function of phatic communion rather than committing to a particular action. A distinguishing bigram of this category is (*as*) *soon as* (n = 30), as shown in the examples above. Interestingly, despite the prototypical completion of this phrase being *as soon as possible*, there is only one occurrence of it in the data, the other, more common completions being: *we/I have/get the info, they are in/available, I hear/learn/know something*. In other words, commitments with this phrase tend to indicate that the speaker's offer or promise is contingent on external factors, understood to be beyond their control, and as such fall in the category of soft commitments as it is impossible for the speaker to give a specific time for their fulfilment. A cynical view of working relations might suggest that this is a convenient strategy for co-workers to show willingness and concern for others' needs without necessarily having to engage in any concrete actions. This impression of distancing one's self from personal responsibility is also supported by the fact that the phrase *as soon as I can*, which implies personal action, only occurs four times.

I will and *as soon as* are not the only typical bigrams of this class; other common phrases which fulfil the function of strengthening working relations, confirm decisions made in other venues, and maintaining a good working flow are phrasal verbs such as *catch up* and *follow up* (six occurrences each):

(cc) I'd be happy to come to your office to meet and catch up.
(dd) I can follow up with details.

Example (cc) also shows another bigram that defines this class, *happy to* (n = 29; also four occurrences of the related term *glad to*). It always occurs in sentences such as *I am/will/would be happy to…* This is an interesting variation on the standard commitment phraseology, which is usually presented with a focus on the use of modal verbs rather than periphrasis. The phrase *happy to* is followed by a variety of main verbs, many of which relate to talking and discussing: *discuss* (5), *talk* (3), *sit* and *provide* (2 each), *input*, *chat*, and *incorporate* (1 each). We suggest that, while

I can and *I will* are used for concrete actions that advance a project or other work tasks, this phrase is more typically used for commitments with a social function. Perhaps the use of the adjective *happy*, as opposed to a more neutral verb, serves to emphasise the positive attitude of the speaker towards the hearer, and makes more explicit the focus on the hearer's wants. We propose to classify *happy to* as an IFID for this category of commitments.

Occasionally, membership of this category is not clear-cut. For example, another frequent commitments bigram is *will forward* (n = 15). This is closely tied to the context of email communication, with typical instances referring to common practices of email use:

(ee) I will forward to you a memo from David.
(ff) I will forward your message to Sam.

The concrete actions, of a transactional nature, indicated by these examples would suggest that they belong to the action category of commitments. However, we can also argue that the phrase *will forward to you* belongs to the same cluster of as *send to you, talk to you, get back to you, keep you posted*: hearer-centred commitments with a focus on the transmission of information and smooth working conditions. This, then, would put examples such as (ee) and (ff) in the social function category. It is not uncommon for actions to have both transactional and interactional components, so it is reasonable to conclude that there can be some types of commitments which encompass both the social and the action function. Indeed, as Handford (2007) notes, in business meetings the distinction between transactional and relational can be hard to draw, and it is possible to claim that "interpersonal language is employed with clear transactional goals in mind" (Handford 2007:43). More detailed analysis of the data, currently underway, will allow us to better understand the characteristics of these dual-function commitments.

7 An Aside on Politeness Markers: *If You*

Politeness markers are usually discussed in speech acts with relation to requests, suggestions, and other actions that involve an imposition on the hearer, and therefore benefit from mitigation. However, the corpus analysis of the commitments data shows that politeness markers appear in this speech act, too, in particular in the form of the bigram *if you* (n = 36). Typical examples are:

(gg) If you agree, I'm going to assign low priority to the issue.
(hh) We can talk about it if you are interested.
(ii) I can make the reservations if you want.

The *if you* phrases, which also include verbs such as *like* and *wish*, are in the main incidental to the commitment and used as a way to show politeness and deference to the hearer's needs and priorities. Although offering to do something is implicitly a show of politeness and face concern, there may be situations where the

Table 5 A 3×2 matrix for the classification of commitments

	SOFT	HARD
ACTION	I can + action + [delimiting clause]	I will/am going to + action + timestamp
	I can call and get the address.	*I will email the report at 9.*
	I can make the reservations if you want.	
INFORMATION	I can + be/make + PP	I am/will be + PP location + time expression
	I can be reached at this number.	*I'll be in at 7 a.m. tomorrow.*
SOCIAL	--- verbs of communication, no timestamp, no simple pattern, ---	–
	I'll let you know when I figure it out.	
	I'll keep you posted.	

relations between co-workers require further use of politeness (cf. example (gg), where the speaker is endeavouring to make his or her decision by seeking the hearer's agreement). The corpus analysis has revealed one such device, and further research will show what other grammatical and lexical strategies speakers use to accomplish politeness in speech acts where politeness is not always studied, such as commitments.

8 Summary of the New Categories of Commitments

The corpus analysis of the vocabulary and phraseology of commitments has shown us that they can be hard or soft, and that the three central functions of this speech act in the business domain are to commit to concrete actions, communicate availability, and maintain good working relations. The 3×2 categorisation discussed in this paper is summarised in Table 5, which is a populated version of the scheme introduced in Table 2.

We can also imagine the functions as placed along a cline indicating the firmness of the commitment, with offers of action at the more firm end, socially motivated promises at the least firm, vaguer, end, and information about availability somewhere in the middle, being less concrete than actions but more easily delimited than social commitments.

9 Speech Act Theory Versus Discourse-Based Views of Pragmatics

One of the central aims of this paper, and the wider research project of which it is part, is to demonstrate how the tools of corpus analysis, together with the availability of real-world corpora, can give us a richer understanding of the way we create and

understand speech acts in everyday communication. We now have the opportunity to revisit long-held beliefs and theories about particular aspects of pragmatics, for example by proposing revised categories of speech acts, as has been done here for commitments. In doing this, the present research contributes to the growing field of corpus pragmatics (cf. Sect. 2).

It could be argued that the speaker-centred view of speech acts discussed here is uninformative, and that a more representative analysis of speech act functions would examine the hearer's response, as that is a clearer indicator of the perceived meaning of the speech act. This discourse-centred approach to speech acts and communication, which examines speech acts not in isolation but within longer sequences of discourse, has been widely discussed in the literature, albeit always with a focus on spoken, interactive communication (see e.g. Schegloff 1988, 1999; Arundale 1999 among many others; but cf. also Van Rees 1992 for a counter-argument). However, in the domain of written business communication, especially one where we do not have much information about the original participants in the exchange, we cannot adopt this approach and have to focus on what we can access from the linguistic form only. Despite being incomplete in some ways, it is the first step towards creating new functional profiles of communication, which can then be expanded and adapted as further sources of data become available.

10 Future Work and Conclusion

10.1 Further Analysis

The analysis of lexicon, bigrams, and particular parts of speech such as adjectives, as well as proper nouns and named entities, has already proved quite fruitful as a basis for a corpus-based view of pragmatics. Within the PROBE project, work is ongoing, and the results of the analysis of other categories such as verb classes and prepositions are likely to shed further informative details. Another major aspect of the analysis is discourse structure: placement within the email discourse is likely to be a further significant factor in the categorisation of the commitments. For example, do commitments tend to occur with other statements, or in combination with requests? Preliminary analysis of a sample of the data has shown that both positions occur, but more research is needed to determine the role these differences play.

10.2 Further Domains

The focus of this work is on business communication, in particular emails. It is expected that other forms of communication – for example spoken language, or non-business emails – might deviate in some ways from the structures found here.

The analytical framework set up in this research project paves the way for similar comparative work on other types of communication, if the data is available, leading to a more comprehensive, corpus-based description of speech acts and pragmatic features.

10.3 To Conclude

A corpus-based study of commitments has shown a different approach to the traditional speech act taxonomy. We have shown that, from an analysis of the lexicon and grammatical properties of real world data, it is possible to arrive at a more detailed and descriptively accurate classification of traditional speech acts. This methodology will also be applied to other speech acts to provide a complete overview of corpus-based pragmatics, with an emphasis on their functions rather than abstract felicity conditions. Stubbs (1983) observed that "[a] final severe restriction on speech act theory, which is due to its self-imposed restriction to invented data, is a strong tendency to study the formulation of trivial speech acts" (Stubbs 1983:490). We now have the tools to overcome this restriction, and enrich speech act theory with empirically derived insights.

Acknowledgments Rachele De Felice would like to gratefully acknowledge the support received by the Leverhulme Trust; this research was undertaken by the Fellowship holder and not on behalf of the Leverhulme Trust. This research was carried out while the author was a research fellow at the Centre for Research in Applied Linguistics, University of Nottingham. Thank you also to Jeannique Darby, Tony Fisher, and David Peplow for their invaluable work in manually annotating the Enron email data and identifying taxonomical issues, and to the anonymous reviewers for their helpful suggestions.

References

Adolphs, Svenja. 2001. Linking lexico-grammar and speech acts: A corpus-based approach. Ph.D. dissertation, The University of Nottingham.

Adolphs, Svenja. 2008. *Corpus and context: Investigating pragmatic functions in spoken discourse*. Amsterdam: John Benjamins.

Aijmer, Karin. 1996. *Conversational routines in English*. London: Longman.

Aijmer, Karin. 2002. *English discourse particles*. Amsterdam: John Benjamins.

Archer, Dawn. 2005. *Questions and answers in the English courtroom (1640–1760): A sociopragmatic analysis*. Amsterdam: John Benjamins.

Arundale, Robert. 1999. An alternative model and ideology of communication for an alternative to politeness theory. *Pragmatics* 9(1): 119–153.

Austin, J.L. 1962. *How to do things with words*. Oxford: Clarendon.

Bargiela-Chiappini, Francesca (ed.). 2009. *The handbook of business discourse*. Edinburgh: Edinburgh University Press.

Bargiela-Chiappini, Francesca, Catherine Nickerson, and Brigitte Planken. 2007. *Business discourse*. Basingstoke: Palgrave Macmillan.

Brinton, Laurel. 1996. *Pragmatic markers in English: Grammaticalization and discourse functions*. Berlin: Mouton de Gruyter.
Brown, Penelope, and Stephen Levinson. 1987. *Politeness*. Cambridge: Cambridge University Press.
Brown, Gillian, and George Yule. 1983. *Discourse analysis*. Cambridge: Cambridge University Press.
Clark, Herbert. 1996. *Using language*. Cambridge: Cambridge University Press.
Clark, Stephen, and James Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics* 33(4): 493–552.
Core, Mark, and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In Proceedings of the Working notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines, Cambridge, MA.
Curran, James and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In Proceedings of the CoNLL Conference, Edmonton, Canada.
Curran, James, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and Boxer. In Proceedings of the ACL 2007 Demonstration Session.
De Felice, Rachele. 2012. Applied Pragmatics: Corpus-based methods and computational tools. Paper presented at "Discourse and Technology: Tools, Methods and Applications", Birmingham, 17–18 May.
De Felice, Rachele, and Paul Deane. 2012. *Identifying speech acts in emails: Toward automated scoring of the TOEIC ® email task*. Princeton: ETS.
De Felice, Rachele, Jeannique Darby, Anthony Fisher, and David Peplow. 2013. A classification scheme for annotating speech acts in a business email corpus. *ICAME Journal* 37.
Georgila, Kalliroi, Oliver Lemon, James Henderson, and Johanna Moore. 2009. Automatic annotation of context and speech acts for dialogue corpora. *Natural Language Engineering* 15(3): 315–353.
Gimenez, Julio. 2006. Embedded business emails: Meeting new demands in international business communication. *English for Specific Purposes* 25(2): 154–172.
Handford, Michael. 2007. The genre of the business meeting: a corpus-based study. Ph.D. dissertation, The University of Nottingham.
Handford, Michael. 2010. *The language of business meetings*. Cambridge: Cambridge University Press.
Holmes, Janet, and Maria Stubbe. 2003. *Power and politeness in the workplace: A sociolinguistic analysis of talk at work*. London: Pearson.
Holmes, Janet, Meredith Marra, and Bernadette Vine. 2011. *Leadership, discourse and ethnicity*. Oxford: Oxford University Press.
Jucker, Andreas, Daniel Schreier, and Marianne Hundt (eds.). 2009. *Corpora: Pragmatics and discourse*. Amsterdam: Rodopi.
Kallen, Jeffrey, and John Kirk. 2012. *SPICE-Ireland: A user's guide*. Belfast: Cló Ollscoil na Banríona.
Koester, Almut. 2002. The performance of speech acts in workplace conversations and the teaching of communicative functions. *System* 30: 167–184.
Koester, Almut. 2004a. *The language of work*. London: Routledge.
Koester, Almut. 2004b. Relational sequences in workplace genres. *Journal of Pragmatics* 36: 1405–1428.
Koester, Almut. 2006. *Investigating workplace discourse*. London: Routledge.
Koester, Almut. 2010. *Workplace discourse*. London: Continuum.
Levinson, Stephen. 1983. *Pragmatics*. Cambridge: Cambridge University Press.
Maynard, Carson, and Sheryl Leicher. 2006. Pragmatic annotation of an academic spoken corpus for pedagogical purposes. In *Corpus linguistics beyond the word: Corpus research from phrase to discourse*, ed. Eileen Fitzpatrick, 107–116. Amsterdam: Rodopi.
Newton, Jonathan, and Ewa Kusmierczyk. 2011. Teaching second languages for the workplace. *Annual Review of Applied Linguistics* 31: 74–92.

O'Keeffe, Anne, Brian Clancy, and Svenja Adolphs. 2011. *Introducing pragmatics in use*. London: Routledge.

Romero-Trillo, Jesús (ed.). 2008. *Pragmatics and corpus linguistics: A mutualistic entente*. Berlin: Mouton de Gruyter.

Rühlemann, Cristoph. 2010. What can a corpus tell us about pragmatics? In *The Routledge handbook of corpus linguistics*, ed. Anne O'Keeffe and Michael McCarthy, 288–301. London: Routledge.

Schegloff, Emanuel. 1988. Presequences and indirection. Applying speech act theory to ordinary conversation. *Journal of Pragmatics* 12: 55–62.

Schegloff, Emanuel. 1999. Discourse, pragmatics, conversation, analysis. *Discourse Studies* 1: 405–435.

Scott, Mike. 2010. *WordSmith tools version 5*. Liverpool: Lexical Analysis Software.

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

Searle, J.R. 1976. A classification of illocutionary acts. *Language in Society* 5(1): 1–23.

Stolcke, Andreas, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, and Dan Jurafsky. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics* 26(3): 339–371.

Stubbs, Michael. 1983. Can I have that in writing, please? Some neglected topics in speech act theory. *Journal of Pragmatics* 7: 479–494.

Styler, Will. 2011. *The EnronSent corpus*. Boulder: University of Colorado at Boulder Institute of Cognitive Science.

van Rees, M.A. 1992. The adequacy of speech act theory for explaining conversational phenomena: A response to some conversation analytical critics. *Journal of Pragmatics* 17: 31–47.

Part III
New Methodologies for the Pragmatic Analysis of Speech Through Corpora

Can English Provide a Framework for Spanish Response Tokens?

Carolina P. Amador-Moreno, Michael McCarthy, and Anne O'Keeffe

1 Introduction

In this chapter we explore a common feature of spoken interaction involving 'small' words (exemplified by words such as *vale, claro* and *bien* in Spanish and *right*, *good* and *okay* in English), which we refer to as *response tokens*. Our aim is to investigate the roles such items play in everyday conversation. We adapt the general corpus methodology used by McCarthy (2002) to illustrate the forms and occurrences of response tokens in British English, along with functional categories and their interpretation as elaborated by O'Keeffe and Adolphs (2008). We apply these English-language-derived frameworks to a corpus of Peninsular Spanish conversation. We examine the Spanish evidence at both the formal and pragmatic levels. We conclude that the frameworks developed for English are sufficiently robust to transfer to Spanish, albeit with certain caveats arising from linguistic and cultural differences between the two languages. Finally, we argue that the use of response tokens is an essential element in being an active and engaged listener in conversation and that they make a significant contribution to fluency.

C.P. Amador-Moreno (✉)
Department of English, Facultad de Filosofía y Letras, Universidad de Extremadura,
Avda. Universidad s/n, 10071 Cáceres, Spain
e-mail: camador@unex.es

M. McCarthy
School of English Studies, University of Nottingham, University Park,
NG7 2RD, United Kingdom
e-mail: JeanneMike11@aol.com

A. O'Keeffe
Department of English Language and Literature, Mary Immaculate College,
University of Limerick, South Circular Road, Limerick, Ireland
e-mail: anne.okeeffe@mic.ul.ie

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_9,

To illustrate our arguments we investigate speaker turns which consist entirely, or mostly, of the tokens under scrutiny, where the speaker is engaged in responding verbally but without taking the floor. Extract (1), from the British English CANCODE corpus (see Sect. 3 below), exemplifies the domain of the present investigation (relevant items in bold):

(1) [Speakers are talking about a new version of a computer operating system]
S1: And it's got rid of some of the bugs so it won't crash.
S2: It's not much different. It's not much different.
S3: **Right. Really?**
S2: But it looks more like a website.
S3: **Right.**

We put the word *small* in quotes above for several reasons: (a) some, but not all, of the items we discuss in this chapter are 'small' morphologically, in that they are monosyllabic (e.g. *good, right, fine*), (b) such words form part of the high-frequency vocabulary of English and, as such, often go unnoticed and remain on the subliminal level of native-speaker consciousness just as common items such as discourse markers have been shown to do (Watts 1989), and (c) although many of the items we examine in this article are 'small', we hope to demonstrate that they have 'big' meanings on the interactive plane of discourse. In this last respect we concur with the stance taken by John Sinclair in relation to high-frequency items as expressed in his plenary address to the American Association for Applied Linguistics annual conference in 2006, the title of which was *Small words make big meanings.*[1]

Informal, casual conversations typically contain response tokens in great number, since participants will often find themselves in the recipient role where they may not wish to assume the floor, or where it may be inappropriate to do so, for example in the midst of a personal anecdote or other report delivered by another speaker, or during the reception of important, extended information. Responses may be simply expressions of body-language (e.g. head-nods, eyebrow-raising) or what have commonly been referred to as backchannel responses, in English typically realised by items such a *mm, uhum, yeah, yes, no* (see Yngve 1970 and further works reviewed below in Sect. 2). Kendon (1967) suggests that speakers rely upon such feedback for guidance as to how the message is being received, while Tottie, (1991: 255) states that such tokens 'grease the wheels of the conversation but constitute no claim to take over the turn'. Here we refer to these listener contributions as indices of *listenership,*[2] and we suggest that good listenership is an aspect of fluency, especially where fluency is considered in relation to the collaborative

[1]The notion of 'small' words having important meanings in interaction is also captured in the title of a paper on the present topic by McCarthy (2003), and a book on oral assessment by Hasselgreen (2005).

[2]We do not claim to have invented the term, which is used by Tannen (1984) to refer to engaged participation in conversation. We adapt the term to our present needs in order to create a distinction between engaged, active participation and 'listening comprehension', which has traditionally focused on message-processing skills.

production of conversation by all participants (McCarthy 2010). We offer the term *confluence* to refer to this way of looking at fluency, and discuss it briefly in Sect. 7. Good listenership is effected without floor-grabbing, through a set of small, but non-minimal, lexical tokens in English and Spanish. By investigating Spanish response items in the present chapter, we especially wish to test the transferability of frameworks designed for English data and to extend the debate to cross-linguistic issues.

2 Previous Studies of Response

2.1 The Back Channel and Beyond

Fries (1952) provides us with an important early study of listener responses in telephone calls. He looked at a range of responses from vocalisations such as *unh* and *hunh*, to the use of *yes* and lexical words and phrases such as *good* and *I see* (Fries 1952: 49). But it was Yngve (1970) who introduced the notion of the 'back channel', which has since become a standard term for short, non-floor-grabbing responses. Yngve looked at items such as *uh-huh*, *yes, okay*, and brief comments (e.g. *Oh, I can believe it*). However, what researchers have included within the notion of back channel in subsequent research has varied greatly from study to study.

A wide variety of communicative behaviour on the part of the speaker, from body language to changes in phonological pitch, pauses, opportunities for syntactic completion, and fully finished turns (e.g. questions, statements with low pitch termination, etc.) may offer the listener a chance to jump in and respond in some way (see Duncan 1972, 1974; Jucker 1986). Especially where the listener's contribution is very brief (often just one or two words), it is often impossible to judge whether the utterance is just backchannel feedback signalling no desire to take the floor, or whether such utterances should be classed as turns which shift the identity of 'current speaker'. As a result, much of the literature on backchannel behaviour has been unable definitively to provide exact and replicable criteria for judging the status of listeners' contributions. Duncan and Niederehe (1974) acknowledge the imprecision of the boundary between brief utterances and proper turns, while accepting the notion that backchannel utterances create an understanding between speaker and listener that the turn has not been yielded. The wide range of options that listeners may exploit, from body language to non-turn-yielding comments, probably explains why the more easily identifiable, non-word vocalisations have become the focus of more extensive research than lexically- or lexico-grammatically-based responses.

Researchers have, over time, expanded the description of response. Duncan (1974) broadened the debate to embrace items such as *right* and *I see*, and included sentence completions, requests for clarification and brief restatements. Öreström (1983) observed features of backchannel response such as degree of overlap with the main speaker's turn and loudness. Öreström also extended the range of items to include lexical tokens such as *quite* and *good*.

Tottie (1991) investigated backchannel phenomena in British and American English corpus data, and placed vocalisations such as *mm, mhm* and *uh-(h)uh*, alongside 'bona fide words and phrases' (Tottie 1991: 255). Tottie also noted cases where an utterance is very short, in the characteristic manner of backchannel feedback, but is responded to by the interlocutor, suggesting that such utterances could be seen as full turns.

Gardner (1997, 1998, 2002) defines backchannels as 'the vocalisation of understandings' and places them as existing 'between speaking and listening' (both quotations from the title of his 1998 paper). Gardner (1997) looks at 'minimal responses', for example *mm-hm*, which he refers to as a 'continuer', encouraging the speaker to go on (see also Schegloff 1982), alongside the 'stronger, more aligning/agreeing' *yeah*. Gardner (1998) classifies listener responses into backchannel items such as acknowledgements and continuers (e.g. *yeah, mm-hm*), newsmarking items (e.g. *oh, really*), evaluative items (e.g. *wow, how terrible*), and clarification requests. The different functions of seemingly similar vocal responses such as *um* and *uh* have been teased out by Clark and Fox Tree (2002).

Stubbe (1998) refers to 'supportive verbal feedback' in the title of her paper, distinguishing between neutral tokens (e.g. *mm, uhuh*) and supportive tokens (e.g. *oh gosh*). Stubbe's study is concerned with cross-cultural issues, and the repudiation of negative evaluations and stereotypes which may arise from differences in realisations of listener feedback across different cultures (see also Holmes and Stubbe 1997, which adds a gender dimension to the study of differences in listener behavior). As in the other works reviewed here, the key point is the acceptance that lexically-based responses need not be turn-grabbing but can be seen as an aspect of listener behaviour.

2.2 Exchange Structure and Adjacency

Sinclair and Coulthard (1975) collected first-language spoken classroom data which led to the construction of a rank-scale for spoken exchanges based on the core notion of speaker 'moves'. The classroom exchanges they observed consisted of *initiating moves* (utterances structurally independent from previous turns), and *answering* or *responding* moves by the recipients of initiating moves (1975: 26–27; see also Sinclair and Brazil 1982: 49). There was also a third move in the classroom data, the *follow-up*, by which teachers acknowledged and evaluated the responding moves of their pupils. Such patterns have been given the shorthand label of *IRF* exchanges, and the R-move is clearly of relevance to the current study. We shall also argue for including the F-move in the notion of response, and comment on the two types further, below, in Sect. 4.

The conversation analysis (CA) literature on adjacency pairs (see Schegloff 2006) has included a focus on 'assessments' (evaluations of people and other entities), and has provided data on how listeners respond to evaluations (see in

particular Pomerantz 1984). Antaki (2000, 2002) investigates what he calls 'high-grade assessments' (emphatic variations of some of the lexical response items we examine here) in recipients' responsive moves. Likewise, Stenström (1990) discusses features which partly overlap with the present study, and other CA researchers have also examined the role of 'third-turn receipts' (a parallel term for the F-move in the IRF sequence; see for example Heritage 1985).

Research into how listeners behave has given strength to the notion of conversation as a joint enterprise, what Erickson (1986: 295) refers to as the 'relationship of intertexuality between speaking and listening'. The notion of listenership in the present paper similarly stresses the jointly occupied territory between speaking and listening. Research into verbal and non-verbal behaviour on the part of listeners (e.g. Goodwin 1981) underscores how listeners respond at appropriate points and in appropriate ways, and also how speakers respond to such verbal and non-verbal feedback and adjust their talk as a result. Duranti (1986) also states the importance of examining how speakers are responded to by their interlocutors, while Erickson (1986) views listening as 'an activity of communicative production as well as one of reception' (Erickson 1986: 297). Erickson and Shultz (1982), in a study of interview data, refer to moments of listening-response relevance (LRRM), after which a speaker may persist with the same point or make a new one. An LRRM is a primary opportunity for the listener to respond, and the role of responses in enabling the discourse to proceed smoothly is seen as crucial. Similarly, oral narrative (see Goodwin 1986) has provided illuminating data for how listeners engage actively to express something more than just 'hearership' (Goodwin 1981: 103). Studies of the joint activity of speakers and listeners point to the importance of listener response, and the ongoing and shifting effects of responses on the way speakers continue their turns (see Bublitz 1988; McGregor and White 1990). Schegloff (1982) states that to neglect the listener and to focus only on the main speaker results in the misleading characterisation of the discourse as 'a single speaker's, and a single mind's, product' (Schegloff 1982: 74). The notion of listenership in the present paper embraces the view of conversation as joint production; good listeners not only acknowledge talk, they offer non-floor-grabbing increments which enable the discourse to flow onwards in a manner satisfactory to all participants.

Within the CA tradition, the turn-taking system is central, and responses are understood as elements of turn construction, allocation and sequencing. Schegloff (1982) posits that the system is fundamentally designed to 'minimize turn size' (Schegloff 1982: 73). There is an inbuilt economy: speakers say no more than what is essential. Above we mentioned the 'smallness' of the response tokens which are our present concern, and this would seem, on the face of it, to support a notion of economy (in that we often find monosyllabic turns). However, it is the additional matter, over and above the bare acknowledgements of vocalisations and *yes/yeah* and *no* which interests us here, and that extra matter is where the interactional engagement takes place and listenership is most clearly displayed. For many of the turns examined in this paper, a simple *yes* would suffice to acknowledge receipt and understanding, and yet listeners so often 'do more', using tokens such as *right, fine,*

vale, ya, venga, claro, anda, which indicate encouragement, engagement, involvement, empathy, enthusiasm, topic-management and a range of other reactions. Communicative 'economy', therefore, seems to be governed by both the propositional and interactional elements of discourse: neither aspect can be threatened or sacrificed simply to keep one's contribution brief. Speakers seem unwilling to economise as regards sociability, human engagement and conversational flow, except in the most pressing circumstances where a purely transactional response suffices. Schegloff (ibid.) observes, interestingly, that repetitive use of the same response item by the same listener over an extended portion of talk may risk being heard as a sign of boredom or inattention; thus listeners normally vary their responses to obviate such risks of misinterpretation. Nonetheless, tight sequences of repeated items could also be heard as a sign of enthusiasm or encouragement; it is only within local contexts that plausible interpretations of the affective intentions of listeners' contributions can be properly assessed and inferred.

Tao (2003), using a spoken corpus, investigated turn-initial items in an attempt to measure their contribution to a turn-construction grammar. Tao regarded as particularly important how speakers start their turns. Turn-initial elements in English, Tao concluded, tend to be syntactically independent items and are mostly lexical. He found that, at the turn-initial slot, items such as *yes, well, right, okay* and pronouns introducing fixed expressions such as *I think, you know, I mean, that's* + adjective (*that's right, that's true*) were dominant. Tao demonstrated clearly how interlocutors attend to the prior turn before they turn to their own transactional concerns, with the turn-initial items being responsible for the creation of much of the interactional side of the communication (see also McCarthy 2010).

2.3 Response Tokens in Spanish: Discourse- and Pragmatic-Marking

In research into spoken Spanish, response tokens have been studied under the umbrella of discourse markers or pragmatic markers (in English too, see Brinton 1996). In recent years, the debate over what counts as a discourse marker has been robust, and mostly emanates from Schiffrin's (1987) seminal work, which is firmly grounded in the contribution of marker-items such as *well*, *oh* and *you know* to a theory of discourse coherence. Fraser (1999) takes a circumscribed view of discourse markers, locating them principally within word- and phrase-classes such as conjunctions, adverbs and prepositional phrases that serve linking functions. However, in his (1996) paper on pragmatic markers, a broader picture is presented which includes sentence adverbials that show stance such as *certainly,* and *frankly*. This broader view of markers encompassing pragmatic acts is that taken by Carter and McCarthy (2006) and Carter et al. (2011: 175), where markers are seen to include response tokens of the type that are the focus of the present study. However, the distinction between markers functioning to contribute to

discourse coherence and markers showing pragmatic stance remains somewhat fuzzy. This is understandable, as a short response (e.g. *right, okay*) may be simultaneously functioning as a non-floor-grabbing backchannel item signalling engagement and as a boundary marker of some sort (see McCarthy 2003 for a discussion of responses at pre-closing and topic-transitional points). Thus, in discussing the autonomy of discourse markers in the context of Spanish, Martín Zorraquino and Portolés (1999) observe that 'ciertos marcadores del discurso – sobre todo, aquellos que denominamos conversacionales (*bien, bueno, hombre*, etc.) – aparecen frecuentemente solos en un turno de palabra' (Martín Zorraquino and Portolés 1999: 4068). This point may be illustrated with the following example from the Spanish COREC (Corpus Oral de Referencia del Español Contemporáneo) corpus used in the present paper (see below):

(2) [From a telephone conversation:]
S1: A qué hora vendrás a comer?
S2: Pues a las tres.
S1: Sobre las tres?
S2: Sí.
S1: **Vale**.
S2: Hasta luego.
S1: Hasta luego, hijo.

[S1: What time will you come for lunch?
S2: I'd say … at three
S1: Around three?
S2: Yes
S1: Right/Okay
S2: See you later
S1: See you later, son.]

Vale is an example of what we call a response token, but likewise here it shares some of the characteristics normally attributed to discourse markers, signaling, in this case, (pre-) closure. Like discourse markers, response tokens may be syntactically optional while nonetheless important from a pragmatic perspective: they are seen as responsive signals and are also a means to achieve conversational continuity and flow. Another parallel is that, without them, the conversation may be grammatically well-formed but will often appear unnatural, dysfluent, sometimes even impolite or unfriendly, epithets often attributed to 'non-nativeness' at a communicative level, and there is always a risk that their absence may result in communicative breakdown or (cross-cultural) 'pragmatic failure' (see Thomas 1983). In the present paper, we make no necessary distinction between response tokens which occur *in medias res* and those which mark boundaries or display other discourse-marking characteristics, but accept their potential for multi-functionality, and comment on this phenomenon where appropriate.

2.4 Research Across Languages and Varieties of Languages

A small, but growing body of comparative research into response tokens exists. A common thread of these studies is that while response tokens have counterparts in other languages, they do not always display direct correlations or transferability. Sorjonen (2001) looks at two responses particles in Finnish, *nii*(n) and *joo*, which in some usages have *yeah* and *yes* as their closest English counterparts. She identifies a number of sequential and contextual uses of these forms, including their use as answers to *yes-no* questions and directives, as responses to a stance-taking by the prior speaker, and during an extended storytelling by the co-participant. She also points to a fine-grained division of how the forms function. She relates this to the epistemic and affective character of the talk and the continuation versus closure-relevance of the activity.

Clancy et al. (1996) look at response tokens in three languages, Mandarin Chinese, English and Japanese. They use the term 'reactive tokens' which seems to equate to 'response token'. They define reactive tokens as 'short utterance[s] produced by an interlocutor who is playing a listener's role during the other interlocutor's speakership … [they] will normally not disrupt the primary speaker's speakership, and do not in themselves claim the floor' (Clancy et al. 1996: 355). They draw on corpora of conversations from each of the three languages and distinguish among several types of reactive tokens: (1) backchannels which in all three languages manifest as non-lexical vocalisations; these carry a 'continuer' function (after Schegloff 1982) and display interest and 'claim of understanding'; (2) reactive expressions which are short, non-grabbing lexical phrases or words (including assessments, Goodwin 1986) uttered by the non-primary speaker. Examples of these in the three languages include *oh really, really,* in English, *sugoi* in Japanese, meaning approximately *great/terrible*, and *dui* in Mandarin, meaning approximately *right*; (3) collaborative finishes, when the non-primary speaker finishes the previous speaker's utterance (see Lerner 1989); (4) repetitions where the non-primary speaker repeats a portion of what the primary speaker has said.

In another contrastive study, Tao and Thompson (1991) look at response tokens in the conversations of Mandarin speakers in Mandarin and in English. They find that, counter to most studies of interference of first language on second language, there is evidence to suggest interference in the opposite direction.

Variation is also found within languages. In an intra-varietal study, O'Keeffe and Adolphs (2008), compare response tokens in British and Irish English. Their findings bring to light a number of points of difference between these two geographically close varieties. Even within a common language, they found variation in the distribution of response tokens. The British English speakers used more response tokens than the Irish English speakers. British speakers were also found to use a broader range of forms. McCarthy (2002) noted a broad range of forms in the British English single-word range that also occur in North American English, but with different frequencies (*right, absolutely, sure, good, lovely, exactly, great, definitely, true, really*). In contrast, the Irish single word forms only have *really, sure* and *right* in common with McCarthy's findings for American English.

2.5 Corpora and the Functions of English Response Tokens

While the present paper deals with Spanish, frameworks derived from studies of English provide a useful benchmark for comparison. O'Keeffe and Adolphs (2008) undertook an analysis of response tokens in British and Irish English, using data from The Cambridge and Nottingham Corpus of Discourse in English (CANCODE), a five million word corpus of spoken British English (McCarthy 1998) and the Limerick Corpus of Irish English (LCIE), one million words of spoken Irish English (Farr et al. 2004). From these, they sampled two sub-corpora of 20,000 words each, consisting of recordings of conversations of young women around 20 years of age. They analysed each response token in the 40,000 words of data and compared their forms and functions. For the purposes of this paper, their findings in terms of the functions of the response tokens will provide the backdrop for our analysis of Spanish. The functions identified across both sub-corpora are summarised as follows:

Continuer responses: These are facilitative in that they maintain the flow of talk. They encourage the current speaker to continue. As noted by Schegloff (1982), Maynard (1989) and Gardner (1997, 1998, 2002), this function is typically realised by a minimal response token, such as *mm*.

Convergence responses: Some response tokens (e.g. *exactly, no*) were frequently found at points of convergence in conversations, that is, where participants agree, or simply converge on opinions or mundane topics facilitating the negotiation of topic collaboratively, so that topic can be shifted or changed. Convergence can also be followed by a conversational closure point.

Engagement responses: These function at an affective level, signalling the addressee's enthusiasm, empathy, surprise, shock etc. at what the speaker is saying, without grabbing the turn. They are typically non-minimal and English items include *brilliant, absolutely, wow, cool, gosh, really* and short phrases, such as *that's tough, that's true, you're not serious, Is that so?*

Information receipt tokens: A small number of response tokens in both datasets did not fit any of the above categories. These seemed to have an organisational function and were usually marked by falling pitch. In the few examples that were found, they seemed to serve a global discourse-marking function (cf. Lenk 1998) within the orientation stage of narratives.

McCarthy (2003) noted that some response tokens are strongly associated with particular contexts. *Fine*, he suggests, most typically occurs in making arrangements and reaching decisions and *certainly* most typically occurs in reply to a request for a service or favour. He also notes that adjectives such as *excellent, fine, great, good, lovely, right*, *perfect* offer positive feedback to the speaker and often mark the boundaries of topics, where speakers express their satisfaction with phases of business such as making arrangements, agreeing on courses of action, and marking the satisfactory exchange of information, goods and services.

3 Data for the Present Study

The focus of this study is on the application of English-language corpus-based frameworks to spoken Spanish, and to this end, we used the Corpus Oral de Referencia del Español Contemporáneo (henceforth referred to as COREC[3]), a corpus of Peninsular Spanish containing 1,100,000 transcribed words which was compiled at the Universidad Autónoma de Madrid in the early 1990s (http://www.lllf.uam.es/corpus/corpus_lee.html). We concentrate on the conversation sub-corpus of COREC, which contains 211,632 running words in total.

4 Analysis

4.1 Identifying Response Tokens

Response tokens are often divided into *minimal* and *non-minimal* tokens, though the distinction is not entirely clear-cut. It is also worth noting that spoken corpora, for the most part, have been based on transcriptions of audio recordings only and usually fail to capture non-verbal responses such as head nods and shoulder shrugs.[4] Usually, minimal responses in English are defined as interjections (for example *yeah, okay*) or non-word vocalisations (such as *mm, umhum*), while non-minimal response tokens are mostly (morphologically speaking) adverbs or adjectives, for example *good, really great, absolutely*, or short phrases/minimal clauses, such as *is that so? by all means, fair enough, that's true, not at all*.

In the present paper we focus only on non-minimal response tokens and restrict our analysis to lexical items taken from the major word-classes. We disregard minimal tokens such as *yes, sí, no, okay*, and vocalisations such as *ah, oh, uhum, ay, oy*. These types of responses are typically already well-covered in the literature on backchannelling (e.g. Drummond and Hopper 1993).

In identifying response tokens, position in the exchange or adjacency pair is obviously important. However, in terms of the exchange structure model proposed by Sinclair and Coulthard (1975), most conversational exchanges consisted of the three moves referred to above (the IRF pattern). Non-classroom conversation requires a slightly different perspective. A typical three-move conversational exchange is illustrated in extract (4), from the British data:

[3]We are grateful to Francisco Marcos Marín for granting us permission to quote from the corpus.

[4]However, recent multi-media corpus projects may be able to obviate this problem by the use of synchronised video records alongside the conversational transcript, see, for example, Knight et al. (2009).

(3) [Speaker 1 is confirming that he will fax something to the listener]

S1:	I'll send that to you in the morning when I've confirmed where it's going.	Initiation
S2:	**Fine. Yeah. Yep that's okay**.	Response
S1:	**Okay**.	Follow-up

S1's *okay* is itself a response to S2's response *Fine. Yeah. Yep that's okay*. Moreover, in multi-party conversation, more than one participant may construct the exchange, making the distinction between response moves and follow-up moves even less obvious:

(4)
S1: We bought a rare record. It's supposed to be worth five hundred pound isn't it.
S2: **Right**.
S3: **Really**?
S2: **Yeah**.
S1: Freddie Mercury when he first started under the name of Larry Lurex.

For this reason, we eschew the distinction between response and follow-up in the present paper and treat all the items in bold in (4) as response tokens.

English non-minimal response tokens can occur in pairs or clusters. Carter and McCarthy (2006: 190–191) note that clustering is particularly evident when a topic is being closed down or at a boundary in the talk when another topic is introduced. Such pairs function to signal a boundary *and* interactive convergence, or else simply to express friendly social support. Occasionally, triple response tokens occur, as in extract (5):

(5)
[Couple asking permission to look at a disused railway line]
S1: It went through, it goes through. Straight, straight on.
S2: **Right. Wonderful. Great**. Can we look round then?
S1: Yes certainly.
S2: Thank you.

In summary, the examples above show that single-word non-minimal response tokens in English may be (morphologically-speaking) adverbs or adjectives, they may occur in clusters or be reduplicated. They may occupy the whole turn, or begin a turn which consists of a small amount of further conversational matter.

4.2 Analysis of the Spanish Data

The present study follows McCarthy's (2002) procedure and applies it to the Spanish data. McCarthy took frequency lists of the British and American corpora he used and scrutinised them manually. The most likely items for consideration as response tokens (based on previous studies and on observation and intuition) were then extracted from the frequency lists. At least 100 occurrences in each corpus was

Table 1 Response tokens in the Spanish data

Item	Frequency	Occurrences as response token	Normalised total (per 1 m words)
vale	270	28	132
claro	911	26	123
anda	99	18	85
joé	39	14	66
exactamente	48	13	61
venga	167	12	57
fíjate	106	12	57
madre	191	10	47
hombre	318	8	38
jolín	26	8	38
hostia	45	7	33
joder	36	7	33
ya	1,575	4	19
bueno	1,241	4	19
bien	617	4	19
vaya	90	4	19
¿ves?	120	3	14

set as the level below which items would be excluded from consideration. Once the initial list was established, a maximum of 1,000 concordance lines from each corpus were isolated for each item in the list (via the random sampling option in the analytical software). These concordance lines were examined to see how many of them actually showed the particular word functioning as a response token. The total number of occurrences of the word used as a response item was then listed and presented as part of the word's lexical profile. McCarthy then discussed various functional aspects of their use.

Based on the same methodology, for the purposes of the present paper, a word-frequency list was generated for the Spanish corpus, using *Wordsmith Tools* (Scott 2011). In this case, the 800 most frequent words were then gleaned manually, and those words considered as potential candidates for response tokens were selected. The same procedure as for the English data was then followed, with concordance lines scrutinised for actual occurrences as response tokens. As Table 1 shows, the Spanish list contains words that fall under different morphological categories (nouns, adjectives, adverbs, verb forms including the imperative, the subjunctive, etc.). Each item was analysed in the context of the conversation in which it appears, which allowed us to single out the instances that functioned as response tokens. The first numerical column shows the total frequency of occurrence of the item; the second column indicates the number of occurrences which function as single-word response tokens for each individual item, and the third column gives the normalised totals per million words.

In the Spanish forms, we see a broad range of items which function as response tokens.[5]

McCarthy's English list consisted entirely of items morphologically classified as adjectives, adverbs, or (in the case of *gosh* and *wow*) interjections. Morphological types in the Spanish data cover adjectives, adverbs and interjections, but also nouns (*madre, hombre*) and verbs (e.g. *vale, vaya, ¿ves?*). Verb responses do occur in English, but they tend to be phrasal, for example *Go on!* and *Get away!*, and since the lists under discussion here are only of single-word items, this potential comparison is put to one side. English does also have noun responses, but principally in the religious and scatological domains (*God, shit*, etc.). While *ya*, *bueno* and *claro* have the highest overall frequencies in the Spanish corpus, *vale* has the largest individual number of occurrences as a response token, closely followed by *claro. Vale* often translates appropriately into English as *right*, which was also the most frequent item in the English list, so there is a neat symmetry in this case. *Vale* is used widely in casual conversation,[6] and, although it used to be a clear sociolinguistic marker of age (i.e. it was initially mostly heard among teenagers) its use is nowadays widespread in Peninsular Spanish, and it is employed by speakers of all ages, as can be observed in the following interaction between S2 (a father) and S1 (his daughter):

(6)
S1: Bueno papá, ¿te sientas ahí mismo?
S2: Aquí, <simultáneo>bueno.
S1: Sí </simultáneo>.
S2: **Vale**.

[S1: Ok, dad, are you sitting there?
S2: Here, <overlapping>Ok.
S1: Yes, <overlapping>
S2: Right.]

Claro can also often translate as *right* (typically with rise-fall intonation), and it is possible that *vale* and *claro*, taken together, occupy a similar pragmatic space to that of *right* in English. Other possible overlaps occur between *right* and *ya*, while

[5]Unlike McCarthy's earlier study, where taboo or religious expressions were deliberately excluded, we include them here in the Spanish list. *Joe* and *jolin* are euphemistic forms of the taboo *joder* (fuck). *Hostia* is a religious reference, which is not translatable into English. In a literal sense, it means *host,* the unleavened bread used in the Catholic mass to represent the body of Christ. While we cannot compare religious references and taboo words in this study, another study which uses CANCODE data and compares it with Irish English response tokens points to more frequent use of religious reference in the Irish data compared with the British data (e.g. *Oh my God, God help us, Jesus, Jesus Christ*), see O'Keeffe and Adolphs (2008). The authors note that religious references are found more in Catholic and post-Catholic contexts such as Ireland, and in this case Spain, where these words' potency as swear words has greater cultural relevance (see Andersson and Trudgill 1990).

[6]For functions in classroom contexts see Amador Moreno et al. (2006).

fine, in its typical use of signaling agreed decisions and arrangements, may overlap with *vale* and *bien*.

A number of the Spanish items, like the English ones, are exclamatives (e.g. *anda*, *vaya*, *hombre, hostia*), expressing affective reactions, and may translate variously as English *wow, gosh, really*, depending on context, though the more taboo-related expressions in Spanish will also have taboo-related equivalents deliberately excluded from McCarthy's original analyses (items such as *God, Christ, shit*, etc.). The precise delineation of pragmatic coverage of the various items, in the final analysis, can only be achieved by examination of their occurrences in context. It is thus to the contextual functions of the Spanish items that we now turn.

5 Functions of Spanish Response Tokens

5.1 Convergence

Functionally, most of the Spanish forms signal convergence, but when we examine them closely we find that there are subtle variations. *Claro,* for example, marks agreement, as an alternative to *sí*, in contrast with which *claro* implies cooperation between speakers. *Claro* reinforces the interlocutor's view, suggesting that no other position than that taken by their co-conversationalist would be possible. It emphasizes solidarity and convergence. Example (7) illustrates this:

(7)

[Speakers are trying to find a space in a car park]
S1: Ya está. Madre mía, se nos ha aparecido la Virgen.
S2: Pues sí. Ha habido suerte.
S1: Es que ha querido quitármelo pero no ha podido.
S2: Ya, ya lo sé. Porque no te has ido.
S1: Es que … no.
S2: Si llegas a ser un poco más blando y te vas.
S1: Sí. No y además es que él no puede aparcar tal y como está y yo sí.
S2: **Claro.**

[S1: That's it. My God, we've been blessed by luck (lit. the Mother of God has appeared to us)
S2: Yes, we were lucky
S1: He tried to take it before me, but he couldn't
S2: Yes, I know. Because you stayed there
S1: Well … no
S2: If you had been softer and went…
S1: Yes. No, and besides, he can't park the way he's facing and I can
S2: You're right.]

Hombre also marks convergence, in a friendly, informal way, projecting a close relationship between speakers:

(8)
S1: También depende de cómo sea la otra persona, ¿no?<simultáneo>El carácter … y todo eso, ¿no?
S2: Sí, sí, sí.
S3: **¡Hombre!**

[S1: It also depends on what the other person is like, doesn't it?<overlapping>Their personality and all that, doesn't it?
S2: Yes, yes, yes.
S3: Absolutely.]

The difference between positive and negative response is sometimes reflected prosodically: with the appropriate intonation, *hombre* can indicate divergence and distancing, as can be observed in example 9.[7] Here, S2's disagreement with S1 is made evident later on, but it is signalled first by the use of *hombre*:

(9)
S1: Es que … es que lo de menos es el dinero, en Harvard
S2: **Hombre…**
S1: <ininteligible>cualquier universidad. Y si no te admiten, por muchos millones<ininteligible>
S2: No; estás equivocada, mamá. Con mucho dinero…
S1: No, (eso es así)

[S1: Well, the least important thing is money in Harvard
S2: Well…
S1: <unintelligible> any university. And if you don't get accepted, regardless of the millions<unintelligible>
S2: No, you're wrong, mum. With a lot of money…
S1: No, that is the way it works.]

As Martín Zorraquino and Portolés point out, "Con *hombre* el hablante atenúa, en las intervenciones reactivas, la expresión de la disconformidad con lo dicho por el oyente e incluso introduce efectos paliativos para calmar su posible enfado" (1999: 4173–4174).

Venga is also used to indicate convergence, as can be seen in example 18. Note how it co-occurs with *vale* in S3's turn, to reinforce the agreement expressed by *venga* (see Sect. 5.5 for more on how items cluster):

(10)
S1: No, bajamos aquí<simultáneo>y yo me voy a aparcar.
S2: **Venga.**

[7] Rising intonation, by contrast, tends to indicate agreement.

S3: **Venga, vale.** <ruido = aullidos de perro>
S2: Descargamos maletas. ¡Quieto, tín! <ruido = ladridos de perro>

[S1: No, we'll get off here and I'll go and park
S2: Ok
S3: Yes, Ok <noise = dog whining>
S2: We'll take out the suitcases. Stay, Tin! <noise = dog barking>]

5.2 Partial/Modified Convergence

Ya and *bueno* have a similar function to the tokens exemplified in Sect. 5.1. They both indicate convergence. However, some of the nuances expressed by them are worthy of mention here. *Ya*, compared with *claro*, for example, indicates a more neutral type of response, to the extent, sometimes, of suggesting a lack of engagement or even disinterest:

(11)
S1: Es que el Chiqui cambia totalmente de casa a estar en el colegio. O sea, en casa le verás revoltoso, le verás que se pega con sus hermanos.
S2: Sí.
S1: Pero en cuanto que sale de la puerta para ir al colegio … o sea, cambia totalmente.
S2: <fático = afirmación>
S1: Digo: "no puede ser". O sea, si no le ve que está sentado en, en la silla, no sabe que hay niño.
S2: Parece que no está.
S1: Entonces a mí no me gusta eso tampoco, Tere, ¿entiendes?
S2: **Ya**.

[S1: Well, the kid behaves completely different at home compared to school. I mean, at home he's hyper, you'd see him fighting with his siblings.
S2: Yes.
S1: But as soon as he goes out the door to go to school … I mean, he changes completely
S2: <phatic = agreement>
S1: And I say: "this can't be". I mean, if you don't see him sitting on the chair, you wouldn't know there's a child there
S2: It's as if he wasn't there
S1: So, I don't like that either, Tere, do you know what I mean?
S2: Yes.]

Note that, if we replace *ya* with *claro* in example (11), S2 seems to show more engagement and greater convergence in the conversation, whereas in the original

version, S2 is simply letting S1 speak. *Ya* can also express other nuances in context, such as irony or incredulity, as in example (12):

(12)
S1: '…' me han dicho <silencio> que está muy difícil. Gente que lleva aquí <silencio> varios años en Madrid y les cuesta mucho trabajo, o sea que…
S2: Sí, no es fácil.
S1: Pero yo te puedo indicar más o menos dos o tres caminos por los que a lo mejor puedes tener suerte.
S2: **Ya.**
S1: Que eso siempre es mejor que nada.

[S1: '…' apparently <pause> it's very difficult. People who have been here <pause> in Madrid for a few years even find it difficult, I mean…
S2: Yes, it's not easy
S1: But I can give you two or three pointers that might help you get lucky.
S2: Oh yeah.
S1: That's always better than nothing.]

Apart from indicating agreement, *bueno* also functions to mitigate those cases when agreement is preceded by disagreement, or when the speaker is trying to avoid giving a more direct answer, as in example (13) (a telephone conversation between a mother and her daughter):

(13)
S1: Pero lo que tienes que hacer es venir aquí.
S2: **Bueno**.
S1: Sí.
S2: Iré para allá. '…' Hoy voy a ir con Papá a … a una exposición y eso.
S1: No; hoy yo no puedo, que tengo que dar un … una charla en alemán.
S2: ¿Sí?
S1: En mi clase. Sí, que es el último día del curso ya.
S2: **Bueno.**

[S1: But what you should do is come here
S2: Ok
S1: Yes
S2: I'll go over '…' Today I have to go with dad to an exhibition and that
S1: No; I can't today, because I have to give a talk in German
S2: Really?
S1: In class. Yes, it's the last day already
S2: Ok.]

In comparison with *claro*, *bueno* is a less enthusiastic way of converging; it shows a lower degree of conviction. As Bauhr (1994: 92 ff.) points out, '[bueno] se utiliza a menudo en situaciones en que el hablante cede ante la insistencia de su

interlocutor o acepta una propuesta, invitación, etc., con desgana; de ahí que su utilización en los contextos en los que podría competir con expresiones alternativas como *sí, claro, muchas gracias* y *con mucho gusto* pueda tacharse de renuente o poco cortés'.

Another token used to indicate convergence is *bien*. Fuentes Rodríguez (1993), whose analysis is based on a corpus of Spanish spoken in Sevilla, looks at the use of *bueno, bien* and *pues bien*. As she indicates, *bien* has a phatic function, and it can be used to convey happiness or annoyance. In comparison to *bueno*, *bien* can be perceived as being a more distancing response, as can be seen in example (14), where speakers 1 and 2 are discussing flat-hunting. The use of *ya* here also indicates that 1 is not fully convinced by 2's advice:

(14)
S1: Ese precio estamos pensando, '…' setenta mil pesetas. '…'
S2: Pues … hay una zona, en el norte de Madrid, en Alcobendas…
S1: ¿Perdón?
S2: Alcobendas<ininteligible>'…' Normalmente, la gente que yo conozco que ha encontrado piso, ha sido gracias a carteles, que ha puesto él mismo.
S1: **Bien.**
S2: Entonces, el … en la Universidad, en la Complutense, '…' en los tablones de anuncios de todas las … o sea facultades, poner anuncios. Eso … eso puede funcionar.
S1: **Ya**.

[S1: That's the price we were thinking of '…' seventy thousand pesetas '…'
S2: Well…there's an area in the north of Madrid, in Alcobendas
S1: Sorry?
S2: Alcobendas<unintelligible>'…' Normally the people I know who have found a flat, have found it through putting ads themselves
S1: Right.
S2: Then the…in the University, in the Complutense … on the noticeboards of all the … I mean, Schools, putting ads. That … that can work.
S1: I see.]

5.3 Convergence and (Pre-)Closure

Some forms are found in the context of conversational closings. In examples (15) and (16) we see *bueno* and *vale* in preambles to the closing of the conversation:

(15)
S1: Ya verás como no me parezco nada a Teresa. Pero nada, ¿eh? Como una patata a un culo.
S2: <risas>Pero ella misma tampoco se parece en su carné, o sea que…

S1: **Bueno**, pues entonces **vale**.
S2: **Bueno.**
S1: **Bueno.**
S2: **Bueno**, que nada, que voy a seguir estudiando.

[S1: You'll see how I don't look like Teresa at all. Not at all, eh? Like a potato to an arse.
S2: <laughter>But she doesn't even look like the picture on her ID card, I mean…
S1: Well, ok then
S2: Ok
S1: Ok
S2: Ok, that's it. I'm going to do some more studying.]

As can be seen in example (16), *vale* (1) seems to be confirming receipt of information, whereas *vale* (2) is signaling a desired (pre-)closure (see McCarthy 2003) which is then reinforced/confirmed by *hasta luego*.

(16)
S1: Bueno, que no te entiendo. Venga, pues a las siete y media bajo al portal y te espero. ¿Vale?
S2: <fático = duda> **Vale (1)**.
S1: Pues nada, hasta luego.
S2: **Vale (2)**, hasta luego.
S1: Hasta luego. Chao.

[S1: Look, I don't understand what you're saying. Ok, at seven thirty I'll go down to the door and I'll wait, Ok?
S2: <hesitating>Ok
S1: Right. See you later, then
S2: Ok, see you later
S1: See you. Ciao!]

5.4 Engagement

Anda, *vaya*, *madre* and *fíjate* are used to express different degrees of surprise. *Anda* and *fíjate* have in common the fact that they are (being second person singular address forms) addressed to the listener directly. *Fíjate*, apart from showing surprise, implies a certain degree of complicity with the listener:

(17)
S1: Si están muy baratos los viajes en avión.
S2: A Londres está barato ahora.
S1: Sale 17000 pelas ida y vuelta.
S2: **Fíjate**.

[S1: Yes, they are very cheap, flights.
S2: To London is cheap now
S1: It works out at 17000 pesetas
S2: Wow!]

Anda is versatile and can appear on its own, expressing surprise, as in (18) below, where two speakers are discussing celebrities; it can also appear, as will be shown in Sect. 5.5, below, in combination with other response tokens, reinforcing affective responses:

(18)
S1: Por cierto </simultáneo>que el único invitado del que se ha dado el nombre, que va a ir al cumpleaños ahora del dieciocho es eh … Kashogui.
S2: **¡Anda!**

[S1: By the way </overlapping>the only guest whose name has been revealed, who is going to the birthday party on the 18th is er … Kashogui
S2: Go away!]

5.5 Other Formal and Functional Features

As well as showing parallel basic forms and functions, other formal features and their functions also generally correspond between the English and Spanish data. Reduplication and clustering occur in the Spanish corpus. When duplicated, *vale* may indicate that the speaker is defending himself/herself against a perceived accusation, or it may simply be a way of making clear for the listener that there is no need to repeat something that has already been understood. This can be observed in example (19). The interaction takes place in what we assume is a solicitor's office. Speakers 1 and 2 are colleagues:

(19)
S1: Ese … ese el más importante que tengo, es el más importante que tengo de todos, Paco.
'…'
S2: Pero ¿el lunes no dijiste que tú no podías?
S1: El lunes … pues lo hago el martes si no.
S2: ¿Lo de Navarro me dijiste que no podía venir?
S1: Esa era el jueves. Que Navarro tiene el juicio
S2: **Vale, vale.**

[S1: That's the most important one I have, the most important of all the ones I have, Paco.
'…'
S2: But didn't you say you couldn't on Monday?

S1: On Monday … but I can do it on Tuesday otherwise
S2: What about Navarro, you told me he couldn't come?
S1: That was Thursday. [I said] that he has to go to court
S2: Right, right.]

S2's use of *vale, vale* here suggests that he does not want to discuss the topic any further.

Reduplication of *anda* may be intended as an expression of disbelief, as can be seen in this conversation between several members of a family, where space distribution is being debated:

(20)
S1: Mamá, ¿qué dices?
S2: Que en la cocina del otro piso decía que había que poner puerta corredera para que cupiera el frigorífico.
S3: Bueno, eso sí. Y lo sigo pensando o quitar un trozo de bañera y poner la bañera pequeña
S4: ¡Sí, hombre!
S3: Hubiera sido la solución.
S4: **Anda, anda.**

[S1: Mum, what are you saying?
S2: That in the kitchen of the other flat we had to put in a sliding door so that the fridge would fit
S3: Yes, Ok, that's right. And I still think that, or cut a bit off the bath and put in a smaller bath
S4: No way!
S3: That would have been the solution
S4: Yeah, right!]

Ya can express impatience when repeated, while reduplication of *venga* may be simply a way of encouraging the interlocutor, as is the case in (21), taken from a TV programme where listeners ring in to participate in a type of raffle. Observe how the repetition of *venga* is first meant to encourage good luck, and is more emphatic than *venga* on its own a few lines further on:

(21)
S1: Vamos a ver, Consuelo, si tenemos mejor suerte esta tarde.
S2: **Venga, venga, venga.**
S1: Del uno al tres. ¿Cuál quieres, Pilar?
S3: El… tres.
S1: El tres. Vamos a ver que le vale el número tres a Pilar, a ver si son cartas maravillosas, Consuelo; ¡que sean buenas, hombre!
S2: <simultáneo> **Venga.**
S1: El dos, </simultáneo el tres, el cuatro, el cinco…

[S1: Let's see, Consuelo, if we have better luck this afternoon

S2: Come on, come on, come on
S1: From one to three, which one do you want, Pilar?
S3: Err … Three
S1: Three. Let's see what number three is worth for Pilar, let's see if they're wonderful cards, Consuelo; let's hope they're good ones!
S2: <overlapping>Let's go
S1: Two, <overlapping>three, four, five…]

Just as in the English data, response tokens cluster, as in (22), where S1 and 2 discuss food they used to have when they were younger (in this case, *anda* reinforces the bonding between speakers and emphasizes the agreement expressed by *claro*):

(22)
S1: Ahí he visto yo hacer muchos chicharros.
S2: Todas … hacía yo las mantecas.
S1: **Claro. ¡Anda!**
S2: Sí.
S1: Menudas estaban de buenas ahí con el pan.
S2: Eso, eso.
S1: Tan<simultáneo>recientitas. ¡Jolín!

[S1: There I saw a lot of "chicharros" (similar to shortbread) being made
S2: All … I used to make the butter
S1: Yes, of course!
S2: Yes
S1: There were so tasty with bread!
S2: They were, they were
S1: So<overlapping>fresh. My God!]

All in all, a reasonably good fit of formal features and interactional functions exists between the Spanish data and those noted by McCarthy (2002) and O'Keeffe and Adolphs (2008) for English. It would seem that both languages possess a repertoire of response tokens which can convey powerful interactional meanings. In both languages these items form part of the high-frequency core vocabulary.

6 Transferability

While the existence of response tokens of some sort is likely to be language-universal, there are equally likely to be problems of transferability and translatability across languages, as the pragmatic analyses of the Spanish items and attempts at ascertaining precise English equivalents above suggests. Here we comment on some of the issues raised by the translatability of the Spanish tokens into English.

At the level of form, Spanish response tokens such as *anda, venga, fíjate*, *vaya, bueno, claro* all display apparent inflection (in this case singular imperative/subjunctive inflexions for *anda, venga, fíjate*, and *vaya,* and masculine gender for *bueno* and *claro*), where their English counterparts do not, a consequence of the differing typologies of the two languages. Although inflected, the inflection is invariable, indicative of their fossilisation as pragmatically specialized tokens. Meanwhile, the noun *hombre* (man), in contrast to *mujer* (woman), can be used as a response token (and, indeed as a discourse marker, see Portolés 1998: 131–132), regardless of whether the interlocutor is a man or a woman, whereas *mujer*, which can also function as a response token, is only used to address a woman.

Another potential problem area is that suprasegmental features are particularly important: the intonation contour that a speaker applies to a particular response token can determine whether the reaction is perceived as convergent or distancing, and whether other nuances are implied. *Vaya*, for example, can indicate amusement, surprise, or pity, depending on the intonation in context. *Anda* can be an expression of surprise, agreement, emphasis, or commiseration. *Venga* can express impatience. *Claro*, with the appropriate intonational contour, can actually indicate distancing (with a note of irony), or reproach, often realised by rising intonation. The same can be said of *bueno*, which, as we indicated above, is a less rotund way of showing convergence (especially if it is accompanied by low pitch). Similar suprasegmental issues attach to English items such as *really, indeed* and *well*, where a variety of pragmatic effects can be achieved by varying intonation in context. This suggests that cross-linguistic comparisons should always be done on the basis of as much linguistic and contextual information as possible.

Reduplication is another area where there is an apparent lack of direct transferability between languages. Some forms in Spanish can be reduplicated, and this can sometimes affect the pragmatic force. For example, *venga* can be used to support agreement, as we saw in the case of the extract from a TV programme, example (21), above. However, there we also noted that when reduplicated, *venga* may be simply a way of encouraging the interlocutor.

A potential problem item is the often over-extended translation of *claro* into English as *of course*, thus endowing the response with an implicature of 'how could you possibly think otherwise?', which may or may not be appropriate. So the English exchange *May I use your bathroom? Of course!* would be pragmatically well-formed, while the sequence *We were at the Louvre on Sunday. Oh, did you see the Mona Lisa? Of course*! may be heard as pompous and brusque. A mis-translation or an over-extension of a translation can potentially generate misunderstanding. As Travis (1998) points out, *well* does not always translate as *bueno* or *bien*, and *really?*, for example, can be equivalent to *¡anda!* in some contexts. Moreover, reduplication of *anda* may be intended as an expression of strong disbelief, as can be observed example (20) above, where space distribution was being discussed. By contrast, in the British data, *really?* does not occur as a reduplicated response token.

These few examples raise some pertinent cross-linguistic issues, not only to do with semantic meanings, but with pragmatic force, and relate not only to individual uses of words, but the effects of reduplication, clustering and intonation too.

7 Response Tokens and Fluency

Given the connection between what has been discussed in the previous sections and the concept of fluency, in this section we return briefly to this issue. Spoken fluency is often seen as related to the solo performance of an individual speaker. Corpus evidence consistently shows that speakers in real conversations support one another and co-construct the talk. Conversation and its ability to flow are the joint responsibility of all co-participants; our perception of fluency is much influenced by the cooperatively created flow of talk, rather than just the talent of any individual speaker. The term 'confluence' may be a more apt label for such joint activity (McCarthy 2010).

The dominant notions of fluency have their roots in linguistic qualities related to lexico-grammatical and phonological flow created by individual speakers, in the ability of participants to converse rapidly, unhesitatingly, coherently and appropriately (see Fillmore 1979 and McCarthy 2010 for further discussion). Here we argue that fluency is enhanced by the degree of interactive support each speaker gives to the flow of talk, by helping one another to be fluent.

8 Conclusion

We have used a corpus-based methodology to investigate response tokens in Spanish, based on frameworks derived from previous studies of English. Corpora not only provide quantitative evidence to make plain aspects of language use which are often difficult to reflect upon via intuition (in this case, everyday uses of some of the most frequent words in the language); they also offer the opportunity for fine-grained analyses of particular items in multiple contexts. The use of corpora for the analysis of banal, everyday conversational phenomena are discussed at length in McCarthy (1998) and O'Keeffe et al. (2007), though even there, cross-linguistic comparisons get little attention. This is hardly surprising, given the dominance, until recently, of corpus studies of English, while other languages (relatively speaking) lagged behind. However, that situation has changed and corpora are now available for both widely-taught and lesser-taught languages. Corpus analysis within and across languages, especially the analysis of spoken data, reveal features of language use of paramount interest to researchers. In the present case, the focus has been on listenership, but one can easily envisage equally fruitful investigations of aspects of spoken language use such as vagueness and approximation, conversational boundary marking, rhetorical moves such as hyperbole and understatement, and a variety of other, similar features which are not easily accessed by intuition alone, whether that of native- or non-native users.

References

Amador Moreno, C.P., A. Chambers, and S. O'Riordan. 2006. Integrating a corpus of classroom discourse in language teacher education: The case of discourse markers. *ReCALL* 18: 83–104.

Andersson, L., and P. Trudgill. 1990. *Bad language*. Oxford: Basil Blackwell.

Antaki, C. 2000. 'Brilliant. Next question…', high-grade assessment sequences in the completion of interactional units. *Research on Language and Social Interaction* 33: 235–262.

Antaki, C. 2002. 'Lovely', turn-initial high-grade assessments in telephone closings. *Discourse Studies* 4: 5–24.

Bauhr, G. 1994. Funciones discursivas de *bueno* en Español moderno. *Lingüística Espanola Actual* XVI: 79–121.

Brinton, L.J. 1996. *Pragmatic markers in English. Grammaticalization and discourse functions*. Berlin/New York: Mouton de Gruyter.

Bublitz, W. 1988. *Supportive fellow-speakers and cooperative conversations*. Amsterdam: John Benjamins.

Carter, R.A., and M.J. McCarthy. 2006. *Cambridge grammar of English*. Cambridge: Cambridge University Press.

Carter, R.A., M.J. McCarthy, G. Mark, and A. O'Keeffe. 2011. *English grammar today*. Cambridge: Cambridge University Press.

Clancy, P.M., S.A. Thompson, R. Suzuki, and H.Y. Tao. 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of Pragmatics* 26: 355–387.

Clark, H., and J. Fox Tree. 2002. Using *uh* and *um* in spontaneous speech. *Cognition* 84: 73–111.

Drummond, K., and R. Hopper. 1993. Backchannels revisited: Acknowledgement tokens and speakership incipiency. *Research on Language and Social Interaction* 26: 157–177.

Duncan, S. 1972. Some signals and rules for taking speaking turns in conversation. *Journal of Personality and Social Psychology* 23: 283–292.

Duncan, S. 1974. On the structure of speaker-auditor interaction during speaker turns. *Language in Society* 2: 161–180.

Duncan, S., and G. Niederehe. 1974. On signaling that it's your turn to speak. *Journal of Experimental Social Psychology* 10: 234–247.

Duranti, A. 1986. The audience as co-author, an introduction. *Text* 6, 239–247 [Introduction to special issue of the journal on *The audience as co-author*, ed. A. Duranti and D. Brenneis].

Erickson, F. 1986. Listening and speaking. In *Georgetown University round table on languages and linguistics, 1985*, ed. D. Tannen and J. Alatis, 294–319. Washington, DC: Georgetown University Press.

Erickson, F., and J. Shultz. 1982. *The counselor as gatekeeper, social interaction in interviews*. New York: Academic.

Farr, F., B. Murphy, and A. O'Keeffe. 2004. The Limerick Corpus of Irish English: Design, description and application. *Teanga* (Yearbook of the Irish Association for Applied Linguistics) 21: 5–29

Fillmore, C.J. 1979. On fluency. In *Individual differences in language ability and language behavior*, ed. C.J. Fillmore, D. Kempler, and W.S.Y. Wang, 85–102. New York: Academic.

Fraser, B. 1999. What are discourse markers? *Journal of Pragmatics* 31: 931–952.

Fries, C.C. 1952. *The structure of English*. New York: Harcourt, Brace & Co.

Fuentes Rodríguez, C. 1993. Comportamiento discursivo de *bueno, bien, pues bien*. *E.L.U.A., 9*: 205–221.

Gardner, R. 1997. The listener and minimal responses in conversational interaction. *Prospect* 12: 12–32.

Gardner, R. 1998. Between speaking and listening, the vocalization of understandings. *Applied Linguistics* 19: 204–224.

Gardner, R. 2002. *When listeners talk: Response tokens and listener stance*. Amsterdam: John Benjamins.

Goodwin, C. 1981. *Conversational organization, interaction between speakers and hearers*. New York: Academic.
Goodwin, C. 1986. Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies* 9: 205–217.
Hasselgreen, A. 2005. *Testing the spoken English of young Norwegians: A study of test validity and the role of 'smallwords' in contributing to pupils' fluency*. Cambridge: Cambridge University Press.
Heritage, J. 1985. Analyzing news interviews, aspects of the production of talk for an overhearing audience. In *Handbook of discourse analysis*, vol. 3, ed. T.A. Van Dijk, 95–117. London: Academic.
Holmes, J., and M. Stubbe. 1997. Good listeners: Gender differences in New Zealand conversation. *Women and Language* 20: 7–14.
Jucker, A. 1986. *News interviews, a pragmalinguistic analysis*. Amsterdam: John Benjamins.
Kendon, A. 1967. Some functions of gaze-direction in social interaction. *Acta Psycholoigia* 20: 22–63.
Knight, D., D. Evans, R. Carter, and S. Adolphs. 2009. HeadTalk, HandTalk and the corpus: Towards a framework for multi-modal, multi-media corpus development. *Corpora* 4(1): 1–32.
Lenk, U. 1998. *Marking discourse coherence: Functions of discourse markers*. Tubingen: Gunter Narr Verlag.
Lerner, G.H. 1989. Notes on overlap management in conversation: The case of delayed completion. *Western Journal of Speech Communication* 53: 167–177.
Martín Zorraquino, M.A., and J. Portolés. 1999. Los marcadores del discurso. In *Gramática de la lengua española 3*, ed. I. Bosque and V. Demonte, 4051–4214. Madrid: Espasa-Calpe.
Maynard, S.K. 1989. *Japanese conversation: Self-contextualization through structure and interactional management*. Norwood: Ablex.
McCarthy, M.J. 1998. *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.
McCarthy, M.J. 2002. Good listenership made plain: British and American non-minimal response tokens in everyday conversation. In *Using corpora to explore linguistic variation*, ed. R. Reppen, S. Fitzmaurice, and D. Biber, 49–71. Amsterdam: John Benjamins.
McCarthy, M.J. 2003. Talking back: "small" interactional response tokens in everyday conversation. *Research on Language and Social Interaction* 36: 33–63.
McCarthy, M.J. 2010. Spoken fluency revisited. *English Profile Journal. Inaugural issue*. Online at: http://journals.cambridge.org/action/displayJournal?jid=EPJ. Accessed 28 April 2013.
McGregor, G., and R. White. 1990. *Reception and response, hearer creativity and the analysis of spoken and written texts*. London: Routledge.
O'Keeffe, A., and S. Adolphs. 2008. Using a corpus to look at variational pragmatics: Response tokens in British and Irish discourse. In *Variational pragmatics*, ed. K.P. Schneider and A. Barron, 69–98. Amsterdam: John Benjamins.
O'Keeffe, A., M.J. McCarthy, and R.A. Carter. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.
Öreström, B. 1983. *Turn-taking in English conversation*. Lund: Gleerup.
Pomerantz, A. 1984. Agreeing and disagreeing with assessments, some features of preferred/dispreferred turn shapes. In *Structures of social action*, ed. J. Atkinson and J. Heritage, 57–101. Cambridge: Cambridge University Press.
Portolés, J. 1998. *Marcadores del discurso*. Barcelona: Ariel practicum.
Schegloff, E. 1982. Discourse as interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In *Analysing discourse. Text and talk*, ed. D. Tannen, 71–93. Washington, DC: Georgetown University Press.
Schegloff, E. 2006. *Sequence organization in interaction: A primer in conversation analysis*. Cambridge: Cambridge University Press.
Schiffrin, D. 1987. *Discourse markers*. Cambridge: Cambridge University Press.
Scott, M. 2011. *Wordsmith tools. Software*. Oxford: Oxford University Press.
Sinclair, J.Mc.H., and D. Brazil. 1982. *Teacher talk*. Oxford: Oxford University Press.

Sinclair, J.Mc.H., and R.M. Coulthard. 1975. *Towards an analysis of discourse*. Oxford: Oxford University Press.

Sorjonen, M.-L. 2001. *Responding in conversation: A study of response particles in Finnish*. Amsterdam: John Benjamins.

Stenström, A.-B. 1990. Lexical items peculiar to spoken discourse. In *The London-Lund corpus of spoken English*, ed. J. Svartvik, 137–175. Lund: Lund University Press.

Stubbe, M. 1998. Are you listening? Cultural influences on the use of supportive verbal feedback in conversation. *Journal of Pragmatics* 29: 257–289.

Tannen, D. 1984. *Conversational style: Analyzing talk among friends*. Norwood: Ablex.

Tao, H. 2003. Turn initiators in spoken English, a corpus based approach to interaction and grammar. In *Corpus analysis, language structure and language use*, ed. C. Meyer and P. Leistyna, 187–207. Amsterdam: Rodopi.

Tao, H.Y., and S.A. Thompson. 1991. English backchannels in Mandarin conversations: A case study of superstratum pragmatic "interference". *Journal of Pragmatics* 16: 209–233.

Thomas, J. 1983. Cross-cultural pragmatic failure. *Applied Linguistics* 4: 91–112.

Tottie, G. 1991. Conversational style in British and American English, the case of backchannels. In *English corpus linguistics*, ed. K. Aijmer and B. Altenberg, 254–271. London: Longman.

Travis, C. 1998. *Bueno*: A Spanish interactive discourse marker. *Berkeley Linguistic Society* 24: 268–279.

Watts, R.J. 1989. Taking the pitcher to the 'well': Native speakers' perception of their use of discourse markers in conversation. *Journal of Pragmatics* 13: 203–237.

Yngve, V. 1970. On getting a word in edgewise. Papers from the 6th Regional Meeting, Chicago Linguistic Society. Chicago: Chicago Linguistic Society.

The Corpus of Language and Nature (CLAN Project)®: A Tool for the Study of the Relationship Between Cognition and Emotions in Language

Jesús Romero-Trillo

1 Introduction

The Corpus of Language and Nature (CLAN)®[1] is a worldwide project based at the Universidad Autónoma de Madrid (Spain), directed by the author of this chapter, whose aim is to analyze the emotional and linguistic responses to the perception of landscapes by speakers of English as a first or second language. The theoretical background of the project is based upon the cognitive linguistic descriptions of the perception of nature described by Romero-Trillo and Espigares (2012). The present chapter intends to describe the principles of the corpus design and its methodology. The description will also include the underlying variables selected for its compilation, and the statistical composition of the personal, linguistic and geographical features of the participants. The most important feature of this corpus is that its main aim is not to compile a massive amount of spoken data of the speech of informants belonging to different regional and first language variables, but the development of a scientific experiment in which certain variables are considered for future linguistic and statistical correlations. In sum, the underlying motive for the corpus compilation is that the speakers of English as a first or foreign/second language express their emotions towards nature on the basis of some universal landscape and cognitive principles, and that their descriptions can be the key to the design of a cognitive map of linguistic features in the relationship between humans and nature.

[1]The Corpus of Language and Nature (CLAN Project) (logos, design, computer platform architecture and data) has the Certificate of Registration No 010091932 issued by the Register of Community Trade Marks of the European Union.

J. Romero-Trillo (✉)
Departamento de Filología Inglesa, Facultad de Filosofía y Letras,
Universidad Autónoma de Madrid, C/Francisco Tomás y Valiente 1,
28049 Madrid, Spain
e-mail: jesus.romero@uam.es

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_10,

2 The Relationship Between Language, Cognition and Nature

Language and nature have been related since the beginning of humankind, when humans had to survive in hostile environments due to the attacks of animals and the dangers of the earth and the weather. In our days, when nature seems to be controlled and mostly appreciated for its aesthetic values, all of us have in one way or another established some relationship with nature and, as a result, have been emotionally impressed after the contemplation of a natural landscape. In this sense, it can be averred that nature has always been inextricably linked to humans and that, to some extent, the presence or absence of humans has been the defining ultimate factor in the current state of nature. Landscapes for humans, therefore, have always been the epitome of the conceptualization of nature in the attempt to capture the essence – both biophysical and aesthetic- of the environment. Bearing in mind all these circumstances, the CLAN Project attempts to link the perception of nature to the emotions and linguistic reactions towards a selection of landscapes that synthesize relevant cultural and ecological features.

Ecologists believe that humans appreciate nature, positively or negatively, and that this appreciation has a direct correlation with the political and social attitudes towards the preservation of nature. In this sense, landscape attraction or rejection is a matter of great concern because, sometimes, the value of a natural spot is not immediately reflected on the specific cultural preferences of a social group. What the CLAN Project tries to study is the extent to which cultural factors, expressed through the preference in the selection of photographs of natural landscapes and through the linguistic comments of the participants, are also at the basis of the emotional responses towards landscapes beyond individual preferences.

The role of culture in the preference of certain natural elements has been the subject of study of ecologists in the second half of the twentieth century. I understand culture in its inheritance fashion insofar as it "denotes a historically transmitted pattern of meanings embodied in symbols, a system of inherited conceptions expressed in symbolic forms by means of which men communicate, perpetuate, and develop their knowledge about and attitudes towards life" (Geertz 1973: 89).

The intention of the CLAN Project is to identify the aesthetic elements that pervade individual choices in the appreciation of nature, which may constitute key elements at a universal cognitive level. In his seminal book, González-Bernáldez (1985) summarized the specific and abstract features that had been studied until that moment by ecologists, psychologists, etc. with a detailed analysis of the experiments that had proven the role of these features in the appreciation of nature. Although the discussion on the findings and interdisciplinary implications for the study of nature and landscapes is not the object of this chapter, I would like to emphasize that González-Bernáldez's international role on the development of human ecology was essential for the understanding of the bio-physical influence of nature on human cognition.

Nevertheless, the CLAN Project intends to look at the relationship between nature and cognition from an angle that has not been explored until now: the

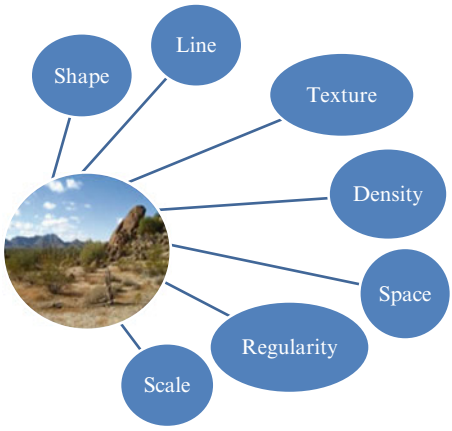


Fig. 1 Taxonomy of visual features (Romero-Trillo and Espigares 2012)

relationship between language and emotion during the contemplation of natural landscapes. The aim is to depart from the personal features of the informants, who belong to different cultures and have different commands of English and other languages, and analyze their comments on the photographs that depict natural landscapes. The theoretical foundation of the study lies on the theory of perception developed by landscape ecology (Zube et al. 1982; Turner et al. 2001; Wiens and Moss 2005), and on the assumed universal parameters of landscape perception and their role in the human adaptation to the environment (Espigares et al. 2008; de Lucio et al. 1996; Romero-Trillo and Espigares 1996).

In order to evaluate the components of landscapes from a linguistic stance Romero-Trillo and Espigares (2012) designed a cognitive taxonomy of visual features that play a role in the description of the components of natural landscapes. The features of the taxonomy function in a systemic format, in Hallidayan terms, as they all appear in all landscapes and viewers identify each of them.

The complete taxonomy of visual features is presented in Fig. 1:

These features are subsequently subdivided into subcategories that are explained following the Natural Semantic Metalanguage theory (NSM) (Goddard and Wierzbicka 2002). This theory accounts for the semantic universals that underlie the linguistic realization of the same concepts in different languages. This approach guarantees the absence of distortion in the linguistic description of the universal natural features identified. For this purpose, the NSM identifies some semantic universals -or 'primes'- i.e., meanings that are semantically simple, that cannot be defined further and are accepted as indefinable. The notion behind these principles is that language, and descriptions in particular, should be based on certain elements with undisputed value. This was one of the concerns of some philosophers in seventeenth Century like, for instance, Leibniz:

> "Amongst the words, some are frequently used and serve as auxiliary to the others" (Leibniz 1678 [1987]: 162). For Leibniz these words were "the alphabet of human thoughts" (Wierzbicka 1972: 6).

In NSM theory certain syntactic qualities of the primes have also been shown to be universal and can be listed as universal canonical combinations of primes. In the present description the selected primes are 'kind', a relational prime, and 'place', a space prime. These primes were combined with the physical and visual features of the landscapes to form a grammar that can explain the objective description of landscapes without distortion. Thus, the emerging metalanguage is capable of representing meanings of more complex concepts and of (shared) cultural attitudes through explications or semantic paraphrases.

Below is the example of the definition of the category 'shape' (Romero-Trillo and Espigares 2012:174):

Shape

The shape in landscapes delimits the volume of what is being observed. It can be two-dimensional, three-dimensional, geometric or complex.

1. Two-dimensional shape: When viewers see this **place**, they can observe some elements of a different **kind** there. They can think about all these elements like this: 'these elements can be well observed and distinguished in horizontal and vertical terms'.
2. Three-dimensional shape: When viewers see this **place**, they can observe some elements of a different **kind** there. They can think about all these elements like this: 'these elements can be well observed and distinguished according to irregular lines in terms of width, height and depth'.
3. Geometric shape: When viewers see this **place**, they can observe some elements of the same **kind** there. They can think about all these elements like this: 'these elements can be well observed and distinguished according to regular lines organized in terms of width, height and depth'.
4. Complex shape: When viewers see this **place**, they can observe some elements of the same **kind** there. They can think of all these elements like this: 'these elements cannot be observed and distinguished according to regular lines organized in terms of width, height and depth'.

As observed, the primes 'place' and 'kind' are pivotal to the four categories and the adjectives 'same' and 'different' are essential in all combinations.

The semantic descriptions are complemented with a set of photographs and figurative sketches that represent the prototypical landscape categories, as shown in Fig. 2:

The complete classification of categories and sub-categories in the present model, with their definition of the seven systemic categories, is shown in Fig. 3:

As can be observed, these categories allow linguists to describe natural landscapes through the use of universal semantic principles that can account for the preferences of speakers in the selection of certain natural features or in the expression of their emotions. The use of these categories in the analysis of the description of landscapes will allow researchers to use unequivocal linguistic tools based on cognitive parameters.

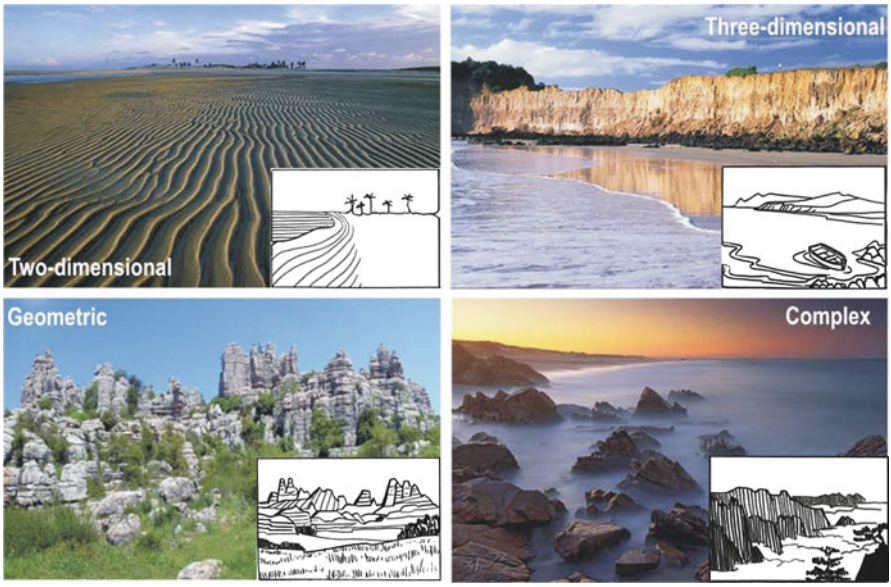


Fig. 2 Prototypical illustrations of the category 'shape' (All photos in the chapter supplied by Jesús Romero-Trillo, with permission of CLAN-Project® research team)

3 The Architecture of the Corpus of Language and Nature (CLAN Project)®

The Corpus of Language and Nature®, as mentioned above, intends to describe the emotional reaction towards natural landscapes by speakers of English as a first or foreign/second language. The corpus takes into consideration the personal and biographical variables of the speakers in its analysis of the spoken description of the photographs.

The corpus organization considers the preference of observers according to the landscape universals described by evolutionary ecology:

- Phytophilia: the preference for places with vegetation.
- Hydrophilia: the preference for places with water.
- Diversity: the preference for places with different kinds of objects.

These general preferences are combined with the expression of the preferences related to the individual cognitive realm:

- Mistery: the like or dislike towards unknown places and situations.
- Risk: the like or dislike towards risky places and situations.
- Domestication: the like or dislike towards places that have been domesticated, i.e., subject to human intervention.

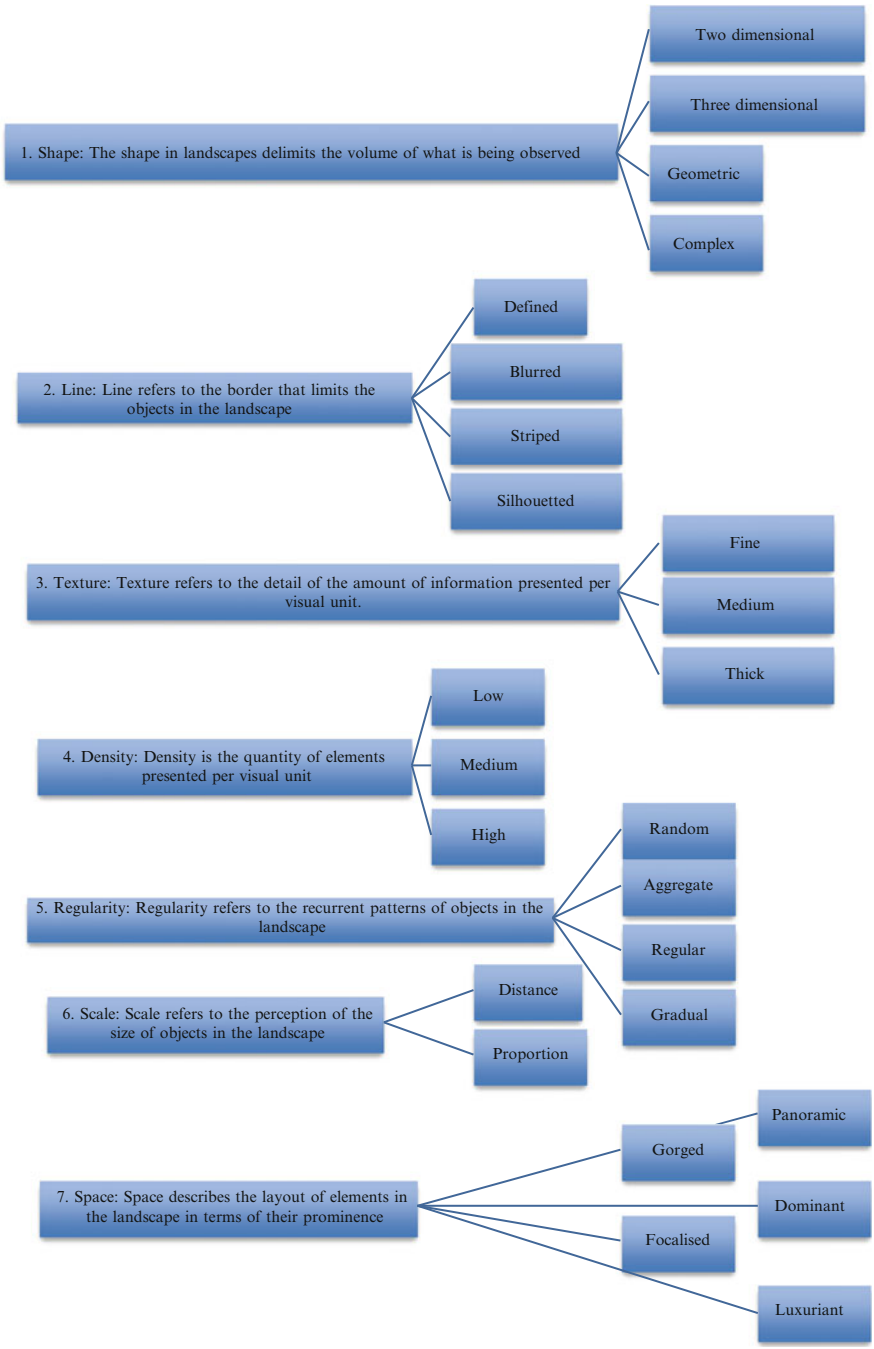


Fig. 3 Definition and classification of visual and cognitive categories

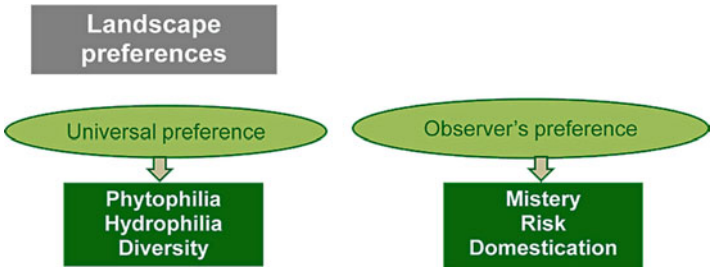


Fig. 4 Classification of universal and personal landscape preferences

Figure 4, below, presents the outline of the universal and the observer's preferences considered in the collection and analysis of the corpus:

The procedure for the collection of the corpus has been the selection of landscape photographs that are shown to all the participants, who are asked to make comments on the basis of prompt questions that would guarantee an equal input of information. In this sense, a good knowledge of the background of the participants is essential in order to evaluate their personal and biographic features. This information is obtained with the aid of a questionnaire that each participant has to answer before commenting on the photographs.

The questionnaire is the following (* indicates an obligatory field):

CLAN Username *

University you attend *

Years you have been enrolled at the University *

1
2
3
4+

Age *

17
18
19
20
21
22
23
24+

Sex *

Female
Male

Nationality *

Country of Residence: *

*In what type of setting did you grow up? **

Urban

Suburban

Rural

*How often do you go camping and/or make trips to the country/mountains? **

1 Never

2

3

4 Very frequently

*Given the two options, would you prefer to live in **

big, urban city

small, suburban town

*How many languages, INCLUDING English, do you speak? **

1

2

3

4

5 or more

*The English language is your Mother/Native tongue **

Yes

No

If English is NOT YOUR MOTHER/NATIVE TONGUE, ALSO ANSWER THE FOLLOWING QUESTIONS AND PLEASE choose the BEST answer from the listed options.

English is

A second language you actively use

A third language you actively use

A language you learn/ed as a foreign language that you use only on occasion

Please rate your fluency in English

1 Low Proficiency

2

3

4 Fluent

Is English the primary language spoken in your home?

Yes

No

Is English the primary language spoken at your University?

Yes

No

If English is not your native language, what was your primary method of English language acquisition?

Courses as part of required school curriculum

Courses as part of elective school curriculum

Language Academy courses

Immersion experiences (e.g. living abroad)

How many years have you spent living in an English speaking country?
1 or less
2–3
4–6
7–9
10 or more
What type of high school did you attend?
Private, American or British
Private, other
Public, American or British
Public, other
Was English the primary language spoken at your high school?
Yes
No

The corpus design, therefore, considers the objective parameters related to the personal variables of the questionnaire such as country of origin, mother tongue, sex, educational background etc. This information allows researchers to correlate landscape descriptions with biographical variables that will be used for the universal analysis of landscape preferences on the basis of countries, first languages, educational backgrounds, etc.

It is important to mention that the questionnaire is filled out online and that informants get access with a unique *CLAN Username*, which then allows them to complete the description of the photographs, also online.

As of November 2012, the corpus has compiled the descriptions made by 597 participants from 20 countries that are geographically distributed as presented in Fig. 5:

Fig. 5 Geographic distribution of speakers in the CLAN Project

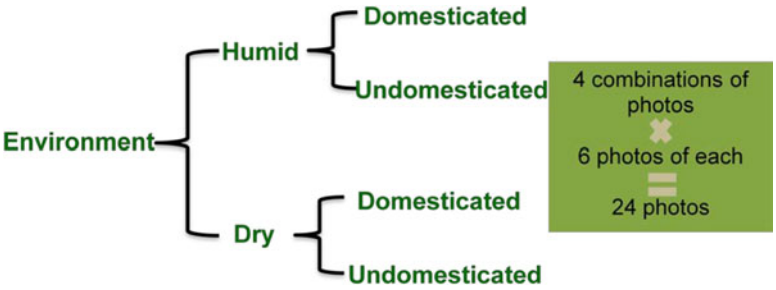


Fig. 6 Photograph selection with specification of the variables

The second essential research issue was the selection of the photographs and the decision was to identify two independent variables related to the landscape in the research design. As the study of all features mentioned above was too large in statistical terms, the decision was to group four possible combinations in terms of the most influential variables according to landscape ecology, i.e. hydrophilia and domestication, with a dual possibility for each of them

1. In relation to the presence of water:
   (a) Humid
   (b) Dry
2. In relation to the presence of humans
   (a) Domesticated
   (b) Not domesticated

For each combination six photographs were selected in order to account for the necessary replicability of the samples, as Fig. 6 shows:

The finally selected 24 photographs, out of a battery of over 1,000, are presented in Fig. 7:

Therefore, the resulting research design in terms of variables for statistical analyses is the following:

1. Independent variables (observer's)
   Culture (Home country, urban or rural origin, sex, etc.)
   L1
2. Independent Variables (of the landscape)
   Environment: Humid/Dry
   Domestication: Yes/No
3. Dependent Variables
   Linguistic realization: lexis, phonology, syntax, pragmatics, etc.

As mentioned above, all participants have a username and a password[2] that are necessary to enter the CLAN computer platform (Fig. 8), whose first task is to fill out the questionnaire.

[2]Passwords can be obtained from the author.

Fig. 7 The 24 photographs of the CLAN Project

After completing the questionnaire, the interface presents the 24 photographs on the screen. The participants are instructed to select the photographs for their comments in whichever order they prefer, and it is important to mention that the order of presentation is different for each participant to guarantee that the sequence does not interfere in the choices. The selection order of the photographs is precisely one of the variables that will be analyzed on the basis of the personal information of the participants.

After a given photograph has been selected, the CLAN platform gives some prompt questions that can help the participants to start their comments. To facilitate the

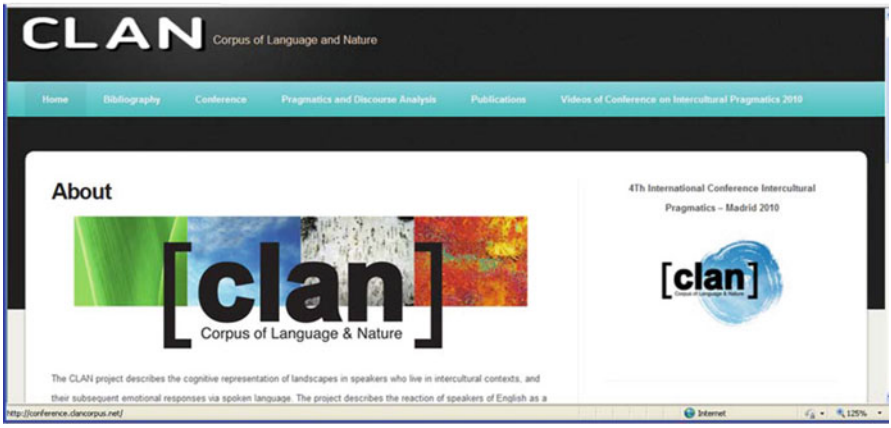


Fig. 8 Visual interface for the corpus collection

process, the computer programme does not allow comments on the same photograph more than once, and indicates the photographs that have been already commented on by placing a red square around them.

Participants are asked to read the prompt questions before their comments, at least the first time, to get a hint on their pronunciation features:

1. *"Imagine a friend of yours just returned from vacation (holiday) and showed you this picture. What would you say to them about the picture?" "What would you want to know about their experiences there?"*
2. *What words come to your mind when you look at this picture?*
3. *"What do you imagine it would be like to live here? Would you like to live there? Why or why not?"*
4. *Imagine you are in this place right now. Describe what you are seeing, feeling, and thinking.*
5. *How is this place similar or different to where you grew up?*
6. *Give a title to the picture*

The interface will ask for permission to use the computer webcam and, after acceptance, the CLAN software will start recording the voice and the video of the participants, the interface saves the recordings and then, in less than 1 min after completion, the video files arrive at the CLAN server. It is also important to highlight that the description of the 24 photographs need not be done in one session, but that the informants can log in the system repeatedly until they finish all the comments. Figure 9 shows the design of the corpus collection:

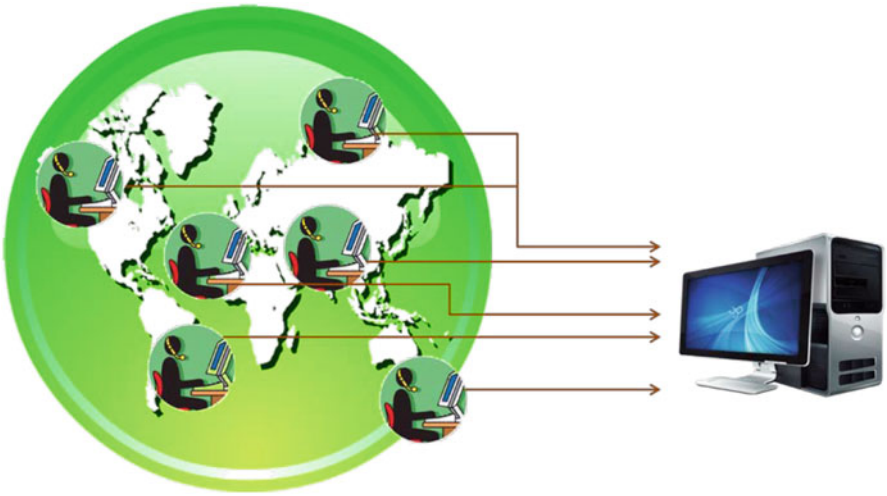
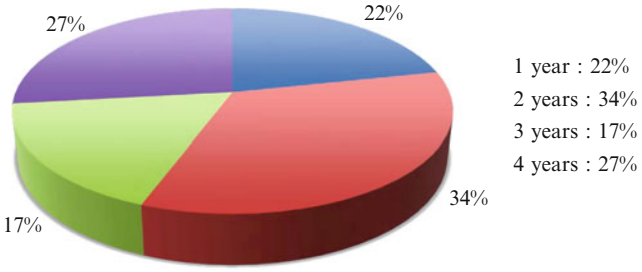


Fig. 9 Corpus collection computer platform

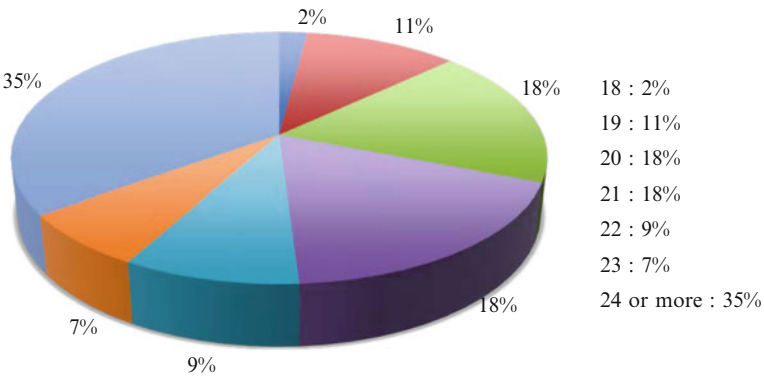
4 Results of the Questionnaire

In this section I will present the results of the biographical questionnaire compiled by the participants. The results show the variety and representative nature of the informants, which shows the representativeness of the corpus descriptions:

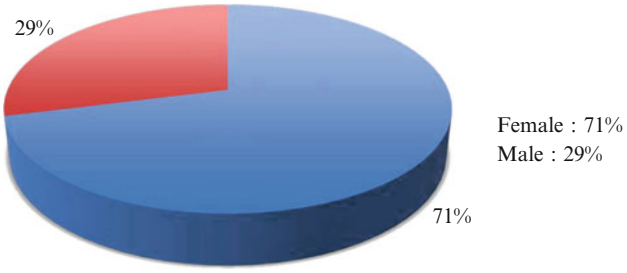
4.1 Years Enrolled at University



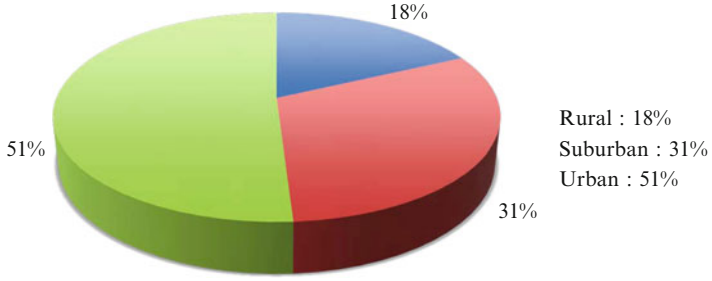
4.2 Age of Participants



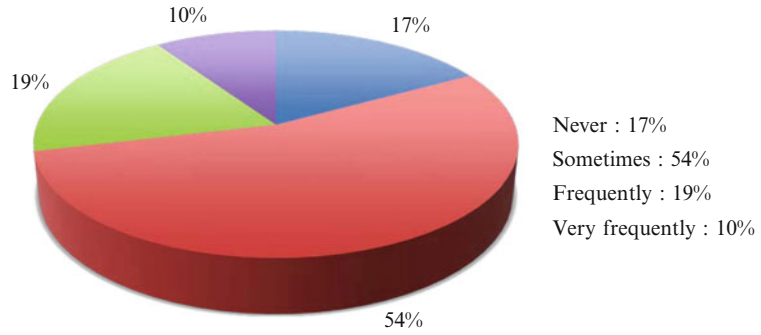
4.3 Sex of the Participants



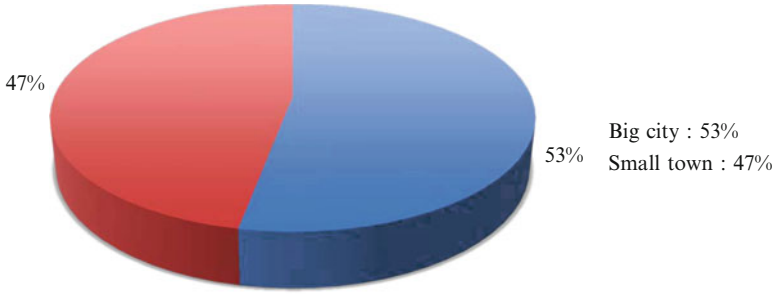
4.4 Type of Setting in Which Participants Grew Up



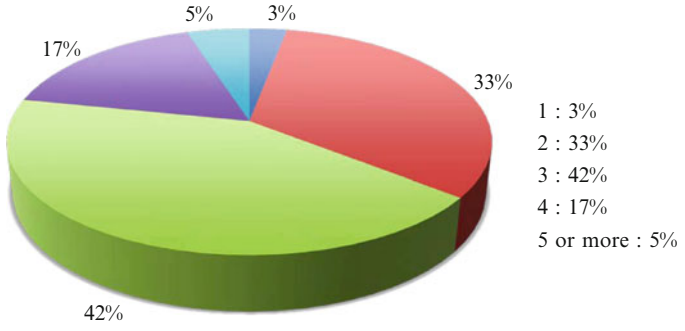
4.5 Frequency with Which Participants Go Camping or Make Trips to Countryside



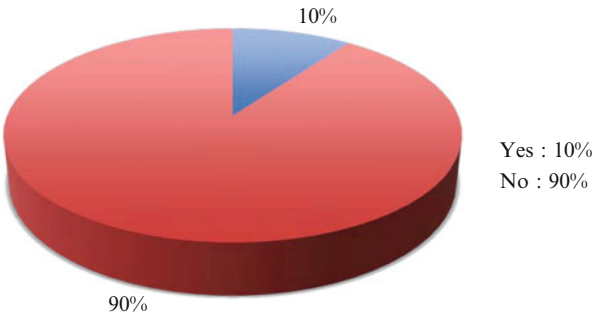
4.6 Preference of Living Environment



4.7 Number of Languages Spoken by the Participants (Including English)

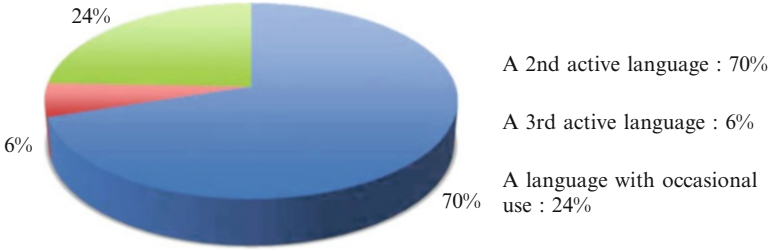


4.8 English as the Mother/First Tongue

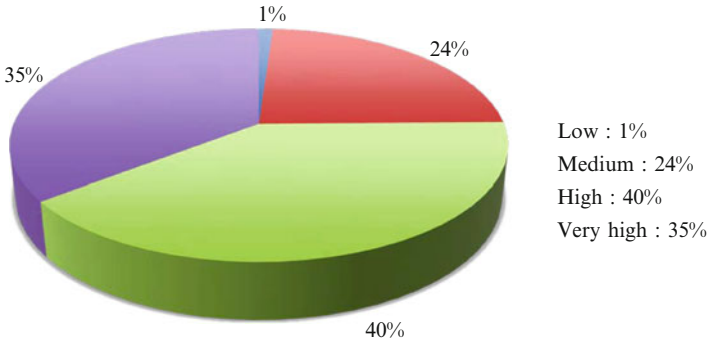


Questions only answered by the participants whose L1 is not English

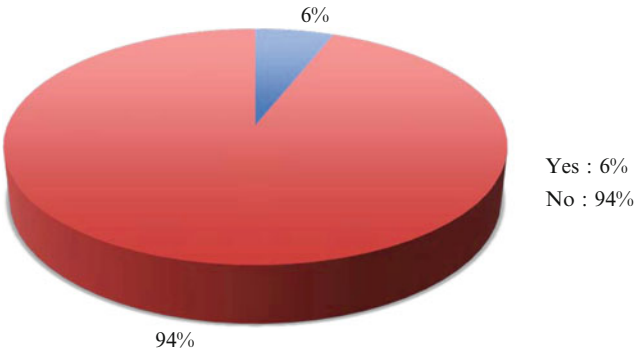
4.9 Use of English



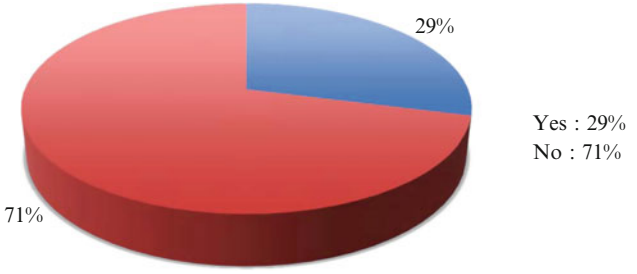
4.10 Fluency in English



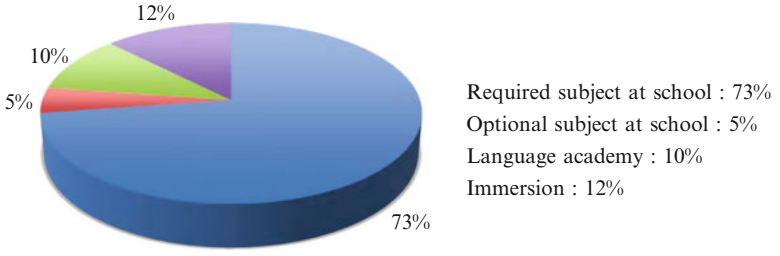
4.11 English as the Primary Language at Home



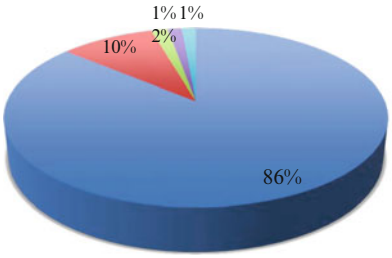
4.12 English Is the Primary Language at University



4.13 Information About the Learning Method of English

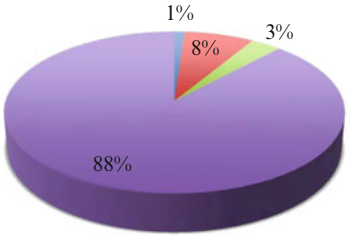


4.14 Years Spent Living in an English Speaking Country



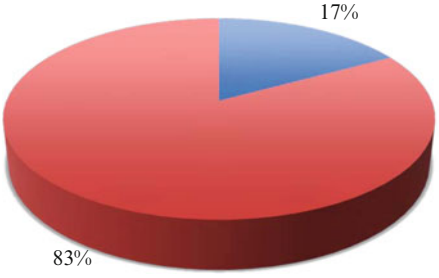
1 or less : 86%
2 or 3 : 10%
4 to 6 : 2%
7 to 9 : 1%
10 or more : 1%

4.15 Type of School Attended by the Participants



Private, American or British : 1%
Private, other : 8%
Public, American or British : 3%
Public, other : 88%

4.16 English as the Primary Language Used at School



Yes : 17%
No : 83%

The pie charts presented above show the general picture of the participants in the corpus based on the individual responses to the initial questionnaire. For research purposes the graphs can be computed differently to identify the personal variables that researchers want to investigate. These features are then contrasted with the actual descriptions of the photographs, archived separately to investigate the cognitive preferences and the realization of the emotions.

5 Conclusions

The present chapter has presented the procedural phases in the design of a corpus, with a particular emphasis on the necessary theoretical support that, in my opinion, has to be present. The theoretical framework developed by Romero-Trillo and Espigares (2012) allows researchers to have cognitive and linguistic tools for the analysis of the corpus, which is fundamental to avoid an impressionistic approach to the data. The chapter has also described in full detail the dependent and independent variables used in the selection of the participants and of the photographs, which will be essential for the statistical analysis of the data. I am convinced that a sound and detailed preparation of the corpus design is a fundamental asset for the validity of the linguistic analysis. Likewise, the questionnaire design and the plurality of the subjects' origins guarantee that the corpus compilation does represent the multifaceted emotional and cognitive responses of the participants.

The results section of the chapter has evidenced with full detail the questionnaire results of the participants, which are essential for the unequivocal evaluation of the linguistic and emotional aspects of the corpus. It is not very frequent that corpora include such a detailed description of the beliefs, attitudes and biographies of the participants in relation to their spoken comments. For the CLAN project analysts, the link between the questionnaire and the comments is essential for the understanding of the general trends in the linguistic and cognitive patterns of use.

In terms of the technical characteristics of the corpus, I have shown the innovative computer platform methodology that enables the recording of the corpus through a webcam, which makes recording easier, more practical, and more dependable in its storage. In this sense, the CLAN centralized server substitutes the individual recordings in different parts of the world with the alleviation of any difficulty of storing and sharing the data.

As a conclusion, I would like to emphasize that the Corpus of Language and Nature does not aim at the recording of large volumes of spoken data without any research question in mind. In fact, I believe that it represents a good example of a corpus that has been designed and compiled to investigate whether the speakers of English as a first or foreign/second language express their emotions towards nature on the basis of universal landscape and cognitive principles. In this sense, the Corpus of Language and Nature aims at the design of a cognitive map of the linguistic features that express the relationship between humans and nature.

References

de Lucio, J.V., M. Mohamadian, J.P. Ruiz, and J. Benayas. 1996. Visual landscape exploration as revealed by eye movement tracking. *Landscape and Urban Planning* 34: 135–142.

Espigares, T., N. Zafra, and M.A. Rodríguez-Fernández. 2008. What do we call adaptive management? A general characterization from a global sample. *Web Ecology* 8: 1–13.

Geertz, C. 1973. *The interpretation of cultures. Selected essays by Clifford Geertz*. London: Hutchinson.

Goddard, C., and A. Wierzbicka (eds.). 2002. *Meaning and universal grammar: Theory and empirical findings 2 vols*. Amsterdam: John Benjamins.

González Bernáldez, F. 1985. *Invitación a la ecología humana. La adaptación afectiva al entorno*. Madrid: Tecnos.

Leibniz, G.W.F. 1678/1987. The analysis of languages. In *Leibniz, Language, signs and thoughts, a collection of essays*, ed. M. Dascal. Amsterdam: John Benjamins.

Romero-Trillo, J., and T. Espigares. 1996. 'Fundamentos ecológico-lingüísticos en la percepción de los paisajes naturales. In *La Interdisciplinariedad en el Discurso Artístico: Realidad o Utopía?* vol. II, ed. J.L. Caramés and J.L. Bueno, 271–285. Spain: Editorial de la Universidad de Oviedo.

Romero-Trillo, J., and T. Espigares. 2012. The cognitive representation of natural landscapes in language. *Pragmatics and Cognition* 20: 168–185.

Turner, M.G., R.H. Gardner, and R. O'Neill. 2001. *Landscape ecology in theory and practice*. New York: Springer.

Wiens, J., and M. Moss (eds.). 2005. *Issues and perspectives in landscape ecology*. Cambridge: Cambridge University Press.

Wierzbicka, A. 1972. *Semantic primitives* (trans: Wierzbicka, A., and J. Besemeres). Frankfurt: Athenäum Verlag.

Zube, Ervin H., James L. Sell, and Jonathan G. Taylor. 1982. Landscape perception: Research, application and theory. *Landscape Planning* 9: 1–33.

System Networks as a Tool for the Pragmatic Analysis of an EFL Spoken Corpus

Silvia Riesco-Bernier

1 Introduction

1.1 Pragmatic Competence

Pragmatic competence refers to the learners' ability to employ their linguistic resources and sociocultural knowledge appropriately to instantiate a particular meaning within a given context. It thus seems essential to explore in what ways meaning and form(s) are related in language. Undertaking such task involves the analysis of how meaning is created in interaction, examine the means for speech act realisation and pay attention to the choices the speaker makes, i.e. how/why meaning is instantiated through an either/or wording (Crystal 1985; Rose and Kasper 2001; Martínez Flor 2004). In addition, given that in the English as a Foreign Language (EFL) teaching environment, students most likely only speak and listen to English in the classroom (Mattioli 2004), that input in the learning context is fundamental to learning (Long 1980, 1981; Ellis 1984; Pica and Long 1986; Coyle 2006) and that classroom interaction is typically dominated by teachers (Allwright 1999; Nystrand and Gamoran 2001), it becomes necessary to focus on the participant who provides the foreign language input in the classroom: the EFL pre-school teacher.

Directives, rather than some other acts, have been the focus for many studies because they are, according to Ervin-Tripp (1976), a substantial proportion of interactional events in young children, they are likely to be relatively sensitive to addressee features and because they often lead to action. More specifically, in classroom interaction, requests and control acts become more salient targets of

S. Riesco-Bernier (✉)
Escuela Oficial de Idiomas de Torrejón de Ardoz, Madrid, Spain

English Department, Universidad Autónoma de Madrid, Madrid, Spain
e-mail: sylvie.riesco@educa.madrid.org

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_11,

investigation than other speech acts such as apologies or compliments that may have been studied in other contexts. Consequently, directives have been examined as the way children engage in activities controlled and influenced by the teacher (Ervin-Tripp 1976, 1982) and as they are typical face-threatening acts, they "serve as a rich illustration of the interpersonal dimension of classroom language" (Dalton-Puffer 2005:126).

Framed within the Systemic Functional Linguistics paradigm (Halliday 1985; Hasan 1985, 1996; Martin 1992), this chapter provides a systematisation of meaning(s) in EFL teachers' regulatory register (cf. Christie 2000; Llinares-García 2002, 2004, 2006). This proposal specifies the semantic options made at the discourse-semantic level by creating a system network (cf. O'Donnell 1995; van Leeuwen 1996; Butt 2002) that would contribute to turn the study of regulatory functions more systematic and thus help future analysts in their ulterior investigations.

1.2 The Use of Networks to Operationalise 'Meaning'

A common denominator to studies focusing on "meaning" is the proposal and explanation of a taxonomy compiling the different types of communicative acts/functions that occur in their analysed data. However, not only do the labels differ across studies but also the criteria followed to define each act which, unfortunately, are not always explicit. Undoubtedly, this hinders comparison and generalisations of results across studies. Against an arbitrary, subjective or unsystematic analysis of meaning where "labels such as command, offer, request, etc. have been treated themselves as semantically invariant" (Hasan 1985:7), the creation of a network draws up the different criteria and variables that define each particular function to enable the analysis of texts.

> A network represents paradigms of options, and their consequences. It encompasses the meaning potential, the relevant 'phase space'. From such elaborated semiotic maps, for any given instance of meaningful behaviour in the context, we can indicate the pattern of selections which that behaviour invokes. (Moore and Butt 2002:4)

Inheritors of Firthian Linguistics, and as its very name indicates, Systemic Functional Linguistics gives priority to the system. Language is conceived as "networks of interlocking options" (Halliday 1994:xiv). A system network of meaning, for instance, presents an inventory of ways in which meaning can be realised and analysed, and where there is an array of choices that will determine which meaning is being instantiated through language. In other words, not only does the network provide the meaning potential but also prompts the researcher to examine which choices have been made in order to convey one or another meaning.

> The network is a tool for establishing what is distinctive, and what is shared, between instances of meaningful behaviour. We are highlighting actual choices and so, unlike rules and "deviations", every case study is in 'the positive'; every observed behaviour changes the probabilities for every feature node (when chosen, or not chosen). (Moore and Butt 2002:4)

1.3 Corpus Linguistics and Second Language Acquisition and Research

L2 acquisition encourages *corpus*-based studies in that it provides quantitative and probabilistic features of language and allows for quick manipulation of data by text retrieval software (Aijmer and Altenberg 1991:2). Indeed, *corpus*-based research nowadays positively influences linguistic theory and its pedagogical implications and shapes the way linguistic studies proceed. Furthermore, c*orpus-based* analyses allow the researcher to observe the learners' linguistic production and enable its comparison to that of other foreign learners' as well as to that of native speakers'. Not only does this lead researchers to draw conclusions as to frequent patterns or mistakes, but also to realise which are the real needs of a specific group of learners (cf. *Lovain International Database of Spoken English Interlanguage*, *LINDSEI*, Granger 1998).

Learner corpora are thus the electronic compilation of second or foreign language data in natural or pseudo-natural contexts, exclusively designed to study how language is acquired and developed, and then to elaborate materials for L2 or FL learning. Llinares-García (2002:164) claims that learner corpora help to describe interlanguage, make progress in second language theory and develop materials to teach foreign languages since linguists and teachers become aware of the learners' real difficulties.

2 Methodology

2.1 The Corpus

The *UAMLESC* (UAM-*Learner English Spoken Corpus*) is a longitudinal *corpus* covering the compilation of the oral interaction in the EFL classroom in different schools in Madrid where the degree of immersion, type of teacher – native vs. non-native speakers of English- and socio-economic background vary in order to investigate the acquisition and development of different linguistic aspects of English as a Foreign Language (Romero-Trillo and Llinares-García 2001, 2004; Llinares-García 2002, 2004, 2006; Ramírez-Verdugo 2003; Riesco-Bernier 2004, 2008, 2011; Riesco-Bernier and Romero-Trillo 2008a, b).

Most of the data compiled embodied natural language in the second language classroom. Teachers were not asked to carry out specific activities or change their methodology. Because the interest of the researchers lay in authentic interaction in the EFL classroom, the data recorded (SONY Handycam Video Hi8 XR) portray free discourse in the classroom.

For the present research, a sub-*corpus* of 17 recorded sessions (51,709 words) was selected from the first year of the compilation (5-year-old children, their pre-school year). This corpus comprises data from schools where non-native

teachers spoke English to children for 30 min daily (26,146 words) and data from schools where children dealt with native teachers with total immersion into English (25,563 words).

2.2 *The Tool of Analysis: The Configuration of a System Network*

Designed from the most general characteristics or features concerning an aspect of language, system networks are developed into more specific options, or subsystems. "Choice" comes into play in that the first option at the level of the most general feature will lead the speaker into a specific contrastive set of features, where only one option is to be selected. In turn, that decision will lead the speaker into a further choice, and so on until there is no further option in the path. Each of these systems or subsystems is concerned with one type of contrast or opposition and they are ordered along a scale of delicacy from left to right, whose extension depends on the researcher's will: "and we go on as far as we need to, or as far as we can in the time available or as far as we know how" (Halliday 1994:xiv).

Following the mechanics of networks (van Leeuwen 1996; Butt 2002), systems are drawn conventionally. Each system is made of a cluster of systems or subsystems which can be identified vertically and that are called "domains of contrast" or "variables". When interpreting a network, the researcher must (as the speaker unconsciously does in discourse) choose within each sub-system, conventionally in angle brackets, one single option, which is in turn indicated by square brackets.

Figure 1 below is the system of speech functions (Halliday 1985), where there are two domains of contrast ("the speaker role" and "the commodity exchanged") the speaker must consider to make a choice. Furthermore, each domain of contrast adds further levels of delicacy in contrasts of meaning, which are represented in the horizontal axis of the network and that will be referred to as "features" throughout this study. As the convention is for them to appear in square brackets, the speaker must make only one choice within the contrastive set of options. Following with the example, the speaker can either "give" or "demand" as far as the role is concerned,

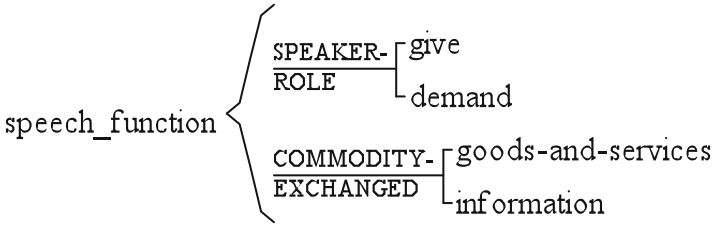


Fig. 1 Systemic network of speech functions (Halliday 1985)

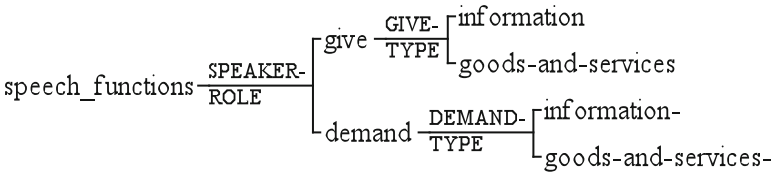


Fig. 2 Primary speech functions

and the commodity exchanged can either be "information" or "goods and services".

Therefore, meaning is the result of the choices that are made at all the levels of domains of contrast manifest within the network. The four primary speech functions result from the interaction of the two main variables and they represents a particular complex of semantic features (Fig. 2).

The speaker first chooses a ***role*** (give vs. demand), a choice that inevitably leads the speaker into a further option: the ***commodity exchanged*** (information vs. goods and services). In this way, if s/he gives information the speech function is *informing*; if the commodity is goods and services, s/he is *offering*; whereas if the speaker demands information, s/he is *questioning* and if s/he is demanding goods and services, the resulting speech function is *commanding*.

For this reason, networks stand as the graphical representation of the different options that the speaker (un)consciously makes in communication at the discourse-semantic stratum of language (instantiated through language). Likewise, networks become a tool of analysis whereby the analyst depicts the different array of choices at the discourse-semantic stratum of language, available to the speaker. This helps the researcher operationalise the study of meaning by analysing the linguistic instantiation of those semantic options at the lexicogrammatical stratum of language. It is this second approach that motivated the creation of the *Regulatory Functions System Network (RFSN)*, a tool that enables the analysis of "regulatory functions" in the EFL classroom. The Systemic Coder (Mick O'Donnell, www.wagsoft.com) was used in order to achieve the technical elaboration of the system network (cf. Fig. 3 below).

2.3 Dynamic Configuration of the Regulatory Functions System Network

System networks are dynamically created as they result from the expansion or modification of previous existing networks. Hence, the present section depicts the gradual configuration of the *RFSN*, which finds its roots in Halliday (1985), Hasan (1985) and Martin (1992). Bearing in mind that this is a *corpus*-based study, the creation of our network as a tool goes hand in hand with the qualitative analysis of

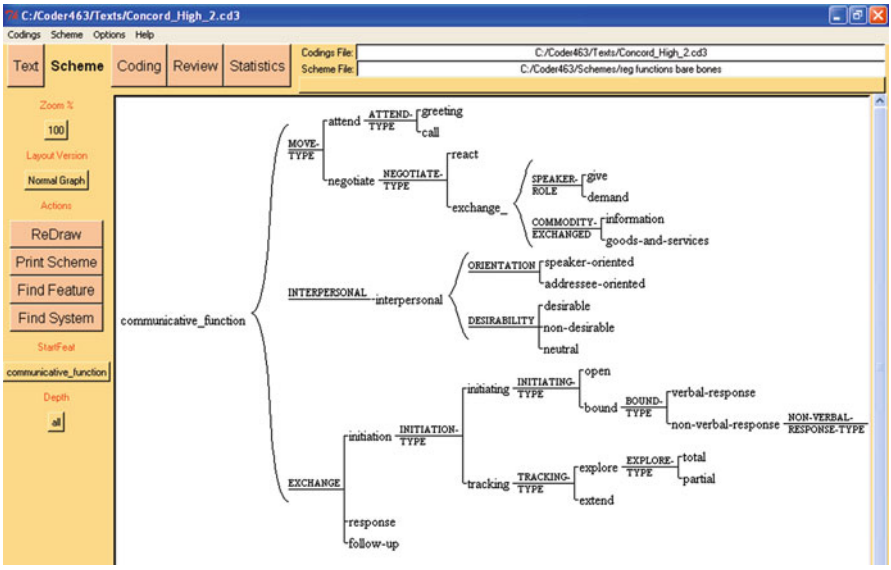


Fig. 3 Illustration of the creation of the *RFSN* by means of the *Systemic Coder Software*

the data. The discourse-semantics system presents the different choices that the speaker may make in order to convey meaning: first, each set of semantic and discursive choices creates a path in the network leading to a different regulatory function conveyed by the speaker at a discourse-semantic level, which is, in turn, instantiated through a linguistic structure at the lower layer of lexicogrammar.

"Speaking is something that might more appropriately be called an interact: it is an exchange" (Halliday 1994:68). The act of speaking thus becomes an interactive process where both participants (speaker and listener) are involved and where their roles depend on each other's, which results in a wide range of different types of "interactions" contingent on the specific context. As seen above, Halliday acknowledges that the four primary speech functions result from the interaction of the two main variables (speaker role and commodity exchanged) and they each represent a particular complex of semantic features instantiated through the Mood options at the lexicogrammatical layer (declarative vs. interrogative vs. imperative) and context (information vs. some goods and services).

Following Halliday (1994:363), two other features come into play in the definition of a vast range of speech functions: the ***orientation*** of the message (speaker-oriented vs. addressee-oriented vs. neutral), and the degree of ***desirability*** (desirable vs. non-desirable). The orientation variable specifies the direction the message follows and towards whom it is addressed, by making the focus of the message explicit (speaker vs. addressee), which is operationalised in the subject and complement choices at the lexicogrammatical stratum. The desirability variable, in turn, accounts for the degree of usefulness, necessity and worth of the message conveyed for the participants and is instantiated through polarity and modality.

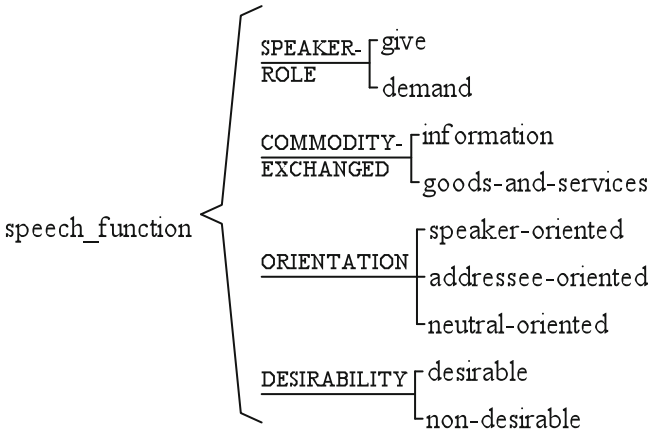


Fig. 4 On-going configuration of speech functions according to Halliday (1985)

As shown in Fig. 4, each domain of contrast implies a level of delicacy leading the speaker to choose among the options in the inventory at this semantic stratum of language: if the speaker gives information that is addressee-oriented and that is desirable for the hearer, s/he might well be praising the hearer, whereas if this is non-desirable, s/he might well be blaming or accusing the hearer.

The interest of a systemic network as a tool such as the one in Fig. 2 above lies in the degree of predictability that the analysis can reach considering the given variables (i.e. *role, commodity, orientation and desirability*). In other words, when the analyst faces an utterance and decides upon the first variable (here, *the role*) and, consequently, on the ulterior choices (in Fig. 2, the degree of delicacy appears in the vertical axis), the set of meanings is progressively more and more reduced until s/he reaches the last choice to make. It thus follows that this path drives the researcher to an explicit and distinct communicative function, which differs, in at least one feature, from the rest of the functions that the system accounts for.

The *RFSN* expanded Halliday's in order to account for the different semantic options met in our data. First, it was felt that the domain of contrast "*orientation*" was restricted to one single variable, namely, the "addressee", as regulatory functions are oriented towards *alter*. Second, the desirable/non-desirable dichotomy was further developed. Instantiated through polarity and modality, desirability is sometimes not explicit in the data, and thus not inferrable. Accordingly, in order to avoid subjectivity as much as possible when interpreting those utterances, a further feature -"neutral"- was inserted within the desirability variable in the *RFSN* (cf. Fig. 5 below).

Since "desirability" involves point of view and this investigation is centred upon classroom discourse as part of a large project where the response and/or reaction of children is of interest (*UAMLESC Corpus*), the analyst here stuck to the linguistic realisation of the message and adopted the child/learner's point of view. Therefore,

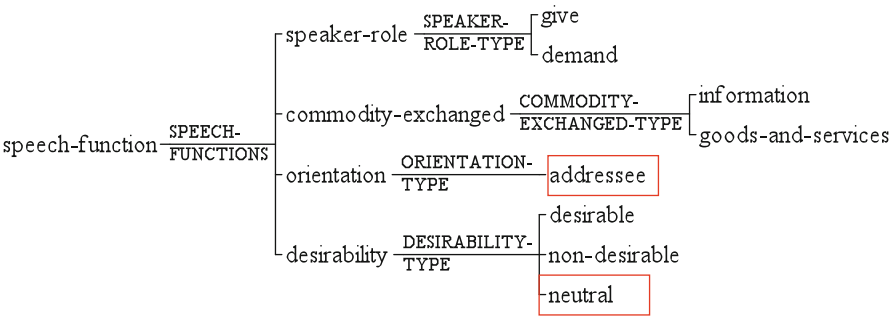


Fig. 5 *Regulatory Functions System Network*: preliminary stage

something "desirable" would mean beneficial for or wished by the learner (a message unmarkedly presented through positive polarity) as opposed to the "non-desirable" feature (unmarkedly conveying negative polarity) and the "neutral" feature (when an utterance did not overtly manifest itself as a "desirable" or "non-desirable" message to the child, see Example 1 from the corpus below).

Example 1

TCH: Ehh.. *Stand up*
everybody!
Turn around!
… Look at the wall
… Hands in front of you, stretched out!
.. Clap three times!
CH: ((They all do, some speak)) One, two, three

The variables enumerated so far belong to the ***semantic*** stratum of language, which constitutes only *one* layer of language, instantiated through lower strata (lexicogrammar and phonology). As language is a complex semiotic system composed of multiple strata, the analysis of meaning inevitably requires the exploration of language within a higher stratum: that which involves *context* as (i) the context of situation (register) and (ii) the context of culture (genre). Since register is the expression form of genre, and language, in turn, is the expression form of register (Martin 1992:495), the study of the context of situation is made feasible by examining language through the articulation of *field*, *tenor* and *mode*. The detailed analysis of the three variables guarantees the depiction of a specific situation, and system networks help in the systematisation of their study.

Whereas semantics refers to clause-size meanings and focuses on the clause, discourse-semantics focuses on text-size meanings and thus bridges text and register. In other words, discourse-semantics implies the exploration of the wording (lexicogrammar) and its meaning (semantics) within a particular context (discourse-semantics). Discourse-semantics is here regarded as the stratum in language that

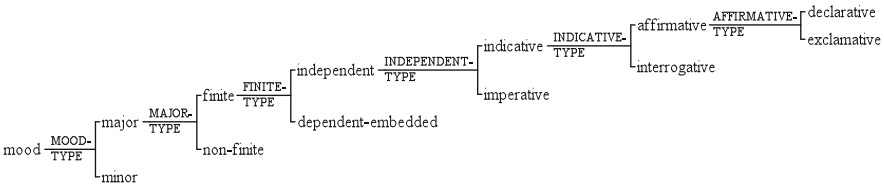


Fig. 6 Mood in English (Martin 1992:44)

focuses on the move within the exchange and that enables the researcher to depict the regulatory functions within the discursive exchange.

A preliminary review of the data, previous to the configuration of the network revealed the existence of a tendency of co-occurrence between the uttering of some words in an immediate discursive context and their association to a certain "regulatory function". For that reason, the analyst considered Martin's discourse-semantic stratum and contributed to its development by expanding the discourse-semantic variables within the *RFSN* in progress.

All the variables and features developed through the scale of delicacy in a network need to have a structural realisation, which relates the system (linguistic surface structure) to processes (meanings). So far, the *speaker role* is realised through the mood choice (declarative vs. interrogative) embodied in turn phonologically (descending tones vs. ascending tones), the *commodity exchanged* is observable in the situation, the *orientation* is made explicit through the choice of subject and complement in the mood structure (first vs. second or third person) and the degree of *desirability* is operationalised through polarity and modality in the mood system (positive vs. negative polarity; inclination vs. obligation, respectively). It thus follows that the discourse variables to be developed in this work also need to respond to a realisation that formalises their instantiation within the system.

Martin (1992) developed the system network of speech functions in discourse, instantiated by a structure at the lexicogrammatical level. Figure 6 portrays the systems of mood in English (Martin 1992) which, as will be seen later, give rise to the basic types of moves.

According to SFL, the unit of analysis for the move is the clause that independently selects for mood. More specifically, there are five different types of clauses depending on the "negotiability" of their content: (i) those whose content can be argued or negotiated about (independent clauses negotiate the content of the message through modalization and modulation), (ii) those whose content has already been negotiated (the dependent and embedded clauses), (iii) those that are in between (the hypotactically dependent clauses), (iv) those whose meaning is non-negotiable (non-finite clauses), and (v) those that, because lacking subject and finite in the mood block, cannot negotiate their meaning (minor clauses). As displayed in Fig. 6, Martin considers (1992:42) that minor clauses initiate different types of adjacency pairs (within the "attending" type of move, e.g. greetings or calls; and "reacting" towards a situation through exclamations within the "negotiating" moves). Major clauses, in turn, initiate the "exchange" moves.

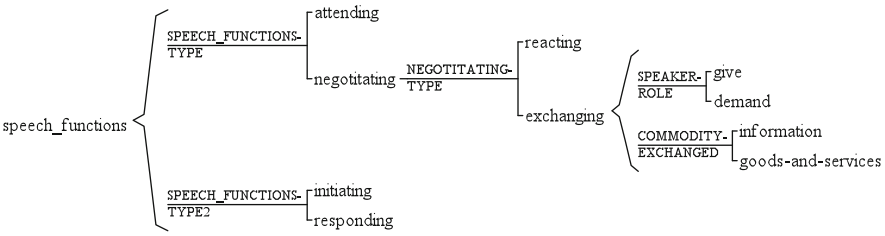


Fig. 7 Speech function network giving rise to seven adjacency pairs (Martin 1992:44)

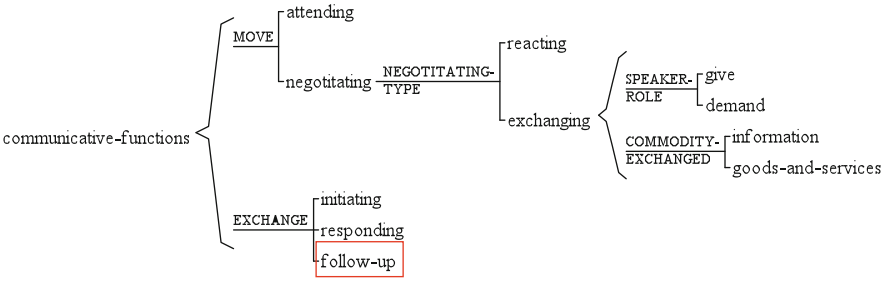


Fig. 8 Bare bones of the *Regulatory Functions System Network* (*post* Martin, *post* Sinclair and Coulthard)

Furthermore, Martin (1992) understands speech functions on a dialogic plane, i.e. in discourse. Hence, following Halliday's (1985) four basic speech functions resulting from the variables *speaker role* and *commodity exchanged* and their expected responses in interaction, Martin instantiates in his diagram the dialogic option "initiate vs. respond" making the discourse option explicit, which can be observed in Fig. 7 above.

Martin (1992) thus advances that a speech function results from the ***move type choice*** (attending vs. negotiating) and its role in the interact (initiate vs. responding).

However, the present chapter considers Sinclair and Coulthard's (1992) rank scaled analysis of discourse (lesson-transaction-exchange-move and act) where the move can be evaluated in its immediate discursive context: the ***exchange*** in classroom discourse. Among their different ranks, the exchange is the minimal interactional unit (as opposed to the interact) and is made of three moves (initiation-response-follow up), which accounts for integrating this move in the network at the exchange level and hence modify Martin's (Fig. 8 above).

Martin (1992) acknowledges two types of moves: those that are adjacent pairs (initiation-response) and those that are non-adjacent, namely the "challenging moves" (refusing attention thus having the potential to abort the exchange (Martin 1992:71)) and "the tracking moves" (interruptions produced in order to negotiate

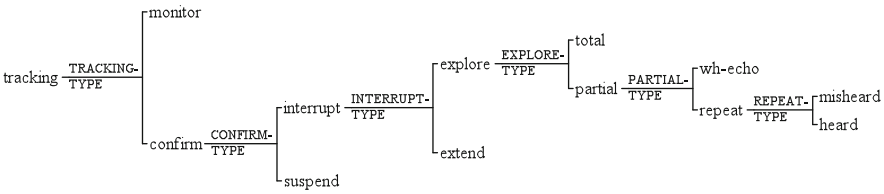


Fig. 9 Tracking moves (Martin 1992:70)

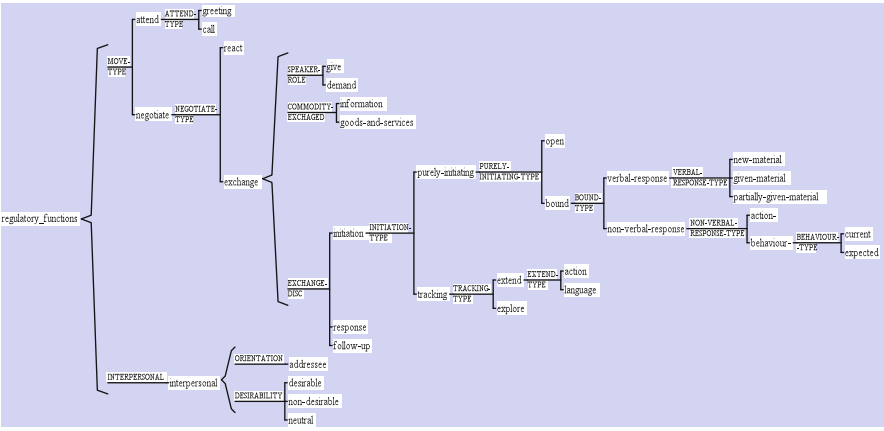


Fig. 10 Bare bones of the *RFSN*: domains of contrast

interpersonal meaning (Martin 1992:67)) either by monitoring the exchange through backchannels or by confirming what has been uttered (Fig. 9).

To adapt it to the classroom context, this study posits that there are two types of initiating moves in the exchange in the EFL classroom: *purely initiating moves* (where the teacher starts an exchange from scratch) and *tracking moves*, which aim at clarifications, replay or repetitions and that, discursively depend on the move that it is tracking (typically the immediately preceding one). It thus follows that the skeleton of the *RFSN* is made up of two domains of contrast: "interpersonal" and "move" (cf. Fig. 10). While the former involves the aforementioned purely semantic traits (*desirability* and *orientation*), the latter results from the combination of two levels that have been modified to suit the analysis of EFL classroom discourse: (i) the *move level* that considers the type of move (attend vs. negotiate) adapted from Martin's work, and (ii) the *exchange* level which considers the role of the move within the classroom discourse pattern (initiation-response-feedback), borrowed from Sinclair and Coulthard (1992), but adjusted in that it distinguishes two different types of initiating moves in the EFL classroom discourse (purely initiate vs. tracking moves (cf. Fig. 10 above)).

2.4 Presentation of the Bare Bones of the RFSN

The bare bones of the *RFSN* presented in Fig. 10 reveal the articulation of the discourse-semantic variables coming into play in the definition of regulatory functions in teacher talk. Regulatory functions are defined through variables (domains of contrast) belonging to the stratum of semantics (*interpersonal*) and discourse (*move type and exchange*). The semantic options are instantiated through the realisations at the lexicogrammatical level (mood system), and the discourse options, in turn, attend to what follows or precedes the move under analysis. What the *RFSN* offers is a systematic analysis of meaning that urges the researcher to consider both discursive and semantic criteria to identify and depict the distinct regulatory functions.

As evidenced in Fig. 10, regulatory functions are defined by making a choice within two main "domains of contrast": (i) ***move*** and (ii) ***interpersonal features***. The ***move domain*** was not modified but faithfully borrowed from those variables Halliday (1985) acknowledged (*speaker role* and *commodity*) and that were later expanded by Martin (1992). At this stage, our task has been the combination of both works into one single network as it was felt that Halliday's criteria were to be found within Martin's categorisation within move types (attend vs. negotiate…). In other words, and as illustrated in Fig. 10 above, the first step the speaker makes in interaction is to select "the move type", i.e. the attend move vs. the negotiate move, an exclusive choice which is realised by a structure at the lexicogrammatical (minor vs. major clauses in the mood system) and phonological levels (prosodic choices).

In turn, once the speaker chooses among attending or exchanging, further levels of delicacy lead the speaker to select one option within those variables: if the speaker "attends", s/he can either call or greet but if s/he "negotiates", s/he can either "react" (exclamations in mood system) or "exchange" which is defined by the speaker role (seen in a mood and phonological choice: declarative vs. interrogative; descending vs. ascending tones) and the commodity exchanged (information vs. goods and services).

As Fig. 10 shows, within *Negotiate Moves*, one can find the *Reacting moves* where the speaker does not properly interact with the interlocutor (usually instantiated by one independent move, not in adjacency pairs, e.g. exclamations) and the *Exchange moves*. The latter are those constituting the main body of an interaction since the speaker thereby makes his/her role explicit and exchanges the basic commodities.

When the speaker actually moves onto the exchange domain of contrast, s/he instantiates his/her move as an initiating, responding or following-up move. The teacher initiates when s/he opens the exchange. In the EFL classroom, it was found that teacher initiations could either purely initiate or belong to what Martin (1992:70) presented as tracking moves. On the one hand, within pure initiations, the system network is expanded taking into consideration that initiations in teacher talk

either expect some kind of response (*bound* option) or do not (*open* option). Within the bound options, and as Fig. 10 illustrates, two major types of responses prevail: *non-verbal* (i.e. action or behaviour change, e.g. to sit, to cut) vs. *verbal* (i.e. linguistic production demanded).

One of the major innovations of the present research results from understanding "language in the L2" as a type of "goods and services" in the EFL classroom context. It thus follows that an utterance bounded to a verbal response can be interpreted as a "request of verbal production" rather than a "demand for information". Indeed, as the analysis of EFL classroom discourse reveals, most of the activities focus on "linguistic" tasks: e.g. making children repeat a new item in the foreign language, eliciting peer conversation in the foreign language, among others. Therefore, the nature of the response, verbal (aiming at language) vs. non-verbal (aiming at action), invites the researcher to further considerations so as to obtain an either/or categorisation of the different types of regulatory functions in the EFL classroom.

Consequently, and as Fig. 10 illustrates, one further level of delicacy was developed in order to discriminate distinct types of verbal responses. The *informational status* constitutes a useful discursive criterion in the definition of functions related to linguistic production. According to Halliday (1967), Prince (1981) and Geluykens (1991) among others, informational status should be understood as the givenness-newness opposition, on the grounds of recoverability at the discourse level.

Accordingly, as Fig. 10 portrays it, the type of discourse in the EFL classroom may be (i) "new" when the teacher obtains a child's L2 production which has not been previously provided by the teacher (Example 2 from the corpus), (ii) "partially given" when the child uses some cue or discourse uttered by the teacher in the immediately preceding discourse (Example 3 below), and (iii) "given" when the child echoes with the identical words what has been produced by the teacher (Example 4 below).

Example 2: [session NrK]

TCH: ***What´s this***
Alejandra?
Alejandra: a fish
TCH: a fish.
And where do they live?
Alejandra: in the sea

Example 3: [session NNcT2]

What are they?
CH: (Alberto) Fingers.
TCH: Not fingers.. These are the fingers and these ((ref. To the gloves)) you put them on, like this ((showing))

CH: (Alberto)<L1 No es mío L1>
TCH: I know it's not yours.. but she can't remember.
CH: <L1 No me acuerdo L1>
TCH: <L1 ¡Ayy! No me acuerdo L1>..
What are they?
Miiii-
CH: ((the girl)) Mittens
TCH: Mittens, mittens.. Very good..

Example 4: [session NskJ]

CH: <L1 Piña L1>
TCH: Okay?
CH: Okay!
TCH: It's a pineapple.
CH: It's a pineapple.
TCH: ***Repeat! Pineapple***
CH: pineapple
((The children do not repeat it very properly))
… Very good, María ..

On the other hand, taking into consideration the two functions which *tracking moves* may have, namely "explore" and "extend" the move that is tracking, the *RFSN* further developed the "tracking: extend" initiation feature.[1] Since the present investigation acknowledges two types of responses (verbal vs. non-verbal), it is here understood that there are two types of extensions: those that would encourage the child to produce further verbal production (Example 5 below) and those that would encourage further actions (Example 6 below).

Example 5: [session: NNncN1]

TCH: and now,
what's this?
CHI: yellow
TCH: ***come on***
aloud
what's this?
blue door?
CHI: nooo
CHI: purple

[1]Note that Martin's option "tracking: explore" is disregarded in the present study as that exclusively applies to the "information" commodity and this investigation focuses on the "goods and services" commodity instead.

Example 6: [session NNncS3]

Come on
go to the blackboard < *DC-a* > *$C-IM-p-Fp-Rp-Radj$.*
Miguel Angel < AS > $MC-V$

Finally, as far as the ***interpersonal domain*** is concerned, the present work has borrowed the *degree of desirability* and *orientation* from Halliday (1994) since it is felt that both contribute to the definition of regulatory functions in classroom discourse. However, and contrary to the way Halliday considers them, these criteria do not appear under the exchange type exclusively but are contemplated at any move type, becoming therefore a second domain of contrast itself (see Fig. 10 above). As it can be observed, they are grouped under the domain of contrast "interpersonal" as I feel they both contribute to the explicitness of the relationship that the message can create between the participants (mainly realised by the structure of polarity and modality at the lexicogrammatical stratum of language). As Fig. 10 portrays with the angle brackets, once the speaker has chosen the ***move type***, s/he enters the ***interpersonal*** domain and makes an option both at the *orientation* and *desirability* of the message, which have been modified and explained above.

3 Results

3.1 The Regulatory Functions System Network

This section presents the *RFSN* expanded and developed through the scale of delicacy. Figure 11 below must be read from left (the most general characteristic where the first choice is made) to right. In other words, the domains of contrast are arranged along a horizontal axis. The analyst (as the speaker in communication) makes a choice at the first level of delicacy, i.e. the move type in this case, then follows the path choosing one option within each variable (signalled through square brackets) and does so until no further choice exists. For presentation purposes, the paths leading to the regulatory functions that obtained in this *corpus* have been developed and the name of each regulatory function appears at the end.

3.2 The System Network Consistency

The goal of educational research is basically to produce descriptions and interpretations of classroom events that will be identified by others as real and meaningful for teachers, learners and the learning process. Consequently, researchers should consider the *reliability* of their instrument, i.e. the consistency to which others agree on the categories and descriptions and the frequencies attributed to them, that is, the degree to which they are free of error of measurement (cf. Brown 1988:98;

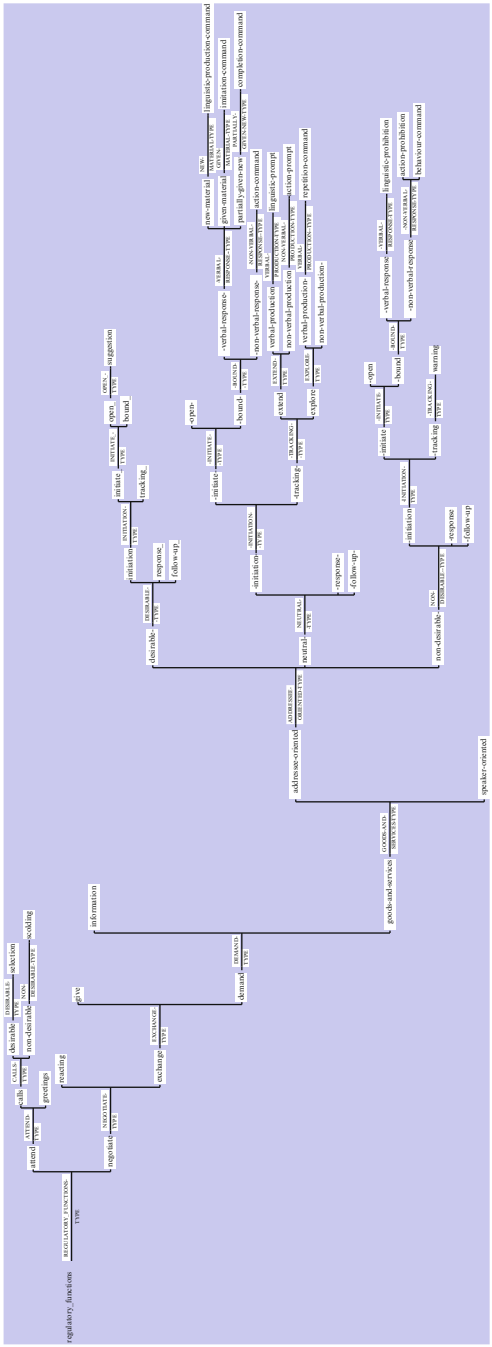


Fig. 11 *Regulatory Functions System Network*: developed scale of delicacy

Chaudron 1988:23). Additionally, linguists are interested in the generalisability of their claims, i.e. the extent to which their conclusions can be meaningful, significant and applicable to future studies in the classroom (namely, *validity*).

3.2.1 Reliability

Among the distinct types of reliability tests, Krippendorff (1980:131) acknowledges "stability" (the degree to which a process is invariant over time), "reproducibility" (the degree to which a process can be recreated under varying circumstances, using different coders) and "accuracy" (the degree to which a process functionally conforms to a known standard, i.e. where the coders' judgements are compared to a standard). Potter and Levine-Donnerstein (1999) agree with Krippendorff that *accuracy* and sometimes *reproducibility* are the strongest procedures, two measures which inextricably call for intercoder reliability.

"Intercoder reliability" commonly arises in language studies (Frick and Semmel 1978; Llinares-García 2002; Murphy 2004) and is used to assess the extent to which independent coders evaluate a characteristic of a message and reach the same conclusion (Lombard et al. 2002) or the extent to which the different judges tend to assign exactly the same rating to each object (Tinsley and Weiss 2000). The degree to which an instrument, in this case the *RFSN*, is reliable is therefore estimated with a reliability coefficient.

Intercoder reliability is obtained by having two or more coders categorise units (in this case, regulatory functions), and then using these categorisations to calculate a numerical index of the extent of agreement between or among the coders (cf. Lombard et al. 2002:590). Several operational considerations provide a guide to design such test (cf. Holsti 1969; Krippendorff 1980; Popping 1988; Potter and Levine-Donnerstein 1999). This study thus considers the issue of how an overlap of coders was designed when setting up the reliability test, examines the degree of reliability and later on adjusts those percentages of agreement for chance so as to get a reliability coefficient.

In order to test the degree of consistency in decision making across coders, there must be some overlap in the coding, that is, at least two coders must make judgements on the same material (cf. Chaudron 1988; Llinares-García 2002). As regards the size for an overlap, Potter and Ware (1987) and Ader (1995) among others, used a 10 % overlap in their analysis of big *corpora* (88 h of news or 2,000 newspapers stories, respectively). Taking into account that the present research analysed 4,259 regulatory functions in a 51,000 word *corpus*, it was decided that both coders would be given a sample that represents 10 % of the total *corpus*.

Admittedly, the sample must be randomly chosen in order for the selected cases in the reliability test to represent the entire *corpus*. However, the selection of the three different sessions (made up of distinct fragments) of the *corpus* was made on the basis of the following criteria in order to guarantee uniformity of coding challenge: (i) since there are 15 distinct regulatory functions, each session contained 10 different functions at least; (ii) each function appeared five times at least; and

(iii) in different lexicogrammatical realisations as those meant different degrees of difficulty in the coding.

A coder-training session was used so as to establish reliability levels for the codings of the sample. That session with the two external coders introduced them to (i) the notion of regulatory functions; (ii) the dynamics of the *RFSN* and (iii) the resulting taxonomy of the distinct regulatory functions. A short extract was analysed together so as to establish the criteria to analyse the data and categorise the distinct regulatory functions. Only then were the external coders given the samples to codify and told to use the *RFSN* as a tool that indicates the path leading to a particular function (ignoring the lexico-grammatical realisation as far as possible).

The external coders worked on their own independently. Later, two meetings were necessary in order to carry out the intercoder reliability test: both coders brought their codings and had to go throughout their analyses to check whether they agreed on the tag that they had assigned to each function individually. Instructions asked them to discuss those instances where no agreement existed so as to (hopefully) reach a common category.

To determine the level of reliability of the *RFSN* instrument, the analysis considered first the degree of agreement between the external coders. A frequent procedure for computing a reliability coefficient is to find the percentage among coders and then correct for chance agreement by employing Cohen's *Kappa*, commonly used in language studies (Dewey 1983; cf. Palmer and Simmons 1995).

Understanding that the present index, as other coefficients, range from 0 to 1 (1 standing for perfect agreement), the obtained coefficient in this study (=.829) reveals a significantly strong agreement between the coders. Neuendorf (2002:145) claimed that "coefficients of .90 or greater would be acceptable to all, .80 or more would be acceptable in most situations, and below that, there exists great disagreement" (*ibid.*). Furthermore, Lombard et al. (2002:593) underline that .70 is often used in exploratory research.

3.2.2 Validity

"The issue of reliability is seen as a necessary but not sufficient condition for validity, that is, reliability is a necessary precondition for validity" (Potter and Levine-Donnerstein 1999:272). Assessing validity is indeed best regarded as a two-step process. The first step is to develop the *RFSN* as a coding scheme that consists of rules that reduce the complexity of all the attributes present in a phenomenon down into a limited and manageable set of regulatory functions.

The second step, in turn, is to assess the decisions made by coders against some standard that serves as a basis to compare codings. The standard is understood as "the correct" or "accurate" set of codes (cf. Folger et al. 1984; Wimmer and Dominick 1991). While codings that deviate from the standard, vary in inaccuracy, the codes that match the standard for correct decision making are regarded as producing valid data (cf. Potter and Levine-Donnerstein 1999).

Working with two qualitative variables (the researcher's analysis vs. the coders' version reaching an agreement after the joint session), the coders-researcher reliability coefficient was calculated. The obtained coefficient (=.880) reveals a significantly strong agreement between the coders and the standard ($p = .001$). Going back to the acceptable level of reliability mentioned above, it can here be stated that the analysis of the coders in relation to the standard is almost acceptable to all.

In summary, it is hereby claimed that the coding system emerging from the RFSN is valid in that the coding scheme has laid out variables, definitions and rules for recognising these variables in the content being coded. Additionally, the coding decisions made by the coders have been compared against the standard established by the researcher and their degree of agreement is almost acceptable to all.

4 Discussion

4.1 The Appropriateness of the RFSN Within the Construct of the Nature of Language

Among the benefits of analysing meaning through a *network approach,* this study highlights its relevance for practice-oriented linguistics since (i) it reflects the nature of language itself, (ii) it constitutes a productive generator of meaning since the metaphor of choice applies to all levels of representation and (iii) it can be tested and validated by practitioners.

The nature of language has been understood in this paper within the framework of *Systemic Functional Linguistics* (Halliday 1985; Martin 1992) and *Cognitive Linguistics* (Langacker 1986, 1987; Radden 1992; Bernárdez 1999; Lakoff and Johnson 1999). Though many differences exist between them, both understand language as a *complex* and *dynamic* entity governed by "constituency", whereby language is made of modules, one inserted into another and where all interact in the process of communication. Adopting the graphical representation of constituency (Martin 1992:496), language is interpreted as a series of concentric circles where the largest circle (here the discourse-semantic layer) comprises the smaller ones (here, the lexicogrammatical layer), the boundaries of which are fuzzy, in constant fluctuation and contact with each other, which again responds to the *naturalness* of language. Further, the inextricable interrelationship of the different strata of language is also graphically suggested by the "network" itself. *CL* uses the metaphor of "connectionism" where the mind is viewed as a network of neurons all engaged in reciprocal interactions via their connections with surrounding neurons and neuronal layers and thus considers the dimensions of human thought, emotion, language and non-verbal behaviour as globally and inextricably correlated:

> In a connectionist approach, such traditional linguistic domains as phonology and semantics operate not as separate modular processes activated serially but concurrently and in parallel, each subject to its own constraints (rules) and to other constraints arising from related dimensions. (Palmer 1996:32)

It thus follows that the analysis of language cannot be achieved at the different levels of description responding to a compartmentalised view of language (Radden 1992:531), but constitutes a *holistic* task where the study of one stratum undoubtedly leads to the consideration of the other counterparts.

The *RFSN* has focused on the discourse-semantic level and constitutes a tool to analyse meaning instantiated through structure, understanding that all strata also play a role (and interplay) in communication. In this sense, the *RFSN* betokens a tool enabling the researchers and practitioners to describe and analyse meaning since the discourse-semantic criteria are materialised through linguistic realisations, which can be observed and objectively studied. Consequently, each domain of contrast developed through the scale of delicacy is linguistically and discursively operationalised, which allows a systematisation of the analysis of meaning and the comparison of results across studies. Although it is here argued that the *RFSN* is the tool bridging the discourse-semantics and lexicogrammar strata, and that the different domains of contrast *must* be instantiated through linguistic structures (Hasan 1985, 1996; Martin 1992; Butt 2002), this does not imply that there exists a determining and unequivocal relationship between the "structure" displayed and the "meaning" conveyed. Instead, the system of Mood (lower stratum of language) provides the resources to instantiate the different domains of contrast existing in the *RFSN* (e.g. "polarity" and "subject" instantiate "desirability" and "orientation", respectively). Recovering the *symbolic* nature of language, it can be claimed that the grammar of a language is "merely providing the speaker with an inventory of symbolic resources, among them schematic templates representing established patterns in the assembly of complex symbolic structures" (Langacker 1986:17), through which meaning can be conveyed.

Additionally, the symbolism of language runs in parallel with two other properties: its *creative* and *productive* potential, which are again reflected in the "network approach". Taking into account that language consists of a finite set of rules and symbols that can be combined to produce a non-finite set of meanings, the network arises as a tool that also generates meaning. A close look at the *RFSN* reveals that the researcher has only developed those pathways in the network that are initiating moves within teacher talk. In so doing, the *RFSN* leaves the way open to explore other discursive options within teacher talk (e.g. responsive or follow-up moves). The network represents the meaning potential since it can be gradually developed by researchers in accordance to their aims and necessities. It could be argued that if the network is regarded as a generator of meaning, its source of energy lies in "choice".

The *RFSN* has been presented as an array of choices at the discourse-semantic level of language where the first choice (move type, i.e. "attend" vs. "negotiate") displays a whole range of communicative options that gradually become more and more restricted when progressive choices are made since the speaker travels throughout the map of meaning through delicacy levels. In other words, each choice leads to further options among which the researcher (as the speaker in communication) must make an exclusive selection, which then reduces the possibilities since the degrees of delicacy imply becoming more and more specific in communication.

Choice does not only apply to the highest layer of language though. In fact, the speaker has at hand the *Mood* system network the speaker might use to shape his/her message. It is because "choice" can be used in all levels of representation that the same regulatory function can be instantiated by two different linguistic realisations. Likewise, bearing in mind that the different strata interact in the communicative act, it is possible to explain why the same linguistic lexicogrammatical structure may convey two different functions, since it is "choice" at the discursive stratum that might have shaped the utterance as a request rather than a question, for instance, despite its interrogative surface structure. Consequently, the present investigation has demonstrated that each defined regulatory function is the result of a selected pathway.

4.2 Appraisal of the Reliability and Validity of the RFSN

Once it has been argued that the design of the tool has been done by confronting foundational/theoretical issues, the challenges of assessing reliability and validity become more manageable:

> Content analyses need not be limited to theory-based coding schemes and standards set by experts. When researchers are clear about *what kind of content they want to analyze and the role of theory in their studies*, they are in a better position to select the most appropriate strategies for demonstrating validity and reliability. (Potter and Levine-Donnerstein 1999:258, my italics)

Statistically speaking, the results have revealed that a significantly strong agreement was reached between the coders' analyses of the data, and between the coders' final joint version and the standard. Thus, the *RFSN* constitutes a reliable and valid tool for the analysis of regulatory functions in teacher talk in the EFL classroom as it provides any analyst with the necessary discourse-semantic criteria to identify regulatory functions in EFL teacher talk. Furthermore, the training session held by the researcher and the external coders helped them learn about how to read the *RFSN* so as to categorise their utterances into the discrete regulatory functions. Since the coders were asked to disregard the lexicogrammatical form at first, attending to the discourse-semantic features exclusively, their agreement (>82 % of the cases) reveals that the *RFSN* provides a systematic way of analysing meaning in that it generates identical analyses of meaning.

Interestingly, the *RFSN* also helped the coders agree over controversial instances. More specifically, some utterances were tagged differently by the two coders (e.g. some coder 1's "linguistic completion commands" were interpreted as "linguistic production command" by coder 2). However, during their joint session, the coders decided to adhere to what they had in common, i.e. the tool of analysis. Hence, they examined each controversial utterance and analysed it by considering the features that are explicit in the system network. In so doing, the coders literally worked with an instrument of analysis that ultimately led them to achieve some consensus.

This may well be a major accomplishment since the *RFSN* opens the door to examine content, invites analysts to become aware of the decisions that are made in their analysis of meaning and helps them reach an agreement by having a common systematic procedure of analysis at hand.

Validity, in turn, tests to what extent the coders' final version of their analyses echoes the standard and thus ensures generalisation of the results. In other words, since the percentage of agreement reaches .880, the *RFSN* can be claimed to be statistically valid in that it implies that the results found by the researcher would also be found by other analysts working with the *RFSN*. Besides, a qualitative analysis of the data also supports the statistical validity. In our view, it is particularly interesting to note that the concurrence obtained in the coders' and researcher's analyses reveals that the criteria stated by the *RFSN* have been adopted and followed by the external coders in almost all the instances. Particularly, it reveals that the coders have not been misled by the versatile linguistic surface structure of those categories and have stuck to the discourse-semantic criteria specified in the *RFSN*.

5 Concluding Remarks

Networks "are a context-sensitive, empirically driven, and relatively direct way of representing these different strands of meaning in a critical context" (Moore and Butt 2002:1) and they respond to the specific necessities of a particular register. While current studies analyse meaning and speech roles/functions at one stratum, (Moore and Butt 2002) or across strata (Van Leeuwen 1996; Martin 2000) in different contexts (Perrett 2000; Tuckwell 2002), no attempt has been made to cover EFL teacher talk, a gap that this study has tried to fill through the elaboration of the *RFSN*. This network has been designed by modelling the existing domains of contrast in previous systems in order to fit classroom discourse. In this sense, neutral desirability and the exchange type features were inserted in the semantic and discursive domains of contrast so as to satisfy the requirements of a particular register: the EFL classroom.

The implications of this study are theoretical and practical in that it is the first time the dynamics of system networks within the Systemic-Functional model is applied to the configuration of an instrument that enables the analysis of spoken data in the EFL pre-school classroom. On the one hand, it has provided a taxonomy of regulatory functions through the explicitation of their inherent characteristics and features, which invites other linguists to consider those criteria in their analysis of regulatory functions, regardless of their nomenclature. On the other hand, it does not present a finite set of options, thus limited to the data analysed in the present work (e.g. regulatory functions). Instead, the RFSN can be expanded and endlessly developed by practitioners in search of new choices and pathways that best portray their object of study.

References

Ader, C.R. 1995. A longitudinal study of agenda setting for the issue of environmental pollution. *Journalism and Mass Communication Quarterly* 72: 300–331.

Aijmer, K., and B. Altenberg (eds.). 1991. *English Corpus linguistics*. London: Longman.

Allwright, D. 1999. Discourse in the language classroom. In *Concise encyclopedia of educational linguistics*, ed. B. Spolsky, 319–323. Oxford: Elsevier.

Bernárdez, E. 1999. Some reflections on the origins of cognitive linguistics. *Journal of English Studies* 1: 9–27.

Brown, J.D. 1988. *Understanding research in second language learning. A teacher's guide to statistics and research design*. Cambridge: Cambridge University Press.

Butt, D.G. 2002. (permission requested). *Conversations on language, brain, culture: Parameters of context*. Centre for Language in Social Life. Macquarie University.

Chaudron, C. 1988. *Second language classrooms: Research on teaching and learning*. Cambridge: Cambridge University Press.

Christie, F. 2000. The language of classroom interaction and learning. In *Researching language in schools and communities. Functional linguistic perspectives*, ed. L. Unsworth, 184–204. London: Cassell.

Coyle, D. 2006. Content and language integrated learning. Motivating learners and teachers. CLIL teachers tool kit: A classroom guide. http://blocs.xtec.cat/clilpractiques1/files/2008/11/slrcoyle.pdf. Accessed 18 Dec 2012.

Crystal, D. 1985. *A dictionary of linguistics and phonetics*, 2nd ed. Oxford: Blackwell.

Dalton-Puffer, C. 2005. Negotiating interpersonal meanings in naturalistic classroom discourse: Directives in content-and-language-integrated classrooms. *Journal of Pragmatics* 37: 1275–1293.

Dewey, M.E. 1983. Coefficients of agreement. *The British Journal of Psychiatry* 143: 487–489.

Ellis, R. 1984. *Classroom second language acquisition*. New York: Pergamon Press.

Ervin-Tripp, S. 1976. Is Sybil there? The structure of some American English directives. *Language in Society* 5: 25–66.

Ervin-Tripp, S. 1982. Structures of control. In *Communicating in the classroom*, ed. L.Ch. Wilkinson, 27–47. New York: Academic.

Folger, J.P., D.E. Hewes, and M.S. Poole. 1984. Coding social interaction. In *Progress in communication sciences*, vol. 4, ed. B. Dervin and M.J. Voigt, 115–161. Norwood: Ablex.

Frick, T., and M.I. Semmel. 1978. Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research* 48(1): 157–184.

Geluykens, R. 1991. Information flow in English conversation: A new approach to the given-new distinction. In *Trends in linguistics 55: Functional and systemic linguistics. Approaches and uses*, ed. E. Ventola, 141–169. Berlin: Mouton de Gruyter.

Granger, S. (ed.). 1998. *Learner English on computer*. London: Longman.

Halliday, M.A.K. 1967. Notes on transitivity and theme in English. *Journal of Linguistics* 3: 177–274 (Part II).

Halliday, M.A.K. 1985/1994/2004. *An introduction to functional grammar*. London: Edward Arnold.

Hasan, R. 1985. Offers in the making: A systemic-functional approach. Revised and enlarged version of a paper presented at the XIIth Systemic Workshop, Michigan, August 1985.

Hasan, R. 1996. Semantic networks: A tool for the analysis of meaning. In *Ways of saying: Ways of meaning. Selected papers of Ruqaiya Hasan*, ed. C. Cloran, D. Butt, and G. Williams, 105–131. London: Cassell.

Holsti, O.R. 1969. *Content analysis for the social sciences and humanities*. Reading: Addison-Wesley.

Krippendorff, K. 1980. *Content analysis: An introduction to its methodology*. Beverly Hills: Sage.

Lakoff, G., and M. Johnson. 1999. *Chomsky's philosophy and cognitive linguistics in philosophy in the flesh*, 469–512. New York: Basic Books.

Langacker, R. 1986. An introduction to cognitive grammar. *Cognitive Science* 10: 1–40.
Langacker, R. 1987. *Guiding assumptions. Foundations of cognitive grammar*, vol. I, 11–55. Stanford: Stanford University Press.
Llinares-García, A. 2002 La interacción lingüística en el aula de segundas lenguas en edades tempranas: análisis de un corpus desde una perspectiva funcional. Dissertation, Universidad Autónoma de Madrid.
Llinares-García, A. 2004. La función reguladora en el lenguaje de profesores y alumnos en el aula bilingüe en edades tempranas. In *Bingualar Unterricht in Spanien and Deutschland*, ed. W. Altmann, 163–178. Berlin: Welter Frey.
Llinares-García, A. 2006. A pragmatic analysis of children's interlanguage in EFL preschool contexts. *Intercultural Pragmatics* 3(2): 171–193.
Lombard, M., J. Snyder-Duch, and C. Campanella-Bracken. 2002. Content analysis in mass communication assessment and reporting of intercoder reliability. *Human Communication Research* 28(4): 587–604.
Long, M.H. 1980. Input, interaction and second language acquisition. Dissertation, University of California at Los Angeles.
Long, M.H. 1981. Input, interaction and second language acquisition. In *Native language and foreign language acquisition*, ed. H. Winitz. Annals of the New York Academy of Sciences 379: 259–278.
Martin, J.R. 1992. *English text: System and structure*. Amsterdam: John Benjamins.
Martin, J.R. 2000. Beyond exchange: Appraisal systems in English. In *Evaluation in text: Authorial stance and the construction of discourse*, ed. S. Hunston and G. Thompson, 142–175. Oxford: Oxford University Press.
Martínez Flor, A. 2004. The effect of instruction on the development of pragmatic competence in the English as a foreign language context: A study based on suggestions. Dissertation, Universitat Jaime I.
Mattioli, G. 2004. On native language intrusions and making do with words: Linguistically homogeneous classrooms and native language use. *English Teaching Forum* 2(43): 20–25.
Moore, A., Butt, D. 2002. Risk and meaning potential: Why a network? Invited colloquium on explaining risk. Sociolinguistics Symposium 14, Ghent, 4–6 Apr 2002.
Murphy, E. 2004. Promoting construct validity in instruments for the analysis of transcripts of online asynchronous discussions. *Educational Media International* 41(4): 346–354.
Neuendorf, K.A. 2002. *The content analysis guidebook*. Thousand Oaks: Sage.
Nystrand, M., Gamoran, A. 2001. Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. CELA Research Report nº 14005. http://www.albany.edu/cela/reports/nystrand/nystrandquestions14005.pdf. Accessed 19 Dec 2012.
O'Donnell, M. 1995. From corpus to coding: Semi-automating the acquisition of linguistic features. In: Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, 27–29 Mar 1995. California: Stanford University.
Palmer, G.B. 1996. The emergence of cognitive linguistics. In *Toward a theory of cultural linguistics*, ed. G.B. Palmer, 27–34. Austin: Texas University Press.
Palmer, M.T., and K.B. Simmons. 1995. Communicating intentions through nonverbal behaviors: Conscious and nonconscious encoding of liking. *Human Communication Research* 22: 128–160.
Perrett, G. 2000. Researching second and foreign language development. In *Researching language in schools and communities. Functional linguistic perspectives*, ed. L. Unsworth, 87–111. London: Cassell.
Pica, T., and M.H. Long. 1986. The linguistic and conversational performance of experienced and inexperienced teachers. In *Talking to learn: Conversation in second language acquisition*, ed. R. Day. Rowley: Newbury House.
Popping, R. 1988. On agreement indices for nominal data. In *Sociometric research. Vol. 1: Data collection and scaling*, ed. W.E. Saris and I.N. Gallhofer, 90–105. New York: St. Martin's Press.

Potter, W.J., and D. Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research* 27: 258–284.

Potter, W.J., and W. Ware. 1987. An analysis of the contexts of antisocial acts on prime-time television. *Communication Research* 14: 664–686.

Prince, E.F. 1981. Toward a taxonomy of given-new information. In *Radical pragmatics*, ed. P. Cole, 223–255. London: Academic.

Radden, G. 1992. The cognitive approach to natural language. In *Thirty years of linguistic evolution*, ed. M. Pütz, 513–541. Philadelphia: John Benjamins.

Ramírez Verdugo, M.D. 2003. Análisis contrastivo de los sistemas entonativos del inglés en la interlengua de hablantes no nativos. Implicación en la organización de la información desde la perspectiva funcional. Estudio basado en un corpus computerizado de aprendices españoles de lengua inglesa". Tesis Doctoral. Universidad Autónoma de Madrid.

Riesco-Bernier, S. 2004. 'There's more to listen to!' Speech functions revealed by tonicity variation in EFL pre-school teacher talk. In *Linguistic perspectives from the classroom: Language teaching in a multicultural Europe*, ed. J. Anderson, J.M. Oro, and J. Varela Zapata, 273–284. Santiago de Compostela: Universidad Santiago de Compostela.

Riesco-Bernier, S. 2008. The discourse-grammar interface of regulatory teacher talk in the EFL classroom. In *Pragmatics and Corpus linguistics: A mutualistic entente*, ed. J. Romero-Trillo, 233–260. New York: Mouton de Gruyter.

Riesco-Bernier, S. 2011. Same but different: The pragmatic potencial of native vs. non-native teachers' intonation in the EFL classroom. In *Pragmatics and Prosody in English language teaching*, ed. J. Romero-Trillo, 171–199. New York: Springer.

Riesco-Bernier, S., and J. Romero-Trillo. 2008a. The acoustics of 'Newness' and its pragmatic implications in classroom discourse. *Journal of Pragmatics* 40: 1103–1116.

Riesco-Bernier, S., and J. Romero-Trillo. 2008b. What does the tonic say in pre-school teacher talk in the EFL classroom? An acoustic-based analysis of tonicity. In *Institutional Discourse in Cross-Cultural Contexts*, Lincom studies in pragmatics 14, ed. R. Geluykens and B. Kraft, 147–169. München: Lincom Europa.

Romero-Trillo, J., and A. Llinares-García. 2001. Communicative constraints in native/non-native preschool settings. *International Journal of Corpus Linguistics* 6(1): 27–46.

Romero-Trillo, J., and A. Llinares-García. 2004. Prosodic competence in reading aloud: an acoustic corpus-based study of native and non-native (Spanish) speakers of English. *Estudios Ingleses de la Universidad Complutense* 12: 63–77.

Rose, K.R., and G. Kasper (eds.). 2001. *Pragmatics in language teaching*. Cambridge: Cambridge University Press.

Sinclair, J.Mc.H., and M. Coulthard. 1975/1992. *Towards an analysis of discourse. The English used by teachers and pupils*. London: Oxford University Press.

Tinsley, H.E.A., and D.J. Weiss. 2000. Interrater reliability and agreement. In *Handbook of applied multivariate statistics and mathematical modeling*, ed. H.E.A. Tinsley and S.D. Brown, 95–124. San Diego: Academic.

Tuckwell, K. 2002. Semantic networks for 'questioning': beyond the clause? Paper presented at the Seminar at Macquarie University, Australia, 16 Aug 2002.

van Leeuwen, T. 1996. The representation of social actors. In *Text and practices. Readings in critical discourse analysis*, ed. R. Caldas and M. Coulthard, 32–70. London: Routledge.

Wimmer, R.D., and J.R. Dominick. 1991. *Mass media research: An introduction*. Belmont: Wadsworth.

A Cultural Semantic and Ethnopragmatic Analysis of the Russian Praise Words *Molodec* and *Umnica* (with Reference to English and Chinese)

Anna Gladkova

1 Introduction

The Russian words *molodec* (roughly, 'fine fellow') and *umnica* (roughly, 'good/clever boy/girl') belong to a class of vocabulary that can be labelled as 'praise words' in English and *poxvala* 'praise' in Russian. Praise can be defined, roughly, as a verbal expression that gives a positive evaluation of someone's action or quality, or of a person in general. Praise embeds the idea that a person's action or quality surpasses the norm or normal expectations. It also involves a component of a positive emotional attitude.

Praise in general is a relatively common way of expressing one's attitude in Russian. Its prevalence can be associated with the 'emotionality' that is characteristic of Russian communicative behaviour. This feature of Russian culture has been identified in numerous linguistic studies (e.g., Wierzbicka 1992, 1998, 2002; Apresjan 1997; Pavlenko 2002; Gladkova 2010a, b, c), as well as in studies by anthropologists, cross-cultural psychologists and cultural historians (e.g., Ries 1997; Pesmen 2000; Visson 2001; Richmond 2009).

The Russian language provides a variety of means to express praise. To praise someone's action one can use adverbs *Prekrasno!* 'Excellent/splendid!', *Zamečatel'no!* 'Splendid!', *Zdorovo!* 'Well done!', *Otlično!* 'Excellent!', *Vosxititel'no!* 'Delightful' or *Bravo!* 'Bravo!'. These expressions provide an evaluation of an action without describing the person in general. In some situations one can praise someone's qualities by using adjectives in exclamatory sentences – for example, *Ona takaja umnaja!* 'She is so clever!'. Another way to express praise is by means of nouns, such as colloquial *Golova*! lit. 'head', roughly 'clever person', *Baška!* lit. 'head', roughly 'clever person', *molodec* roughly 'fine fellow' or *umnica*

A. Gladkova (✉)
Linguistics, School of Behavioural, Cognitive and Social Sciences,
The University of New England, Armidale, NSW 2351, Australia
e-mail: anna.gladkova@une.edu.au

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_12,

roughly ‘good girl/boy’. This way of expressing praise combines an evaluation of a person’s action with an evaluation of the person him- or herself.

Among these means of praise, the words *molodec* and *umnica* are the most intriguing from the point of view of cultural information because their meanings embody certain scenarios of behaviour which are positively evaluated in Russian culture. Therefore, studying the meaning and use of these words can provide us with understanding of such modes of behaviour. The significance of these words can be explained by the fact that they provide an evaluation of a person in general rather than just one quality and, therefore, have a wider range of use. They are also relatively frequent. According to the Russian National Corpus data, *molodec* has the frequency of 46 uses per 1 million words and *umnica* – 9 uses per 1 million of words. For the sake of comparison, the English expressions of praise *good boy* and *good girl* have the frequency of 1 use per 1 million of words each (according to Collins Wordbanks online data).

Molodec and *umnica* are language- and culture-specific words and, for example, they do not have exact translational equivalents in English. The Oxford Russian-English Dictionary (*ORED* 1980) glosses *molodec* as “fine fellow..; as interj. (coll.) well done!” and *umnica* as “1. f. good girl; 2. (m. & f.) clever person”. However, this translation is very rough and even inaccurate because it presents these words as gender specific. Even though like any other noun in Russian these words belong to a certain grammatical gender (masculine and feminine respectively), syntactically they can both be masculine or feminine. It is possible to say *on takoj umnica* ‘he is such.MASC *umnica*’ and *ona takaja molodec* ‘she is such.FEM *molodec*’. Therefore, glossing these terms as gender-specific is not accurate. Moreover, the use of English language- and culture-specific expressions obscures the explanations of the meaning of these words.

Praise is a significant means of pragmatic influence in childrearing, at least in some cultures (cf. Quinn 2003, 2005). It is so for Russian culture. For example, Ispa (1994) confirms that praise is a common means applied in childrearing in Russian. In particular, she observes “generous use of praise” in kindergartens of the early 1990s in Russia, which she thought was even greater than in the USA, a country widely known for its emphasis on praise. She comments, “I was quite struck with the amount of praise children heard each day in the centers I visited; I had a strong sense that it was used there more frequently than in the American centers in which I had observed” (Ispa 1994: 171). Ispa (2002) confirms this observation based on the results of her later visit in the early 2000s. The words *molodec* and *umnica* are commonly used as words of appreciation and praise in childrearing. Therefore, studying meaning and use of these words has an implication of unravelling models of positively evaluated behaviour in Russian childrearing (cf. Wierzbicka 2004; Wong 2006). While this study will examine examples of use of these words in interaction with children, it is not limited to this use only.

This study aims to implement a cultural semantic and ethnopragmatic approach to the study of cultural words with the help of universal concepts and to investigate the cultural features of *molodec* and *umnica* in Russian. In this chapter I aim to achieve three main goals: (1) explore the semantics of these words; (2) elicit cultural

information about the modes of behaviour that they evaluate and the kind of positive evaluation that is attached to them; and (3) provide a cultural interpretation of the meanings of these words by relating them to important themes in Russian culture and comparing them to the similar expressions *good boy* and *good girl* in English and *guāi* in Chinese on the basis of studies by Wierzbicka (2004) and Wong (2006). I will also provide a comparison of the results of my study with the results of a sociological study of child-rearing beliefs among Russian and American students done by Williams and Ispa (1999). Our approach goes beyond merely classifying praise as an 'expressive speech act' (Searle 1975) and allows us to decode important cultural information associated with praise.

2 Methodology and Data

There is a growing understanding in the field of pragmatics that to adequately interpret human interaction one needs to adopt an 'insider's perspective' (Wierzbicka 2003[1991]; Goddard 2006; Haugh 2007; Wong 2010; Ide and Ueno 2011). Following this view, this study aims to demonstrate how native speakers of Russian conceptualize selected praise terms and how this conceptualization affects communication. This task requires a methodology that can successfully study the semantics and pragmatics of the terms. Such methodology can be found in the Natural Semantic Metalanguage (NSM) and the associated technique of cultural scripts.

The Natural Semantics Metalanguage is an empirically established inventory of 63 semantic universals (or primes) and their universal combinatorial properties (grammar) (see Table 1). NSM was developed on the basis of empirical analysis of over 30 languages (Goddard and Wierzbicka 1994, 2002; Peeters 2006). NSM represents a mini-language that lies at the core of every language. At the same time NSM has sufficient expressive power to represent meaning and pragmatic aspects of use.

NSM was initially developed as a metalanguage to represent reductive paraphrase formulae. Later its capacities were extended to formulate cultural scripts in the area of studies known as ethnopragmatics (Goddard and Wierzbicka 2004; Gladkova 2011). The notion of cultural scripts is consistent with the idea that dominating speech practices and ways of speaking in a particular language and culture are reflective of and influenced by dominating cultural understandings. As formulated by Geertz (1973: 452), "the culture of a people is an ensemble of texts". Such 'texts' or 'rules' are embedded in the way people speak and they can be extracted and formulated to unravel cultural influence in speech.

For example, Wierzbicka (2006) associates the cultural rules of using *thank you* and the avoidance of phrases like *you must* in suggestions in English with the prevalence of the value of 'personal autonomy' (cf. Culpeper 2011; Culpeper and Demmen 2011). She argues that the cultural idea that 'it is not good to impose and force other people to do certain things' is shared by English speakers, and that it finds its realisation in language. Wierzbicka (2006: 52) formulates this cultural rule as follows:

Table 1 Semantic primes: English exponents (Goddard 2011: 66)

Substantives:	I, YOU, SOMEONE, SOMETHING~THING, PEOPLE, BODY
Relational substantives:	KIND, PART
Determiners:	THIS, THE SAME, OTHER~ELSE
Quantifiers:	ONE, TWO, MUCH~MANY, SOME, ALL
Evaluators:	GOOD, BAD
Descriptors:	BIG, SMALL
Mental predicates:	THINK, KNOW, WANT, FEEL, SEE, HEAR
Speech:	SAY, WORDS, TRUE
Actions, events, movement, contact:	DO, HAPPEN, MOVE, TOUCH
Location, existence, possession, specification:	BE [SOMEWHERE], THERE IS, HAVE, BE [SOMEONE/SOMETHING]
Life and death:	LIVE, DIE
Time:	WHEN~TIME, NOW, BEFORE, AFTER, A LONG TIME, A SHORT TIME, FOR SOME TIME, MOMENT
Space:	WHERE~PLACE, HERE, ABOVE, BELOW, FAR, NEAR, SIDE, INSIDE
Logical concepts:	NOT, MAYBE, CAN, BECAUSE, IF
Augmentor, intensifier:	VERY, MORE
Similarity:	LIKE~WAY

[people think like this:]
no one can say to another person:
"I want you to do this
you have to do it because of this"

Semantic explications and cultural scripts have a great pedagogical value in that due to the universal language they employ they can be applied in language education and cross-cultural training (Goddard and Wierzbicka 2007; Karimnia and Afghari 2010).

NSM is a highly appropriate tool for the purpose of cross-linguistic and cross-cultural comparison. The universal meanings it is composed of serve as *tertium comparationis* in such studies. NSM has been successfully implemented in contrastive studies of emotions, values, speech acts and communicative styles, among others (e.g., Wierzbicka 2003[1991]; Gladkova 2010a, b; Goddard 2012).

The study relies on a thorough investigation of the contextual use of the terms in question as they are represented in the Russian National Corpus. This Corpus is an online electronic resource of Russian texts with a size of 190 million words. The Corpus includes texts of the period from the end of the eighteenth century to the beginning of the twenty-first century. The Corpus is a representative linguistic source that comprises both written texts (including literary, academic, journalist, and educational works) and transcripts of spoken Russian (including data from a number of dialects). The Corpus is relatively balanced; different types of texts are represented in proportion to their occurrence within the target period. English examples were drawn from Cobuild Wordbanks Online. This corpus uses data from eight varieties of English (British, American, Australian, Canadian, Indian, South

African, New Zealand and Irish). It comprises around 550 million words dating between 2001 and 2005.

3 *Molodec*

The noun *molodec* is derived from the adjective *molodoj*, which means 'young'. *SRJ* dictionary distinguishes five meanings of *molodec*: (1) a young man in the prime of life, strong and stately; (2) (used as praise) approval of a person who has shown boldness, resourcefulness, and acted in a worthy manner; (3) (as adverb) in a youthful manner, cheerfully, dashingly; (4) (obsolete) a servant; (5) (usually plural) an accomplice or a member of some reactionary, evil groups or organisations. This polysemy is fully justified because these meanings are clearly distinguished by the different syntactic functions of *molodec*. However, an analysis of corpus data suggests that only meanings 2, 3 and 5 are actively used in Russian these days. *Molodec*$_1$ is a meaning that is recognised by contemporary users of Russian because it is commonly used in folklore and fairy tales that are read to and by children, but it is not commonly used.

The most commonly used meaning of *molodec* – *molodec*$_2$ – will be discussed in this study. It is a praise word used in a predicative function and can be addressed to a person (1), as well as to talk about a person (2):

(1) *Vse-taki ty molodec, ja by ne smog tak, – govorit on.*
"Well done (*molodec*)! I wouldn't have been able to do it," he said.

(2) *Mama molodec, ona ničego ne skazala bratu i sestre…*
Mother is great (*molodec*)! She didn't say anything to either brother or sister…

Molodec$_2$ is used in a wide range of situations to praise someone who has done something good that surpasses the norm or normal expectations. For this reason, *molodec*$_2$ is a common way to praise someone whose achievement is publicly recognised, as is in a competition or a public performance:

(3) *U ženščin naši segodnja serebro i bronzu vyigrali. – Molodcy. A skol'ko plyli? – 25 kilometrov. – O-go!*
"Our women today won silver and bronze." "Well done! (*molodec*.PL.) And what distance did they swim?" "25 kilometres." "Wow!"

(4) *Vošel Venecianov i spokojno, budto i ne volnovalsja za prem'eru, skazal: Nu čto že, pozdravljaju, molodcy! … Prekrasno vas prinimajut. Dlja Leningrada ėto xorošo.*
Venecianov came in and calmly, as if he had not been worrying about the premiere, said, "Well, congratulations, well done (*molodec*.PL)! … You are getting a wonderful reception. It is good for Leningrad."

(5) *Pozdravljaem tebja, Dina, s roždeniem ešče odnoj kaliningradki. Molodec! … Redakcija "NK".*

We are congratulating you, Dina, on the birth of another Kaliningrad female citizen. Well done (*molodec*)! … Editorial board of "NK".

In example (3) two sportsmen discuss the achievements of their fellow women athletes who won silver and bronze medals in a swimming marathon. One of them, after learning about the outstanding results, praises the women using the word *molodec*$_2$ in the plural. In this situation the achievements of the women are publicly acknowledged because they received medals for them. In example (4) a director praises young actors for their good performance during a premiere in Leningrad. He bases his conclusions on the reaction of the audience; therefore it can be considered an objective achievement. In example (5) *molodec*$_2$ is used in an article where an editorial board congratulates one of the employees of the newspaper who has had a baby girl. Giving birth can also be considered an outstanding achievement.

Attaining a better social or financial status – also a noticeable achievement – can be a reason to praise someone as *molodec*$_2$, as in (6) and (7).

(6) *Molodec, Fedor, xot' odin iz rodni v bol'šie ljudi vybilsja.*
Good on you (*molodec*), Fedor, at least one person from the family has become an important man.

(7) *… molodec, molodec, devočka! Iz nikudyšnyx obstojatel'stv, sovsem iz ničego, postroila ved' očen' neploxo: obrazovanie, svoja kvartira, daže vnešnost' svoju nevygodnuju oblagorodila, imeet stil', v konce koncov.*
… Well done (*molodec, molodec*) girl! From terrible conditions, almost from nothing, she built up her life not badly at all: education, her own apartment, she even made her plain appearance nobler, created her own style after all

In example (6) a young man is praised in this way for his being the only person in the family to achieve a high social status. Example (7) illustrates a similar situation with reference to a young woman.

Molodec$_2$ is a common word to praise someone for a good performance due to outstanding physical ability. As Levontina (2004: 543) writes, "only *molodec*, not *umnica*, is possible if a person is praised for brave conduct in combat, sports achievements, etc." Numerous examples from the Corpus support this argument:

(8) *"Molodec! Silač"; – kriknuli v odin golos i staryj, i malyj.*
"Well done (*molodec*)! Strong man!" the old and the young shouted unanimously.

(9) *Ona molodec, posle zatjažnoj polosy neudač pokazyvaet sejčas fantastičeskoe katanie.*
She is great (*molodec*). After a long period of failures she now demonstrates fantastic skating.

(10) *My naučim vas igrat'. Vidite sosnu? Lezem naperegonki! Kto na veršine budet pervyj, tot molodec.*
We'll teach you to play. Do you see the pine tree? Let's compete in climbing it! The one who gets to the top first will be a good boy (*molodec*).

In example (8) the audience praises and encourages a man who is demonstrating his strength. In example (9) a woman is praised for her outstanding skating after a lengthy period of unsuccessful performances. Example (10) is taken from a children's story where baby bears want to teach baby hedgehogs to climb trees. In their understanding, the one who climbs the tree first, displaying outstanding physical ability, can be called *molodec*.

Similarly, *molodec*$_2$ can be used to praise people for their intellectual achievements. In the majority of such examples from the Corpus, *molodec*$_2$ is addressed to children who are praised for good and persistent studies, as in (11–15). In example (11) a girl is praised *molodec* for completing her first ABC book; in examples (12) and (13) young people are praised for studying well:

(11) *Davajte vse vmeste pozdravim Marinu. Ona zakončila svoju pervuju knižku-bukvar'. Molodec, Kroxa. Pozdravljaem tebja! Teper' ty čelovek gramotnyj.*
Let's congratulate Marina. She's finished her first ABC book. Well done (*molodec*), Little One. Congratulations! Now you are a literate person.

(12) *Molodec, čto učiš'sja xorošo.*
It's great (*molodec*) that you study well.

(13) *Čem zanimaeš'sja v nastojaščee vremja? A? Uroki! Molodec! Papa tvoj učilsja, čelovekom stal…*
What are you doing now? Well? Homework! Well done (*molodec*)! Your father studied and became a real man.

Example (14) refers to a student praising his fellow student for solving a difficult mathematical problem:

(14) *Molodec, Alik, – skazal ja tixo Komarovu, – takuju trudnuju zadaču rešil.*
"Well done (*molodec*), Alik," I said quietly to Komarov, "you solved such a difficult problem."

In example (15) a teacher comments on students' essays and praises one girl for her clean and neat work, which he evaluated as excellent:

(15) *Čerez pjat' dnej učitel' prines tetradki. – Galine Grebenkinoj – "pjat'". Molodec! Akkuratno i čisto, bez edinoj ošibki.*
In 5 days the teacher brought back the exercise books. "Galina Grebenkina got an A. Well done (*molodec*)! Neat and tidy, without a single mistake."

More examples can be quoted to show that *molodec*$_2$ can be used to praise someone for good mental skills. In example (16) *molodec*$_2$ is used to praise a person's witty answer:

(16) *Postoj, govorit Stalin, – otec u tebja kto? – Evrej. – Mat' kto? – Evrejka. – A ty kto? – A ja kommunist! – gordo skazal Kaganovič. – Aj molodec, – skazal Stalin, – nastojaščij internacionalist.*
"Hold on," said Stalin, "who is your father?" "A Jew." "Who is your mother?" "A Jew." "And who are you?" "I am a communist!" Kaganovič said proudly. "Well, good on you (*molodec*)," said Stalin, "you are a real internationalist."

Example (17) describes a situation when a boy praises his friend who recited his own poems in front of the whole school. In this case the boy admires his friend's creativity and courage.

(17) *I Miška poklonilsja i polez so sceny. I vse emu zdorovo xlopali, potomu čto, vo-pervyx, stixi byli očen' xorošie, a vo-vtoryx, podumat' tol'ko: Miška ix sam sočinil! Prosto molodec!*
And Miška bowed and climbed down from the stage. And everyone applauded him enthusiastically because, firstly, the poems were very good, and secondly, it was impossible to believe it, but Miška wrote them himself! Simply a fine fellow (*molodec*)!

Example (18) shows that *molodec*$_2$ can extend to praising people for their moral qualities. In this case a speaker praises another person for a caring attitude to people:

(18) *A ėto ty xorošo delaeš', čto o ljudjax zabotiš'sja tak. Molodec.*
It's good that you care about people. Well done (*molodec*).

As all these examples from the Corpus suggest, *molodec*$_2$ can be used to praise a person who has shown his or her outstanding physical, mental, or moral abilities. Therefore, the range of situations when *molodec*$_2$ can be used as a praise word is very wide. The characteristic that unites these situations is that the person performed something that surpasses the norm, which not many people can do. There is also a hint of this in the stem of the word *molod-* (from *molodoj* 'young') – not all people are "young", or "young men".

I will now discuss the attitude of the person who gives praise that is reflected in the meaning of *molodec*$_2$. The person who uses the word *molodec*$_2$ admires the action of another person. This action reveals something positive about this person and raises the opinion of the speaker about him or her. The following example shows that *molodec*$_2$ can express the pleasant surprise of the speaker about the person whom he or she praises. In this example a boy expresses surprise and admiration for his father's skill in helping another boy who has stuck a coin into his nose. The boy says the following about his father:

(19) *Nu, tovarišči, ja i ne znal, čto moj papa takoj molodec.*
Well, comrades, I didn't know that my Dad was such a good fellow (*molodec*).

In example (20) a school principal praises as *molodec*$_2$ one student from her school who used to have a poor record. She says that extra-curricular activities allowed this boy to reveal his good moral qualities. She presents this information in such a way that with this new activity she discovered that this student was actually a good person.

(20) *Vzjat', k primeru, togo že Sašu Plotnova. Ešče nedavno ne znali, čto s nim delat'. A on, von kakoj molodec! Klassnyj rukovoditel' Saši – Tat'jana Sergeevna Burkova – srazu zametila v nem peremenu.*

Take, for example, Saša Plotnov. Not long ago we did not know what to do with him. And look at him now – such a fine fellow (*molodec*)! Saša's class teacher Tatiana Sergeevna Burkova immediately recognised a change in him.

Therefore, the behaviour of another person who is praised as *molodec*$_2$ can often be a pleasant revelation about the abilities of that person, as in the following example:

(21) *Molodcom, paren'! Ja s toboj teper' v ljubuju razvedku pojdu.*
Well done (*molodec)*, lad! I will now go to any reconnaissance with you.

In the example above a person calls another person *molodec*$_2$ and expresses his new degree of confidence in this person by saying that he can go to any reconnaissance with him. A similar kind of attitude is expressed in example (22) where the speaker says that one won't get into trouble when being with such a person:

(22) *Vot tak Van'ka, molodec! S takim čelovekom ne propadeš', on vsegda znaet, čto nado delat'.*
Vanka is such a great guy (*molodec*)! One will not get into trouble with such a man; he always knows what to do.

Molodec$_2$ is an egalitarian way of praising someone. As numerous examples have shown, it is a common way to praise someone younger or of the same age using the 'egalitarian' pronoun *ty* (the pronoun which is used to a person of the same or younger age), not the 'respectful' *vy* (the pronoun which is addressed to older people or people of higher status). However, a person can call someone older *molodec*$_2$ if he or she uses the *ty* form when speaking to this person (as in 23). In such cases older people are usually praised for their youthful behaviour and spirit, and are in a way considered similar to young people. In the following examples (23) and (24) *molodec*$_2$ is used to older people:

(23) *Ne vyderžal, podošel k odnomu 106-letnemu tancoru Ruslanu Džogija, govorju, ty molodec, slušaj.*
I couldn't help it, I went up to a 106-year old dancer Ruslan Džogija and told him: Look, you are an awesome guy (*molodec*).

(24) *Vse-taki naši babuški molodcy. Pokolenie nesgibaemyx.*
Our grandmas are great women (*molodec.*PL), though; a generation that will not bend.

In example (23) the person calls a 106-year old dancer *molodec*$_2$ and admires his skill and youthful spirit. Example (24) is taken from a context in which the author praises her grandmother's generation for their endurance and optimism. In this example *molodec*$_2$ is used in the plural as a reference term, not as an address form.

Having all these characteristics in mind, the following explication can be proposed for *molodec*$_2$

molodec$_2$
(a) I think like this now:
(b) you did something very good
(c) not many people can do something like this

(d) because you did this, I know something very good about you
(e) when I think like this about you, I feel something very good
(f) I want to say something good to you because of this

In this explication component (a) shows that, as a praise word, *molodec*$_2$ has a cognitive basis. Components (b–d) capture the reasons for praise: (b) – performance of a very good action, (c) – the outstanding character of such performance, which is explained by comparison of the action of a person with the majority of people who are not capable of performing such an action. Component (d) reflects the new view of another person as a good person. Component (e) captures a very good feeling which is caused by such a way of thinking. The desire to express this attitude verbally to the other person is captured in component (f).

4 *Umnica*

The word *umnica* is polysemous. *SRJ* gives two meanings for *umnica*: (1) clever, sensible, intelligent person; (2) sensible, obedient child. I agree with *SRJ* that *umnica* has two meanings. *Umnica*$_1$ means "a clever person who is liked by the speaker" and is used to characterise a person in general (cf. Levontina 2004), as in (25):

(25) *On ved' byl umnica, talantlivyj čelovek, oficer…*
He, after all, was a clever guy (*umnica*), a talented person, an officer…

The second meaning – *umnica*$_2$ – which is discussed in this chapter, is used to praise a person for a certain action, not for a general characteristic. However, it is unjustified to say that *umnica*$_2$ can only be used to refer to children. As I will demonstrate with numerous examples from the Corpus, it is also used when speaking to adults.

The word *umnica* derives from the adjective *umnyj* – a culture- and language-specific word that incorporates the qualities of being clever, wise, intelligent, and sensible. The adjective *umnyj* derives from the noun *um* – something like mind/intellect – which is a significant component of the Russian folk model of a person. Uryson (2004: 1203) describes *um* as the ability of a person to think, as well as an invisible organ. The meaning of the word *umnyj* is important in order to understand the meaning of *umnica*$_2$. *Umnyj* is someone who can think well about many things and because of this behaves and does things well. Boguslavskaja (2004: 1207) explains the meaning of *umnyj* via *um* and notes that *umnyj* "gives a holistic characteristic of a person whose *um* surpasses the norm and is revealed in ideas, words, actions and in the ability to understand other people". The fact that the quality of being *umnyj* is related to the ability of a person to act wisely in certain situations can be found in the derivational properties of this word. The adjective *umnyj* and the noun *um* have the same root as the verb *umet'* 'be able to/know how to' and the noun *umenie* 'ability/skill' (or its obsolete form *umen'e*). Interestingly, in Dal's (1957

[1862]) collection of Russian proverbs, which represents ethnographic material collected by the author in the first half and the middle of the nineteenth century, we find instances of *um* and *umen'e* being used interchangeably:

(26) *Budet imen'e, budet i umen'e. S bogatstvom um prixodit.*
If there is property, there will be *umen'e*' (skill). With wealth comes *um* (reason).
(Dal' 1957[1862]: 442)

(27) *Ne kop'em pobivajut, a umom (to est' umen'em).*
One kills another person not with a spear, but with *um* (mind/reason) (that is *umenie* (skill)).
(Dal' 1957[1862]: 431)

These examples show that at some point of time in the past in the Russian folk model of a person the concepts of *um* and *umen'e* were closely related. This link, however, is not so obvious in modern Russian. Nevertheless, the derivational relationship between *umnyj* and *umenie* supports the idea that the quality *umnyj* is related to the ability to act wisely in certain situations.

The explication of *umnyj* can be represented as follows:

Kto-to X – umnyj čelovek (someone X – *umnyj* person)
(a) someone X can think well about many things
(b) because of this, someone X can do many things well
(c) people think: it is good if someone can be like this

The semantic part that is shared by the words *umnyj* and *umnica*$_2$ is that a person gets praised for good thinking, which results in performing good actions.

Semantically *umnica*$_2$ and *molodec*$_2$ differ in the reasons for praise and the speaker's attitude. While both words can be used to praise someone for a good achievement, *umnica*$_2$ praises a person for good results achieved due to good thinking. *Molodec*$_2$, as shown in the previous section, can be used to praise people for physical as well as intellectual achievements. Levontina (2004: 534) rightly notes that fans when praising their favourite team for scoring a goal can scream *Molodcy!* (plural of *molodec*). It is impossible to use *umnicy* (*umnica*.PL) in this situation because *umnica*$_2$ cannot be used to praise positive results achieved due to physical ability without thinking about the action. Examples (28) and (29) illustrate how *umnica*$_2$ is used to praise someone for providing a witty reply or suggestion which has a practical application:

(28) *A čto vy delaete, esli obožžetes'? Kakoe est' narodnoe sredstvo? – Sljuna? – predpoložil ja. – Umnica! Vot imenno, sljuna! … moja dogadlivost' nasčet sljuny očen' ponravilas' staruxe, i ona rasxvalivala menja …*
'And what do you do if you burn yourself? What folk remedy is there?' 'Saliva,' I suggested. 'Well done (*umnica*)! Right, saliva!' … the old woman liked my shrewdness about the saliva very much, and she praised me …

In example (28) an old woman praises a boy for suggesting the right answer to her question. In example (29) an army officer praises a young man who thought of a place to hide when they were in a situation of danger:

(29) *Sprjatat'sja tut legko. Tam vsego metrax v pjatidesjati bylo to li ozero, to li boloto. V takuju žaru ono vysoxlo, no ostalis' kamyši, očen' vysokie. – Umnica, – vpervye poxvalil Zudina Devjatkin.*
'It is easy to hide here. There used to be a lake or a swamp fifty metres away from here. It is dried out in such hot weather, but there are very tall weeds left.' 'Well done (*umnica*),' Devjatkin praised Zudin for the first time.

However, *umnica*$_2$ is not a word that is used to praise intellectual abilities only. As the derivational link between the words *umnyj* and *umnica*$_2$ suggests, it extends to some kind of worldly wisdom. The good thinking of *umnica*$_2$ cannot be abstract; it should be applied to some performance. Also *umnica*$_2$ can be used to characterise actions of a person that were reasonable and somewhat expected from the point of view of a more experienced person. As Levontina (2004: 543) writes, "*umnica* is applied to a person who behaves well or in the right way… *umnica*, most likely, implies that a person is praised for correspondence to a norm or expectations." To illustrate this point, example (30) is taken from a story for children in which mother-hedgehog praises her babies for clever behaviour in a situation of danger. The results were very good results, but it was most obviously safe behaviour.

(30) *Zaletela k ežam osa. Ežata slepen'kie: nedelja ot rodu, no uslyšali nedobryj gud – v komočki svernulis'. – Ax, umnicy moi! – obradovalas' ežixa-mama. – Kolju̇čki mjagon'kie, a za sebja ežiki uže gotovy postojat'.*
A wasp flew into the hedgehogs' lair. The baby hedgehogs were blind, just 1 week old. But having heard a hostile buzz, they rolled themselves into balls. 'My good girls (*umnica.*PL)!' happily exclaimed Mother-Hedgehog. 'The spikes are still soft, but the little hedgehogs are ready to stand up for themselves.'

Other examples can be quoted to show that *umnica*$_2$ is used to praise a person who behaves in a way that another person wants. In the following examples *umnica*$_2$ is used between adults: in example (31) a woman praises her friend for not getting angry; in example (32) a husband praises his wife for guessing that it is necessary to lay the table for unexpected guests; in example (33) a woman praises her lover in her thoughts for expressing his love in the way she anticipates; in example (34) a man praises his sister for coming to visit him.

(31) *Krepko uxvativ Strigunkova za lokot', Maja skazala…: – Ne rasserdilsja? Vot i unmica.*
Having grasped Strigunkov firmly by the elbow, Maja said, "You aren't angry, are you? You are a good boy (*umnica*)."

(32) *Ty nam v gostinoj nakryla? – sprosil Maksim, celuja ženu v ščeku. – Umnica.*
'Have you laid the table in the drawing room for us?' asked Maksim kissing his wife on the cheek. 'Good girl (*umnica*).'

(33) … *Vse bylo uže ustroeno v ee žizni, vse cvelo, byl porjadok, obrazovalos' glavnoe: vsegda vmeste s Samsonom. Čerez mnogo časov on priexal pod okno i prosvistal: "Ax, net sil snesti razluku", umnica.*
… Everything was already decided in her life, everything was blossoming, was in order, and the main thing emerged: always to be together with Samson. In a few hours he came under her window and whistled: "I can't bear the separation", good boy (*umnica*).

(34) *Medeja … ne uspela ešče vstat', kogda on sgreb ee v oxapku, podnjal, prižal k sebe, kak rebenka: – Sestrušen'ka, umnica, priexala!*
Medeja … hadn't yet had time to get up, when he swept her up in his arms, and cuddled her like a child: Dear sister, good girl (*umnica*), you've come!

The use of *umnica*$_2$ in situations when a person expects a certain behaviour from another person shows that *umnica*$_2$ is often used by an older person or one of a higher status. For this reason, *umnica*$_2$ is usually not used to someone older. In this respect *umnica*$_2$ differs from *molodec*$_2$, which is a rather egalitarian way of praising somebody.

At the same time *umnica*$_2$ is a very affectionate way to praise another person. As the previously quoted examples suggest, the use of *umnica*$_2$ signifies a loving attitude between two people who are respectively parents and children, close friends, spouses, lovers, or brothers and sisters. *Umnica*$_2$ is commonly used with the possessive pronoun *moj/moja* 'my.MASC/my.FEM'. By the use of this pronoun the speaker shows that the addressee belongs to his or her private sphere. It emphasises the positive emotional attitude of the speaker and shows a close relationship between them. Examples (35–40) illustrate how *umnica*$_2$ is used with possessive pronouns.

(35) *Mam, smotri, ja pervoe mesto zanjal! – voskliknet Aleša. – Smotri, kakaja krasivaja gramota! – Umnica moj, – skažet mama, poceluet Alešu v lob.*
"Mum, look, I came first!" Aleša would exclaim. "Look, what a beautiful certificate!" "My boy (*umnica*)," mother would say and kiss Aleša on the forehead.

(36) *Devočka iz igrušečnoj lejki polivala cvety. – Umnica ty moja! – govorila babuška.*
A girl was watering flowers from a toy watering-pot. "You are my good girl (*umnica*)!" said grandmother.

(37) *Raj u tebja, moja umnica, – govorila … njanja.*
"Your place is heaven, my dear (*umnica*)," said … nanny.

(38) *Dočen'ka premiju polučila, takuju rubašečku podarila, umnica moja.*
My dear daughter received a bonus, and bought me this shirt, my good girl (*umnica*).

(39) *Muročka, umnica moja … Ona malen'kix žaleet. … Ona ego ne obidit.*
Muročka, my good girl (*umnica*) … She pities small ones … She will not hurt him.

(40) *Ty moja umnica, – skazal Ivan Dmitrievič. – Ty lučše vsex.*
"You are my good girl (*umnica*)," said Ivan Dmitrievič. "You are better than anyone."

Unlike *umnica*$_2$, *molodec*$_2$ is not used with a possessive pronoun, which suggests that it is a more 'distant' and 'detached' form of praise and that it does not imply that the person conforms to any norms. Levontina (2004: 543) says that *molodec*$_2$ expresses admiration (*vosxiščenie*) and triumph (*toržestvo*) while *umnica*$_2$ expresses affection (*umilenie*) and tenderness (*nežnost'*).

Using simple universal concepts the explication of *umnica*$_2$ is as follows:

umnica$_2$
(a) I think like this now:
(b) you did something good
(c) someone can't do something like this if this someone doesn't think very well
(d) when you do something, I want you to do it in this way
(e) because you did this, people can know that you can think very well
(f) when I think like this about you, I feel something very good
(g) I feel something good towards you
(h) I want to say something good to you because of this

In this explication component (a) stands for the cognitive basis of the praise word and presents the scenario from the point of view of the speaker. Component (b) shows that another person did something good, and component (c) explains that it was done due to good thinking. The 'parental' attitude of the speaker is reflected in component (d), which shows that the speaker wants another person to behave in such a way. Such phrasing of this component explains why *umnica*$_2$ cannot be applied to someone older or of a higher status. Component (e) shows that this action of the person reveals to other people his or her ability to think well, and it explains the social acceptance of such behaviour. Realisation of the success of the action of another person leads to the state of emotional satisfaction (component f) and a positive attitude towards this person (component g). The speaker wants to say something good about this realisation and the experience of these feelings to this person (component g).

5 *Molodec* and *Umnica* Compared

Now that I have explicated *molodec*$_2$ and *umnica*$_2$ using the same set of universal concepts, it is easy to identify semantic similarities and differences between them. Both terms contain the component 'I think like this' that explains the cognitive and pragmatic basis of praise words. Both concepts also share the component of acknowledging a good action of the referent: 'you did something good'. In the case of *molodec*$_2$, this component is intensified by the element 'very' and is presented as 'you did something very good'. They both also have a component that reflects the speaker's positive feeling, which is caused by the realisation of the outstanding behaviour of the referent – 'when I think like this about you, I feel something very good', as well as the component indicating the desire to verbally express praise – 'I want to say something good to you because of this'. Presumably, these three

components will be found in any praise words which can be addressed to another person because they reflect the essence of praise: thinking about another person in a certain way, a good feeling caused by this way of thinking about this person, and the desire to verbalise this good attitude.

Nevertheless, the explications of *molodec*$_2$ and *umnica*$_2$ differ in the evaluation of the action of the referent and the emotional attitude that is associated with this way of thinking about this person. Firstly, components (b) differ due to the presence of the element 'very' in the explication of *molodec*$_2$. This fact explains that from the point of view of the speaker *molodec*$_2$ commends someone for an outstanding action, and this line in the explication is supported by component (c), which explains the outstanding character of this action – 'not many people can do something like this'. Component (c) in *umnica*$_2$ – 'a person can't do something like this if this person doesn't think very well' – explains that *umnica*$_2$ is used to praise people whose achievements result from a certain way of thinking. This line excludes achievements that are of a physical character only.

Molodec$_2$ and *umnica*$_2$ also differ in terms of the impact of the action of another person on the speaker. *Molodec*$_2$ has component (d) – 'because you did this, I know something very good about you'. This component captures the fact that the action of the referent makes the speaker think about him or her as a good person in general; it also explains that there is some degree of unexpectedness in the action of the other person. The explication of *umnica*$_2$ does not contain such a component because it is a praise word for more anticipated actions. This kind of expectation is reflected in component (d) – 'when you do something, I want you to do it in this way'. This component also indicates the 'parental-like' attitude of the speaker, which is why *umnica*$_2$ is usually used by older people to younger people or by people with more knowledge or expertise in the field to less experienced ones.

Umnica$_2$ is a way to praise someone for conformity with social norms achieved due to the right way of thinking. This idea is captured in component (e) of *umnica*$_2$, – 'because you did this, people can now know that you can think very well'. This component reflects the 'social' orientation of the value of the action of another person (the element 'people'). The presence of the components of 'correspondence of behaviour to social norms and expectations' and 'acting in accordance with the wish of the speaker' in the semantic structure of *umnica*$_2$ explains the possibility of its use in an imperative construction *bud' umnicej* 'be *umnica*', as in the following examples:

(41) *Bud'te umnicej i ne pejte ničego sliškom na svad'be.*
Be a good boy (*umnica*) and don't drink too much of anything at the wedding.

(42) *A teper', – skazal on, podxodja k dočeri i gladja ee po golove, – perestan' plakat', uspokojsja i bud' umnica.*
'And now,' he said approaching his daughter and stroking her head, 'stop crying, calm down, and be a good girl (*umnica*)'.

Bud' umnicej sounds like a request: the speaker asks the addressee to behave in a way he or she wants. ?*Bud' molodcom* 'be *molodec*' is anomalous because in

molodec$_2$ there are no semantic components which can be considered to be a reason for a request.

Molodec$_2$ and *umnica*$_2$ also differ in the emotional attitude of the speaker. While both words signify the positive emotional state of the praising person, *umnica*$_2$ has an additional component, which reflects the positive attitude of this person, that is, component (g) – 'I feel something good towards you'. The presence of this component makes *umnica*$_2$ a more affectionate form of praise than *molodec*$_2$.

6 Cultural Elements of Meaning in *Molodec*$_2$ and *Umnica*$_2$

Cultural elements of meaning in the words *molodec* and *umnica* can be discussed from two points of view – from within Russian culture and in comparison with other languages and cultures. I will first relate these words to some important Russian cultural values. Then I will compare them with the similarly used English praise expressions *good boy* and *good girl* and the Chinese word *guāi* on the basis of the semantic explications proposed by Wierzbicka (2004) and Wong (2006) respectively.

6.1 Molodec$_2$ *and* Umnica$_2$ *in Relation to Some Russian Cultural Themes*

Justification for the relation of these words to other Russian cultural words can be found in their common collocations. The word *molodec* can be used in such collocations as *udal' molodeckaja* 'youthful daring' and *molodec-udalec* '*molodec* who performs *udal'* – expressions which are virtually untranslatable into English. Both these collocations relate the word *molodec* (or the adjective *molodeckij* which derives from it) to the word *udal*' 'daring/boldness' which is regarded as a significant cultural word in Russian (cf. Šmelev 2005). Šmelev (2005: 57) writes the following about *udal'*:

> This word is not used to talk about doing one's duty. It is suitable to describe someone who acts against reason and by doing so performs actions that are impossible for other people to perform … At the same time the word *udal'* in Russian has a distinct positive evaluation.

Like *udal'*, *molodec*$_2$ positively evaluates actions performed for good which require courage and which may even involve acting against norms and rules. *Molodec*$_2$ refers to someone acting not like others, but whose actions are good. In this word the Russian language expresses a positive attitude towards ingenuity,

courage, and doing good. This kind of behaviour can be unreasonable, yet lead to good results.

Umnica$_2$ evaluates positively a somewhat different scenario of behaviour. *Umnica*$_2$ is someone who behaves in a clever way and within socially accepted norms. It is linked with the idea of harmony among people achieved by thought. Therefore, *umnica* is related to the concepts *um* and *razum* 'reason/mind/intellect' and is used in the compounds *umnica-razumnica* 'clever and reasonable girl' and *umnica-krasavica* 'clever and beautiful girl', which are used for females only.

Thus, by conducting a detailed semantic analysis and relating these words to other important cultural values I have shown that *molodec*$_2$ and *umnica*$_2$ positively evaluate quite different types of behaviour: the first values goodness reached by boldness and courage, and the latter – goodness reached by conforming to expectations of a more experienced person and preserving harmony. Interestingly enough, both these words are equally important in terms of serving as guiding words for children.

On the basis of the conducted analysis it is possible to formulate several Russian cultural scripts generalising culturally-values ways of talking and behaving. Firstly, the corpus data suggests that praise in general is common in Russian interaction. This practice can be formulated as follows:

[many people think like this:]
if I think about someone else like this:
    this someone did something very good
it is good if I say something like this to this someone:
    you did something very good
    I feel something good because of it

Secondly, there are also scripts emphasizing the value of two types of behaviour (related to the semantics of the words *molodec* and *umnica*):

[many people think like this:]
at many times it is good if someone does something like this:
    someone does something very good
    not many people can do something like this
people think something good about this someone because this someone does something like this

[many people think like this:]
at many times it is good if someone does something like this:
    someone does something good
    someone can't do something like this if this someone does not think very well
people think something good about this someone because this someone does something like this

6.2 *A Comparison of* Molodec$_2$ *and* Umnica$_2$ *with Their Cultural Counterparts in English and Chinese*

Although neither *good boy/girl* in English, nor *guāi* in Chinese can be regarded as full semantic equivalents of either *molodec*$_2$ or *umnica*$_2$, it is nevertheless interesting to compare the meanings of these words due to their salience in the three cultures. Wierzbicka (2004) argues that the expressions *good boy* and *good girl* are important cultural words which reflect an Anglo culture-specific model of childrearing which has its roots in England's and America's Puritan past. In particular, she demonstrates that the specificity of these expressions lies in the fact that they establish a link between "the value of something that a child has done and the 'value' of the child him- or herself" (Wierzbicka 2004: 253). She also shows that these expressions are preferred by parents when the child acts in accordance with their wishes. Wierzbicka (2004: 253–254) proposes the following explication of *good boy* and *good girl* using simple universal concepts:

Good boy (girl)!
(a) I think now: you did something very good
(b) I want you to do things like this
(c) because of this I say: you are a good boy (girl)
(d) I feel something good because of this

Wierzbicka (2004: 254) comments on the components of this explication as follows:

> component (a) accounts for the spontaneous character of these exclamations, which imply a parent's current thought (usually, but not necessarily, related to the child's current action); component (b) refers to parental wishes; component (c) expresses a positive evaluation of the child, linked with the child's actions; and component (d) indicates an element of emotion.

In sum, the expressions *good boy* and *good girl* praise children for the behaviour that satisfies parents and teaches them to submit to norms and observe certain rules.

Collins Wordbanks online corpus offers sufficient evidence for the idea that the expressions *good boy/girl* imply the existence of some rules (usually set or verbalised by parents) that the child is expected to follow:

(43) *If you don't eat your vegetables, you're not going to rugby. If you're not going to be a good boy, we're not taking you to rugby.*
(44) *I'm a good boy 'cause I cleaned my room, aren't I?*
(45) *I never did sit there. I was a good boy.*
(46) *You'll have to earn it though. You'll have to go to bed like a good girl then I'll give it to you in the morning.*
(47) *Now be a good girl and do as you're told.*

At the same time the emotional aspect of praise is evident from the following examples:

(48) *You sit down first, ask politely if it's okay for you to share the table, make some little comment about the eclipse. I'll come over maybe thirty seconds after you. All clear? Good boy. Go ahead, now.*

(49) *And it wasn't too long before Zoe was floating. "My water-baby," Alma said. "Good girl, Zoe, I'm proud of you."*

I will identify two main differences between the English and Russian expressions. Russian *molodec*$_2$, unlike the English expressions, values courageous and bold behaviour rather than behaviour that complies with certain norms. Of course, the expressions *good boy* and *good girl* sound enthusiastic, but they do not encode admiration for boldness or courage. *Umnica*$_2$ differs from the English expressions by a strong emotional element of affection towards the person who is praised. The lack of affection in the expressions *good boy* and *good girl* is perhaps an implication that 'this is how such things should be done' (while expressing goodness and necessity of goodness and conformity to rules). For an English speaker the actions for which a person is praised as *molodec*$_2$ might seem too bold and unreasonable, whereas *umnica*$_2$ might sound too soft. If we try to relate the meanings encoded in these Russian and English expressions to a wider picture of cultural attitudes to child rearing, we might suggest that there is preference for warmth and, possibly, indulgence in Russian childrearing and perhaps less expectation that rules will be followed – in contrast to the implications of the English expressions *good boy* and *good girl*.

The Chinese word *guāi* presents another model of a 'good' child, which reflects Chinese cultural norms. According to Wong (2006), *guāi* describes a child who is, from an adult's perspective, well behaved and obedient. Wong (2006: 115) explains the cultural significance of *guāi* as follows:

> A child who is *guāi* does what is culturally expected of him or her, such as showing respect for seniority in age among his elders by using the appropriate address form. … In Chinese culture, children are expected to be *guāi* and compliant to people older than him/her, rather than to exercise personal autonomy.

Wong (2006: 115) provides the following example when *guāi* can be used in the situation of adult-child interaction: "One could well imagine a prototypical exchange like the following when a mother and child start interacting with another adult female they have just met or run into:

Mother (to child): *Call Aunty.*
Child (to adult female): *Aunty.*
Adult female (to child): *Guāi.*"

Wong (2006: 115) proposes the following semantic explication for *guāi*:

guāi
(a) people think like this:
a child [M] has to do some things at some times
it is good if a child [M] knows what these things are
it is good if a child [M] does these things

(b) I think about this child [M] like this:
"this child [M] is like this, this is good"
(c) I feel something good towards this child [M] because of this

As this explication shows, in the Chinese cultural model a child is praised for doing what an adult wants him or her to do and for knowing the cultural expectations in observing the social hierarchy. As Wong (2006: 116) comments further, *guāi* does not mean the same as *obedient*:

> an *obedient* child refers to one who does what he or she is told without questioning, a *guāi* child knows what to do without necessarily being told every time because he or she can understand what older people expect according to cultural norms.

Guāi can only be compared to *umnica*$_2$, because both these words are usually used by a person of an older age to a person of a younger age. As my analysis has shown, the use of *umnica*$_2$ in Russian is not restricted to its use by adults to children, but is used among adults as well. This fact tells us that there is a more 'egalitarian' attitude encoded in *umnica*$_2$ than in *guāi.* To add to this, *umnica*$_2$ stresses that the person thinks well. *Umnica*$_2$ is also a more affectionate and outgoing form of praise. It contains components of a positive feeling caused by the realisation of the goodness of the child's action, a positive feeling and a desire to say something good which are expressed *towards* the child (in the explication they are marked as 'to you'). The explication of *guāi* has only one component capturing a positive feeling of the speaker – 'I feel something good towards this child because of this', but it is stated in a way that it shows that it is not addressed *to* the child. *Guāi* mainly stresses complying with norms and the social hierarchy as well as *knowing* what to do, rather than *thinking* about it (something that *umnica*$_2$ embeds in its meaning).

To conclude, a cross-linguistic and cross-cultural comparison of the Russian praise words *molodec*$_2$ and *umnica*$_2$ with the English expressions *good boy* and *good girl* and the Chinese word *guāi* shows that *molodec*$_2$ and *umnica*$_2$ are language- and culture-specific words and that their meanings embody a culture-specific positive attitude towards two modes of behaviour. *Molodec*$_2$ refers to courageous and daring behaviour which surpasses normal expectations, while *umnica*$_2$ is an affectionate way of praising someone who acts in a way that meets the expectations of someone who is more experienced. This good action is achieved by applying good thinking about the situation.

7 Discussion

This study demonstrates that linguistic material acquired from a linguistic corpus can be regarded as sound evidence in describing cultural specificity. Given that language is the main human communication tool, it can be used as a reliable and very informative source of cultural studies. However, sometimes the findings of linguistic cultural analysis can go against existing stereotypes. In this chapter I have

considered the words *molodec*$_2$ and *umnica*$_2$ as a reflection of culturally significant ideas about good behaviour that are prevalent in Russian cultural models of behaviour. I discovered value attached to bold and courageous behaviour in *molodec*$_2$ and to behaviour that brings social harmony achieved by thought in *umnica*$_2$. A comparison with the English expressions *good boy* and *good girl* highlighted the features mentioned. It also showed that the English expressions put more value on the necessity to conform to rules and doing what one is being told to do. These findings go against commonly expressed stereotypes that say that rule conformity is more valued in Russian culture than in Anglo culture (cf. e.g., Lewis 1999). However, a comparison with the results of a sociological study of childrearing ideals among Russian and American students by Williams and Ispa (1999) supports my findings.

Williams and Ispa (1999) studied the importance of four child-rearing goals among Russian and US university students: obedience, inquisitiveness, peer orientation, and neatness/cleanliness. They report the following results (Williams and Ispa 1999: 544):

> The follow-up tests showed significant differences between Russian and U.S. students with regard to rule conformity, … peer orientation, … and neatness/cleanliness …. Compared to U.S. students, Russian students rated rule conformity as less important …, peer orientation as more important …, and neatness/cleanliness as more important. …
>
> … the order of importance of the four goals was different for Russian and U.S. students. … Russian students gave inquisitiveness their highest importance ratings, peer orientation their second highest, neatness/cleanliness their third highest, and rule conformity their lowest importance ratings.

Although the results of this study should be treated with caution because we do not get a clear description of the exact words that were used in the questionnaire in both Russian and English, and also because they reflect the ideals of a very small proportion of the population of the respective societies, it can be tentatively suggested that my results are consistent with Williams and Ispa's findings. My study has shown a more significant preference for rule conformity encoded in the English expressions *good boy/girl* than in the Russian words *molodec*$_2$ and *umnica*$_2$. It also showed a value attached to boldness and courage in the Russian words. These characteristics can be associated with 'inquisitiveness'.

Williams and Ispa (1999: 545), however, call their findings "intriguing" because they show that American university students value rule conformity higher than Russian university students. The authors provide the following explanation for this fact: "This difference appears to reflect both the conservative swing among U.S. students and anti-authoritarian attitudes in Russia that helped to topple the Soviet government." Therefore, they regard this difference as recent and not culture-specific. However, from the point of view of my research, this finding is not that intriguing. The study of the Russian language corpus shows that the words *molodec* and *umnica* have been used as praise words throughout the nineteenth and twentieth centuries with a relatively stable level of frequency. This fact indicates that the values reflected in these words have been important in Russian culture throughout this time. Similarly, as Wierzbicka (2004) shows, the expressions *Good boy!* and *Good girl!* have been in use in English for the last two centuries. Of course, praise words

even though they are commonly used in childrearing, cannot be regarded as the only source in the analysis of cultural models. However, they do provide us with clues about positively evaluated modes of behaviour. Thus, we can regard language as a valuable source of cultural information.

8 Concluding Remarks

The praise words *molodec*$_2$ and *umnica*$_2$ contain intriguing cultural information. Their meanings embody types of behaviour that are positively evaluated in Russian culture and they also contain a culture-specific attitude towards these types of behaviour.

A detailed semantic analysis demonstrates that *molodec*$_2$ is an egalitarian form of praise and that its meaning incorporates the idea that a person has performed an action which surpasses normal expectations. This idea is worded in the explication by comparing this action with actions of other people who cannot normally do something like this. Normally, such action is associated with a person's excellent physical performance, endurance, optimism, persistence or courage. *Molodec*$_2$ also expresses an element of unexpectedness and pleasant surprise. As a praise word it contains a component of a positive emotional attitude. The cultural salience of *molodec*$_2$ lies in the fact that it positively evaluates goodness achieved by boldness and courage. Its meaning can be related to another cultural word *udal'*, which also embodies a culturally-specific positive attitude towards unreasonably bold behaviour which benefits people.

Compared to *molodec*$_2$, *umnica*$_2$ is a milder and more affectionate form of praise. It is preferred for use towards children rather than adults, however it can also be used towards adults when the person who is praising has more experience and authority. *Umnica*$_2$ positively evaluates a type of behaviour which is achieved due to thinking about the matter and which meets the expectations of someone who is more experienced. Overall, it is related to the idea of harmony among people that can be achieved by thought. The word *umnica* contains a derivational link with the words *um* and *razum*, which highlights the cultural significance of its meaning.

A comparison of the words *molodec*$_2$ and *umnica*$_2$ with the English expressions *good boy* and *good girl* and the Chinese word *guāi*, which are frequently used in childrearing in the respective cultures, emphasises the cultural specificity of *molodec*$_2$ and *umnica*$_2$. A semantic comparison shows that the cultural significance of *molodec*$_2$ is found in the appreciation of bold and courageous behaviour that surpasses the norm. *Umnica*$_2$ is culturally-specific due to the affectionate attitude that it embodies and the positive attitude it shows to socially accepted behaviour which is achieved by thought.

The analysis conducted emphasises the value of linguistic corpus material in the interpretation of culture.

References

Apresjan, Valentina. 1997. 'Fear' and 'pity' in Russian and English from a lexicographic perspective. *International Journal of Lexicography* 10(2): 85–111.

Boguslavskaja, O. 2004. Umnyj, neglupyj smyšlennyj, mudryj, glubokij, pronicatel'nyj, prozorlivyj. In *Novyj ob"jasnitel'nyj slovar' sinonimov russkogo jazyka* [The new explanatory dictionary of Russian synonyms], ed. J. Apresjan, 1206–1213. Moskva-Wien: Jazyki slavjanskoj kul'tury-Wiener Slawistischer Almanach.

Collins Wordbanks Online. http://www.collinslanguage.com/content-solutions/wordbanks. Accessed Jan 2013.

Culpeper, Jonathan. 2011. English politeness: The 20th century. Paper presented at the 12th International Pragmatics Conference, Manchester, UK, 3–8 July 2011.

Culpeper, Jonathan, and Jane Demmen. 2011. Nineteenth-century English politeness Negative politeness, conventional indirect requests and the rise of the individual self. *Journal of Historical Pragmatics* 12(1–2): 49–81.

Dal', Vladimir. 1957[1862]. *Poslovicy russkogo naroda* [Proverbs of the Russian people]. Moskva: Izdatel'stvo Xudožestvennoj Literatury.

Geertz, Clifford. 1973. *The interpretation of cultures. Selected essays by Clifford Geertz*. London: Hutchinson.

Gladkova, Anna. 2010a. 'Sympathy', 'compassion', and 'empathy' in English and Russian: A linguistic and cultural analysis. *Culture & Psychology* 16(2): 267–285.

Gladkova, Anna. 2010b. A linguist's view of 'pride'. *Emotion Review* 2(2): 178–179.

Gladkova, Anna. 2010c. *Russkaja kul'turnaja semantika: ėmocii, cennosti, žiznennye ustanovki* [Russian cultural semantics emotions, values, attitudes.]. Moscow: Jazyki slavjanskoj kul'tury.

Gladkova, Anna. 2011. Cultural variation in language use. In *Pragmatics of society*, Handbooks of pragmatics, vol. 5, ed. K. Aijmer and G. Andersen, 567–588. Berlin: Mouton de Gruyter.

Goddard, Cliff. 2006. Ethnopragmatics: A new paradigm. In *Ethnopragmatics: Understanding discourse in cultural context*, ed. C. Goddard, 1–30. Berlin: Mouton de Gruyter.

Goddard, Cliff. 2011. *Semantic analysis: A practical introduction*, 2nd ed. Oxford: Oxford University Press.

Goddard, Cliff. 2012. 'Early interactions' in Australian English, American English, and English English: Cultural differences and cultural scripts. *Journal of Pragmatics* 44: 1038–1050.

Goddard, Cliff, and Anna Wierzbicka (eds.). 1994. *Semantic and lexical universals: Theory and empirical findings*. Amsterdam: John Benjamins.

Goddard, Cliff, and Anna Wierzbicka (eds.). 2002. *Meaning and universal grammar: Theory and empirical findings. Vols. I, II*. Amsterdam: John Benjamins.

Goddard, Cliff, and Anna Wierzbicka. 2004. Cultural scripts: What are they and what are they good for? *Intercultural Pragmatics* 1–2: 153–166.

Goddard, Cliff, and Anna Wierzbicka. 2007. Semantic primes and cultural scripts in language learning and intercultural communication. In *Applied cultural linguistics: Implications for second language learning and intercultural communication*, ed. F. Sharifian and G. Palmer, 105–124. Amsterdam: John Benjamins.

Haugh, Michael. 2007. Emic conceptualisations of (im)politeness and face in Japanese: Implications for the discursive negotiation of second language learner identities. *Journal of Pragmatics* 39(4): 657–680.

Ide, Sachiko, and Kishiko Ueno. 2011. Honorofics and address terms. In *Pragmatics of society*, Handbook of pragmatics; 5, ed. G. Andersen and K. Aijmer, 439–470. Berlin: Walter de Gruyter.

Ispa, Jean. 1994. *Child care in Russia: In transition*. Westport/London: Bergin and Garvey.

Ispa, Jean. 2002. Russian child care goals and values: From Perestroika to 2001. *Early Childhood Research Quarterly* 17(3): 393–413.

Karimnia, Amin, and Akbar Afghari. 2010. On the applicability of cultural scripts in teaching L2 compliments. *English Language Teaching* 3(3): 71–80.

Levontina, I. 2004. Molodec, umnica. In *Novyj ob"jasnitel'nyj slovar' sinonimov russkogo jazyka* [The new explanatory dictionary of Russian synonyms], ed. J. Apresjan, 543–545. Moskva-Wien: Jazyki slavjanskoj kul'tury-Wiener Slawistischer Almanach.

Lewis, R. 1999. *When cultures collide: Managing successfully across cultures*. London: Nicholas Brealey Publishing.

ORED – M. Wheeler (ed.). 1980. *The Oxford Russian-English dictionary*. Oxford: Clarendon Press.

Pavlenko, Aneta. 2002. Emotions and the body in Russian and English. *Pragmatics and Cognition* 10(1–2): 207–241.

Peeters, Bert (ed.). 2006. *Semantic primes and universal grammar: Empirical evidence from the romance languages*. Amsterdam: John Benjamins.

Pesmen, D. 2000. *Russia and soul: An exploration*. Ithaca: Cornell University Press.

Quinn, N. 2003. Cultural selves. *Annals of the New York Academy of Sciences* 1001: 145–176.

Quinn, N. 2005. Universals of child rearing. *Anthropological Theory* 5(4): 477–516.

Richmond, Yale. 2009. *From Nyet to Da: Understanding the new Russians*, 4th ed. Boston/London: Intercultural Press.

Ries, Nancy. 1997. *Russian talk: Culture and conversation during Perestroika*. Ithaca/London: Cornell University Press.

Russian National Corpus. http://www.ruscorpora.ru. Accessed Jan 2013.

Searle, John R. 1975. A taxonomy of illocutionary acts. In *Language, mind, and knowledge*, Minneapolis studies in the philosophy of science, vol. 7, ed. K. Günderson, 344–369. Minneapolis: University of Minneapolis Press.

Šmelev, Alexey. 2005. Širota russkoj duši [The breadth of the Russian soul]. In *Ključevye idei russkoj jazykovoj kartiny mira* [Key ideas of the Russian linguistic picture of the world], ed. Anna A Zalizniak, I. Levontina, and A. Šmelev, 51–63. Moskva: Jazyki slavjanskoj kul'tury.

SRJ – A. Evgen'eva (ed.). 1981–1984. *Slovar' russkogo jazyka* [Russian dictionary], 4 Vols. 2nd ed. Moscow: Russkij Jazyk.

Uryson, E. 2004. Um, razum, rassudok, intellekt. In *Novyj ob"jasnitel'nyj slovar' sinonimov russkogo jazyka* [The new explanatory dictionary of Russian synonyms], ed. J. Apresjan, 37–44. Moskva-Wien: Jazyki slavjanskoj kul'tury-Wiener Slawistischer Almanach, 1203–1206.

Visson, Lynn. 2001. *Wedded strangers: The challenges of Russian-American marriages*. New York: Hippocrene Books.

Wierzbicka, Anna. 1992. *Semantics, culture, and cognition: Universal human concepts in culture-specific configurations*. Oxford: Oxford University Press.

Wierzbicka, Anna. 1998. Russian emotion expression. *Ethos* 26(4): 456–483.

Wierzbicka, Anna. 2002. Russian cultural scripts: The theory of cultural scripts and its applications. *Ethos* 30(4): 401–432.

Wierzbicka, Anna. 2003[1991]. *Cross-cultural pragmatics*, 2nd ed. Berlin: Mouton de Gruyter.

Wierzbicka, Anna. 2004. The English expressions "good boy" and "good girl" and cultural models of child rearing. *Culture and Psychology* 10(3): 251–278.

Wierzbicka, Anna. 2006. *English: Meaning and culture*. Oxford: Oxford University Press.

Williams, D., and J. Ispa. 1999. A comparison of the child-rearing goals of Russian and U.S. university students. *Journal of Cross-Cultural Psychology* 30(4): 540–546.

Wong, Jock. 2006. Social hierarchy in the "speech culture" of Singapore. In *Ethnopragmatics: Understanding discourse in cultural context*, ed. C. Goddard, 99–125. Berlin: Mouton de Gruyter.

Wong, Jock. 2010. The "triple articulation" of language. *Journal of Pragmatics* 42: 2932–2944.

Part IV
Book reviews

Corpus Linguistics: Methods, Theory and Practice by Tony McEnery and Andrew Hardie

Dawn Knight

'Corpus Linguistics: Methods, Theory and Practice' provides the reader with a good balance of detailed and interesting facts, figures and findings from the history and use of corpus analysis as well as in-depth discussions of the theoretical underpinnings of corpus linguistics. It documents how corpus linguistics perhaps 'lives' and 'breathes', and how the diverse nature of its utility and scope have helped to contribute to its ubiquitous appeal in modern day academia. Corpus linguistic approaches are ever-increasingly being seamlessly integrated with other methodologies used in applied linguistics (and beyond), and this textbook functions to present and discuss such developments in an accessible yet thought-provoking way. Building on seminal, entry-level works that exist in this field (such as those by Biber et al. 1998; Kennedy 1998; McEnery et al. 2006; Adolphs 2006), this volume manages to reflect a certain level of maturity in its content, maturity that is reflective of advances witnessed in corpus linguistics over the past half century or so.

This book does not simply present 'hands-on' step-by-step guidance of building, accessing and analysing corpora, material that is already extensively documented in other textbooks and is typically covered as part of standard BA and MA level corpus linguistics modules at British Universities. Instead, it focuses on promoting a certain level of critical engagement and reflective thinking from its readers; exposing them to the past and present in the corpus linguistics landscape and paving the way for discussions about its far-reaching potential applications for the future.

The textbook can be particularly recommended for those who are either about to or are currently embarking on PhD studies as well as providing a solid, more general, resource for those already working in the field. So those wanting to brush up their skills and knowledge of the different facets of modern corpus linguistics. It will also,

D. Knight (✉)
Lecturer in Applied Linguistics and TESOL School of Education, Communication and Language Sciences, Newcastle University, Newcastle upon Tyne, UK
e-mail: Dawn.Knight@ncl.ac.uk

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_13,

perhaps most importantly, prove to be useful for those who are tempted to dip their big toes into the world of corpus linguistics, as it is likely to overcome preliminary reservations about the usefulness of corpus linguistics and provide some good scaffolding for own their studies. For such reasons, I feel, this volume is extremely well-positioned within the Cambridge Textbooks in Linguistics series.

The content of the book is presented in a clear and persuasive way, covering all of the key topics one would expect to be addressed within a volume on corpus linguistics, as well as some welcome surprise additions on lesser-documented particulars. A specific focus on, for example, English Corpus Linguistics and particular schools of thought (Chapter 4); the different generations of concordancing tools (Chapter 2) and the multi-dimensional (MD) approach (Chapter 5) all provide an enhanced depth and additional perspective to the rest of the content that is presented.

The glossary is extensive and offers users with clear and informative definitions of key terminology often encountered in corpus linguistic research. It is clear to readers, throughout, that this textbook has been written by leaders in the field, and the thorough and impartial accounts of the methods, theories and practices that are presented provide the readers with a high level of confidence and reassurance that the content is something that is to be revered.

Discussions of the notion of corpus linguistics and a theory vs. method are particularly useful for people embarking on the field for the first time as this provides them with a well-balanced insight into the origins and current debates that are relevant to corpus linguistics. Undertones of this particular line of debate can be found to resonate throughout the book, but are paid particular attention in chapter 6. Corpus analysis is arguably more than a method, although the specific characteristics of this theory/method are ever-changing and difficult to definitively categorise. So discussions such as these are vital to foreground in any seminal corpus linguistics publication and are made even more persuasive by the additional inclusion of detailed accounts of how corpus analysis 'fits' in to the landscape of linguistics in general. The chapter on the convergence of corpus linguistics, psycholinguistics and functionalist linguistics is particularly powerful to this regard (Chapter 8).

Each chapter provides the reader with some practical activities and questions for discussion. The latter of which is likely to be particularly useful for MA and PhD level students, to get them thinking about how corpus linguistics methodologies can be adapted and utilised for the purposes of their own research, and to inspire the next generation of people working with corpus linguistics to push-the-boundaries in even more directions than already is the case.

A word of warning must be regarded, however, before embarking on these activities because, even from chapter one, it appears that it is essential for readers to have a certain level of practical experience of corpus analysis and using concordancers before the begin. This is perhaps not a text for complete beginners for this reason. This textbook does not, nor does it claim to, provide a step-by-step guide to these particular processes so, as already commented, it best functions to enhance the basic content that is already being presented in academic modules at various institutions, the latter of which will undoubtedly provide the more rudimentary stepping stones as a precursor to this textbook.

While the content of this textbook is both insightful and relevant throughout, the structure could perhaps be presented more logically than it currently is. So, with a basic overview (as with chapter 1) first; comments on the history of (English) Corpus Linguistics second (Chapter 4); the methodological, practical and ethical issues third (Chapters 2, 3 and 7) and then more detailed discussions on corpus analysis fourth (as with chapters 5, 6 and 8). Having said this, such a revised structure may have relinquished a certain level of the originality that the textbook contains, as this is perhaps the typical structure that the authors contemporaries utilise, myself included. Given that it is a textbook, it is not perhaps intended to be read from cover to cover, so the present structure may indeed more indicative and appropriate to this nature and need to be 'dipped in' to at any point. The reader can pick up this text and, from any point, derive a clear sense of narrative from the start to the end of individual chapters – perhaps what goes before and after the chapter is less important that this fact.

Overall, this refreshing textbook both deserves, and is likely to attract, a wide readership from within the field of applied linguistics and beyond. This is a recommended addition to any Higher Education module or programme on language and a key reference text for any discerning bookshelf.

References

Adolphs, S. 2006. *Introducing electronic text analysis*. London: Routledge.

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Kennedy, G. 1998. *An introduction to corpus linguistics*. London: Longman.

McEnery, T., R. Xiao, and Y. Tono. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.

Cyberpragmatics. Internet-Mediated Communication in Context by Francisco Yus

Francisco Javier Díaz-Pérez

It is an undeniable fact that Internet-mediated communication has had an enormous and increasing impact on our society in the last 20 years. The internet has become a pervasive and almost indispensable tool not only in our work but also in our leisure. It would be difficult to imagine our lives today without the Internet and without the different forms of communication available on the web, such as the electronic mail, the virtual conversation or the increasingly present and influential social networks. These new forms of interaction have influenced human communication in general and also the type of language human beings use to interact with other human beings. More than a decade ago, Francisco Yus initiated his research into the use of language on the web with the publication of his book *Ciberpragmática. El uso del lenguaje en internet*, in 2001. Afterwards, in 2007, he published a book entitled *Virtualidades reales. Nuevas formas de comunidad en la era de Internet*.

The monograph under review is an expansion on earlier work by the author and, particularly, on the first book mentioned above, which in 2010 had a second edition with the title *Ciberpragmática 2.0. Nuevos usos del lenguaje en internet*. The theoretical framework adopted in those publications is cognitive pragmatics and, more specifically, Relevance Theory.

Cyberpragmatics. Internet-Mediated Communication in Context consists of eight chapters preceded by an introduction and followed by a 46 page bibliography. As mentioned in the introduction, cyberpragmatics, a coinage by the author himself, "aims at applying pragmatics to Internet users' interactions" (xi). In this introductory chapter, the author also anticipates and justifies the theoretical framework he will use in the subsequent chapters, Relevance Theory. In the author's opinion, Relevance Theory has proved to be useful to explain face-to-face interaction as well as asynchronous communication. The only difference involved in Internet-mediated communication, Yus claims, is the way communication is achieved and the means

F.J. Díaz-Pérez (✉)
English Department, Faculty of Arts and Education, University of Jaén, Jaén, Spain
e-mail: fjdiaz@ujaen.es

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_14,

created by human beings to be used in interactions. A chapter-by-chapter summary of the contents of the book is also included in the introduction.

After providing a summary of the main concepts and theoretical assumptions of Relevance Theory – such as ostension, mutual manifestness, cognitive effects, or the principle of relevance –, chapter 1, "Pragmatics, context and relevance", focuses on the notion of cyberpragmatics and makes some claims about Internet-mediated communication from the perspective of Relevance Theory. Thus, for instance, it is claimed that context plays a decisive role in the production as well as in the interpretation of information on the web. There is no difference, in this respect, between Internet-mediated communication and face-to-face interaction. Certain features of interaction on the web, as well as some typical aspects of Internet users' behaviour are explained from the perspective of Relevance Theory. The resort to techniques to endow the text with oral connotations, the excess of information, or spam e-mail messages are some of the aspects on which this chapter focuses.

The second chapter, "The presentation of the self in everyday web use", addresses the topic of the presentation of self identity in virtual settings. As put forward by Yus, there is a growing tendency towards hybridization between traditional physical communication and virtual Internet-supported interactions. This tendency to hybridization has an effect on the notion of identity, which is constructed in a variety of physical-virtual groupings and interactive environments. Other forms of self-presentation on the Internet which are considered in this chapter are the personal web page and the nickname or nick.

In chapter 3, "Relevance on the web page", the author applies Relevance Theory to the analysis of web environments. The author refers to three perspectives which can be adopted to undertake a cognitive pragmatics analysis of web pages, namely, from the author's point of view, from the textual or discursive point of view, and from the reader's point of view. The relevance or irrelevance of communication through websites is analysed, which involves paying attention to the roles of addresser user and addressee user, the levels or patterns of interactivity and the availability of information on the Internet. With respect to this latter aspect, Yus introduces the concept *infoxication*, blending of information and intoxication which refers to a mental intoxication due to an excess of information. This excess of information may have negative consequences for eventual relevance, since it may require an increase in processing effort without an offset of cognitive effects. A section is devoted to the notion of usability, which is considered from the perspective of Relevance Theory. The chapter finishes with a discussion of the transference of two offline types of discourse to the Net. In particular, the chapter focuses on the transformation of printed newspapers into cybernewspapers and on the change of printed advertisements into banners and pop-up advertisements.

Chapter 4, "Social networks on the Internet: The Web 2.0", focuses on asynchronous web environments, such as blogs, social networking sites and *Twitter*, and analyses how interaction occurs in the context of the Web 2.0. These relatively new forms of interaction, which put an emphasis on sociality, interactivity and mutuality of information, have represented a revolution in the world of Internet communication. Blogs, or weblogs, are analysed from the perspective of author, content, reader,

and interactivity. The difference between blogs and personal web pages, to which chapter 2 is devoted, is the social orientation of blogs and the possibility of interaction with other users they offer. One of the inherent features of social networking sites, such as *Facebook, MySpace*, or the Spanish *Tuenti*, is the inclusion of the profile as the basic unit for content sharing and interaction among users. By means of these profiles, as put forward by Yus, users make a self-presentation of themselves and make manifest potentially relevant information. The chapter closes with a section devoted to the microblog *Twitter*, a short-messaging service with a clear social orientation and accessible from several devices (the Net, mobile phones, or PC tablets).

In Chapter 5, "The virtual conversation", attention is focused on virtual synchronous conversations. These virtual conversations take place in chat rooms and through instant messaging with messaging software such as *Skype* and *Messenger*. According to Yus, the most outstanding feature of this type of conversation is what he himself calls *oralized written text*. In other words, participants in virtual conversations tend to resort to certain strategies which make the written text a more expressive type of discourse and more similar to face-to-face interactions. A new type of virtual conversation is introduced in the last part of the chapter, namely, that in which users interact by means of 3D avatars or alter-egos that display non-verbal behaviour. The 3D virtual world on which the chapter focuses is *Second Life*, which according to Yus has interesting attributes for a pragmatic analysis of Internet-mediated communication. Videoconferencing, to which the last section of the chapter is devoted, represents the highest rank of the contextualization scale, as it is very similar to face-to-face dialogues.

The sixth chapter – "You've got mail" –, which borrows its title from a well-known film, concentrates on the e-mail genre, which covers private e-mail, newsgroups, and e-mail distribution lists. After describing the most important features of the genre, the main elements of an e-mail message (sender, addressee, e-mail address, subject line, body of the message, and signature) are analysed, paying particular attention to their role in the final interpretation. Electronic mail, like the virtual conversation, presents features of oral and written communication. Those features are dealt with in the third section of this chapter following the four dimensions of analysis proposed by Baron (1998), namely *social dynamics, format, grammar*, and *style*. As Yus observes, one of the most important pragmatic characteristics of the electronic mail is that it is, using his own words, "an ostensive technological medium" (238). In that sense, e-mail messages have the characteristics of any act of ostensive communication, which implies that they carry the presumption of their eventual relevance.

Chapter 7 addresses the topic of how politeness is expressed on the Net. Different theories and approaches to the study of politeness are considered, such as Brown and Levinson's (1987) Politeness Theory and maxim-based approaches to politeness (Lakoff 1973; Leech 1983). As stated by Yus, the lack of physical co-presence has resulted in the existence of uncontrolled conversational strategies associated with rudeness, which receive the name of *flaming*. The chapter finishes with an attempt to couple politeness and Relevance Theory. After Escandell Vidal (1998, 2004) it is

concluded that in order to ensure effective communication with interactants from other countries, a *default level of politeness* should be assumed, which, as a consequence of the use of English as a lingua franca on the Internet, has been identified with the Anglo-Saxon use of politeness.

The final chapter of the book, "Conclusion: Prospects for cyberpragmatic research", is, as its title indicates, devoted to general conclusions and future research suggestions. Yus goes back to his initial assumption that the ways of processing information and obtaining the intended interpretation in a given context are the same in Internet-mediated communication and in offline communication. However, the availability of contextual information as well as the message attributes may have an effect on how the balance between cognitive effects and processing effort is assessed while trying to obtain a relevant interpretation. In this sense according to Yus, "a central goal of cyberpragmatics is to analyse the role of this contextualization in the interpretation of utterances transferred through the Net and will remain central in the future" (289). Current technological advances are taken into account, such as the use of mobile telephones to access the Net. In this sense, this chapter explores the differences in the way information is presented and processed in computer screens and on mobile phone screens, as well as the pragmatic consequences derived from those differences.

This volume is an excellent contribution to the study of human communication in general and to Internet-mediated communication in particular. Any future study in this field will inevitably have to refer to this work, which will certainly be of interest for any researcher in pragmatics, computer-mediated communication, or Relevance Theory. Due to its depth of analysis, some background knowledge is required to be able to fully understand this text. Therefore, it is probably not an accurate reference work for undergraduate students. However, the topic is particularly appealing to young people, and consequently, some extracts could be used in undergraduate courses. In particular, chapter 1 contains a very good summary of the basics of Relevance Theory, which could be really useful in Pragmatics undergraduate courses. In addition, the book includes a comprehensive review of existing literature.

Chapter 7, which deals with politeness, could probably have reviewed other more recent theories of verbal politeness, such as those by Watts (2003) or Locher and Watts (2005). There is no reference to a special issue of the *Journal of Politeness Research* devoted to politeness in computer-mediated-communication (Locher 2010) or to the volume edited by Bousfield and Locher (2008) on impoliteness in language. Instead, the book only relies on widely criticised theories, such as those based on maxims (Lakoff 1973; Leech 1983) and Brown and Levinson's (1987) Politeness Theory, which, though very influential in subsequent research in politeness phenomena, has not escaped criticism either. In addition, Yus's claim about the adoption of Anglo-Saxon politeness as a global norm on the Net should probably be supported by further research to be confirmed.

The inclusion of several figures or diagrams throughout the book to illustrate different topics in a graphic way is a very good support to the written text. In this sense, something which could also have helped to improve the monograph would be

the insertion of screenshots displaying the interfaces of the different communication tools dealt with. This could be particularly useful for those readers who are not familiar with the interfaces of some of these interaction tools. In fact, figure 3.3., on page 84, contains the interface of the Spanish newspaper *El País* for reading printed news online, but it is the only case in which the interface of a website or communication tool is presented. The reason might be that while in this case permission was granted by *El País* to reproduce that figure — as the author acknowledges in a footnote —, in other cases permission may not have been achieved.

As mentioned above, the book contains a comprehensive bibliography as well as a good review of studies carried out in the field of Internet-mediated communication. Not only does it build on a previous research by the author but it also opens up new avenues for future research which will surely be explored by the author himself and by other researchers.

References

Baron, Naomi. 1998. Letters by phone or speech by other means: The linguistics of email. *Language and Communication* 18: 133–170.

Bousfield, Derek, and Miriam A. Locher (eds.). 2008. *Impoliteness in language: Studies on its interplay with power in theory and practice*. Berlin: Mouton de Gruyter.

Brown, Penelope, and Stephen Levinson. 1987. *Politeness: Some universals in language use*. Cambridge: Cambridge University Press.

Escandell Vidal, Victoria. 1998. Politeness: A relevant issue for relevance theory. *Revista Alicantina de Estudios Ingleses* 11: 45–57.

Escandell-Vidal, Victoria. 2004. Norms and principles. Putting social and cognitive pragmatics together. In *Current trends in the pragmatics of Spanish*, ed. Rosina Márquez-Reiter and María Elena Placencia, 347–371. Amsterdam/Philadelphia: John Benjamins.

Lakoff, Robin. 1973. The logic of politeness; or, minding your P's and Q's. Papers from the Ninth Regional Meeting of the Chicago Linguistic Society, 292–305.

Leech, Geoffrey. 1983. *Principles of pragmatics*. London: Longman.

Locher, Miriam A. (ed.). 2010. *Journal of Politeness Research* 6 (Special issue on politeness and impoliteness in computer-mediated communication).

Locher, Miriam A., and Richard J. Watts. 2005. Politeness theory and relational work. *Journal of Politeness Research* 1(1): 9–33.

Watts, Richard J. 2003. *Politeness*. Cambridge: Cambridge University Press.

Yus, Francisco. 2001. *Ciberpragmática. El uso del lenguaje en Internet*. Barcelona: Ariel.

Yus, Francisco. 2007. *Virtualidades reales. Nuevas formas de comunidad en la era de Internet*. Alicante: University of Alicante.

Yus, Francisco. 2010. *Ciberpragmática 2.0. Nuevos usos del lenguaje en Internet*. Barcelona: Ariel.

Author Index

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3, © Springer Science+Business Media Dordrecht 2013

Subject Index

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3, © Springer Science+Business Media Dordrecht 2013

Q

R

S