New Frontiers in Translation Studies

Michael Carl
Srinivas Bangalore
Moritz Schaeffer *Editors*

New Directions in Empirical Translation Process Research

Exploring the CRITT TPR-DB

Springer

New Frontiers in Translation Studies

Series editor
Defeng Li
Centre for Translation Studies, SOAS, University of London, London, United Kingdom

Centre for Studies of Translation, Interpreting and Cognition, University of Macau, Macau SAR

More information about this series at
http://www.springer.com/series/11894

Michael Carl • Srinivas Bangalore •
Moritz Schaeffer
Editors

New Directions in Empirical Translation Process Research

Exploring the CRITT TPR-DB

Springer

Editors
Michael Carl
Center for Research and Innovation in
Translation and Translation Technology
Department of International
Business Communication
Copenhagen Business School
Frederiksberg, Denmark

Srinivas Bangalore
Interactions Corporation
New Providence
New Jersey, USA

Moritz Schaeffer
Center for Research and Innovation in
Translation and Translation Technology
Department of International
Business Communication
Copenhagen Business School
Frederiksberg, Denmark

ISSN 2197-8689 ISSN 2197-8697 (electronic)
New Frontiers in Translation Studies
ISBN 978-3-319-20357-7 ISBN 978-3-319-20358-4 (eBook)
DOI 10.1007/978-3-319-20358-4

Library of Congress Control Number: 2015945979

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

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

General Editor's Preface

New Frontiers in Translation Studies, as its name suggests, is a Series which focuses on new and emerging themes in Translation Studies. The last four decades have witnessed a rapid growth of this fledgling discipline. This Series intends to publish and promote these developments and provide readers with theories and methods they need to carry out their own translation studies projects.

Translation Studies is now expanding into new or underexplored areas both in theories and research methods. One recent development is the keen interest in translation theories that transcend Eurocentrism. Translation Studies has for decades been dominated by Western modes of understanding and theorizing about translation and closed to models of other traditions. This is due to, as many have argued, the "unavailability of reliable data and systematic analysis of translation activities in non-European cultures" (Hung and Wakabayashi 2005). So in the past few years, some scholars have attempted to make available literature on translation from non-European traditions (Cheung 2006). Several conferences have been held with themes devoted to Asian translation traditions. Besides, rather than developing translation theories via a shift to focusing on non-Eurocentric approaches, efforts have been directed towards investigating translation universals applicable across all languages, cultures, and traditions.

Modern Translation Studies has adopted an interdisciplinary approach from its inception. Besides tapping into theories and concepts of neighboring disciplines, such as linguistics, anthropology, education, sociology, and literary studies, it has also borrowed research models and methods from other disciplines. In the late 1970s, German translation scholars applied Think-aloud Protocols (TAPs) of cognitive psychology in their investigation of translators' mental processes, and more recently, process researchers have incorporated into their research designs lab methods, such as eye-tracker, EEG, and fMRI. In the early 1990s, computational and corpus linguistics was introduced into Translation Studies, which has since generated a proliferation of studies on the so-called translation universals, translator style, and features of translated language. Studies on interpreting and translation education have also taken a data-based empirical approach and yielded interesting and useful results.

As Translation Studies seeks further growth as an independent discipline and recognition from outside the translation studies community, the interest to explore beyond the Eurocentric translation traditions will continue to grow. So does the need to adopt more data- and lab-based methods in the investigations of translation and interpreting. It is therefore the intent of this Series to capture the newest developments in these areas and promote research along these lines. The monographs or edited volumes in this Series will be selected because of either their focus on non-European translation traditions or their application of innovative research methods and models, or both.

We hope that translation teachers and researchers, as well as graduate students, will use these books in order to get acquainted with new ideas and frontiers in Translation Studies, carry out their own innovative projects, and even contribute to the Series with their pioneering research.

Defeng Li

References

Cheung, M. (2006). *An anthology of Chinese discourse on translation, volume one: From earliest times to the Buddhist project*. Manchester/Kinderhook: St. Jerome Publishing.

Hung, E., & Wakabayashi, J. (2005). *Asian translation traditions*. Manchester/Northampton: St. Jerome.

Foreword

The appearance of the present volume coincides with the 10th anniversary of CRITT, the Center for Research and Innovation in Translation and Translation Technology, which was inaugurated on 10 June 2005. As it happens, the publication of the book also coincides with the 20th anniversary of the development of the first version of Translog, which my son Lasse Schou programmed for me towards the end of 1995 (when he was 15). The idea of the program came to me because I had become interested in knowing about the mental processes involved in translating and had learnt elementary programming. I had become somewhat frustrated with my own attempts at analyzing verbal data from think-aloud experiments, which was the dominant methodological paradigm at the time following the publication of Ericsson and Simon's influential *Protocol Analysis: Verbal Reports as Data* (1984; 2nd ed. 1993). Therefore I was trying to think of a way of getting harder, less subjective data as a corrective to, or control on, inferences based on think-aloud data. I first intended Translog only as an instrument I would use to log timed keystrokes in my personal research, but the program quickly generated broader interest and soon colleagues were contributing ideas for additional features, primarily in the CBS TRAP project (1996–2002) and in the international Translation Expertise group of researchers generously funded by the University of Oslo (1999–2005). One important outcome of meetings and publications in the context of the Translation Expertise group activities was the idea of the CRITT center at CBS. Another major outcome was our successful application for the EU *Eye-to-IT* project (2006–2009), which made it possible to thoroughly re-program Translog (2006) so that it would accept UTF8-encoded characters, present output in xml, accept data from an eye-tracker via a gaze-to-word mapping program developed at the University of Tampere, and have many other new features. The *Eye-to-IT* project also made it possible for CRITT to recruit researchers, among them Michael Carl (in 2008), the present director of the Center. With his machine translation background and his blend of computational expertise and complete dedication to research, he gave the Center's research a new direction and was the main driver in working out the successful EU CASMACAT project proposal. In the context of this project (2011–2014), apart from developing an interactive post-editing workbench and contributing the promised deliverables,

he oversaw the development of Translog II, made sure that key and gaze data were properly integrated and recorded, and organized the construction of the TPR-DB, now probably the largest database anywhere of key and gaze process data from translation and post-editing sessions, so that both legacy recordings and recent recordings would all have the same data formats. All of the contributions in the present book are based on recordings stored in the TPR-DB and testimony to its huge value as a TPR resource.

For the past 5 years, CRITT has offered a 1-week summer course for PhD students with an interest in TPR. This has been enormously gratifying both on professional and a personal level. Some students have come back; more have stayed in contact and now constitute a network ranging from Brazil to Canada and from China and India to most of the countries in Western Europe from north to south. Some of the participants in the "early" days (2011) have returned as co-instructors. It is truly gratifying to see that at least one of the authors of each of the 14 contributions to this book has attended one of the TPR summer courses.

Following a chance meeting at a conference in India between Michael Carl and Srinivas Bangalore, then at AT&T in New Jersey, they agreed to run an 8-week workshop called SEECAT at CBS in 2013. Most of the participants came from leading universities and IT institutions in India. The aim of the workshop was to implement voice recognition as well as gaze control of certain screen operations in a translation workbench solution. In the course of 8 weeks, this aim was achieved and prototypes recognizing not just English but Hindi and (less successfully) Danish were produced.

A 4-week follow-up workshop aimed at developing a new TPR subdiscipline to be called Translation Data Analytics (TDA) was run by Michael Carl, Srinivas Bangalore, and Moritz Schaeffer in July–August 2014. Here, participants worked in teams on developing the appropriate computational, statistical, and other analytical tools that would constitute TDA and make it possible, by applying TDA to large-scale process data of the kind stored in the TPR-DB, to produce reliable descriptions of and predictions about, e.g., translator profiles, the use of default translation strategies, and the occurrence of revision patterns and of predictable delays triggered by syntactic and word order rearrangements, all of which will contribute to generating a process-oriented model of human translation.

All of this, in much more detail, is what the reader will find in this most welcome celebration of 10 years of CRITT.

Frederiksberg, Denmark Arnt Lykke Jakobsen

List of Contributors

Vicent Alabau Sciling SL, Valéncia, Spain

Fabio Alves Laboratory for Experimentation in Translation (LETRA), Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil

Marceli Aquino Laboratory for Experimentation in Translation (LETRA), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

Laura Winther Balling Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

Srinivas Bangalore Interactions Corporation, New Providence, NJ, USA

Bergljot Behrens Department of Literature, Area studies and European Languages, University of Oslo, Oslo, Norway

Michael Carl Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

Francisco Casacuberta Pattern Recognition and Human Language Technology research center (PRHLT), Technical University of Valencia, Valencia, Spain

Lidia S. Chao Department of Computer and Information Science, University of Macau, Macau, China

Igor Antônio Lourenço da Silva Universidade Federal de Uberlândia (UFU), Uberlândia, Brazil

Joke Daems Department of Translation, Interpreting and Communication, Language and Translation Technology Team, Ghent University, Ghent, Belgium

Arthur de Melo Sá Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil

Barbara Dragsted Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

Norma Fonseca Linguistic Studies Department, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

Ulrich Germann Machine Translation Group, School of Informatics, University of Edinburgh, Edinburgh, UK

Maheshwar Ghankot Indian Space Research Organisation, Hassan, India

José Luiz Gonçalves Universidade Federal de Ouro Preto (UFOP), and Laboratory of Experimentation in Translation (LETRA/UFMG), Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Jesús González-Rubio Unbabel Lda, Samora Correia, Portugal

Robert Hartsuiker Department of Experimental Psychology, University of Ghent, Ghent, Belgium

Arndt Heilmann English Linguistics Department, RWTH Aachen, Aachen, Germany

Kristian Tangsgaard Hvelplund Department of English, Germanic and Romance Studies, University of Copenhagen, Copenhagen, Denmark

Arnt Lykke Jakobsen Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

Arlene Koglin Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Samuel Läubli Machine Translation Group, School of Informatics, University of Edinburgh, Edinburgh, UK

Ana L. V. Leal Department of Portuguese, University of Macau, Macau, China

Lieve Macken Department of Translation, Interpreting and Communication, Ghent University, Ghent, Belgium

Mercedes García Martínez Computer Laboratory, University of Maine, Le Mans, France

Bartolomé Mesa-Lao Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

Jean Nitzke Department for Language, Culture and Translation Studies in Germersheim (FTSK), University of Mainz, Mainz, Germany

Daniel Ortiz-Martínez Pattern Recognition and Human Language Technology research center (PRHLT), and Statistics Department, Technical University of Valencia, Valencia, Spain

Katharina Oster Department for Language, Culture and Translation Studies Germersheim (FTSK), University of Mainz, Mainz, Germany

Adriana Pagano Laboratory for Experimentation in Translation, Federal University of Minas Gerais, Belo Horizonte, Brazil

Dagmara Płońska University of Social Sciences and Humanities, Warsaw, Poland

Paulo Quaresma Department of Informatics, Universidade de Évora, Évora, Portugal

Germán Sanchis-Trilles Sciling SL, Valencia, Spain

Moritz Schaeffer Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

Institute for Language, Cognition and Computation University of Edinburgh, Edinburgh, UK

Márcia Schmaltz Department of Portuguese, University of Macau, Macau, China

Kyoko Sekino Laboratory for Experimentation in Translation (LETRA), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

Annegret Sturm University of Geneva, 36, rue Prévost-Martin, 1205 Geneva

Karina Sarto Szpak Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil

Sonia Vandepitte Department of Translation, Interpreting and Communication, Ghent University, Ghent, Belgium

Derek F. Wong Department of Computer and Information Science, University of Macau, Macau, China

Julian Zapata School of Translation and Interpretation, University of Ottawa, Ottawa, Canada

Contents

Part I
Empirical TPR

Chapter 1
Introduction and Overview

Michael Carl, Srinivas Bangalore, and Moritz Schaeffer

Abstract *New Directions in Empirical Translation Process Research* is a continuation of the development which originates in descriptive translation studies as conceived by Holmes (1972) and Toury (1995). This introduction shows how this volume is a documentation of a technological development which makes it possible for translation research to go beyond the description. As the various chapters in this volume argue, the analysis of records from keyloggers and eye-trackers enable us to "explain and predict" (Holmes, 1972:71) translators' behaviour on various levels of granularity. All contributions are centered around the CRITT TPR-DB, a unique resource of more than 500 h of recorded translation process data augmented with over 200 different annotations. The chapters describe aspects of computational, statistical and psycholinguistic models of the translation process that are facilitated by the TPR-DB. This chapter gives an overview of the contributions and provides a background for the work reported in the volume.

Keywords Predictive translation process studies • Computational • Statistical and psycholinguistic modelling of the translation process

For centuries, the discourse on translation was rather *prescriptive*, as Horace, in his *Ars Poetica* (circa 20 BC) may exemplify: "do not strive, as a literal translator, to render texts word for word" (in Hardison and Golden 1995: 11). For centuries, scholars debated how a translator should or should not translate. Many translators

M. Carl (✉)
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark
e-mail: mc.ibc@cbs.dk

S. Bangalore
Interactions Corporation, New Providence, NJ, USA

M. Schaeffer
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

Institute for Language, Cognition and Computation, University of Edinburgh, Edinburgh, UK

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_1

and theorists after Horace were equally prescriptive in their writing on translation (cf. Robinson 1997).

In 1972, Holmes (1972) made the case for research on translations to be *descriptive*. He produced a map of what is now called translation studies and applied standard scientific methods to the study of translations by arguing that "... translation studies thus has two main objectives: (1) to describe the phenomena of translating and translation(s) ..., and (2) to establish general principles by means of which these phenomena can be explained and predicted." (71) Following Holmes, Toury (1995) turned scholars' attention away from questions regarding whether a translation is equivalent or not by setting out the methods and theoretical framework for what became known as descriptive translation studies. One of Toury's central hypotheses was that translation is a norm governed activity, and with the availability of large quantities of translated text in electronic form, corpus-based translation studies set out to find empirical evidence for these norms (e.g. Mauranen and Kujamäki 2004). While corpus-based translation studies has been prolific in the production of hypotheses regarding norms found in target texts, it has not been easy to draw inferences from these regarding the translation process.

We are now at a stage in the development where translation research becomes *predictive*. The records from keylogging software and eye-trackers make it possible to address Holmes' (1972) second main objective, to "explain and predict" translators' behaviour: at present, we have all the necessary tools to address the challenge of building a model of human translation which makes specific, falsifiable predictions regarding the process and the product of translation. Perhaps the most fundamental question in this regard is to determine the mechanisms underlying the production of translations which are common to all translators. This babelian question attempts to find, on the one hand, the cognitive processes which are shared among all translators during the translation of diverse language combinations; it is the quest for linguistic and cognitive universals of translation. On the other hand, from a utilitarian viewpoint, having a model which can predict translators' behaviour makes it possible to design translator assistance just when it is needed. It will allow us to automate those aspects of the translation process that save mechanical effort, so that the translator can dedicate their full attention to those aspects which cannot be automatized.

Some 20 years ago it was very difficult to base any investigation of human translation processes on empirical observations. Any finding regarding cognitive processes during translation was either based on an analysis of the final product, i.e. the target text itself, or on Think Aloud Protocols (TAPs) (Krings 1986; Lörscher 1991). In TAPs, participants are asked to verbalize their thoughts during a concurrent task such as translation. While studies using TAPs have been highly valuable in the investigation of the cognitive processes during translation, the very act of verbalizing thoughts has been shown to have a considerable effect on the cognitive processes during translation (Jakobsen 2003). However, given the technology used in modern Translation Process Research (TPR), and as exemplified in this volume, it is possible to have "... a structured record of the exact temporal succession of translators' eye and hand activity ... "(Jakobsen 2011: 47)and it is

therefore possible to "... ask old questions in a new way and begin to formulate tentative answers to them ..." (ibid).

An important landmark for empirical TPR was set up in 1995 by a group of researchers at the Copenhagen Business School when developing a data-acquisition software, Translog (Jakobsen and Schou 1999)with which translators' keystroke could be recorded, replayed and analysed. In contrast to previous TAP elicitation methods, a keylogger runs in the background so as not to interfere with the writing or translation process. In a replay mode the translation processes can be visualized and analysed. Since 2009, this program has been extended with an eye-tracker interface, so that gaze activities can also be logged (Carl 2012). If connected to an eye-tracker, Translog-II records gaze-sample points, computes gaze fixations and maps the fixations to the closest character on the screen. The Translog tool and the emerging research activities around it have given rise to the foundation of the Center for Research in Translation and Translation Technology (CRITT) in 2005, and has resulted in considerable research which has been reported, amongst others, in a number of edited volumes published within the Copenhagen Studies in Language series, in volumes 24, 27, 35–39 and 41 (Hansen 1999, 2002; Pöchhacker et al. 2007; Göpferich and Jakobsen 2008; Göpferich et al. 2010; Mees et al. 2010a,b; Sharp et al. 2011).

Since then, three developments have given rise to the research reported in this volume. The first development is related to the extension of Translog for languages with different scripts and a tighter integration of eye-trackers; second to apply empirical TPR methods to investigate and predict processes of human-machine interaction in computer aided translation and third the collection of a large amount of translation process data in a translation process research database, (TPR-DB), so as to arrive at generalizable results. The large set of language combinations in the TPR-DB and multiple translation modes have made it possible to arrive at statistically reliable results. To this end, a consistent and transparent representation for logging the diverse input modalities across different languages, and scripts was needed.

Within Translog-II the first requirement was addressed by replacing Translog's initial keyboard logging method with *text-diff logging*[1] method that records differences in the emerging texts, rather than memorizing the pressed keystrokes. For languages written in the Latin script, there is an isomorphism between the produced keystrokes and the modifications in the text, which does not exist for some other scripts, such as, e.g., Chinese or Japanese. These logographic scripts make use of special input methods, such as e.g. *SoGou* (see Chap. 11), with the effect that the relation between the pressed keys and the characters that appear on the screen cannot be reproduced from the keystroke log only. Switching from keystroke logging to text-diff logging in Translog-II was triggered by the requirement for language and script independency, so that now, irrespectively of the script, Translog-II encodes the text modifications in UTF-8 and stores it in an XML file. At the same time a tight

[1] Most papers which use the TPR-DB, not only in this volume, still refer to this as *keylogging*, even though, strictly speaking, this is actually not correct.

integration with eye-trackers was achieved, which is now functional for TOBII, SMI and eyelink eye-trackers. As a consequence, data records are compatible and can be compared across different languages.

The second development concerns the increasing interest of TPR to study the interaction of computer assisted translation and human translation processes. The importance of human-computer interaction in translation has been acknowledged since the early days. The ALPAC report (ALPAC 1966) suggested that studies and computer applications should be supported "for speeding up the human translation process" and for "the production of adequate reference works for the translator, including the adaptation of glossaries ..." (ALPAC 1966: 34), In 1980, concrete suggestions were made how such systems could be implemented (Kay 1998), although, until recently, the investigation of cognitive processes in computer-assisted translation has not been a topic of concern for TPR. As some of the chapters in this volume describe, TPR has practical implications when investigating how translation assistance is used in translator's every-day applications, and what technologies are suitable to support the underlying cognitive processes—a field of research labelled translator-computer interaction (TCI), or, as proposed in Chap. 7, translator-information interaction (TII).

In order to study cognitive processes underlying the task of post-editing machine translation, the Translog-II system that was originally designed to investigate reading, writing and translation processes, was extended with an operation mode to record sessions of post-editing machine translation. The machine-translated text would appear in an editable text box, which a post-editor would edit to create the final translation of the text. Text modifications would be recorded, in addition to the gaze data, if an eye-tracker were to be connected. However, Translog-II does not provide an experimental environment similar to real working conditions. Translog-II presents two running texts in a source and target window, while modern translation aides, such as translation memories, segment the texts into fragments and present each source segment with its translation in a more structured manner.

In order to obtain a more realistic picture of professional translators' working styles and to assess how to support their translation processes with advanced machine translation technology, the CASMACAT project (see Chaps. 3–8, but also Sanchis-Trilles et al. 2014; Alabau et al. 2013; Koehn et al. 2013) has implemented an advanced state-of-the-art, browser-based post-editing environment and combines this with Translog-II style keyboard logging and eye-tracking possibilities. In this way, detailed empirical data can be collected from a realistic translation environment, with the hope that the assessment of this data would lead to a more complete picture and better predictive models of human cognitive processes during computer-aided translation.

The third development concerns the creation of a large database of TPR data. Given the compatible representation of Translog-II in its various languages, scripts and operation modes, it became possible to collect data from different studies into one single repository and to process them in a generic and consistent manner (see Chap. 2). The TPR-DB stores Translog-II data from reading, writing, translation, copying and post-editing experiments, as well as CASMACAT translation sessions

in a single format, with common metrics, which make it possible to analyse the data from different language combinations and translation modes consistently. The TPR-DB is therefore ideally suited as a resource to answer questions regarding cognitive processes during translation and post-editing, reading and copying which are shared across different individuals and different language combinations. It facilitates the generation and validation of hypotheses regarding translation processes across different language combinations and different translation modes. Since the database contains a large number of different languages and many language-agnostic features, it is now possible to verify these predictions, as illustrated by the range of studies reported in this volume.

This volume is, hence, centered around the CRITT TPR-DB, a unique resource of more than 500 h recorded translation process data, augmented with over 200 different rich annotations. Chapter 2 introduces the CRITT TPR Database, which is a publicly available database of recorded text production (writing, copying, translation) sessions for TPR. It contains user activity data (UAD) of translators, editors, post-editors and authors' behaviour recorded with Translog-II and with the CASMACAT workbench. In addition to the raw logging data, the TPR-DB consists of tables with rich features set that can be easily processed by various visualization and analysis tools.

The remaining 12 chapters make up part II and part III of this book, which describe the diverse directions in translation process research, including computational, statistical and psycholinguistic modelling that is facilitated by the TPR data. The second part of this book is dedicated to the CASMACAT post-editing workbench, outlining implementation details and usability issues of interactive machine translation, the usage of external resources and translator-information interaction. The third part contains studies modeling the human translation process.

Chapter 3 describes the integration of online and active learning techniques in the CASMACAT. The foundations of current phrase-based statistical machine translation (SMT) model, the mathematical basis for interactive translation prediction (ITP), and the use of online and active learning for translation are discussed in this chapter. During online learning (OL), modifications by the translators are immediately learned by the system with the aim of preventing the same errors in the machine generated translations. During active learning (AL), only a subset of the machine generated translations with worst quality are post-edited, the SMT model is re-trained with the new translation example, and finally, the improved SMT system returns the remaining (presumably correct) translations. The chapter also presents a pilot evaluation with translators using the system. Results showed that translators using the ITP systems incorporating OL required less typing effort and had increased post-editing speed for 60 % of the translators.

Chapter 4 investigates the CASMACAT ITP post-editing mode with 'traditional' MT post-editing (PE) for the language pair English → Brazilian Portuguese using metrics to quantify the temporal, technical and cognitive post-editing effort. Two

medical texts from the EMEA corpus[2] were post-edited by 16 participants with recordings of their gaze and keyboard activity. The measured effort was correlated with an objectively computed score, Translation Edit Rate (TER) that was designed to compare translations of a text. While the authors found that the technical effort is higher for ITP than in the PE mode, the cognitive effort in ITP is lower than for post-editing due to shorter fixation durations.

Based on the assumption, that interactive post-editing (ITP) is a new technology that post-editors need to get acquainted with, Chap. 5 compares the CASMACAT ITP and traditional post-editing modes (PE) in a longitudinal study (LS14), to investigate whether and how the performance of professional post-editors improved over time when working with ITP. Five post-editors used both modes over a period of 6 weeks in which their activity data was recorded. In a second experiment (CFT14), the translators' learned behaviour was compared with a control group of post-editors who did not have experience with ITP. It was found that the technical post-editing effort, as measured by the ratio of coherent production time divided by the overall post-editing time, was lower after the 6 weeks period of using the ITP than the technical effort measured in the control group in CFT14 study who had not worked with ITP before.

Chapters 6 and 7 highlight the use of external resources during translation, post-editing and post-editing with online learning. As these contributions show, usage of external resources is an important aspect, which can account for more than 50 % of the total translation time (CITE).

Chapter 6 discusses "the effectiveness of consulting external resources during translation and post-editing of general text types" by analysing 40 from-scratch translation sessions and 40 post-editing sessions of 10 master's level translation students, using the CASMACAT workbench. The usage of external resources was recorded with Inputlog, and 'infused' into the CASMACAT logfile. In this way, the authors were able to go beyond previous studies which were restricted to manual assessment of external resource usage or on only one type of external resource. The study found that translation students spend significantly more time in external resources when translating from scratch, compared to post-editing. No statistically confirmative evidence was found to suggest that different types of resources were used during translation compared to post-editing. However, longer consultation of external resources during from-scratch translation correlated with higher translation quality, while consultation of external resources during post-editing correlated with lower translation quality.

Chapter 7 concludes the first part of this volume with a broader view on "translator-information interaction" (TII), that is, translators' interaction with (digital) information and information tools. The study is based on the CFT14 data, mentioned in Chap. 5, and investigates the interaction of post-editors with the CASMACAT BiConc tool (biconcordancer). On the basis of screen recordings and a total of 55 instances of BiConc usage, it was found that four of the seven participants

[2] http://opus.lingfil.uu.se/EMEA.php

in this study did not use BiConc. Participants who used BiConc also used other Internet resources, such as term banks, dictionaries and corpora, to complement their information retrieval efforts, and those who did not use the CASMACAT BiConc also used fewer external resources overall. Factors such as relevance and trust seem to play an important role in the usage of external resources, since only 47 % of the CASMACAT BiConc searches were adopted by participants.

The third part of the volume is concerned with cognitive and statistical modeling of human translation processes, the investigation of multilingual co-activation and priming effects at the lexical, syntactic and discourse levels of granularity, translation literality and syntactic annotation schemata.

Chapter 8 starts with the assumption that there are three human translation processes (HTPs) during post-editing of machine translation output: orientation, revision and pausing. Since these processes are not directly observable in the logging data, the authors conceptualize the recognition of these phases as a Hidden Markov process. The logging data is automatically segmented into fragments of 3–10 s and transformed into vectors of observations O. The observations are automatically clustered, and Hidden Markov models trained with the observations where the cluster labels serve as output symbols of the Hidden Markov models. The aim of the model is to yield the most probable HTP for each observation o *in* O, taking into account (1) the feature values (dimensions) of the current observation and (2) the HTPs assigned to the preceding observations $o_1, o_2, \ldots, o_n$. In a final step the cluster labels are mapped on the three HTPs: orientation, revision and pause. The authors show that the system reaches as high an accuracy to predict the times spent on orientation, revision and pause as some of the human annotators.

There has been a long tradition of studying priming effects in comprehension and production models of human sentence processing. More recently, effects of lexical priming in translation tasks have been observed.

Chapter 9 shows that translators are primed in terms of semantics and syntax already during very early stages of the reading process. Two features of the TPR-DB, i.e., relative word order (*Cross*) and word translation entropy (*HTra*), are used to predict first fixation durations, among other early eye movement measures. A first fixation duration is the time a reader spends on a word, before either re-fixating that same word or before moving the gaze to a different word. This chapter shows that reading of a source text word leads to the automatic activation of shared semantic and structural representations. This chapter further shows that these primed representations serve as the basis for later, conscious processes during which the source text is regenerated in the target language. The results presented in this chapter further suggest that word recognition is essentially non-selective, i.e., during the early stages of reading, the reader makes no distinction regarding the language to which a word belongs and both linguistic systems are co-activated. Implications for models of the bilingual lexicon are discussed.

In Chap. 10 the authors provide evidence of priming at the level of syntactic structure. By introducing a concept of syntactic entropy—a measure of uncertainty for a translator to pick syntactic structures for a target sentence given a source sentence—the authors correlate syntactic entropy with the observable measurements

found in the TPR database such as the time spent reading either the source text or the target and typing speed. They demonstrate positive correlations between syntactic entropy and the durations for translation activities, in translation tasks across a few language pairs. In a monolingual copy task these correlations between syntactic entropy and behavioural measures are not observed, lending support to the claim that not only the lexicon but also syntactic structures might be co-activated for the two languages.

Chapter 11 investigates translation and post-editing processes of cohesive chains in translations from Portuguese to Chinese. One group of participants translated and another group of participants post-edited the same text. Eye movements and keyboard activity for two cohesive chains were analysed. Establishing a semantic relationship between the words in one of these chains relied on the general lexicon of the language, while doing the same for the other chain required text-local relationships. It was hypothesized that establishing text-local semantic relationships was more difficult than establishing semantic relationships on the basis of the general lexicon. The authors find that the type of chain has an effect on eye movements on the target text and on keyboard activity, suggesting that cohesion is established mainly during target text production. The task had no effect on the processing of cohesive chains, suggesting that cohesive chains are processed similarly in post-editing and translation.

Typing Chinese texts involves using a graphical user interface which converts sequences of Alphabetic letters into Chinese characters. This chapter describes also how process data from this different input method for text is captured.

Chapters 12, 13 and 14 re-consider and underpin some of the basic units and annotations, on different levels of granularity and on the level of the translation product and translation process data that were assumed in the previous of the assumptions in the previous chapters, the notion of basic human translation processes activity unit orientation, revision and pausing used in (HTPs) during post-editing of machine translation output:

Chapter 12 discusses the merits of three possible ways of operationalizing restructuring of source material in the target text. The first of these possibilities is the one reported in the context of Chap. 10. The author points at the fact that the annotations which were used for the analysis in Chap. 10 were relatively shallow and, by analysing small number of examples in minute detail, suggests how the annotation could be improved in order to better capture the variation in the alternative translations. In addition, the author discusses the merits of an annotation system used in a large product-based corpus and argues that this annotation system would most likely capture more fine-grained details which are not covered by the annotation system used for the analyses presented in Chap. 10. However, this corpus does not contain process data. Finally, the most promising annotation schema which might best capture restructuring effort is discussed in the final sections of the chapter. This annotation schema uses relevance theoretical notions applied to translation.

Chapter 13 presents an experiment which investigates the claim that novices translate more literally than professionals. Previous research suggests that novices

translate more literally than professional translators, because novices focus less on a representation of the whole text at a discourse level than professionals do, who rely less on linguistic equivalence and take into account more world knowledge and pragmatic considerations. Three groups of twenty (non-professional bilinguals, student translators and professional translators) took part in the experiment which had two conditions: translating after a first reading of the source text in addition to producing a summary of the source text in the target language versus translating straight away without a first reading and summary. Results showed that students translated freer than professionals, but initial reading and summary of the source text had a different effect on professionals and students: students translated more literally after a first reading and professionals translated freer. The definition of literality used in this chapter is different to the one used in Chap. 9. The target texts in Chap. 13 are annotated manually, while the definition of literality used in Chap. 9 is generated automatically. A comparison between these two measures shows, however, that they are significantly correlated.

Chapter 14 introduces an alternate annotation of the user activity data and suggests methods that provide visualizations that may be easier for visual analytics of the translation process data. The chapter goes on to discuss and quantify the differences in translation-from-scratch and post-editing activities for general purpose texts as compared to domain-specific texts. As it might be expected, the time for post-editing is shorter than for translation-from-scratch independent of the domain of the texts, the keystroke activity is less and the gaze on the target text is more for post-editing domain texts.

The volume assembles a number of studies that explore possibilities for predictive modelling of human translation processes, which, we believe, opens prospectives for new directions in empirical translation process research.

References

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., et al. (2013). CASMACAT: An open source workbench for advanced computer aided translation. *The Prague Bulletin of Mathematical Linguistics, 100*, 101–112.

ALPAC. (1966). *Languages and machines: Computers in translation and linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council (124 pp.). Washington, D.C.: National Academy of Sciences, National Research Council (Publication 1416).

Carl, M. (2012). Translog-II: A program for recording user activity data for empirical reading and writing research. In *Proceedings of the 8th international conference on language resources and evaluation (LREC)* (pp. 4108–4112), Istanbul, Turkey.

Göpferich, S., & Jakobsen, A. L. (Eds.). (2008). *Looking at eyes* (Copenhagen studies in language, Vol. 36). Frederiksberg: Samfundslitteratur.

Göpferich, S., Jakobsen, A. L., & Mees, I. M. (Eds.). (2010). *Behind the mind* (Copenhagen studies in language, Vol. 37). Frederiksberg: Samfundslitteratur.

Hansen, G. (Ed.). (1999). *Probing the process in translation: Methods and results* (Copenhagen studies in language, Vol. 24). Denmark: Samfundslitteratur.

Hansen, G. (Ed.). (2002). *Empirical translation studies: Process and product* (Copenhagen studies in language, Vol. 27). Denmark: Samfundslitteratur.

Hardison, O. B., & Golden, L. (1995). *Horace for students of literature. The 'Ars Poetica' and its tradition*. Miami: University Press of Florida.

Holmes, J. S. (1972). The name and nature of translation studies. In *Translation section of the third international congress of applied linguistics*, August 21–26 (pp. 66–79). Copenhagen.

Jakobsen, A. (2003). Effects of think aloud on translation speed, revision and segmentation. In *F. Alves (Ed.), Triangulating translation: Perspectives in process oriented research* (pp. 69–95). Amsterdam: Benjamins.

Jakobsen, A., & Schou, L. (1999). Translog documentation. In *G. Hansen (Ed.), Probing the process in translation: Methods and results* (pp. 1–36). Frederiksberg: Samfundslitteratur.

Jakobsen, A. L. (2011). Tracking translators' keystrokes and eye movements with Translog. In *C. Alvstad, A. Hild, & E. Tiselius (Eds.), Methods and strategies of process research integrative approaches in translation studies* (pp. 37–55). Amsterdam: John Benjamins Publishing

Kay, M. (1998). The proper place of men and machines in language translation. In *Readings in machine translation* (pp. 221–232), MIT Press.

Koehn, P., Carl, M., Casacuberta, F., & Marcos, E. (2013). CASMACAT: Cognitive analysis and statistical methods for advanced computer aided translation. In A. Way, K. Sima'an, M. L. Forcada, D. Grasmick, & H. Depraetere (Eds.), *Proceedings of the XIV Machine Translation Summit* (p. 411). Allschwil: European Association for Machine Translation.

Krings, H. P. (1986). *Was in Den Köpfen von Übersetzern Vorgeht: Eine Empirische Untersuchung Zur Struktur Des Übersetzungsprozesses an Fortgeschrittenen Französischlernern*. Tübingen: Günter Narr Verlag.

Lörscher, W. (1991). *Translation performance, translation process, and translation strategies. A psycholinguistic investigation*. Tübingen: Günter Narr Verlag.

Mauranen, A., & Kujamäki, P. (2004). *Translation universals: Do they exist?* Amsterdam; Philadelphia: John Benjamins.

Mees, I. M., Alves, F., & Göpferich, S. (Eds.). (2010a). *Methodology, technology and innovation in translation process research* (Copenhagen studies in language, Vol. 38). Frederiksberg: Samfundslitteratur.

Mees, I., Göpferich, S., & Alves, F. (Eds.). (2010b). *New approaches in translation process research* (Copenhagen studies in language, Vol. 39). Frederiksberg: Samfundslitteratur.

Pöchhacker, F., Jakobsen, A. L., & Mees, I. M. (Eds.). (2007). *Interpreting studies and beyond. A tribute to Miriam Shlesinger* (Copenhagen studies in language, Vol. 35). Denmark: Samfundslitteratur.

Robinson, D. (1997). *Western translation theory: From Herodotus to Nietzsche*. Manchester: St. Jerome Publishing.

Sanchis-Trilles, G., Alabau, V., Buck, C., Carl, M., Casacuberta, F., & Martinez, M. G. (2014). Interactive translation prediction versus conventional post-editing in practice: A study with the CasMaCat workbench. *Machine Translation, 28*(3-4), 217–235.

Sharp, B., Zock, M., Carl, M., & Jakobsen, A. L. (Eds.). (2011). *Human-machine interaction in translation* (Copenhagen studies in language, Vol. 41). Frederiksberg: Samfundslitteratur.

Toury, G. (1995). *Descriptive translation studies and beyond*. Amsterdam; Philadelphia: John Benjamins.

Chapter 2
The CRITT Translation Process Research Database

Michael Carl, Moritz Schaeffer, and Srinivas Bangalore

Abstract Since its existence 10 years ago, the Center for Research and Innovation in Translation and Translation Technology (CRITT) at the Copenhagen Business School has been involved in Translation Process Research (TPR). TPR data was initially collected by the Translog tool and released in 2012 as a Translation Process Research Database (TPR-DB). Since 2012 many more experiments have been conducted and more data has been added to the TPR-DB. In particular, within the CASMACAT (Sanchis-Trilles et al. 2014) project a large amount of TPR data for post-editing machine translation was recorded and the TPR-DB has been made publicly available under a creative commons license. At the time of this writing, the TPR-DB contains almost 30 studies of translation, post-editing, revision, authoring and copying tasks, recorded with Translog and with the CASMACAT workbench. Each study consists of between 8 and more than 100 recording sessions, involving more than 300 translators. Currently, the data amounts to more than 500 h of text production time gathered in more than 1400 sessions with more than 600,000 translated words in more than 10 different target languages.

This chapter describes the features and visualization options of the TPR-DB. This database contains recorded logging data, as well as derived and annotated information assembled in seven kinds of simple and compound process—and product units which are suited to investigate human and computer-assisted translation processes and advanced user modelling.

Keywords Empirical translation process research • Translation process research database

M. Carl (✉)
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark
e-mail: mc.ibc@cbs.dk

M. Schaeffer
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

Institute for Language, Cognition and Computation, University of Edinburgh, Edinburgh, UK

S. Bangalore
Interactions Corporation, New Providence, NJ, USA

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_2

2.1 Introduction

Empirical translation process research requires the availability of suitable process data. Thus, in order to allow for empirically grounded translation process research, Jakobsen and Schou (1999) have devised—in 1995—a keyboard logging tool, Translog, with which translation sessions could be recorded, the data visualized and statistically analyzed. Since then, Translog, the data acquisition tool, and the format and representation of the collected process data have undergone a number of changes (Jakobsen 2011) so as to allow for more powerful analyses of the data: The current Translog-II (Carl 2012a) has been complemented with the CASMACAT workbench (Sanchis-Trilles et al. 2014, see also Chap. 3 in this volume) as a browser-based machine translation post-editing tool and the raw logging data gathered at the output of the recorded translation sessions can be enriched with annotations and converted into a Translation Process Research Database (TPR-DB). As of now, the TPR-DB has accumulated a large amount of process data, with the aim to:

1. Represent activity data for TPR in a consistent manner, so as to facilitate research across hundreds of translation sessions, different languages and different translation modes.
2. Implement and make available a large number of features across the entire collected dataset which would be difficult or nearly impossible to compute individually for each session separately.

The aim of the TPR-DB is thus to stimulate and lower the barrier of entry for large-scale translation process research facilitated by a consistent database format and a well-defined set of features.[1]

The TPR-DB is organized in studies and sessions. As described in various chapters in this volume, a study is a collection of sessions that are conducted in the same experimental context. Translog and CASMACAT generate a single log file for each session. This raw logging data is subsequently annotated and processed into a set of *tables* that contain a rich set of features and attributes.

This chapter describes the tables and the features that are extracted from logged and annotated data. Section 2.2 provides an overview of the TPR-DB; it describes the process of annotating the data logged from a translation session, their mapping into the TPR-DB, and gives an overview over the TPR-DB tables. Sections 2.3, 2.4 and 2.5 describe the tables in more detail. Section 2.3 tackles the tables that encode single keystrokes and fixations. Section 2.4 illustrates tables of production and fixation units. A special property of those units is parallel and alternating reading and typing behavior that indicates the workload of a translator. Section 2.5 describes the tables of translation product units, i.e. units that are derived from

[1]The database is freely available under a creative commons license, and can be downloaded free of charge from https://sites.google.com/site/centretranslationinnovation/tpr-db

the final translation product: source tokens, target tokens and alignment units. Section 2.6 shows visualization possibilities of the process data and Sect. 2.7 points to possibilities for adding externally generated features to the TPR-DB. Three appendixes complement this chapter, Appendices 1 and 2 give an overview of the studies in the TPR-DB. An exhaustive list of features is given in the Appendix 3.

2.2 Overview of the TPR-DB

The CRITT TPR Database is a publicly available database of recorded translation (and other text production) sessions. It contains user activity data (UAD) of translators behaviour collected in almost 30 studies of translation, post-editing, revision, authoring and copying tasks, recorded with the CASMACAT workbench (Sanchis-Trilles et al. 2014) and with Translog-II (Carl 2012a). Each study consists of between 8 and more than 100 recording sessions. Currently, the data amounts to more than 500 h of text production time gathered in more than 1400 sessions and more than 600,000 translated words in more than 10 different target languages. The TPR-DB website[1] makes available all the data logged during a translation process (>20 GB), as well as an annotation enriched translation process research database (TPR-DB, zipped 170 MB), both under a creative commons license. In this section we describe how the data logged during a translation session is transformed into the TPR-DB.

2.2.1 TPR-DB Compilation

The raw User Activity Data (UAD), which includes the translation process data, such as keystrokes, fixations, mouse movements, as well as the translation product data, i.e. the source text and the final translation product is stored and maintained in a subversion[1] repository. Within a TPR-DB compilation process[2] (Carl 2012b), a number of tables are generated from the raw UAD, which can then be used as a basis for further analysis and visualization, as shown in the various chapters in this volume.

Figure 2.1 shows a processing float chart of the TPR-DB compilation process. The logged UAD data (labeled *Translog-II* in Fig. 2.1) is processed in two independent streams, to annotate the product data (top) and process data (bottom). Annotations of the product data, i.e. the source and the target texts, include tokenization, sentence and token alignment and (optionally) lemmatization, PoS tagging among others. Translog-II also offers the possibility to adjust and annotate

[2]While the Translog-II and CASMACAT logged UAD is slightly different, the structure of the generated tables is identical.

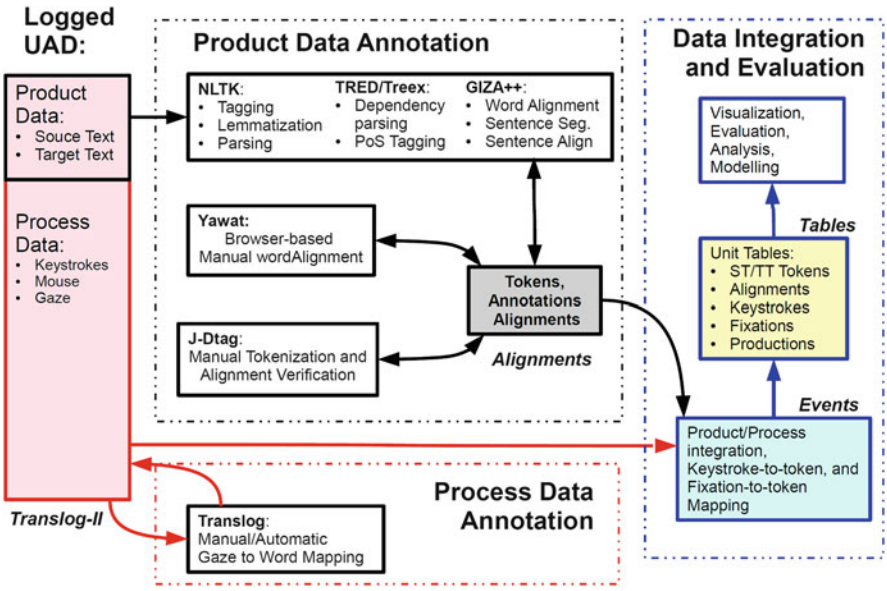


Fig. 2.1 Architecture of the TPR-DB compilation process

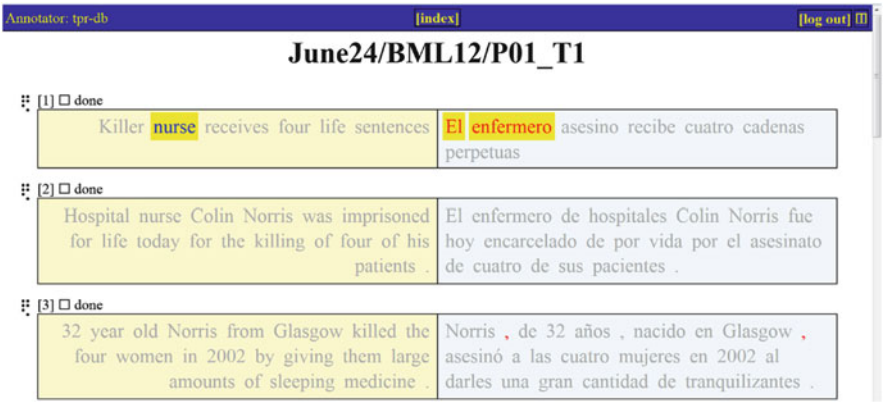


Fig. 2.2 Screenshot of YAWAT, a browser-based word alignment tool

process data, such as manually gaze to word re-mapping. Further, a data integration step computes keystroke-to-token and fixation-to-token mappings, as described in (Carl 2012a). Finally, a number of different *tables* are produced, which contain a large number of features, describing various kinds of product and process units, as described in Sect. 2.2.3.

The TPR-DB compilation process is fully automatic, but provides a GUI in which word alignments can be semi-automatically adjusted. Figure 2.2 shows the YAWAT

(Germann 2008) GUI in which word translation alignments can be highlighted, verified and amended.

2.2.2 TPR-DB Studies

Each study is a coherent collection of translation or text production sessions, which may invoke different tasks. All studies in the TPR-DB contain recorded key logging, and a large number also contains eye-tracking data. Each study in the TPR-DB was conducted with a (set of) research question(s) in mind, which can be roughly summarized as follows:

- The TPR-DB contains nine studies conducted with the different CASMACAT workbenches to test and evaluate its different functionalities. Chapters 3–8 in this volume report on details of these studies.
- Seven studies are part of a multilingual experiment to compare from-scratch translation, post-editing and monolingual post-editing, for English into six different target languages by more than 120 different translators. Chapters 9 and 10 analyze in detail these studies.
- In addition, the TPR-DB contains a few individual experiments that were conducted with Translog-II, the purposes of which are described in Appendix 1. Chapters 11 and 13 in this volume report on two different studies.

Appendix 1 gives a detailed overview of the various studies collected in the TPR-DB, their purposes, participants and durations. Table 2.1 is an excerpt from Appendix 1. It shows the summary information of the CFT14 study, which is also the basis of investigations in Chaps. 5, 7 and 8.

Each study consists of one or more recording sessions (*Sess*), with a number of different participants (*Part*), texts, and translation direction with a source language (*SL*) and a target language (*TL*).

For instance, the CFT14 study in Table 2.1 has three different tasks R, P and PIO.[3] Two English source texts (*SL* = en) were post-edited into Spanish (*TL* = es).

Table 2.1 Excerpt from table in Appendix 1

Study	Task	Part	Sess	Texts	SL	TL	FDur	KDur	PDur	SLen	TLen
CFT14	R	4	14	14	en	es	4.54	1.21	0.28	2901	40,614
CFT14	P	7	7	2	en	es	16.51	7.90	3.41	2901	20,273
CFT14	PIO	7	7	2	en	es	15.68	7.98	3.49	2901	20,341

[3]The example is taken from a CASMACT study. Tasks are R: revision, P: post-editing and PIO: Interactive post-editing with online learning. A full list of task descriptions is in Appendix 1.

Seven participants produced seven translations (sessions) for each of the P and the PIO tasks, and 4 participants subsequently reviewed the 14 post-edited texts.

The total production time is given in terms of *FDur*, *KDur*, and *PDur*, which represent the sum of the durations for all sessions, excluding pauses before the first keystroke and after the last keystroke, as well as pauses between successive keystrokes depending on the pause length:

- *Dur* provides the entire session duration
- *FDur* excludes pauses longer than 200 s between successive keystrokes
- *KDur* is the session time excluding inter keystroke pauses longer than 5 s
- *PDur* typing time with no keystroke pauses longer than 1 s

In Table 2.1, the total duration (*FDur*) for post-editing (P) of the 7 texts by the 7 post-editors took 16.51 h. Two additional duration values indicate typing durations. According to the *KDur* value post-editors were typing roughly 50 % of that time (7.90 h) while based on *PDur* it was only 3.41 h, or approximately 20 % of the post-editing time.[4] Table 2.1 shows also average source text length (*SLen*) and the total number of produced target language words (*TLen).*

2.2.3 TPR-DB Summary Tables

The TPR-DB compilation process projects the raw UAD into various different product and process units that are gathered in TPR-DB *tables*. Each line in a table describes a particular unit with a number of attributes that will be described in detail in Sects. 2.3, 2.4 and 2.5[5]:

Basic product unit tables are:

1. Source tokens (ST): this table lists ST tokens, together with TT correspondence, number of insertions and deletions needed to produce the translation, micro unit information section etc. *(Sect. 2.4.6)*
2. Target tokens (TT): this table lists TT tokens together with their ST correspondence, number of insertions and deletions to produce the TT, micro unit information, amount of parallel reading activity during typing, etc. *(Sect. 2.4.6)*

[4] A large number of different pause thresholds have been suggested and are used. Vandepitte et al. (2015) segment keystroke sequences at 200 ms, while Lacruz and Shreve (2014: 250) find that "complete editing events are separated by long pauses (5 s or more.) They normally contain short pauses (more than 0.5 s, but less than 2 s,) and more effortful complete editing events will often include multiple short pauses. Post-editors may make intermediate duration pauses (more than 2 s, but less than 5 s) during a complete editing event". Jakobsen (2005) suggests 2.4 s for his definition of "key performance".

[5] The letters in brackets in the list represent the file extensions in the TPR-DB. The section in italics points to the section where the table is described in more detail.

Composed product unit tables are:

3. Session (SS): this table describes session-related properties, such as source and target languages, total duration of session, beginning and end of drafting, etc. (*Sect. 2.3.1)*
4. Segments (SG): this table lists properties of the aligned source and target text segments, including duration of segment production, number of insertions and deletions, number and duration of fixations, etc. (*Sect. 2.3.2)*
5. Alignment units (AU): this table lists ST-TT alignment units, together with the number of keystrokes (insertions and deletions) needed to produce the translation, micro unit information, amount of parallel reading activity during AU production, etc. (*Sect. 2.4.1)*

Basic process unit tables are:

6. Keystroke data (KD): this table enumerates text modification operations (insertions and deletions), together with time of keystroke, and the word in the final text to which the keystroke contributes (*Sect. 2.5.1)*
7. Fixation data (FD): this table enumerates fixations on the source or target text, defined by the starting time, end time and duration of fixation, as well as the offset of the fixated character and word in the source or target window (*Sect. 2.5.2)*

Composed process unit tables are:

8. Production units (PU): this table lists units of coherent sequences of typing activity of the session, defined by starting time, end time and duration, percentage of parallel reading activity during unit production, duration of production pause before typing onset, as well as number of insertion, deletions. (*Sect. 2.5.3)*
9. Fixation units (FU): this table lists coherent sequences of reading activity, characterized by a starting time, end time and duration, as well as scan path indexes to the fixated words *(Sect. 2.5.4)*
10. Activity Units (CU): this table provides a list of fragments of the session, where each fragment is defined by activities of typing, reading of the source or reading of the target text *(Sect. 2.5.5)*

In addition, usage of external resources is summarized in

11. External resources (EX): this table lists keylogging data that was recorded with Inputlog (Leiten and Waes 2013) *(Sect. 2.7.1)*

2.3 Session and Segment Summary Information

Depending on its design, a study consists of one or more sessions. During a session, a text is translated, copied, edited or revised, where each text has several sentences (i.e. segments). This section describes the study and the session summary information.

Table 2.2 General session information including length of the source and the target text in terms of token and characters

Study	Session	SL	TL	Part	Text	Task	SegST	SegTT	TokS	LenS	TokT	LenT	...
BML12	P01_E5	en	es	P01	5	E	6	6	139	788	153	840	...
BML12	P01_P4	en	es	P01	4	P	5	5	110	668	131	763	...
BML12	P01_T1	en	es	P01	1	T	10	10	160	838	180	964	...

Table 2.3 Session duration information

...	Dur	TimeD	TimeR	Pause	Fdur	Kdur	Pdur	Pnum	...
...	310,234	114,140	232,656	0	167,110	80,374	23,366	29	...
...	268,328	71,234	264,765	0	193,531	29,407	14,485	15	...
...	757,281	92,016	290,391	0	654,812	314,378	210,415	72	...

Table 2.4 Session processing information for Study BML12

...	FixS	TrtS	FixT	TrtT	Scatter	Mins	Mdel	Ains	Adel
...	3	167	661	68,214	17	85	93	0	0
...	551	78,224	236	18,668	9	77	62	0	0
...	1122	115,692	392	26,605	30	1152	186	0	0

2.3.1 Session Summary Information

Session summary information is contained in the session table (SS), as shown in Tables 2.2, 2.3, and 2.4, and can be divided into:

(a) General session information includes:

- The S*tudy* and the S*ession* name, that is, the directory and the log-data file name
- The source and target languages (*SL* and *TL*)
- A study-unique participant identifier (*Part*)
- A study-unique text identifier (*Text*)
- The *Task* type (as discussed in Table 2.1)
- A session-unique segment identifier *SegST* and *SegTT* which refer to the source and the target texts, as discussed below.
- *TokS*, *TokT* give the number of tokens (words) in the source- and target texts
- *LenS* and *LenT* are the length of source and target texts in characters.

(b) Session duration information indicates how long it took to process the including

- The total duration of the session (*Dur*) and the Pause duration (*Pause*) in case the session was interrupted.
- The beginning of the drafting time (*TimeD*) revision time (*TimeR*). *TimeD* indicates the time offset from the beginning of the session until the first

keystroke, which coincides with the end of the orientation phase. *TimeR* indicates the time when the drafting phase ended and the revision phase started. This is defined as the end of the first micro unit (see below) in which the last token of the source text was translated (cf. Jakobsen 2002).
- The durations *FDur*, *KDur*, and *PDur* were already discussed previously. The *PDur* interval fragments the UAD into production units (PUs), which will be discussed in Sect. 2.5. *Pnum* provides the number of PUs within a session.

(c) Session processing information provides keystrokes and gazing behaviour:

- *FixS* and *FixT* are the number of fixations on the source token(s) and on the target token(s), while *TrtS* and *TrtT* represents the total reading time, i.e. the sum of all fixation durations on the source and target text respectively.
- *Mins* and *Mdel* are the number of manually inserted and deleted characters, while *Ains* and *Adel* are the automatically inserted and deleted characters. *Ains* and *Adel* account for post-editing in CASMACAT where the edited text can be programmatically changed in the interactivity mode.
- The *Scatter* feature indicates how often the typing was not in a sequential order, i.e. how often the translator or editor typed successive keystrokes which were part of two or more different words.

Tables 2.2, 2.3, and 2.4 show three sessions from the *BML12* study, conducted by participant *P01*. Text 5 was edited (*Task* = *E*), text 4 was post-edited (*Task* = *P*) and text 1 was translated from scratch (*Task* = *T*). Translation took longest in terms of all available duration measures, *Dur*, *FDur KDur* and *PDur*, whereas post-editing was quicker than editing with respect to *Dur, KDur* and *Pdur*, but slower with respect to *FDur*. Not that editing was a more scattered activity than post-editing as many more PUs were produced.

2.3.2 Segment Summary Information

The session name *P01_T1* codes that participant *P01* translated (*T*) text 1. This text consists of 11 source text segments (*STseg*) that were translated into 10 target text segments (*TTseg*). The properties of these segments are given in more detail in the segment summary tables (SG), as shown in Table 2.5. Segment summary tables contain very similar information as the session tables, but each line in the table refers to a segment, instead of a session. Thus *Dur*, *FDur KDur* and *PDur* indicate the segment translation time, rather than the session translation times, as in Table 2.5. The columns *STseg* and *TTseg* indicate segment alignment information. Table 2.5 shows that all but the last two segments are aligned one-to-one. That is, the source segments (sentences) 10 and 11 were translated into one target sentence 10.

Table 2.5 Alignment unit process information

STseg	TTseg	Study	Session	Nedit	Dur	...	Scatter	Literal	HTra	HSeg	CrossS	CrossT
1	1	BML12	P01_T1	2	20,028	...	2	27.93	2.16	1.18	2	1.29
2	2	BML12	P01_T1	3	38,951	...	5	48.24	1.23	0.67	2	1.29
3	3	BML12	P01_T1	2	83,452	...	5	67.41	1.7	0.95	1.57	1.08
4	4	BML12	P01_T1	4	73,292	...	4	29.45	1.74	0.8	1	1.29
5	5	BML12	P01_T1	3	24,373	...	3	31.67	1.84	0.79	1.14	1.5
6	6	BML12	P01_T1	2	14,030	...	2	33.3	2.43	1.36	1.3	1.09
7	7	BML12	P01_T1	2	58,966	...	4	19.65	0.97	0.46	1.47	0.94
8	8	BML12	P01_T1	2	40,779	...	4	151.9	2.9	1.59	2.94	1.19
9	9	BML12	P01_T1	1	32,812	...	1	31.6	1.38	0.72	1.21	1.1
10 + 11	10	BML12	P01_T1	3	61,326	...	6	29.11	1.24	0.61	1.67	1.28

The *Nedit* attribute indicates how often the segment was revised. A number >1 indicates that the translator first drafted the translation and then came back later to revise it. For instance, segment 4 was drafted and then three times revised, whereas only *STseg* 9 was not revised during the translation process.

The features *Literal*, *HTra*, *HSeg*, *CrossS* and *CrossT* will be discussed in detail in Sects. 2.4.4 and 2.4.6 and there are many examples of their application throughout this volume. *CrossS* and *CrossT* measure the amount of syntactic similarity between the source and the target text. *HTra* and *HSeg* give the average word translation and average segmentation entropy while *Literal* is the sum of the product of *HTra* and *CrossS*.

2.4 Word Level Summary Information

This section introduces lower level word-based alignment units (AUs), source text tokens (ST) and target text tokens (TT). As most of the AU attributes appear also in the TT and ST units, we start by presenting AUs in Sect. 2.4.1. In Sect. 2.4.2, we discuss representations of micro units, and Sect. 2.4.3 introduces a typing (in)efficiency metric. Section 2.4.4 presents the *Cross* feature, which quantifies syntactic distortions between source and target texts. ST and TT tokens are introduced in Sect. 2.4.5 together with more detailed gaze information (Sect. 2.4.6) and word translation entropy in Sect. 2.4.6.

2.4.1 Alignment Units

Source and target tokens correspond to sequences of characters, usually separated by a blank, while AUs are to m-to-n source-to-target token correspondences. The unit tables provide a similar kind of information for these three different kinds of units. These tables contain:

- General information: *study* and *session* name, task type (*Task*), participant (*Part*), text number (*Text*), numbers of source and target segments (*STseg* and *TTseg)*, source and target languages
- Product information: including the source and target language strings (*SAU* and *TAU*), and the number of tokens of these strings (*SAUnbr* and *TAUnbr*), as well as the relation between the source and the target in terms of *Cross* values (see Sect. 2.4.5)
- Process information: number of keystrokes (insertions and deletions), production duration, gazing behaviour in terms of number of fixations on the source and on the target strings of the AU (*FixS* and *FixT*), total reading time (*TrtS* and *TrtT*) and first pass duration (*FPDrS* and *FPDurT*). These features will be explained in Sect. 2.4.6

Table 2.6 Alignment unit general information

AUid	STseg	TTseg	Study	Session	SL	TL	Task	Text	Part	SAU	TAU	SAUnbr	TAUnbr
44	3	3	BML12	P01_T1	en	es	T	1	P01	of	de	1	1
45	3	3	BML12	P01_T1	en	es	T	1	P01	sleeping_ medicine	tranquili- zantes	2	1

Table 2.7 Alignment unit process information

Ins	Del	Dur	FixS	FPDurS	TrtS	FixT	FPDurT	TrtT
24	21	11,407	2	167	167	18	50	1232
15	0	1610	27	631	1896	8	465	615

Table 2.8 Alignment unit process information

Cross	InEff	Munit	Edit
1	15	2	de_medicinas_para_dormir[rimrod_arap_sanicidem]
2	0.94	1	tranquilizantes

Tables 2.6, 2.7, and 2.8 show the English → Spanish translation in two AU_{44} and AU_{45} of "*of* ↔ *de*" and "*sleeping medicine* ↔ *tranquilizantes*". As indicated in the columns *SAUnbr* and *TAUnbr,* AU_{44} is a one-to-one correspondence, whereas AU_{45} is a two-to-one correspondence. The *Edit* column traces the sequence of keystrokes which were typed to produce the translation. It shows for AU_{44} that first "*de medicinas para dormir*" was typed but later "*medicinas para dormir*" was deleted, so that only "*de*" remained from that initial typing activity, while for AU_{45}, the translation "*tranquilizantes*" was typed with no revision. The table shows the overall number of keystrokes produced: for AU_{44} there were 24 insertions, of which 21 characters (the string in square brackets) were later deleted. Note that deletions are to be read in the reverse direction, so that reading "[*rimrod_arap_sanicidem*]" from right-to-left results in the deleted string. Even though "*medicinas para dormir*" and "*tranquilizantes*" are paraphrases, the former deleted string is part of AU_{44}, while the latter is part of AU_{45}. The assignment of multi-word deletions to words in the final text to which they contribute can only be approximated, so that an error margin to neighboring words should be expected. In line with Alves and Vale (2011), we refer to these revisions as micro units that will be discussed in Sect. 2.4.2.

The time needed to type the translation is given by the duration feature (*Dur*). In the example above, more than 11 s (11,407 ms) were needed for all the typing activities in AU_{44} while 1610 ms were needed to type AU_{44} "*tranquilizantes*".

Table 2.8 shows the total reading time (*TrtS* and *TrtT*) and number of fixations (*FixS* and *FixT*) on the source token(s) and on the target token(s). According to this information, the SAU word "*of*" in AU_{44} was fixated twice with a total reading time of 167 ms, while the translation *"de"* was fixated 12 times with a total reading time of 1232 ms. The source string in AU_{45} was fixated 27 times with a *TrtS* of 1896 ms and the target string received 8 fixations with a *TrtT* of 615 ms.

Table 2.9 First micro unit tranquilizantes

AUid	Edit1	Time1	Dur1	Pause1	FixS1	ParalS1	FixT1	ParalT1
44	de_medicinas_para_dormir	225,703	11,110	187	10	716	2	116
45	tranquilizantes	570,250	1610	172	0	0	9	536

Table 2.10 Micro unit 1 and micro unit 2

AUid	Edit2	Time2	Dur2	Pause2	FixT2	ParalS2	FixT2	ParalT2
44	[rimrod_arap_sanicidem]	569,781	297	22,937	0	0	4	214
45	–	0	0	0	0	0	0	0

2.4.2 Micro Units

Alves and Vale (2011) refer to recurring editing activities of the same word translations as micro units. For them, "a micro TU is defined as the flow of continuous TT production ... separated by pauses during the translation process" (Alves and Vale 2011: 107). A macro unit, then, is a collection of micro units "that comprises all the interim text productions that correspond to the translator's focus on the same ST segment" (Alves and Vale 2011: 107).

The TPR-DB computes units of "continuous TT production" as production units (see Sect. 2.4.5), and lists details of the first two micro units contributing to the production of a translation in the tables. The column *Munit* in Table 2.8 indicates how many micro units have contributed to the production of an AU. While there can be, in principle, any number of micro units—a translator may revise a piece of text very often—detailed information of the first two micro units are indicated as follows.

Tables 2.9 and 2.10 show the micro unit information for AU_{44} and AU_{45}. A micro unit has a starting Time and duration (*Dur*) of the typing activity, a pause preceding the typing activity (*Pause*), and the amount of concurrent reading activity in the source text (*ParalS*) and in the target text (*ParalT*). Most importantly, a micro unit is characterized by the actual typing activity, *Edit* string.

Tables 2.9 and 2.10 decompose the production activity in Table 2.8 into two micro units: at Time 225,703 the translator produces a first micro unit in AU_{44} by typing "*de medicinas para dormir*". During a revision phase more than 4 min later, at time 569,781 in micro unit 2 (Table 2.10), the string "*medicinas para dormir*" is deleted and replaced by "*tranquilizantes*" at Time 570,250 which is part of AU_{45}, micro unit 1 (Table 2.9). The duration of those activities is given, together with the pause following it and the concurrent gaze activity. Given the information in Table 2.3, we know that the revision phase (*TimeR*) started in this translation session at time 290,391. We hence see that micro unit 1 in AU_{44} takes place during translation drafting, while micro unit2 of AU_{44} and AU_{45} micro unit 1 emerge both as revision events.

2.4.3 Typing Inefficiency

The editing inefficiency (*InEff*) measures the ratio of the number of produced characters divided by the length of the final translation, which is approximately equivalent to the number of insertions and deletions divided by their difference as in Eq. (2.1):

$$\begin{aligned} InEff &= number\ of\ typed\ characters/length\ of\ final\ translation \\ &\approx Insertions + Deletions/Insertions–Deletions + 1, \end{aligned} \tag{2.1}$$

In most of the cases, the length of a word equals the number of character insertions minus character deletions + 1. We add 1 since the white space following the word is counted as being part of it. However, in some cases no white space follows a word, in which case the *InEff* value may be smaller than 1. Thus, for AU_{44} in Table 2.8 the number of the insertion and deletion keystrokes amounts to 45 which, divided by the length 3 of the final word "*of*" (including a white space), results in an editing inefficiency of 15, while the number of keystroke string to produce "*tranquilizantes*" in AU_{45} amounts to the length of the final translation, and thus the editing effort is 0.94. Note that for post-editing the *InEff* can be 0 if an MT proposal was accepted without any modifications, while it would be 2 if the word was deleted and another word of identical length was retyped.

2.4.4 Cross Feature

The *Cross* feature represents word translation alignment information as a local cross-lingual distortion. This distortion is direction dependent and can be conceived from the source to the target side (*CrossS)*, or from the target to the source side (*CrossT*). *Cross* values can perhaps be best thought of as a method to generate a sentence—from left-to-right—through the alignment links, by counting how many words have to be skipped in the one language in order to produce the next word in the other language.

Figure 2.3 gives an example from an English → Spanish translation. The figure shows two aligned sentences, the *Cross* value, and an enumeration of the tokens in the two sentences, in addition to the actual ST-TT links.

In order to produce the first Spanish TT word ("*El*"), two English words ("*Killer*" and "*nurse*") have to be skipped in the input text, which results in a *Cross* value of 2. Since the second input word ("*nurse*") produces two adjacent TT words, no further ST word has to be skipped to produce "enfermero", which results in a *Cross* value of 0. To produce the third Spanish word, "*asesino*", one ST word to the left of "*nurse*" has to be processed, leading to the *Cross* value −1. The next Spanish word "*recibe*" is the translation of two words to the right of the current ST cursor

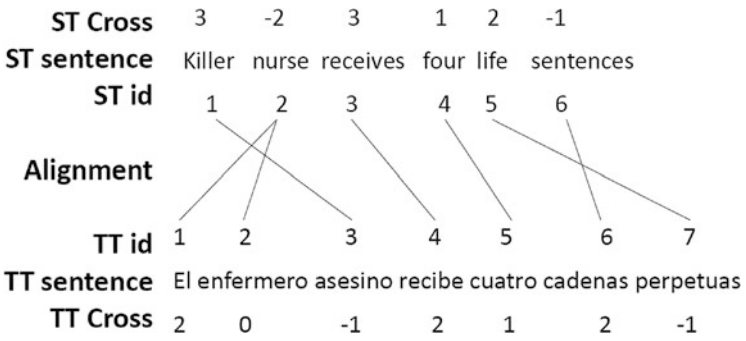


Fig. 2.3 Cross values for ST and TT units

position leading to a *Cross* value of 2, etc. In this way, the *TT Cross* values indicate the relative reordering of ST words to arrive at the TT translation.

A Cross value is also computed for the source text. The *ST Cross* values assume that the ST text is the output text and the TT text is the input. Accordingly *ST Cross* indicates the relative reordering of TT words to arrive at the ST.

Languages with similar word order will have low average Cross values. In a monotonous 1-to-1 translation all Cross values are 1. The more syntactic reordering between source and target text takes place the higher the average Cross value. See also Chap. 8, Sect. 2.3 for extended discussion on the *Cross* feature.

2.4.5 Source and Target Text Tokens

Summary tables of source text tokens (ST) and target text tokens (TT) contain essentially the same information as those in AUs. In particular, general information, such as *Study* and *Session* names are identical. However, instead of the *SAU* and *TAU* attributes, as in Table 2.6, ST and TT tables provide *SToken* and *TToken*, as well as a *Lemma* and *PoS* tag for the ST tokens and TT tokens respectively. Table 2.11 shows three ST token units. The *Prob1* and *Prob2* attributes are log10 probabilities of uni- and bigrams. In the case of English, the BNC[6] was used as a reference corpus: Thus there is a chance of $10^{-3.4339} = 1$ out of 2715 that "*four*" occurs in the BNC, while the occurrence of "*life sentences*" is only $10^{-6.1669} = 1/1{,}468{,}588$.

In addition, there is more detailed gaze information in the ST and TT tables, and ST tables also contain information on word translation entropy.

[6]British National Corpus http://www.natcorp.ox.ac.uk/

Table 2.11 Source text token information (ST)

STid	STseg	Study	Session	...	SToken	Lemma	Prob1	Prob2	PoS	TToken	TTid
5	1	BML12	P01_T1	...	four	four	−3.4339	−50	CD	cuatro	5
6	1	BML12	P01_T1	...	life	life	−3.3508	−50	NN	perpetuas	7
4	1	BML12	P01_T1	...	sentences	sentence	−4.64	−6.1669	NNS	cadenas	6

2.4.6 Gaze Information

Several new reading time measures have been added in the TPR-DB 2.0. We follow suggestions as proposed by Kertz Lab[7] which have been adopted in the following manner:

- *FFTime*: first fixation time is the time offset (in ms) of the first fixation on the token
- *FFDur*: first fixation duration is the duration of the first fixation on the token
- *FPDurS*: first pass source token reading duration is the sum of all fixation durations on the source token from the first fixation until the participants looks at a different token
- *FPDurT:* first pass target token reading duration is the sum of all fixation durations on the target token from the first fixation until the participants looks at a different token
- *RPDur*: regression path duration is the amount of time it took from *FFTime* until the eyes move on to the right in the text. It includes all regressions to the left.
- *Regr*: a boolean value indicating whether a regression followed the first pass reading
- *FixS*: total number of fixations on the source token
- *FixT*: total number of fixations on the target token
- *TrtS*: total reading time on source token is the sum of all fixations on the source token for the duration of the entire session.
- *TrtT*: total reading time on target token is the sum of all fixations on the target token for the duration of the entire session.

Table 2.12 shows examples for these gaze measures. According to the definitions, *FFDur* is always smaller than *RPDur* or *Trt*. Chapter 9 (Sect. 2.1) in this volume provides a more detailed discussion of these features.

2.4.7 Word Translation Entropy and Perplexity

Word translation perplexity indicates how many translation choices a translator has at a given point of the source text, i.e. how many equally likely words can be produced for a source word in a given context. We assume that the choice of translations follows a certain distribution of translation probabilities $\boldsymbol{p}$ and estimate these probabilities from a corpus of aligned translations. The word translation probabilities $\boldsymbol{p(s \rightarrow t_i)}$ of an ST word $\boldsymbol{s}$ and their possible translations $\boldsymbol{t_{i \ldots n}}$ are

[7] We follow suggestions as proposed by Kertz Lab as in https://wiki.brown.edu/confluence/display/kertzlab/Eye-Tracking+While+Reading

Table 2.12 Gaze information in ST and TT units

STid	Study	Session	...	FFTime	FFDur	RPDur	Regr	FixS	FPDurS	TrtS	FixT	FPDurT	TrtT
4	BML12	P01_T1	...	280	50	1317	1	9	50	1567	3	133	183
5	BML12	P01_T1	...	1843	650	650	0	8	650	1600	24	149	1945
6	BML12	P01_T1	...	2515	283	866	1	12	666	1498	0	0	0

computed as the ratio of the number of alignments $s \rightarrow t_i$ counted in TTs over the total number of observed TT tokens, as in Eq. (2.2):

$$p\,(s \rightarrow t_i) = \frac{count\,(s \rightarrow t_i)}{\#translations} \tag{2.2}$$

The information of a distribution with equal probability p is defined as $I(p) = -log_2(p)$. While the probability expresses the expectation for an event, the information indicates the minimum amount of bits with which this expectation can be encoded. The entropy H indicates the expectation of that information, also across unequal probability distributions, as shown in Eq. (2.3):

$$H\,(s) = \sum\nolimits_{i=1}^{n} p\,(s \rightarrow t_i)^* - \log_2\,(p\,(s \rightarrow t_i)) \tag{2.3}$$

Word translation entropy $H(s)$ is the sum over all observed word translation probabilities (i.e. expectations) of a given ST word s into TT words $t_{i\ldots n}$ multiplied with their information content. It represents the average amount of information contained in a translation choice. Thus, if a given source word s has only one possible translation t in a given context, its word translation probability is $p(s \rightarrow t) = 1$, its information $I(p(s \rightarrow t)) = 0bit$ and thus the entropy $H(s) = 0$ is minimal. The more different equally probable translations a source word has, the higher is its word translation entropy $H(s)$. Chapter 10, Sect. 10.2 in this volume gives a more in depth background on word translation entropy.

Perplexity (PP) is related to entropy H, as an exponential function as shown in Eq. (2.4):

$$PP\,(s) = 2^{H(s)} \tag{2.4}$$

The higher the perplexity, the more similarly likely choices exist and hence the more difficult is a decision to make.

The ST tables provide some of this information: *CountT* represents the number of observed $SToken \rightarrow TToken_i$ alignments count($s \rightarrow t_i$), and *AltT* the number of different $TTokens_i$. *ProbT* is the probability of that token and *HTra* is the word translation entropy of *SToken*. For instance, consider $STid_4$ in Table 2.13. The translation "four → cuatro" occurred 25 times in the corpus with a probability of 0.8. With this we can reconstruct the total number of translations in the corpus to be $31 \approx 25/0.8$, and the remaining six translations (31–25) were distributed over three different word forms.

HSeg indicates the entropy of the word alignment segmentation. For instance, an expression like "life sentences" could be aligned as a multi-word unit, or compositional as two different units. The number of source and target language words of the alignment unit (AU) of which "life" is part, is reflected in the *SAUnbr* and *TAUnbr* values respectively. The *HSeg* attribute takes into account this alignment segmentation context, and is calculated in a similar way as *HTra* with the difference that it relies on counting identical *TAUnbr*, instead of *TToken*.

Table 2.13 Gaze information in ST and TT units

STid	Study	Session	...	SToken	TToken	SAUnbr	TAUnbr	AltT	CountT	ProbT	HTra	HSeg
4	BML12	P01_T1	...	four	cuatro	1	1	4	25	0.8065	0.9511	0.7088
5	BML12	P01_T1	...	life	perpetuas	1	1	8	17	0.5484	1.9385	0.6595
6	BML12	P01_T1	...	sentences	cadenas	1	1	8	18	0.5806	1.899	0.4587

The *Literal* feature in Table 2.5 is then simply the average word translation literality, computed as $Literal = \frac{1}{n} * \sum_{j}^{n} \boldsymbol{abs}\left(\boldsymbol{cross} * \boldsymbol{HTra}\right)$, where $\boldsymbol{n}$ is the length of the source text sentence.

2.5 Processing Units

This section starts with describing the basic processing units, single keystrokes (KD) and fixations (FD). Section 2.5.3 introduces production units (PUs) and Sect. 2.5.4 fixation units (FUs). Section 2.5.5 presents a notion of activity units (CU) which exhaustively fragments the translation process into eight types of segments.

2.5.1 Keystroke Data

Within the TPR-DB, each keystroke, as produced by a human translator, is characterized by the following seven criteria:

1. *Time*: the delay in time (ms) after which the keystroke is produced
2. *Type*: whether the keystroke is an insertion or a deletion
3. *Cursor*: at which offset in the target text the keystroke is produced
4. *Char*: which character (UTF8) is produced (inserted or deleted)
5. *TTseg*: the target segment (sentence) that is being produced
6. *STid*: the source text word id of which the produced target word is a translation
7. *TTid*: the id of the target text word that is being produced by the keystroke

The example in Table 2.14 shows the processed keylog data for the production of two Spanish words "El enfer[e]mero ", as a translation of source word $STid_2$. These are the first two words of the first segment in the translation. The table records only text modifying keystrokes, insertions and deletions—navigation information such as mouse clicks etc. are ignored. Insertions and deletions can be produced manually (*Mins* and *Mdel*) or automatically (*Ains* and *Adel*). An example for a manual deletion is in line 9 in Table 2.14.

2.5.2 Fixation Data

During a fixation, the gaze remains on a single location for several milliseconds. Within the TPR-DB, the center of a fixation is mapped onto the closest character on the screen and connected to the following 10 attributes:

1. *Time*: at which the fixation starts
2. *Dur*: duration of the fixation in ms

Table 2.14 Keystroke information as extracted from session P01_T1 of study BML12

KDid	Time	Type	Cursor	Char	TTseg	STid	TTid
0	92,016	Mins	0	"E"	1	2	1
1	92,172	Mins	1	"l"	1	2	1
2	92,313	Mins	2	"_"	1	2	1
3	92,375	Mins	3	"e"	1	2	2
4	92,563	Mins	4	"n"	1	2	2
5	92,828	Mins	5	"f"	1	2	2
6	92,938	Mins	6	"e"	1	2	2
7	93,047	Mins	7	"r"	1	2	2
8	93,266	Mins	8	"e"	1	2	2
9	93,610	Mdel	8	"e"	1	2	2
10	93,797	Mins	8	"m"	1	2	2
11	93,875	Mins	9	"e"	1	2	2
12	93,938	Mins	10	"r"	1	2	2
13	94,078	Mins	11	"o"	1	2	2
14	94,203	Mins	12	"_"	1	2	2

3. *Win*: source window (1) or target window (2) in which the fixation is observed
4. *Cursor*: mapping of the fixation center on the closest character in the window
5. *STid*: id of the source text token that is being looked at
6. *TTid*: id of the target text word that is being looked at
7. *Seg*: segment id of the source text word (*STid*) that is being looked at
8. *ParalK*: amount of concurrent keyboard activity, i.e. production unit (PU, see Sect. 2.5.3)
9. *Edit*: character(s) that have been typed during fixation
10. *EDid*: the target segment id that is being produced by the typed characters

Table 2.15 shows a sequence of 13 fixations, *FDid* 507–519, which are part of the P01_T1 session, introduced above. All fixations take place in window 1, on the first segment and *STid* tokens 4, 6, 3 and 5, which are translated into *TTids* 5, 6 4 and 7, respectively. Some of the fixations show concurrent typing activity: as the amount of parallel keyboard activity (*ParalK)* equals the fixation duration time (*Dur*), the first seven fixations (*FDid* 507–513) overlap to 100 % with text production. No keyboard activity took place during fixations *FDid* 515–517, and a partial overlap of 16 % (124 ms/750 ms) typing activity is recorded for fixation *FDid* 518. During fixations 507–510, for instance, was typed (*Edit*) the sequence "eno", which is part of the production of "as**e**si**no**". The column *EDid* indicates the *STid* of the produced translation, i.e. "as**e**si**no**" is a translation of STid 3. In Sect. 2.5.3, we show that the keyboard sequence is part of one production unit PU_0 while the fixations are part of FU_{14}. Section 2.6 visualizes the data in a larger context.

Table 2.15 Fixation information (.fd file)

FDid	Time	Dur	Win	Cursor	Seg	STid	TTid	ParalK	Edit	EDid
507	94,530	150	1	25	150	4	5	150	e	3+
508	94,749	67	1	24	1	4	5	67	–	–
509	95,077	67	1	25	1	4	5	67	n	3+
510	95,218	67	1	26	1	4	5	67	o	3+
511	98,952	50	1	36	1	6	6	50	i	4+
512	99,015	167	1	37	1	6	6	167	b	4+
513	99,202	50	1	36	1	6	6	50	–	–
514	99,265	83	1	25	1	4	5	1	e	4+
515	99,499	100	1	16	1	3	4	0	–	–
516	99,624	83	1	16	1	3	4	0	–	–
517	99,718	50	1	17	1	3	4	0	–	–
518	99,780	750	1	24	1	4	5	124	_	4+
519	100,546	250	1	30	1	5	7	250	–	–

Table 2.16 Three production units from session P01_T1 of study BML12

PUid	Study	Session	Time	Dur	Pause	Ins	Del	Edit
0	BML12	P01_T1	92,016	7250	92,016	34	7	El_enfere[e]mero_ asesiono_re[er_ono]no_ recibe
1	BML12	P01_T1	100,406	1313	1140	8	0	_cuatro_
2	BML12	P01_T1	103,594	4187	1875	23	3	sentencias_de_vida. __[__.]__

2.5.3 Production Units

Production units (PUs) are sequences of coherent typing activity. According to a definition in (cf. Carl and Kay 2011), a PU boundary is defined as a pause of 1000 ms or more without keyboard activity. Beyond this pause duration, it is assumed that coherent typing is interrupted, with a likely shift of attention towards a different text segment. As a coherent temporal/textual segment PUs have a temporal beginning (*Time*) and a duration (*Dur*), and they may cover one or more insertion or deletion keystrokes (*Edit*) that contribute to building up one or more target text tokens (*TTid*). The edit sequence of PU_0 in Table 2.16 is shown in example (2.5):

$$\text{El\_enfere [e] mero\_asesiono\_re [er\_ono] no\_recibe} \tag{2.5}$$

started at time 92,016 and was typed within 7250 ms, with no inter-key delay of more than 1000 ms. It was preceded by a pause of 92,016 ms. The next PU_1 starts with a *Pause* of 1140 ms. Follows this pause the typing sequence starts at Time 100,406 ms and lasts for 1313 s. Table 2.16 indicates the number of insertions and deletions of the PUs. PU_0 contains 34 insertions (*Ins*) and 7 deletions (*Del*). The

latter are within square brackets in the *Edit* column and must be read in the reverse direction. Thus, the substring "[er_ono]" reflects actually the deletion "ono_re", as shown in example (2.6):

$$\text{asesiono\_re} \rightarrow \text{asesino\_recibe} \tag{2.6}$$

Table 2.17—which is a continuation of Table 2.16—contains additional process and product information for the three PUs. *STseg* and *TTseg* indicate that the three PUs are part of the first segment translation. *STid* and *TTid* show the source and target words covered by the translation. Note that *TTid* refers to the word numberings in the final translation. Thus, the word numeration in an intermediate version of the text may not coincide with that in the final text if words are inserted or deleted. As can be seen in the succession of *STid*, the translation evolves successively in the order of the source text words. PU_1 "_cuatro_" accounts for two source and two target words ($STid_{3+4}$ and $TTid_{4+5}$), as the blank—represented by an underscore "_"—already counts as part of the next word. PU_2 also covers two words $TTid_{5+7}$, even though the PU consists of three words "sentencias de vida". This compound noun was later re-written into "cadenas perpetuas" which make up $TTid_6$ and $TTid_7$. Note that this discontinuity is also the reason for the *Scatter* value to be one: there is one sequence of two successive keystrokes in this PU that produces translations more than one word apart.

FixS and *FixT* represent the number of fixations counted on the source and target side of the PUs respectively. Note that, due to poor eye-tracking quality, no fixations were recorded on the target strings.

The feature *ParalS* and *ParalT* give the amount of time the translator was looking at the source and the target window respectively while producing the translation. That is, during the 7250 ms that it took to produce PU_0, the translator looked almost 1 s (900 ms) at the source text window.

CrossS and *CrossT* give the average local distortion between the source side and the target side of the PUs. The calculation of the Cross features is discussed in detail in Sect. 2.4.4. *PosS* and *PosT* indicate the part-of-speech tags for the source and the target words involved in the PUs.

2.5.4 Fixation Units

Fixation Units (FUs) describe sequences of coherent reading behavior. Based on experimental evidence (Carl and Kay 2011), we define a boundary between two successive FUs as a gazing pause longer than 400 ms. For instance, as the gaze directs away from the screen for more than 400 ms, thus interrupting coherent reading activity, we assume a boundary of a fixation unit. An FU has a start *Time*, a duration, followed by a pause (of more than 400 ms), before the next FU starts.

Path describes the sequence of words looked at, in the source window (1) or the target window (2). A gaze path consists of one or more fixations indicated by

Table 2.17 Three production units from session P01_T1 of study BML12

STseg	TTseg	STid	TTid	FixS	ParalS	FixT	ParalT	Scatter	CrossS	CrossT	PosS	PosT
1	1	1 + 2 + 3	1 + 2 + 3 + 4	10	735	0	0	0	2.67	1.25	NNP + VBP + NNS	ART + NC + NC + VLfin
1	1	3 + 4	4 + 5	4	504	0	0	0	2	1.5	NNS + CD	VLfin + CARD
1	1	4 + 5	5 + 7	4	216	0	0	1	1.5	1	CD + NN	CARD + ADJ

Table 2.18 Four fixation units

FUid	Time	Dur	Pause	ParalK	Path
14	94,530	755	5293	755	1:4 + 1:4 + 1:4 + 1:4+
15	98,952	1844	3667	704	1:6 + 1:6 + 1:6 + 1:4 + 1:3 + 1:3 + 1:3 + 1:4 + 1:5+
16	101,577	1272	781	142	1:5 + 1:5 + 1:6 + 1:6 + 1:6 + 1:5 + 1:5+

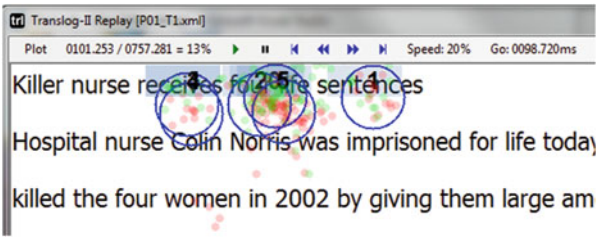


Fig. 2.4 Screen shot of replay situation of FU_{13}

a tuple"*Win*:*WordID*" where successive fixations are separated by a "+". FU_{14} in Table 2.18 has a path of four fixations (1:4+ 1:4+ 1:4+ 1:4+), on source word "four" (1:4). FU_{15} is plotted in Fig. 2.4 and represents a reading pattern of the words in bold in the title "Killer nurse **receives four life sentences"**. It shows how the gaze goes back and forth between the four words which took 1844 ms.

ParalK in an FU table indicates the amount of parallel gaze and keyboard activity. During FU_{14} the translator is writing at the same time as reading, while there is an overlap of 11 % keyboard activity during FU_{16}.

Note that the sum of all FU durations may be longer than the sum of all fixation durations, since FUs include inter fixation delays shorter than 400 ms that may not be part of any fixation.

2.5.5 Activity Units

Activity units (CUs) segment the recorded session exhaustively into sequences of activities that are slightly different from PUs and FUs. In contrast to the latter ones, CUs segment a session exhaustively into typed segments: For translation tasks we distinguish between the following three basic types of translator activities:

- **Type 1**: source text reading
- **Type 2:** target text reading
- **Type 4**: translation typing

Since source or target text reading and typing can occur in parallel (see Sect. 2.4.3), we also have the following additional concurrent activities:

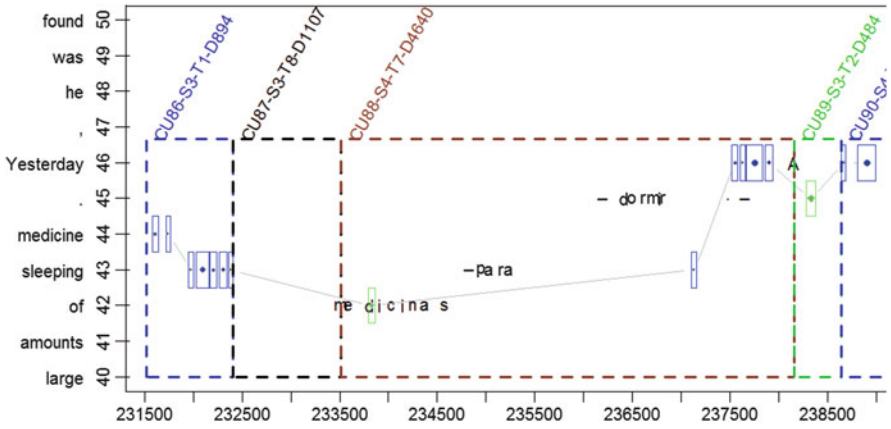


Fig. 2.5 Segmentation into successive activity units

- **Type 5**: translation typing while reading the source text
- **Type 6**: translation typing while reading the target text
- **Type 7**: translation typing while reading the source and the target text

A coherent typing activity is defined as coherent keyboard activities (similar to PUs) with no more than 1 s pause between two successive keystrokes. If neither gaze nor keyboard activity is recorded for more than 1 s, an idle segment is assigned:

- **Type 8**: no activity was recorded

A CU is described by its start time (*Time*), duration (*Dur*), and the segment (*Seg*) in which it takes place. Figure 2.5 shows a sequence of four activity units involved in the translation shown in (7):

$$sleeping\ \ medicine \rightarrow medicinas\ \ para\ \ dormir \tag{2.7}$$

Figure 2.5 shows boundaries of successive CUs and their labels: the first CU with time stamp 231,500–232,500 ms is a source text reading activity of 894 ms, followed by an "idle" unit (Type 8) of 1107 ms in which no activities were recorded. Then follows a typing CU (Type 7) at time stamp 233,500 ms of 4640 ms in which concurrent ST reading and TT reading can be observed. During this time span "*medicinas para dormir A"* is produced. This is followed by a target text reading activity (Type 2, Duration 484 ms) in which the just typed word (*dormir*) is monitored. The figure represents a translation progression graph (TPG) which will be discussed in Sect. 2.6.

2.6 Visualizing Product and Process Data in a Translation Progression Graph

Information from various tables can be analyzed, evaluated and visualized in many different ways. One method of visualization that is part of the TPR-DB, is the Translation Progression Graph (TPG). TPGs visualize how translations emerge in time, plotting partial information of several unit tables at one time. Figure 2.6 shows a TPG visualizing a post-editing session of a CASMACAT post-editing session.

The graph traces post-editing activities of six consecutive segments. The vertical axis enumerates the source text words (0.. 140) with horizontal dotted lines separating the segments, whereas the horizontal axis shows the time at which the translations of the source text were produced. The various symbols in the graph represent:

- Blue diamonds represent fixations on the source text
- Green diamonds represent fixations on the target text
- Black characters represent insertions
- Grey characters represent automatic insertions
- Red characters represent deletions

The graph shows when segments loaded into the target buffer, when and where translators read the source and the target segments, and when the text was modified. TPGs are thus a useful means to assess the TPR-DB data qualitatively.

Another TPG is shown in Fig. 2.7. This graph puts into relation the translation product on the source text (vertical axis, left) and the target text (vertical axis,

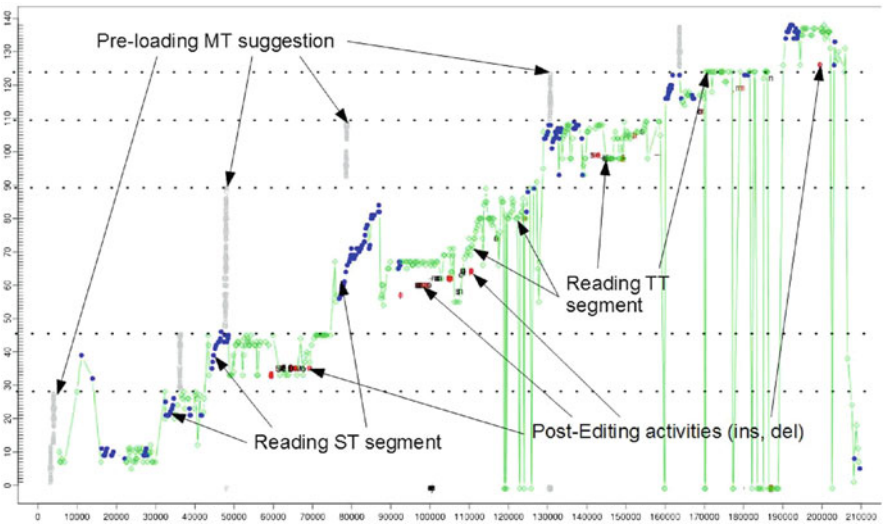


Fig. 2.6 A translation progression graph plotting keystroke and gazing information

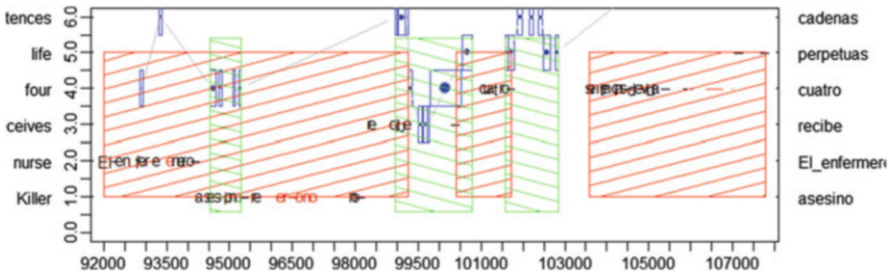


Fig. 2.7 The progression graph shows product and process information from Tables 2.14, 2.15, 2.16, 2.17, and 2.18

right) and the translation process data on a time line on the horizontal axis. It visualizes how the translation emerges in time. Insertions are represented in black letters, deletions are red, and fixations are blue dots in rectangular boxes that stretch their duration in time. The TPG in Fig. 2.7 plots the keystroke data of Table 2.14, the fixation data from Table 2.15, as well as the three FUs from Table 2.18 and three PUs of Table 2.16 and 2.17. The red horizontally striped boxes indicate PUs while the green boxes represent FUs. The first part (approx. Time 92,000 ms to 94,000 ms) reproduces the production of words 1 and 2 ("El enfermero") as plotted in Table 2.14.

As discussed in Sects. 2.5.3 and 2.5.4, reading and writing activity can occur in parallel. For instance, FU_{14} around time stamp 95,000 takes place while the translator produces "asesino", the translation of "Killer", while FU_{15} and FU_{16} at time stamps 99,000 and 101,500 respectively only partially overlaps with two adjacent PU_0 and PU_1. Progression graphs, illustrate in a graphical manner the relation between reading and writing activities.

2.7 External Resources

2.7.1 Infusion of External Inputlog Data

Translog-II and CASMACAT only log the Keystroke data that is produced within the GUI. However, in many cases, translators use external resources, such as e-dictionaries, collocation tools, they search for expressions on the web among others. These activities are not recorded in Translog-II or CASMACAT, but external search behaviour may be interesting to investigate and correlate with Translog-II UAD.

Inputlog (Leijten, and Van Waes 2013) is a windows-based logging tool that logs all types of input modes: keyboard, mouse and speech recognition. In contrast to Translog-II and CASMACAT, Inputlog is not application dependent. That is, it can log keyboard activities independent of the (windows-based) application that receives the input. Inputlog knows which application is on focus, and stores this information

together with the actual key pressed and the time stamp of a keystroke (or mouse movement) in its IDFX log file.

A script,[8] InfuseIDFX.pl, can be used to integrate Inputlog IDFX logging data into Translog-II files. The 'InfuseIDFX.pl' script first synchronizes the Inputlog and the Translog-II logging data based on common keystrokes and then inserts the data that was collected outside the Tranlog-II (or CASMACAT) GUIs into the Translog-II log file. The TPR-DB compilation process subsequently generates an EX table indicating usage of the external resources.

For instance in a browser-based application, Inputlog knows which window is on focus. Successive keystrokes can accordingly be associated with the web page in focus. In this way web searches can be tracked and reconstructed. On the one hand, Inputlog is universally deployable in different windows-based applications. On the other hand, Inputlog has no possibility to know where the typed characters occur in a text. From Inputlog we know which keystrokes were pressed, but not necessarily which characters are produced or which characters are deleted and we also do not know where in a text these operations would take place—unless produced within MS word.

As an example, Table 2.19 plots an excerpt from a converted Inputlog table showing that Google Chrome was used as a main external resource in a Translog-II session. At the time instant 33,453 ms, an application with the name "TASKBAR" was activated for approximately ½ s, followed by a search query in "Google Chrome", which lasted slightly more than 32 s. The user then went via an application "Menu Iniciar" back to the "Translog-II User" program, which he left again after 14,297 s. The *Edit* column contains the concatenation of the typed keystrokes that occur during the time in focus. It is empty if no keystroke was produced. Examples of the **EDIT** string in the 32 s between time stamp 34,000 and 66,818, when "Google Chrome" was in focus may consist of:

1. bring

2. emit[..]otional trra[..]adução

3. in [.] the arr[.]ticle
 tradução
presented

A query is usually terminated by a return, which is here encoded as "
", and deletions are in square brackets [..]. Thus in (1) the translator typed *bring* and then pressed the return key. In (2) the translator deleted twice two characters in the input string. From the Inputlog IDFX file we do not know which characters were deleted, but it is most likely that first "it" and then "ra" was deleted using backspace so as to produce the search string "emotional tradução". In example (3) three search strings were produced "in the article", "tradução" and "presented". While we can re-construct the query which the translator produced in the external resource, we do not know what the results of these queries were. However, we can trace the translators reaction from within Translog-II. The attribute

[8] InfuseIDFX.pl script is part of the TPR-DB and can be downloaded from the TPR-DB website, https://sites.google.com/site/centretranslationinnovation/tpr-db

Table 2.19 Usage of external resources in the TPR-DB

EXid	Study	Session	Focus	Time	Dur	STsegN	STsegL	STidN	STidL	KDidN	KDidL	Edit
0	N1	P02_P1	TASKBAR	33,453	547	2	1	18	8	27	26	
1	N1	P02_P1	Google Chrome	34,000	32,813	2	1	18	8	27	26	**EDIT**
2	N1	P02_P1	Menu Iniciar	66,813	812	2	1	18	8	27	26	
3	N1	P02_P1	Translog-II User	67,625	14,297	2	1	18	8	27	26	

KDidL indicates the last keystroke (*KDid)* before Translog-II was left, *KDidN* gives the next keystroke after s/he came back. Similarly, *STidL* and *STidN* indicate the source word id of the translation of those keystrokes and *STsegL* and *STsegN* the source segments. Thus, the last keystroke before the translator left Translog at time 33,453 was *KDidL* = 26 and the first keystroke after coming back into Translog-II User was *KDidN* = 27. These two keystrokes are part of the production for the translation of *STidL* = 8 and *STidN* = 18 which belong to two successive segments 1 and 2. While we thus do not know what exactly a translator may have learned from visiting the external resource, we have a means of re-constructing the effect by investigating the behavior that precedes and follows its consultation. A usage of this tool is described in Chap. 6 in this volume.

2.7.2 Adding Columns to TPR-DB Summary Tables

In some cases, users would like to further add columns to some of the TPR-DB tables with their own annotations. For instance, in an experiment on syntactic entropy (see Chap. 10) each segment was manually annotated with a set of Triplets, describing the syntactic structure of the sentence. Such annotations can be automatically added to the appropriate summary table, by means of a script that is part of the TPR-DB.[9] A file with the extension of the unit table specifies the *Study*, *Session* and unit ID in addition to the columns to be added to the table, as shown in Table 2.20.

Table 2.20 Annotations extending segment information with 5 additional columns

Study	Session	STseg	SynH	STriplet	TTriplet	PrimeDiff	Prime Prob
default	default	default	0	–	–	DIFF	0
BML12	P03_P1.sg	2	0	TPI	TPI	PRIME	1
BML12	P06_P1.sg	2	0	TPI	TPI	PRIME	1
BML12	P03_P1.sg	3	0.721	TAI_DAD	TAI_DAD_IAD	DIFF	0.2
BML12	P06_P1.sg	3	0.721	TAI_DAD	TAI_DAD_IAD	DIFF	0.2
ML12	P28_P1.sg	3	0.721	TAI_DAD	TAI_DAD	PRIME	0.2
BML12	P32_P1.sg	3	0.721	TAI_DAD	TAI_DAD	PRIME	0.2
BML12	P03_P1.sg	4	0	TPI_TAD	MPI	DIFF	0

[9]The script *AddExtColumns.pl* can be downloaded from https://svn.code.sf.net/p/tprdb/svn/ and called with the parameters AddExtColumns.pl -C ExtraColumnsFile –S Study_name

Acknowledgement This work was supported by the CASMACAT project funded by the European Commission (7th Framework Programme). We are grateful to all contributors to the database and for allowing us to use their data.

Appendix 1

Overall the TPR-DB contains more than 580 h of text production time in terms of *Fdur* duration. In the 1689 sessions were involved 132 different translators producing all together more than 660,000 words in 9 different languages.

The language pair en → es is the by far the largest language represented in TPR-DB, with 660 sessions, 500,000 target words and more than 320 h of *Fdur* production time. The second most represented language pair is en → hi with 161 sessions, more than 20,000 tokens in the Hindi translations and more than 46 h of *Fdur* production time. The third language pair is en → de with 146 sessions, more than 24,000 tokens in the German translations and more than 24 h of *Fdur* production time production time, followed by en → da with 127 sessions, more than 18,000 tokens in the Danish translations and 12 h of *Fdur* production time. The rest of the language pairs in the TPR-DB involve more than 20 translation directions in 7 different source and 16 target languages (This includes language directions not shown in Table 2.21). Please consult the TPR-DB website for an updated version of the database contents.

Each study in the TPR-DB was conducted with a (set of) research question(s) in mind, which can be roughly summarized as follows:

(A) The TPR-DB contains ten studies conducted with the three different CASMACAT workbenches as follows:

1. ALG14: This study compares professional translator and bilinguals while post-editing with the third prototype of the CASMACAT workbench featuring visualization of word alignments.
2. CEMPT13: This study contains post-editing recordings with the second prototype of the CASMACAT workbench, featuring interactive machine translation.
3. CFT12: This study contains data of the first CASMACAT field trial from June 2012, comparing post-editing with from-scratch translation.
4. CFT13: This study contains data of the second CASMACAT field trial from June 2013, comparing post-editing and interactive machine translation.
5. CFT14: This study contains data of the second CASMACAT field trial from June 2014, comparing interactive machine translation and online learning.
6. EFT14: The study compares active and online learning during interactive translation prediction

Table 2.21 Summary table for TPR-DB studies: continuation below

Study	Sess	SL	TL	Task	Texts	Part	Fdur	Kdur	Pdur	Stok	Ttok
ACS08	30	en	da	T	4	17	4.6776	2.9704	1.9332	5085	5075
ACS08	30	en	en	C	4	17	2.0436	1.8316	1.6013	5099	5109
ALG14	8	en	es	P	2	8	2.6018	0.4854	0.1747	4460	4807
ALG14	8	en	es	PA	2	8	2.7954	0.4437	0.1692	4460	4801
BD08	10	en	da	T	1	10	1.4575	0.7493	0.448	1100	1056
BD13	8	en	da	T	2	8	0.8079	0.5368	0.3213	786	751
BD13	10	en	da	P	2	10	0.4412	0.1074	0.0569	970	1014
BML12	64	en	es	P	6	32	4.6394	0.9079	0.4418	9012	10,216
BML12	63	en	es	T	6	32	9.8032	5.9308	3.8062	8936	10,102
BML12	60	en	es	E	6	30	3.7009	0.9657	0.4729	8468	9594
CEMPT13	20	en	pt	PIA	2	20	6.634	1.823	0.5387	6706	6840
CEMPT13	20	en	pt	P	2	20	5.5943	1.2678	0.5732	6494	6585
CFT13	27	en	es	R	26	4	8.3388	0.9733	0.4413	26,919	28,738
CFT13	27	en	es	PI	9	9	30.0923	10.2351	3.3044	31,752	33,871
CFT13	27	en	es	P	9	9	28.167	8.0677	3.51	31,294	33,770
CFT13	27	en	es	PIA	9	9	35.5658	11.2626	3.9125	31,838	34,047
CFT14	7	en	es	RE	7	3	3.8435	0.2465	0.0586	20,341	22,015
CFT14	7	en	es	R	7	4	3.2497	0.3687	0.1485	20,273	22,251
CFT14	7	en	es	P	2	7	16.8321	7.9316	3.418	20,273	22,067
CFT14	7	en	es	PIO	2	7	15.8297	8.1574	3.4917	20,341	22,284
DG01	60	fr	pl	T	2	60	33.8564	17.5784	11.2075	25,380	20,329
EFT14	11	en	es	PIVO	3	11	10.221	5.2041	2.2521	12,437	13,549
EFT14	11	en	es	PI	3	11	11.9495	6.8647	3.2755	12,437	13,696
EFT14	10	en	es	PIVA	3	10	10.7885	5.1993	2.3594	11,327	12,472
GS12	8	es	en	P	4	4	4	4	2.1901	0.3586	0.1909
HLR13	15	en	et	T	3	5	2.5457	1.1214	0.673	1535	1186
JIN15	18	en	zh	S	1	18	2.0227	0.2641	0.0455	1947	1728
JIN15	18	en	zh	P	1	18	4.5318	0.8192	0.1442	1998	1845
JIN15	17	en	zh	R	1	17	2.594	0.3451	0.0567	1946	1833
JLG10	10	en	pt	T	3	5	5.6048	2.1218	1.2302	2577	2781
JLG10	10	pt	en	T	3	5	5.6391	2.0787	1.1718	2611	2621
JN13	4	en	de	PIA	2	4	2.7428	0.7284	0.2735	2590	2668
JN13	4	en	de	P	2	4	2.3311	0.6374	0.2189	2590	2571
KTHJ08	69	en	da	T	3	24	7.4469	5.6824	3.8183	10,571	10,667
LS14	60	en	es	PI	24	5	53.3764	22.0971	9.5166	72,109	80,278
LS14	60	en	es	P	24	5	51.7256	17.3211	7.4178	72,126	80,454
LWB09	40	da	en	T	3	18	3.7061	2.8926	2.0511	5652	6206
MS12	19	en	zh	P	6	11	2.6953	0.4817	0.0497	2708	2562
MS12	15	en	zh	T	5	10	3.7369	1.0512	0.1088	2061	1916
MS12	10	en	zh	E	5	8	0.7714	0.1564	0.0183	1295	1203

(continued)

Table 2.21 (continued)

Study	Sess	SL	TL	Task	Texts	Part	Fdur	Kdur	Pdur	Stok	Ttok
MS13	16	zh	pt	P	2	16	2.7139	0.9211	0.4443	1410	1648
MS13	16	pt	zh	T	2	16	2.3327	0.7687	0.1161	1386	1378
MS13	22	zh	pt	T	2	22	4.1631	2.1803	1.2265	1938	2216
MS13	18	pt	zh	P	2	18	2.555	0.6698	0.0934	1555	1507
NJ12	39	en	hi	T	6	20	14.4697	7.5368	3.3156	5505	5784
NJ12	61	en	hi	P	6	20	17.4402	6.8654	3.0615	8581	9365
PFT13	9	en	es	P	1	9	2.0861	0.3154	0.1406	3035	3144
PFT13	19	en	es	PI	1	19	5.2058	1.5351	0.4267	6689	7437
PFT13	16	en	es	PIC	3	16	2.7853	0.744	0.1518	5396	5147
PFT13	15	en	es	PIO	3	15	2.4784	0.4741	0.0669	4611	4666
PFT13	16	en	es	PIL	3	16	2.7226	0.6761	0.1511	5572	5344
PFT14	3	en	es	PIVO	2	3	2.1558	0.6775	0.1622	3245	3150
PFT14	2	en	es	PIVA	1	2	2.0228	0.7255	0.1843	2286	2184
PFT14	2	en	es	PIV	2	2	1.987	0.7667	0.1905	2161	2077
RH12	2	es	es	A	2	2	2.9849	0.9786	0.6398	1207	1207
ROBOT14	40	en	nl	P	8	10	10.8706	3.2467	1.5417	7375	7527
ROBOT14	40	en	nl	T	8	10	12.2457	5.1006	3.1753	7375	7329
SG12	46	en	de	E	6	23	7.0716	1.8571	0.9342	6522	6741
SG12	45	en	de	P	6	23	8.027	1.9976	1.055	6352	6470
SG12	47	en	de	T	6	24	11.7259	4.7344	2.9421	6632	6777
TDA14	48	en	en	C	6	8	3.8335	3.5653	2.6617	6792	6779
WARDHA13	34	en	hi	T	6	18	15.2298	3.6917	0.5553	4832	4790
WARDHA13	31	hi	hi	C	6	18	11.49	5.3097	0.7569	4365	4104
WARDHA13	27	en	hi	P	6	15	8.0582	1.9611	0.4418	3780	4016
ZHPT12	12	zh	pt	T	1	12	3.5244	1.4856	0.851	1104	1603
Total	**1689**	**7**	**9**	**15**	**132**	**418**	**586.769**	**217.2386**	**100.2227**	**702,701**	**660,595**

The table shows summary information of the TPR-DB for each session: task, language direction, number of different texts, number of different participants, production duration (*Fdur*, *Kdur*, *Pdur*) as well as total source text length (*STok*) and total produced target language (*TTok*) in words (tokens)

7. JN13: This study is recorded with the second prototype of the CASMACAT workbench featuring interactive machine translation and word alignments.
8. LS14: This study investigates learning effects with interactive post-editing over a period of 6 weeks (longitudinal study) with the third prototype of the CASMACAT workbench.
9. PFT13: This study is a pre-field trial test prior to the second CASMACAT field trial.
10. PFT14: This study is a pre-field trial test prior to the third CASMACAT field trial.

(B) The aim of the MultiLingual experiment is to compare from-scratch translation (T), post-editing (P) and monolingual post-editing (E), for different translators and for different languages. The six English source texts are translated by

student and experienced translators; three texts (1–3) are news, three texts (4–5) sociological texts from an encyclopedia. Texts were permuted in a systematic manner so as to make sure that each text was translated by every translator and every translator translated two different texts in each translation mode.

11. BML12: This study contains translating, post-editing and editing data of six texts from English into Spanish.
12. KTHJ08: This study contains only translation data for the news text 1–3.
13. MS12: This study contains translating, post-editing and editing of the six texts English into Chinese.
14. NJ12: This study contains translating, post-editing and editing of the six texts English into Hindi by professional translators.
15. SG12: This study contains translating, post-editing and editing of the six texts English into German.
16. TDA14: In this study participants were asked to copying the six English texts.
17. WARDHA13: This study contains translating, post-editing and editing of the six texts English into Hindi by students.

(C) In addition, the TPR-DB contains a few individual experiments that were conducted with Translog-II:

18. ACS08: This study explores the way in which translators process the meaning of non-literal expressions by investigating the gaze times associated with these expressions.
19. BD08: This study involves Danish professional translators working from English into Danish.
20. BD13: This study involves secondary school students translating and post-editing from English into Danish.
21. DG01: The study compares students, professional and non-professional translators with and without a representation of the text.
22. GS12: This study contains post-editing data of four pieces of news from Spanish into English.
23. HLR13: This is a translation study from English into Estonian (5 participants translating 3 different texts).
24. JLG10: This study investigates L1 and L2 translations from/to English and Brazilian Portuguese.
25. LWB09: This study reports on an eye tracking experiment in which professional translators were asked to translate two texts from L1 Danish into L2 English.
26. MS13: This study is an investigation of translator's behaviour when translating and post-editing Portuguese and Chinese in both language directions.
27. RH12: This is an authoring study for the production of news by two Spanish journalists.
28. ROBOT14: This study investigates usage of external resources during translation and post-editing.

29. ZHPT12: This study investigates translator's behaviour when translating journalistic texts. The specific aim is to explore translation process research while processing non-literal (metaphoric) expressions.

Appendix 2

During each session a particular Task is conducted, as follows:

- **A**: Authoring of a journalistic text. Source and target languages are identical.
- **C**: Copying a text (manually) from the source window into the target window. Source and target languages are identical.
- **E**: Editing of post-editing of MT output without access to the source text (monolingual post-editing).
- **P**: Traditional post-editing of MT output (no additional help is provided during the process).
- **R**: Review of post-edited text.
- **T**: Translation 'from-scratch'.

Within the CASMACAT context, a large number of different post-editing settings were investigated:

- **PA**: Traditional post-editing visualizing source (ST) and target (TT) alignment links (triggered by mouse or cursor).
- **PI**: Advanced post-editing through interactive translation prediction (ITP) / interactive machine translation.
- **PIA**: Advanced post-editing through ITP showing ST-TT alignments (visualization option).
- **PIC**: Advanced post-editing through ITP showing ST-TT alignments (visualization option).
- **PIO**: Advanced post-editing through ITP and online learning techniques.
- **PIL**: Advanced post-editing through ITP showing the post-edited text (suffix) in grey (visualization option).
- **PIV**: Advanced post-editing through ITP showing Search&Replace bar, alignments and mouse-triggered alternative ITP options.
- **PIVA**: Advanced post-editing through ITP and active learning techniques.
- **PIVO**: Advanced post-editing through ITP and online learning techniques.

Appendix 3

This appendix lists all features that are used in the TPR-DB v2 to describe the unit tables. There are in total 275 features and 111 different features describing 11 different unit tables discussed in this chapter. These features are clustered here into

12 types, according to whether they describe a session, segment, token, keyboard or gaze behaviour, etc. In parenthesis are indicated the unit tables in which the features appear.

1. Session data: these features describe the sessions of a study:
   - **Study**: Study name as in the TPR-DB (AU, EX, PU, SG, SS, ST, TT)
   - **Session**: Session name, a composite of Participant, Text and Task (AU, CU, EX, PU, SG, SS, ST, TT)
   - **Text**: Text identifier in the study (AU, SS, ST, TT)
   - **Task**: Type of task, see Appendix 2 (AU, SS, ST, TT)
   - **Part:** Participant ID of study (AU, ST, TT, SS)
   - **SL**: Source text language (AU, SS, ST, TT)
   - **TL**: Target text language (AU, SS, ST, TT)
   - **Break:** Duration of session break (SS)
   - **TimeR**: Starting time of revision phase (SS)
   - **TimeD**: Starting time of drafting phase (SS)
2. Segment: information related to segments:
   - **Seg**: Source or target segment identifier, depending on **Win** feature (FD)
   - **STseg**: Source segment identifier (AU, PU, SG, SS, ST)
   - **Nedit:** Number of times the segment was edited (SG)
   - **TTseg**: Target segment identifier (AU, CU, KD, PU, TT, SG, SS)
   - **LenS**: Length in characters of the source segment (SG, SS)
   - **LenT**: Length in characters of the target segment (SG, SS)
   - **LenMT**: Length in characters of the pre-filled MT segment (SG)
   - **TokS**: Number of source tokens in segment (SG, SS)
   - **TokT**: Number of target tokens in segment (SG, SS)
   - **Literal:** Degree of segment literality (SG)
   - **Nedit**: Number of times the segment has been edited (SG)
3. Tokens: information concerning source and target text tokens in the translation product
   - **STId:** unique identifier of source text token (FD, KD, PU, ST, TT)
   - **TTId:** unique identifier of target text token (FD, KD, PU, ST, TT)
   - **SAU**: Source text segment string (AU)
   - **TAU**: Target text segment string (AU)
   - **SAUnbr:** Number of tokens in source side of alignment unit (AU, ST, TT)
   - **TAUnbr:** Number of tokens in target side of alignment unit (AU, ST, TT)
   - **SToken**: Source text token (ST, TT)
   - **TToken**: Target text token (ST, TT)
   - **Lemma**: Lemma of token (ST, TT)
   - **PoS**: Part-of-Speech of token (ST, TT)
   - **PosS:** Part-of-Speech of source token sequence (PU)
   - **PosT:** Part-of-Speech of target token sequence (PU)

- **Prob1:** Probability of uni-gram occurrence (ST, TT)
- **Prob2:** Probability of bi-gram occurrence (ST, TT)

4. Translation literality metric

- **AltT:** number of different translation alternatives (ST)
- **CountT:** number of observed current translation choice (ST)
- **ProbT:** Probability of current translation choice (ST)
- **HTra:** Word translation entropy (SG, ST)
- **HSeg:** Translation segmentation entropy (SG, ST)
- **Cross:** Cross value of token (AU, ST, TT)
- **CrossS:** Cross value for source tokens (PU, SG)
- **CrossT:** Cross value for target tokens (PU, SG)
- **Literal:** Degree of segment literality (SG)

5. Keystrokes: information concerning keystroke activities

- **KDid:** keystroke ID (KD)
- **Del**: Number of manual and automatic deletions (AU,PU,ST, TT)
- **Ins**: Number of manual and automatic insertions (AU,PU,ST, TT)
- **Adel**: Number of automatically generated deletions (SG, SS)
- **Ains**: Number of automatically generated insertions, (SG, SS)
- **Mdel**: Number of manually generated deletions (SG, SS)
- **Mins**: Number of manually generated insertions (SG, SS)
- **Char**: UTF8 character typed or deleted (KD)
- **Munit**: Number of micro units (AU, ST, TT)
- **Edit**: Sequence of keystrokes producing TT string (AU, EX,FD,PU,ST, TT)
- **Edit1**: Sequence of keystrokes of the first micro unit (AU, ST, TT)
- **Edit2**: Sequence of keystrokes of the second micro unit (AU, ST, TT)
- **InEff**: Inefficiency measure for segment generation (AU, ST, TT)
- **Scatter**: Amount of non-linear text production (PU, SG, SS)

6. Gaze on source and target window

- **Path**: Sequence of fixations on source or target window (FU)
- **FFTime**: Starting time of first fixation (ST, TT)
- **FFDur**: Duration of first fixation (ST, TT)
- **FPDurS**: First pass duration on source text unit (AU, ST, TT)
- **FPDurT**: First pass duration on target text unit (AU, ST, TT)
- **FixS**: Number of fixations on source text unit (AU, PU, SG, SS, ST, TT)
- **FixT**: Number of fixations on target text unit (AU, PU, SG, SS, ST, TT)
- **TrtS**: Total gaze time on source text unit (AU,SG, SS, ST, TT)
- **TrtT**: Total gaze time on target text unit (AU, SG, SS, ST, TT)
- **FixS1**: Number of fixations on source text unit during production of first micro unit (AU, ST, TT)
- **FixS2**: Number of fixations on source text unit during production of second micro unit (AU, ST, TT)

- **FixT1**: Number of fixations on target text unit during production of first micro unit (AU, ST, TT)
- **FixT2**: Number of fixations on target text unit during production of second micro unit (AU, ST, TT)
- **RPDur**: Regression path duration (ST, TT)
- **Regr**: Boolean value indicating whether regression started from token (ST, TT)

7. Concurrent keyboard and gaze activities:

- **ParalK**: Parallel keyboard activity during gaze activity (FU, FD)
- **ParalS**: Parallel source text gaze activity during typing (PU)
- **ParalT**: Parallel target text gaze activity during typing (PU)
- **ParalS1**: Parallel source text gaze activity during typing micro unit one (AU, ST, TT)
- **ParalS2**: Parallel source text gaze activity during typing micro unit two (AU, ST, TT)
- **ParalT1**: Parallel target text gaze activity during typing micro unit one (AU, ST, TT)
- **ParalT2**: Parallel target text gaze activity during typing micro unit two (AU, ST, TT)

8. Starting times and durations of units and phases:

- **Dur**: Duration of unit production time (AU, CU, EX,FD, FU, PU, SG, SS, ST, TT)
- **Dur1**: Duration of first micro unit production time (AU, ST, TT)
- **Dur2**: Duration of second micro unit production time (AU, ST, TT)
- **Fdur**: Duration of segment production time excluding keystroke pauses ≥200s (SG, SS)
- **Kdur**: Duration of coherent keyboard activity excluding keystroke pauses ≥5 s (SG, SS)
- **Pdur**: Duration of coherent keyboard activity excluding keystroke pauses ≥ s (SG, SS)
- **Pnum:** Number of production units (SG, SS)
- **Time:** Starting time of unit (CU, EX, FD, FU, KD, PU)
- **Time1:** Starting time of first micro unit (AU, ST, TT)
- **Time2:** Starting time of second micro unit (AU, ST, TT)
- **TimeR:** Starting time of revision phase (SS)
- **TimeD:** Starting time of drafting phase (SS)

9. Pausing before the starting time of a unit:

- **Pause**: Pause between end of previous and start of current unit (FU, PU)
- **Pause1**: Pause between end of previous unit and start of first micro unit (AU, ST, TT)
- **Pause2**: Pause between end of previous unit and start of second micro unit (AU, ST, TT)

10. GUI related information:
    - **Win**: Window in which gaze activity was recorded,1: source text, 2: target text window (FD)
    - **Cursor**: Character offset on which activity, keystrokes, fixations, was recorded, (FD, KD)
11. External resources
    - **Focus:** Name of the window in focus (EX)
    - **KDidL:** ID of last keystroke before leaving Translog-II (EX)
    - **KDidN:** ID of next keystroke after returning to Translog-II (EX)
    - **STidN:** ID of next source token after returning to Translog-II (EX)
    - **STidL:** ID of last source token before leaving Translog-II (EX)
    - **STsegL:** Source segment identifier of last event (EX)
    - **STsegN:** Source segment identifier of next event (EX)
12. Miscellaneous features:
    - **Type:** Type of keystroke: [AM]ins, [AM]del (KD)
    - **Type:** Type of activity unit, as discussed in Sect. 2.4.5 (CU)
    - **Label:** Label for activity units (CU)

References

Alves, F., & Vale, D. C. (2011). On drafting and revision in translation: A corpus linguistics oriented analysis of translation process data. *Translation: Corpora, Computation, Cognition. Special Issue on Parallel Corpora: Annotation, Exploitation, Evaluation, 1*(1), 105–122. http://www.t-c3.org/.

Carl, M. (2012a). Translog-II: A program for recording user activity data for empirical reading and writing research. In *The eighth international conference on language resources and evaluation* (pp. 2–6). May 21–27, 2012, Istanbul, Tyrkiet. Department of International Language Studies and Computational Linguistics.

Carl, M. (2012b). The CRITT TPR-DB 1.0: A database for empirical human translation process research. In S. O'Brien, M. Simard, & L. Specia (Eds.), *Proceedings of the AMTA 2012 workshop on post-editing technology and practice (WPTP 2012)* (pp. 9–18). Stroudsburg, PA: Association for Machine Translation in the Americas (AMTA).

Carl, M., & Kay, M. (2011). Gazing and typing activities during translation: A comparative study of translation units of professional and student translators. *Meta, 56*(4), 952–975.

Jakobsen, A. L. (2002). Translation drafting by professional translators and by translation students. In G. Hansen (Ed.), *Empirical translation studies: Process and product* (pp. 191–204). Copenhagen: Samfundslitteratur.

Jakobsen, A. L. (2011). Tracking translators' keystrokes and eye movements with translog. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research: Integrative approaches in translation studies* (Benjamins translation library, Vol. 94, pp. 37–55). Amsterdam: John Benjamins.

Jakobsen, A. L., & Schou, L. (1999). Translog documentation. In G. Hansen (Ed.), *Probing the process in translation methods and results* (pp. 1–36). Copenhagen: Samfundslitteratur.

Jakobsen, A. L. (2005). Instances of peak performance in translation. *Lebende Sprachen, 50*(3), 111–116.

Germann, U. (2008). Yawat: Yet another word alignment tool. In *Proceedings of the ACL-08: HLT demo session (Companion Volume)* (pp. 20–23). Columbus, OH: Association for Computational Linguistics.

Lacruz, I., & Shreve, S. (2014). Pauses and cognitive effort in post-editing. In post-editing of machine translation: Processes and applications. In S. O'Brien, M. Simard, L. Specia, M. Carl, & L. W. Balling (Eds.), *Expertise in post-editing: Processes, technology and applications* (pp. 246–274). Cambridge: Scholars Publishing.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392.

Sanchis-Trilles, G., Alabau, V., Buck, C., Carl, M., Casacuberta, F., Martinez, M. G., et al. (2014). Interactive translation prediction versus conventional post-editing in practice: A study with the CasMaCat workbench. *Machine Translation, 28*(3–4), 217–235.

Vandepitte, S., Hartsuiker, R. J., & Van Assche, E., (2015). Process and text studies of a translation problem. In A. Ferreira, & J. W. Schwieter (Eds.), *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting.* (pp. 127–143).

Part II
Post-editing with CASMACAT

Chapter 3
Integrating Online and Active Learning in a Computer-Assisted Translation Workbench

Daniel Ortiz-Martínez, Jesús González-Rubio, Vicent Alabau, Germán Sanchis-Trilles, and Francisco Casacuberta

Abstract This chapter describes a pilot study aiming at testing the integration of online and active learning features into the computer-assisted translation workbench developed within the CASMACAT project. These features can be used to take advantage of the new knowledge implicitly provided by human experts when they generate new translations. Online learning (OL) allows the system to learn from user feedback in real time by incrementally adapting the parameters of the statistical models involved in the translation process. On the other hand, active learning (AL) determines those sentences that need to be supervised by the user so as to maximize the final translation quality minimizing user effort and, at the same time, improving the statistical model parameters. We investigate the effect of these features on translation productivity, using interactive translation prediction (ITP) as a baseline. ITP is a computer assisted translation approach where the user interactively collaborates with a statistical machine translation system to generate high quality translations. User activity data was collected from ten translators using key-logging and eye-tracking. We found that ITP with OL performs better than standard ITP, especially in terms of typing effort required from the user to generate correct translations. Additionally, ITP with AL provides better translation quality than standard ITP for the same levels of user effort.

D. Ortiz-Martínez (✉) • F. Casacuberta
Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, Camino de Vera s/n, 46021 Valencia, Spain
e-mail: dortiz@prhlt.upv.es; fcn@prhlt.upv.es

J. González-Rubio
Unbabel Lda., 1000-201 Lisboa, Portugal
e-mail: jesus@unbabel.com

V. Alabau • G. Sanchis-Trilles
Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, Camino de Vera s/n, 46021 Valencia, Spain

Sciling S.L., Valencia, Spain
e-mail: valabau@sciling.com; gsanchis@sciling.com

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*, New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_3

Keywords Active learning • Computer assisted translation • Interactive translation prediction • Online learning • Post-editing • Statistical machine translation

3.1 Introduction

The use of machine translation (MT) systems for the production of post-editing drafts has become a widespread practice in the industry. Many language service providers use post-editing workflows due to a greater availability of resources and tools for the development of MT systems, as well as a successful integration of MT systems in well-established computer-assisted translation (CAT) workbenches.

This chapter reports on the CAT workbench developed within the CASMACAT project.[1] This study is focused on one of the different features implemented in this workbench, more specifically, the *interactive translation prediction* (ITP) approach (Langlais and Lapalme 2002; Casacuberta et al. 2009; Barrachina et al. 2009). Within the ITP framework (see Sect. 3.2 for more details), the user collaborates with a statistical machine translation (SMT) system so as to generate high quality translations with less effort.

Conventional translation systems are not able to learn from user feedback, repeating the same errors when translating the same or similar sentences contained in a given document. One of the main goals of the CASMACAT project is to design and implement techniques to effectively deal with this problem. For this purpose, the ITP approach is extended by introducing two new features, namely, online and active learning. These two new features (see Sect. 3.3 for more details) are designed to allow the system to incrementally update the model parameters in real time from the target translations validated by the user. After the models have been updated for a specific sentence, the system will generate better translations not only for that sentence but for similar ones, improving the productivity of the users. Despite the strong potential of these features to improve the user experience (Ortiz-Martínez et al. 2010; González-Rubio et al. 2012; Bertoldi et al. 2013; Denkowski et al. 2014), they are still not widely implemented in CAT systems. To the best of our knowledge, the only exception is Ortiz-Martínez et al. (2011), where the authors describe the implementation of online learning within an ITP system.

This chapter reports the results obtained during an evaluation of the CASMACAT workbench with human users under three different conditions[2]: (1) basic ITP, (2) ITP with online learning, and (3) ITP with active learning (see Sects. 3.4 and 3.5). The ultimate aim of testing these different configurations was to assess their potential in real world post-editing scenarios for the benefit of the human translator.

---

[1] CASMACAT: *Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation*. Project co-funded by the European Union under the Seventh Framework Programme Project 287576 (ICT-2011.4.2).

[2] The logging data for this study can be found in the TPR-DB as the ETF14 study: https://sites.google.com/site/centretranslationinnovation/tpr-db.

3.2 Background

In this section, we briefly describe the statistical approach to machine translation, as well as its application to ITP.

3.2.1 Statistical Machine Translation

Given a sentence $\mathbf{s}$ in a source language, the *machine translation* problem can be stated as finding its translation $\mathbf{t}$ in a target language of maximum probability (Brown et al. 1993):

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} \Pr(\mathbf{t} \mid \mathbf{s}) \tag{3.1}$$

$$= \arg\max_{\mathbf{t}} \Pr(\mathbf{t}) \cdot \Pr(\mathbf{s} \mid \mathbf{t}) \tag{3.2}$$

The terms in Eq. (3.2) are the *language model* probability $\Pr(\mathbf{t})$ that represents the well-formedness of $\mathbf{t}$ and the *translation model* $\Pr(\mathbf{s} \mid \mathbf{t})$ that represents the relationship between the source sentence and its translation. The reader should note that, if we had perfect models, the use of Eq. (3.1) would suffice. Given that we have only approximations, the use of Eq. (3.2) allows the language model to correct deficiencies in the translation model.

However, in practice we often estimate $\Pr(\mathbf{t} \mid \mathbf{s})$ directly by combining all these models (and possibly others) into a *log-linear model* (Och and Ney 2002):

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} \left\{ \sum_{n=1}^{N} \lambda_n \cdot \log(f_n(\mathbf{t}, \mathbf{s})) \right\} \tag{3.3}$$

where $f_n(\mathbf{t}, \mathbf{s})$ can be any model that represents an important feature for the translation process, N is the number of models (or features), and λ_n is the weight of the nth model in the log-linear combination.

Currently, most popular MT systems are based on the use of n-gram[3] models (see for instance Chen and Goodman 1996) to implement language models and phrase-based models (Koehn et al. 2003) as translation models. The so called n-gram models assign probabilities to individual words of the target language taking into account the last n-1 words. On the other hand, the basic idea of phrase-based translation is to segment the source sentence into phrases, then to translate each source phrase into a target phrase, and finally to reorder the translated target phrases in order to compose the target sentence. If we summarize all the decisions made

[3] Sequences of n consecutive words in the translation.

during the phrase-based translation process by means of the hidden variable $\tilde{a}_1^K$, we obtain the expression:

$$Pr(\mathbf{s}|\mathbf{t}) = \sum_{K,\tilde{a}_1^K} Pr(\tilde{s}_1^K, \tilde{a}_1^K \mid \tilde{t}_1^K) \tag{3.4}$$

where each $\tilde{a}_k \in \{1 \ldots K\}$ denotes the index of the target phrase $\tilde{t}$ that is aligned with the k-th source phrase $\tilde{s}_k$, assuming a segmentation of length K.

3.2.2 Statistical Interactive Translation Prediction

Unfortunately, current MT technology is still far from perfect. This implies that, in order to achieve good translations, manual post-editing is needed. An alternative to this decoupled approach (first MT, then manual correction) is given by the ITP paradigm (Barrachina et al. 2009). Under this paradigm, translation is considered as an iterative left-to-right process where the human and the computer collaborate to generate the final translation.

Figure 3.1 shows an example of the ITP approach. There, a source Spanish sentence $\mathbf{s}$ = "Para ver la lista de recursos" is to be translated into a target English sentence $\hat{\mathbf{t}}$. Initially, with no user feedback, the system suggests a complete translation $\mathbf{t}_s$ = "To view the resources list". From this translation, the user marks a prefix $\mathbf{p}$ = "To view" as correct and begins to type the rest of the target sentence.

source (s): Para ver la lista de recursos
desired translation ($\hat{\mathbf{t}}$): To view a listing of resources

IT-0	**p**						
	$\mathbf{t}_s$	*To*	*view*	*the*	*resources*	*list*	
IT-1	**p**	To	view				
	k			**a**			
	$\mathbf{t}_s$				*list*	*of*	*resources*
IT-2	**p**	To	view	a	list		
	k				**i**		
	$\mathbf{t}_s$				*ng*	*resources*	
IT-3	**p**	To	view	a	listing		
	k					**o**	
	$\mathbf{t}_s$					*f*	*resources*
END	**p**	To	view	a	listing	of	resources

Fig. 3.1 ITP session to translate a Spanish sentence into English. The desired translation is the translation the human user wants to obtain. At iteration zero (IT-0), the system suggests a translation ($\mathbf{t}_s$). At IT-1, the user moves the mouse to accept the first eight characters "To view " and presses the a key (k), then the system suggests completing the sentence with "*list of resources*" (a new $\mathbf{t}_s$). Iterations 2 and 3 are similar. In the final iteration, the user accepts the current translation

Depending on the system or the user's preferences, the user might type the full next word, or only some letters of it (in our example, the user types the single next character "a"). Then, the system suggests a new suffix $\mathbf{t}_s$ = "list of resources" that completes the user-validated prefix and the input the user has just typed ($\mathbf{p}$ = "To view a"). The interaction continues with a new prefix validation followed, if necessary, by new input from the user. This process continues until the user considers the translation to be satisfactory.

The crucial step of the process is the production of the suffix. Again, decision theory tells us to maximize the probability of the suffix given the available information. Formally, the best suffix of a given length will be:

$$\hat{\mathbf{t}}_s = \arg\max_{\mathbf{t}_s} \Pr(\mathbf{t}_s \mid \mathbf{s}, \mathbf{p}) \tag{3.5}$$

which can be straightforwardly rewritten as:

$$\hat{\mathbf{t}}_s = \arg\max_{\mathbf{t}_s} \Pr(\mathbf{p}, \mathbf{t}_s \mid \mathbf{s}) \tag{3.6}$$

$$= \arg\max_{\mathbf{t}_s} \Pr(\mathbf{p}, \mathbf{t}_s) \cdot \Pr(\mathbf{s} \mid \mathbf{p}, \mathbf{t}_s) \tag{3.7}$$

Note that, since $\mathbf{p}\,\mathbf{t}_s = \mathbf{t}$, this equation is very similar to Eq. (3.2). The main difference is that now the search process is restricted to those target sentences $\mathbf{t}$ that contain $\mathbf{p}$ as prefix. This implies that we can use the same MT models (including the log-linear approach) if the search procedures are adequately modified (Och et al. 2003). Finally, it should be noted that the statistical models are usually defined at a word level, while the ITP process described in this section works at a character level. To deal with this problem, during the search process it is necessary to verify the compatibility between $\mathbf{t}$ and $\mathbf{p}$ at a character level.

3.2.3 Search

In conventional SMT, the best translation for a given source sentence is produced by incrementally generating the target sentence from left to right. This problem can be solved by means of *dynamic programming* (DP) techniques (Bellman 1957). Due to the great complexity of the search process in SMT, DP-based search is typically restricted by introducing the beam-search heuristic (Jelinek 1997).

Due to the demanding temporal constraints inherent to any interactive environment, performing beam-search each time the user validates a new prefix is unfeasible. The usual approach is to rely on a certain representation of the search space that includes the most probable translations of the source sentence. The computational cost of this approach is much lower, since for each source sentence,

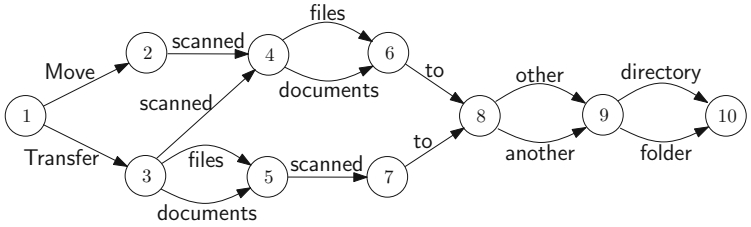


Fig. 3.2 Example of a word graph encoding different English translations for the Spanish source sentence "Transferir documentos explorados a otro directorio"

the translation representation can be generated only once when obtaining the initial translation and reused for further completion requests.

The representation usually chosen to implement ITP is known as *word graph*. A word graph is a weighted directed acyclic graph, in which each node represents a partial translation hypothesis and each edge is labeled with a word of the target sentence and is weighted according to the language and translation model scores. Ueffing et al. (2002) give a detailed description of word graphs and how to produce them easily as a sub-product of the SMT search process. An example of word graph is displayed in Fig. 3.2. During the interaction process, the system makes use of this word graph in order to complete the prefixes accepted by the human translator. First, the system looks for a node in the word graph that represents the partial translation validated by the user. Then, the system follows the most probable path from such a node to an end node, and returns the target language suffix defined by this path.

For a fixed source sentence, if no pruning is applied, the word graph represents all possible sequences of target words for which the posterior probability is greater than zero according to the models. In practice, however, the pruning needed to render the problem computationally feasible implies that the resulting word graphs only represent a subset of the possible translations. Therefore, it may happen that the user sets a prefix not encoded in the word graph. To circumvent this problem, we introduce the use of error correction techniques. First, we look for the node that represents a partial translation with minimum edit distance (Levenshtein 1966) to the prefix. Then, we select the completion path which starts with the last word of the prefix and has the best backward score.[4] This scoring mechanism based on edit distance can be introduced in the statistical formalization of the system by using probabilistic finite-state machines (see Ortiz-Martínez 2011 for a detailed description).

[4]This is the score associated with a path going from the node representing the partial translation with minimum edit distance to the final node.

3.3 Online and Active Learning for SMT

The proposed CAT workbench has been extended by incorporating online and active learning, which are targeted at optimizing the quality of the final translations and speeding the post-editing process by taking advantage of user feedback in real time.

3.3.1 Online Learning

Online learning (OL) allows us to efficiently re-estimate the parameters of the SMT model with the new translations generated by the user. As a result, the SMT system is able to learn from the translation edits of the user preventing further errors in the machine generated translations.

Conventional batch learning techniques establish a strict separation between model training and the subsequent use of the estimated parameters for prediction. As a result, SMT systems implementing batch learning require to retrain on the whole corpus whenever a new training example is available, spending days or even weeks of computation depending on the size of the training set. In contrast, OL techniques process the training examples one at a time or in small batches. This approach allows the re-estimation of the parameters of an SMT model in constant time, regardless of the number of training examples previously processed.

The application of OL to the SMT framework requires the definition of incremental update rules for the statistical models involved in the translation process. For this purpose, first it is necessary to identify a set of sufficient statistics for such models. A sufficient statistic for a statistical model is a statistic that captures all the information that is relevant to estimate this model. If the estimation of the statistical model does not require the use of the EM algorithm (Dempster et al. 1977), e.g. language models, then it is generally easy to incrementally extend the model given a new training sample. By contrast, if the EM algorithm is required, e.g. alignment models, the estimation procedure has to be modified, since the conventional EM algorithm is designed for its use in batch learning scenarios. To address this problem, the incremental version of the EM algorithm defined in Neal and Hinton (1999) can be used.

Here we adopt the online learning techniques described in Ortiz-Martínez et al. (2010) and substantially extended in Ortiz-Martínez (2015). In these works, the authors define an incrementally updateable log-linear model for SMT. This log-linear model is composed of a set of seven feature functions, including a n-gram language model using interpolated Kneser-Ney smoothing, an inverse sentence-length model implemented with gaussian distributions, inverse and direct phrase-based models, a target phrase-length model, a source phrase-length model, and a distortion model. The authors of the above mentioned works define sufficient statistics for the different models, providing the inverse and direct phrase models a special treatment, since their estimation involves the use of HMM-based alignment

models to generate word alignment matrices for the sentence pairs contained in the training corpus (see Koehn et al. 2003). The parameters of such alignment models are obtained by means of the incremental version of the EM algorithm.

3.3.2 Active Learning

Active learning (AL) applied to ITP aims at optimizing translation quality when the available resources (e.g. manpower, time, money, etc.) are limited (González-Rubio and Casacuberta 2014). In particular, the user is asked to post-edit only a subset of the machine generated translations with worse quality. After each translation is post-edited, we re-train the SMT model with the new translation example, which is immediately available for the next sentence to post-edit. Finally, the translation system, which has been improved with all the post-editings performed, returns the SMT outputs for the rest of the sentences.

This AL framework has several potential advantages over conventional ITP technology. On the one hand, asking the user to only translate a subset of the sentences allows us to limit the amount of effort to be invested in the translation process and, by focusing human effort on those sentences for which the investment of user effort is estimated to be more profitable, we also maximize the utility of each user interaction. On the other hand, the underlying SMT model is continually updated with new examples which allows the system to learn new translations and to adapt its outputs to match the preferences of the user. As a result, the subsequent machine generated translations will be closer to those preferred by the user thus reducing the human effort required to translate them. Additionally, all these technicalities are transparent to the user who interacts with the system in the same way she does with a conventional ITP system.

An important practical challenge is the strict bound on the response time imposed by the interaction with the user. This fact constrains the models and techniques that can be used to implement AL. Particularly, we select which sentences should be post-edited by the user according to an uncertainty criterion (Lewis and Gale 1994). Sentences to be post-edited would be those for which the system is more uncertain about its translation. Then, given a new translation example, the parameters of the SMT model are re-estimated via the OL techniques described above.

Under the assumption that the "certainty" of a model in a particular translation is correlated with the quality of that translation, we measure the uncertainty of a translation with a sentence-level quality measure based on statistical translation lexicons (González-Rubio et al. 2012). Given a translation, we first compute a quality score for each of its words as the maximum word-to-word translation probability respect to any of the words in the source sentence. In our experiments, we used an IBM-1 alignment model (Brown et al. 1993) to measure translation probabilities between words. Then, the uncertainty score for the translation is computed as one minus the geometric average of such word-quality scores.

In the experimentation, we used the incremental version of the EM algorithm (Neal and Hinton 1999) to update the word-to-word translation probability model with the new sentence pairs available. We thus maintain an updated version of the probability distribution over translations so that the user is not repeatedly asked to supervise translations that provide similar information.

3.4 Experimental Design

Online and active learning were assessed through a series of experiments with ten professional translators. In this section, we describe the workbench, corpus, SMT engine, participants, methodology, and assessment measures employed in the experiments. The logging data that was used to carry out these experiments can be found in the TPR-DB as the ETF14 study.[5]

3.4.1 The CASMACAT *Workbench*

The CASMACAT workbench (Alabau et al. 2013) has been developed on top of the MATECAT post-editing interface (Bertoldi et al. 2012). The user is presented with a GUI in which the left-hand window displays the source text while the right-hand one contains the target text. Texts are split into segments (corresponding to sentences and headings in the text) so that the translator post-edits one translation segment at a time. The user can see several segments on the screen at the same time and can scroll back and forth to choose which segment to translate. The workbench contains a fully-fledged MT engine with interactivity which can search for alternative translations whilst the user is post-editing the machine translation. Figure 3.3 shows a screenshot of the CASMACAT workbench.

Moreover, the workbench includes facilities for logging system configuration and user activity data including keystrokes and gaze obtained using an eye-tracking device. However, these logging features were not exploited in the present experimentation.

3.4.2 Corpus

Our experiments were based on ITP systems with models initialized using the well known Europarl corpus (Koehn 2005). Europarl is extracted from the proceedings of the European Parliament, which are written in the different languages of the

[5] https://sites.google.com/site/centretranslationinnovation/tpr-db.

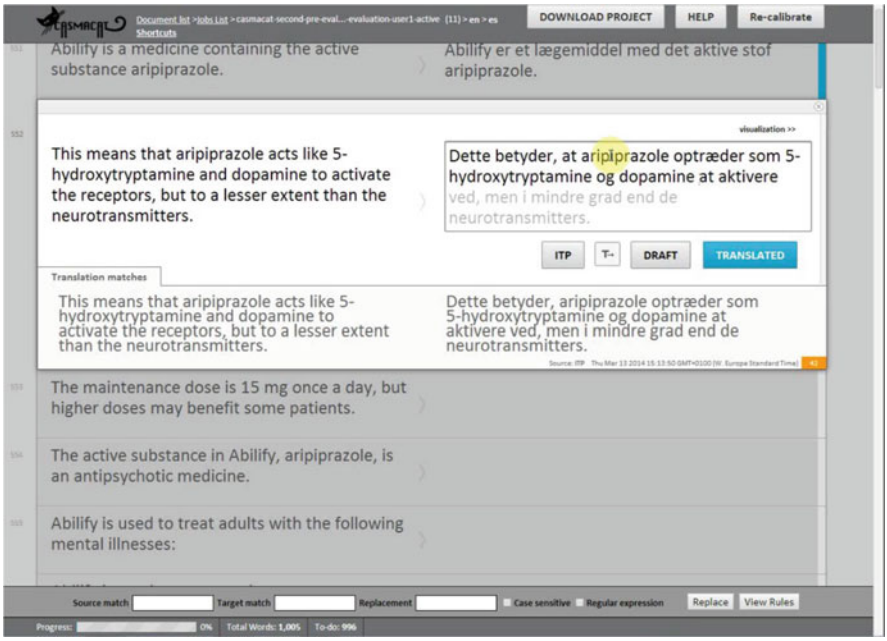


Fig. 3.3 Screenshot of the CASMACAT workbench showing the user's point of gaze in *yellow*

Table 3.1 Main figures of the corpora used in the experiments

	Europarl		EMEA		
			Test		
	Training	Development	d1	d2	d3
Segments	1.9 M	3003	53	55	55
Words (En/Es)	51.3 M/53.7 M	73.0 K/78.8 K	989/–	958/–	979/–

European Union. In our experiments we have used the version that was created for the shared task of the ACL 2013 Workshop on Statistical Machine Translation (Bojar et al. 2013). This version includes a training set, used to estimate the parameters of the language and translation models, as well as a development corpus that has been used to adjust the weights of the log-linear model underlying our ITP systems.

The test texts involved in this experimentation were documents from the European Medicines Agency as compiled in the EMEA corpus (Tiedemann 2009). From the English-Spanish (En-Es) partition of EMEA, we created three different documents (d1, d2 and d3) containing consecutive sentences and being roughly of the same size. Table 3.1 shows the main figures of the corpora used in the experiments.

3.4.3 SMT Engine

The SMT engine providing ITP functionalities integrated with the OL and AL techniques described in Sect. 3.3, has been implemented using the Thot toolkit (Ortiz-Martínez and Casacuberta 2014). Thot is an open source toolkit for SMT incorporating a fully-fledged machine translation decoder as well as tools to train state-of-the-art log-linear translation models.

The Thot toolkit is fully integrated into the CASMACAT workbench. In the experimentation reported here, Thot was used to train models for the above mentioned Europarl corpus. Once the models were trained, each English source text belonging to the test sets extracted from EMEA was automatically translated into Spanish using Thot and then automatically loaded into the CASMACAT workbench for the participants to post-edit.

3.4.4 Participants

We conducted our experiments in cooperation with Celer Soluciones, a language service provider (LSP) based in Madrid, Spain. The experiments involved ten freelance translators, all native speakers of Spanish offering translation and post-editing services on a regular basis for this LSP.

In an attempt to unify post-editing criteria among participants, all of them were instructed to follow the same post-editing guidelines aiming at a final high-quality target text (publishable quality). The post-editing guidelines distributed in hard copy were[6]:

- Retain as much raw MT as possible.
- Do not introduce stylistic changes.
- Make corrections only where absolutely necessary, i.e. correct words and phrases that are clearly wrong, inadequate or ambiguous according to Spanish grammar.
- Make sure that there are no mistranslations with regard to the English source text.
- Publishable quality is expected.

Additionally, before starting their tasks, participants were introduced to the CASMACAT workbench and the ITP post-editing protocol. They were given time to familiarize themselves with the workbench, and asked to start the translation tasks only after they consider themselves comfortable working with the tool.

[6]Similar instructions were also used in other CASMACAT studies (Chaps. 4, 5, 7, 8 and 13).

Table 3.2 Task assignments in the experiments

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
ITP	d1	d3	d2	d1	d3	d2	d1	d3	d2	d1
PIVO	d2	d1	d3	d2	d1	d3	d2	d1	d3	d2
PIVA	d3	d2	d1	d3	d2	d1	d3	d2	d1	d3

Each user translated the three test documents (d1, d2 and d3) in the three different conditions analyzed in this study: ITP, ITP with online learning (PIVO) and ITP with active learning (PIVA). The number contained in the document name indicates the order in which the participant translated such document

3.4.5 Methodology

Three different setups of the CASMACAT workbench were evaluated in the experiments: conventional ITP (ITP), ITP with online learning (denoted PIVO according to the TPR-DB naming conventions), and ITP with active learning (PIVA). Each participant translated one test document in one of the conditions. The assignment between documents and conditions were randomized for each participant, Table 3.2 gives an overview of such assignments. The number in the document name indicates the order in which the participant translated that document. Note that both the assignment between document and condition, and the order of the conditions were randomized for the ten participants. Keyboard and mouse activity was logged for each individual task.

3.4.6 Assessment Measures

The main goal of this user study was to assess and compare online learning and active learning against conventional ITP. Specifically, we wanted to study the impact of these two techniques in the performance of ITP workbenches. In our case, we evaluated the performance of the ITP workbench using two different measures of the translation process:

- Speed: total number of source words translated by the participant divided by time in seconds.
- Effort: total number of edits (i.e., key-strokes) needed to generate the translations divided by the total number of source words.

The speed measure is a coarse measure of the productivity of a CAT system. It is easy to interpret but often highly noisy. The effort measure aims at evaluating the typing effort invested by the participant to generate the translations. In this case, a better system would be the one requiring less effort from the user.

Additionally, we are also interested in evaluating the quality of the final translation of each test document; particularly in the case of PIVA that aims at optimizing the trade-off between translation quality and human post-editing effort.

In our experiments, we used the widespread BLEU score (Papineni et al. 2002) that measures the quality of a candidate translation by comparing it against a reference translation. Specifically, BLEU is calculated as the precision of n-grams weighted by a brevity penalty designed to penalize short translations. BLEU results vary between zero and one although usually it is represented as a percentage where 100 % denotes a perfect translation.

3.5 Results

The results of the experiments carried out are presented in the following two subsections. First, we present results related to online learning and its impact in the performance of ITP. Then, we present the corresponding results for active learning.

3.5.1 Impact of Online Learning

Online learning is expected to save typing effort to users of ITP systems, since it allows to take advantage of user feedback, avoiding the necessity of correcting recurrent errors. In order to assess the performance of online learning, we have conducted experiments measuring the number of edits per each translated source word as well as the number of source words per second for the different participants.

Table 3.3 shows the typing effort measured in number of edits per each translated source word for conventional and online ITP systems. In all cases, the online ITP system outperformed the conventional ITP system, obtaining a 26 % reduction in the typing effort on average. This reduction was equal or close to 50 % for some specific users (P1, P8 and P9).

On the other hand, Table 3.4 shows a comparison of participant translation speed measured in terms of number of translated source words per second for conventional and online ITP systems. In the same way that was observed for the effort measures, the system with online learning was able to outperform the results obtained by means of conventional ITP. However, in this case the measures for individual ITP system users were mixed, with six participants that were able to type faster by means of the online ITP system, and four participants that were faster using conventional ITP. It should be noted that the number of edits was always smaller for the online system, even in those cases in which the translation speed

Table 3.3 Effort required from the user measured in terms of number of edits per each source word to be translated for conventional ITP and ITP with online learning (best results are shown in bold face)

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Avg
ITP	10.6	9.1	11.8	10.9	8.5	11.2	5.4	10.1	8.6	5.3	9.1
PIVO	**6.3**	**8.0**	**7.3**	**10.5**	**7.9**	**9.1**	**4.5**	**4.8**	**4.8**	**4.7**	**6.7**

Table 3.4 Typing speed measured in translated source words per second for conventional ITP and ITP with online learning (best results are shown in bold face)

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Avg
ITP	0.37	0.18	0.21	**0.34**	**0.39**	0.24	0.37	**0.26**	0.17	**0.46**	0.29
PIVO	**0.51**	**0.21**	**0.28**	0.30	0.36	**0.30**	**0.49**	0.18	**0.24**	0.34	**0.32**

with respect to conventional ITP was worse. Despite surprising at first, these results were coherent with previous research. In fact, in Alabau et al. (2014) it was found that post-editors tend to spend more time outside the CASMACAT workbench when working with the online learning approach. The authors of that work hypothesized that the participants felt the necessity to do Internet searches so as to double check correct translations generated by the ITP system with OL. This hypothesis was confirmed by inspecting the recorded videos of the translation process of selected source segments, obtaining significant gains in translation speed when the time spent by the participants making Internet searches was removed from the study. This same explanation can be applied to our experiments.

3.5.2 Impact of Active Learning

As described in Sect. 3.3.2, the main potential advantage of AL in the ITP framework (PIVA) is its ability to optimize the quality of the final translations per unit of human effort. In this scenario, the participant only post-edits a subset of the translations while the system returns automatic SMT outputs for the rest of the sentences. Thus, we were interested in studying the trade-off between translation quality and post-edit workload that can be achieved in the ITP and PIVA scenarios.

Figures 3.4 (participants 1–5) and 3.5 (participants 6–10) display translation quality obtained for the test documents as a function of the workload involved in generating the translations. Each row displays the results for one participant. Each point denotes the quality of the translated document for a given level of workload. The different workload levels are given by the number of sentences actually post-edited by the participant. That is, the leftmost point in each effort plot (rightmost for speed plots) represents the quality of the translated document when the participant post-edits zero sentences, the next one represents translation quality when the participant post-edits one sentence, and so forth up to the rightmost point (leftmost for speed) that represents the quality of the translated document when the participant post-edits all sentences. For those sentences not post-edited by the user, we return their automatic SMT translations.[7] The sentences to be post-edited each time were selected according to the active learning scoring function described in

[7]Note that the leftmost and rightmost points in each plot are equivalent to the SMT and PIVO scenarios respectively.

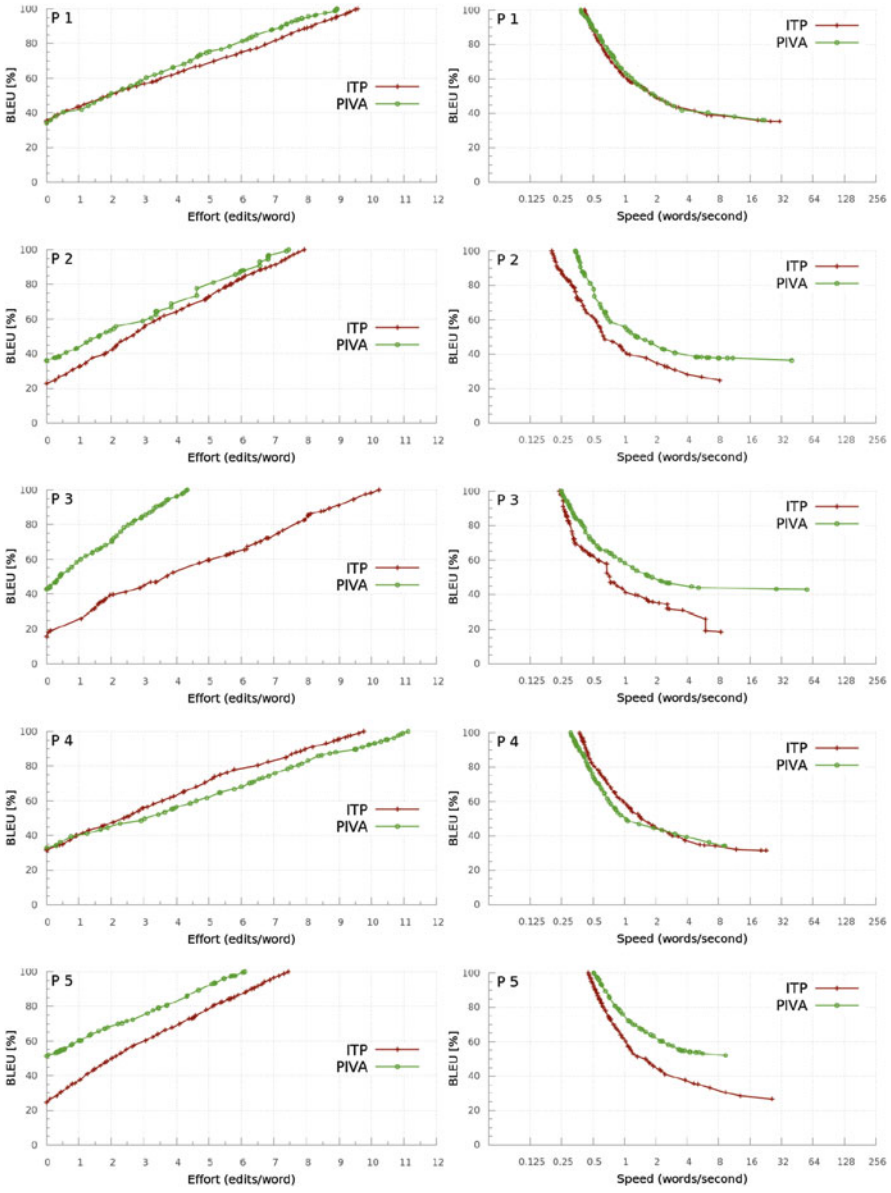


Fig. 3.4 Quality of the final translations (BLEU) generated by participants 1–5 (one per row) as a function of the translation effort (*left*) or the translation speed (*right*)

Sect. 3.3.2. We used the whole post-edited document generated by each participant as the reference translation of this participant for that document. Workload was measured both in terms of typing effort (left column) and translation speed (right

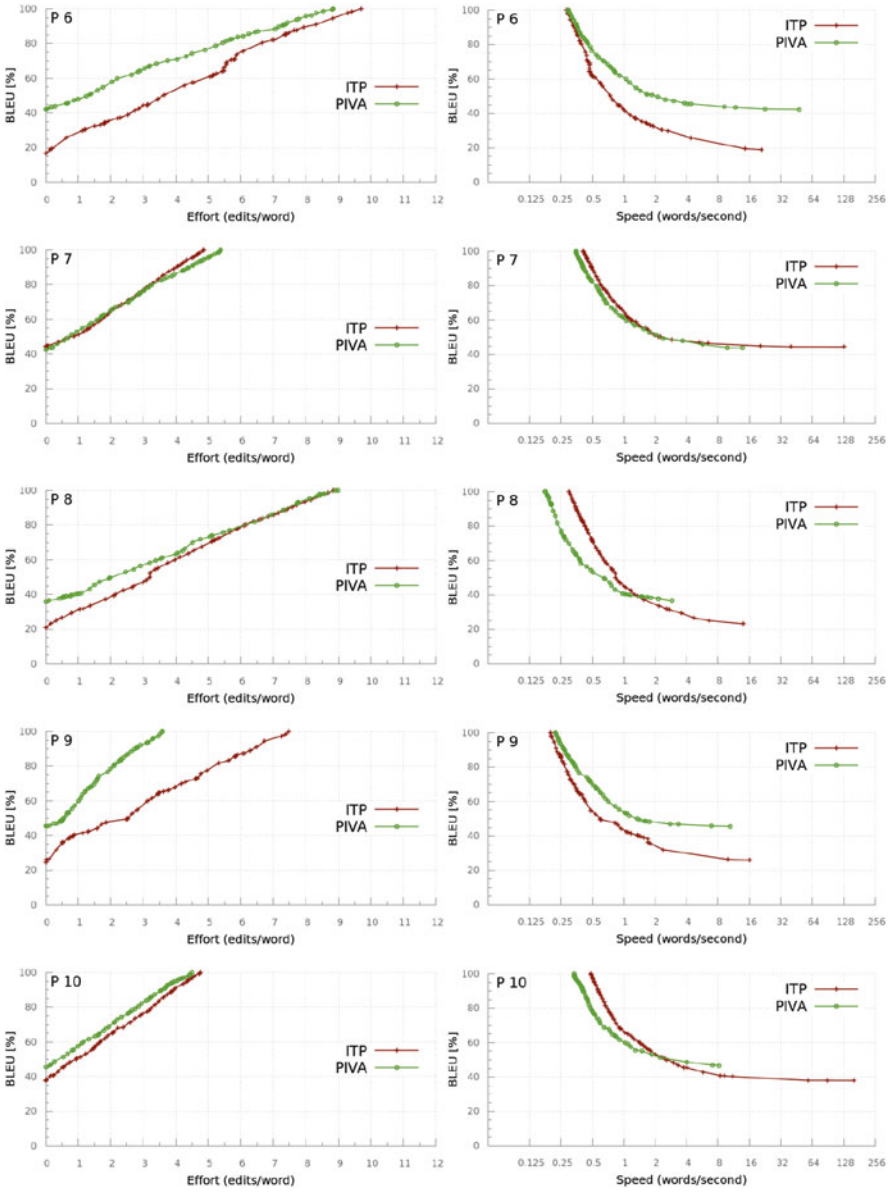


Fig. 3.5 Quality of the final translations (BLEU) generated by participants 6–10 (one per row) as a function of the translation effort (*left*) or the translation speed (*right*)

column). The difference between the translation quality of ITP and PIVA with zero workload can be explained by the different document translated, e.g. participant P1 translated document d1 with ITP and document d3 with PIVA, and the different order in which translations were performed.

Results for the different participants were quite varied and noisy. As a general result, we can say that there was a quite clear tendency of PIVA obtaining better translation quality than ITP at the same level of typing effort (left column). Nevertheless, results varied greatly between participants, as it can be seen when comparing for example the plots of participants P1 (slight improvement), P3 (clear improvement), and P4 (slight deterioration). This tendency was less clear when we measured post-editing workload in terms of translation speed (right column). In this case, there is a number of participants (P4, P7, P8, and P10) that post-edited at a lower speed using PIVA.

In order to achieve a more robust conclusion, we grouped together post-edit results for all participants and documents. Figure 3.6 displays the quality of the translated documents as a function of the typing effort (top), or the translation speed (bottom) of the post-edit process. Additionally, we also display the least-squares fit for the results of the ITP and PIVA scenarios. These fitted lines show the tendencies of the data that were shadowed by the noisy results of the individual participants. When measuring the post-edit workload by the number of edits performed by

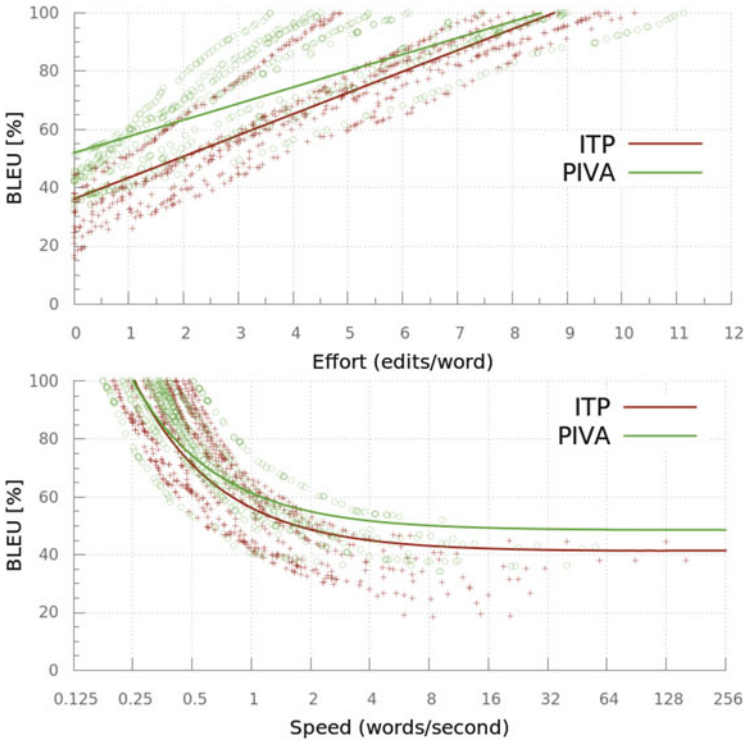


Fig. 3.6 Quality of the final translations for all participants (*dots*) as a function of the translation effort (*top*) or the translation speed (*bottom*). We also include least-squares fits (*solid lines*) to show average tendencies of the data

the participant (typing effort), results indicate that AL allowed to improve the performance of the ITP framework. That is, participants employing the same effort generated translations of higher quality when working with the PIVA approach. Similar results were obtained for post-edit speed. However, in this case, differences tend to be smaller between ITP and ITP with AL.

As we have said before, this difference between the post-edit effort and post-edit speed measures can be explained by the tendency of the users to double check translations proposed by re-trained SMT systems. However, we consider an additional complimentary explanation. The uncertainty measure used to implement active learning (see Sect. 3.3.2) is based on an estimation of the quality of the individual words in the translation. Since low-quality words are more prone to be edited by the user (González-Rubio et al. 2010), our uncertainty measure is closely related to the amount of edits required to post-edit the translation. Given these considerations, it is intuitively clear why results as measured by post-edit effort may be better than those measured by post-edit speed.

3.6 Conclusions

We have presented the results of a pilot study involving real users concerning the implementation of online and active learning within a CAT workbench with ITP functionalities. The main goal of the study was to determine whether the use of OL and AL allows to improve the performance of a conventional ITP system or not. For this purpose, the typing effort measured as the number of edit operations per each source word, as well as the speed calculated as the number of translated source words per second were obtained for ten different users translating a test set extracted from the EMEA corpus, a real translation task belonging to the medical domain.

Results showed that the users of ITP systems incorporating OL consistently required less typing effort than those using regular ITP. OL also increased the translation speed for 60 % of the users. The rest of the users were faster using the conventional system, despite requiring a greater typing effort. As it was explained in Sect. 3.5.1, a previous study using the CASMACAT workbench with OL capabilities showed that the participants felt the necessity to do Internet searches so as to double check correct translations generated by means of the ITP system with OL, substantially decreasing the translation speed. We think that this circumstance is also the explanation for the greater translation times using OL observed for particular users in the work presented here.

On the other hand, the translation quality obtained using ITP with AL was consistently better than that obtained by means of conventional ITP at the same level of typing effort. The differences in translation speed between ITP and ITP with AL were smaller and more dependent on the particular user. Again, we think that these observations are due to the tendency of the users to double check the translations generated by the updated systems, which will be reduced as the user's trust in the system learning capabilities improves with time.

Acknowledgements Work supported by EU's 7th Framework Programme (FP7/2007-2013) under grant agreement 287576 (CASMACAT).

References

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., García-Martnez, M., et al. (2013). Casmacat: An open source workbench for advanced computer aided translation. *The Prague Bulletin of Mathematical Linguistics, 100*, 101–112.

Alabau, V., Carl, M., García-Martínez, M., González-Rubio, J., Mesa-Lao, B., Ortiz-Martínez, D., et al. (2014). *D6.3: Analysis of the third field trial*. Technical report, CasMaCat project.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., et al. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics, 35*(1), 3–28.

Bellman, R. (1957). *Dynamic programming* (1st ed.). Princeton, NJ: Princeton University Press.

Bertoldi, N., Cattelan, A., & Federico, M. (2012). Machine translation enhanced computer assisted translation. First report on lab and field tests. Available from: http://www.matecat.com/wp-content/uploads/2013/01/MateCat-D5.3-V1.2-1.pdf.

Bertoldi, N., Cettolo, M., & Federico, M. (2013). Cache-based online adaptation for machine translation enhanced computer assisted translation. In *Proceedings of the MT Summit* (pp. 35–42).

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., et al. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation* (pp. 1–44). Sofia: Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19*(2), 263–311.

Casacuberta, F., Civera, J., Cubel, E., Lagarda, A. L., Lapalme, G., Macklovitch, E., et al. (2009). Human interaction for high quality machine translation. *Communications of the ACM, 52*(10), 135–138.

Chen, S. F., & Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In A. Joshi & M. Palmer (Eds.), *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics* (pp. 310–318). San Francisco: Morgan Kaufmann.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*(1), 1–38.

Denkowski, M., Dyer, C., & Lavie, A. (2014). Learning from post-editing: Online model adaptation for statistical machine translation. In *Proceedings of the EACL* (pp. 395–404). Gothenburg: Association for Computational Linguistics.

González-Rubio, J., & Casacuberta, F. (2014). Cost-sensitive active learning for computer-assisted translation. *Pattern Recognition Letters, 37*, 124–134.

González-Rubio, J., Ortiz-Martínez, D., & Casacuberta, F. (2010). On the use of confidence measures within an interactive-predictive machine translation system. In *Proceedings of the EAMT*.

González-Rubio, J., Ortiz-Martínez, D., & Casacuberta, F. (2012). Active learning for interactive machine translation. In *Proceedings of the EACL* (pp. 245–254).

Jelinek, F. (1997). *Statistical methods for speech recognition*. Cambridge: MIT.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit* (pp. 79–86).

Koehn, P., Och, F., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of Human Language Technologies: The 2003 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 48–54).

Langlais, P., & Lapalme, G. (2002). TransType: Development-evaluation cycles to boost translator's productivity. *Machine Translation, 17*(2), 77–98.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

Lewis, D., & Gale, W. (1994). A sequential algorithm for training text classifiers. In *Proceedings of the ACM SIGIR conference on Research and development in information retrieval* (pp. 3–12).

Neal, R., & Hinton, G. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models* (pp. 355–368). MIT press.

Och, F. J., & Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the ACL* (pp. 295–302).

Och, F. J., Zens, R., & Ney, H. (2003). Efficient search for interactive statistical machine translation. In *Proceedings of the European chapter of the Association for Computational Linguistics* (pp. 387–393).

Ortiz-Martínez, D. (2011). *Advances in fully-automatic and interactive phrase-based statistical machine translation*. Ph.D. thesis, Universidad Politécnica de Valencia, Valencia (Spain). Advisors: F. Casacuberta and I. García-Varea.

Ortiz-Martínez, D. (2015, submitted). Online learning for statistical machine translation.

Ortiz-Martínez, D., & Casacuberta, F. (2014). The new thot toolkit for fully automatic and interactive statistical machine translation. In *Proceedings of the EACL* (pp. 45–48).

Ortiz-Martínez, D., García-Varea, I., & Casacuberta, F. (2010). Online learning for interactive statistical machine translation. In *Proceedings of the NAACL-HLT* (pp. 546–554).

Ortiz-Martínez, D., Leiva, L. A., Alabau, V., García-Varea, I., & Casacuberta, F. (2011). An interactive machine translation system with online learning. In *ACL (System Demonstrations)* (pp. 68–73).

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02 (pp. 311–318). Association for Computational Linguistics.

Tiedemann, J. (2009). News from opus—a collection of multilingual parallel corpora with tools and interfaces. In *Proceedings of the RANLP* (Vol. V, pp. 237–248).

Ueffing, N., Och, F., & Ney, H. (2002). Generation of word graphs in statistical machine translation. In *Proceedings of the EMNLP* (pp. 156–163).

Chapter 4
Analysing the Impact of Interactive Machine Translation on Post-editing Effort

Fabio Alves, Arlene Koglin, Bartolomé Mesa-Lao, Mercedes García Martínez, Norma B. de Lima Fonseca, Arthur de Melo Sá, José Luiz Gonçalves, Karina Sarto Szpak, Kyoko Sekino, and Marceli Aquino

Abstract The combination of temporal, technical and cognitive effort has been proposed as metrics to evaluate the feasibility of post-editing on machine-translation (MT) output (Krings, 2001). In this study, we investigate the impact of interactive machine translation on the post-editing effort required to post-edit two specialized texts under experimental conditions and correlate it with Translation Edit Rate (TER) scores. Using the CasMaCat workbench as a post-editing tool in conjunction with a Tobii T60 eye tracker, process data were collected from 16 participants with some training on postediting. They were asked to carry out post-editing tasks under two different conditions: i) traditional post-editing (MT) and ii) interactive

F. Alves • K. Sekino • M. Aquino
Laboratory for Experimentation in Translation (LETRA), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil
e-mail: fabio-alves@ufmg.br; kyokosekino@ufmg.br; marceliaquinoufmg@gmail.com

A. Koglin (✉) • A.de.M. Sá • K.S. Szpak • N.B.de.L. Fonseca
Laboratory for Experimentation in Translation (LETRA), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil
e-mail: arlenekoglin@yahoo.com.br; arthurdemelosa@gmail.com; kszpak@ufmg.br; normafonseca@ufmg.br

B. Mesa-Lao
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark
e-mail: bm.ibc@cbs.dk

M.G. Martínez
Computer Laboratory, University of Maine, Le Mans, France
e-mail: Mercedes.Garcia_Martinez@univ-lemans.fr

J.L. Gonçalves
Laboratory for Experimentation in Translation (LETRA), Universidade Federal de Ouro Preto (UFOP), Belo Horizonte, Brazil
e-mail: zeluizvr@ichs.ufop.br

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_4

post-editing (IMT). In the IMT condition, as the user types, the MT system suggests alternative target translations which the post-editor can interactively accept or overwrite, whereas in the traditional MT condition no aids are provided to the user while editing the raw MT output. Temporal effort is measured by the total time spent to complete the task whereas technical effort is measured by the number of keystrokes and mouse events performed by each participant. In turn, cognitive effort is measured by fixation duration and the number of eye fixations (fixation count) in each task. Results show that IMT post-editing had significantly lower fixation duration and fewer fixation counts in comparison to traditional post-editing.

Keywords Post-editing effort • Interactive post-editing • Traditional post-editing • TER scores • CASMACAT workbench

4.1 Introduction

First investigations of post-editing effort go back to Krings's (2001) seminal work and relate mainly to his separation and categorization of temporal, technical and cognitive effort. More recently, however, with the advent of interactive machine-translation technologies, human-machine interaction has come to the centre stage. Consequently, empirical-experimental research has included interactivity as a main component of research designs. One might say that Krings (2001) was ahead of his time. He suggested a strategy that entails a crucial motivation for carrying out research on post-editing with a focus on the impact of interactivity on post-editing process. According to Krings, this strategy implies that one should benefit from interactive systems architecture when he recommends

> renunciation of fully automatic machine translation in favour of interactive system architectures where the computer relies on the knowledge of a human translator in specific situations in order to achieve better machine translation results (Krings 2001, p. 24)

In this chapter, we have taken Krings's suggestion as the starting point of our research which aims at achieving two complementary goals, namely, to compare the impact of interactive machine translation on post-editing effort and to correlate post-editing effort with Translation Edit Rate (henceforth TER) scores (Snover et al. 2006).

Building on the existing literature, we assume that Krings's distinction between temporal, technical and cognitive effort still holds true. Thus, if temporal effort refers to the amount of time needed to post-edit the MT output, a shorter task time for tasks involving interactive translation prediction (ITP) in post-editing would be a positive indicator that the ITP condition should be favoured in comparison with traditional post-editing. In other words, if post-editors spent less time when carrying out an ITP post-editing task compared to the time spent on traditional MT

post-editing, then the ITP condition should be recommended as a good practice. On the basis of such an assumption, we formulate our first hypothesis:

Hypothesis 1 Interactivity will contribute to a significant decrease in the time spent on post-editing tasks. Therefore, all variables remaining equal, temporal effort will be lower for ITP post-editing.

However, one must not forget that the output of post-editing processes should be also compared to other post-edited versions as reference to measuring the amount of editing performed on each of the segments. Translation Edit Rate, also known as TER scores, measures the minimum number of edits (insertions, deletions, substitutions or reordering) that are needed to transform the MT output into the post-edited segment used as reference. Therefore, the higher the TER score, the higher the number of modifications in the MT output. On the basis of such an assumption, we formulate our second hypothesis:

Hypothesis 2 TER scores will be lower for ITP post-editing compared to standard PE which can be measured by a significant difference in the number of edits in tasks performed under PE and ITP conditions,

In order to test these hypotheses, this chapter is divided into six sections, including this Introduction. Section 4.2 presents the theoretical framework and reviews the literature on machine translation, post-editing processes and post-editing effort. Section 4.3 introduces the method and procedures used in the experimental design whereas Sect. 4.4 presents the data analysis and the results of the experiment. Next, Sect. 4.5 discusses the results in the light of the literature review and our methodological framework. Finally, Sect. 4.6 highlights the main results as well as the shortcomings of the current study and points to future research avenues.

4.2 Theoretical Framework

4.2.1 Machine Translation

The use of machine translation (MT) systems for the production of post-editing drafts has become a widespread practice in the industry. The reasons for that are a greater availability of resources and tools for the development of MT systems, a change in the expectations of MT users, as well as a successful integration of MT systems in already well-established computer-assisted translation (CAT) workbenches.

Recent studies (Koehn 2009; Plitt and Masselot 2010; Federico et al. 2012; Flournoy and Duran 2009; Green et al. 2013) have concluded that post-editing is, on average, more efficient than translating from scratch. However, the exact design

of a more efficient form of interaction between humans and machines in the context of computer-assisted translation is still an open research question.

Traditionally, post-editing workflows only take into account the human component in a serial process (Isabelle and Church 1998). First, the MT system provides complete translations which are then proofread by a human translator. In such a serial scenario, there is no actual interaction between the MT system and the human translator, making it impossible for the MT system to benefit from overall human translation skills and preventing the human translator from making the most out of the adaptive ability of some MT systems.

The interactive framework constitutes an alternative to fully automatic MT systems in which the MT system and the human agent interact to generate translations according to different degrees of quality. The system proposes a new translation whenever the user edits a word, trying to guess the correct auto-completion for the text that the user inputs. The user can then accept or partially accept the ITP proposal.

Our research is focused on a study where the set of data was generated with the aid of the CASMACAT workbench featuring interactive translation prediction (ITP). For a description of the CASMACAT system, see Chap. 3 (Langlais and Lapalme 2002; Casacuberta et al. 2009; Barrachina et al. 2009).

CASMACAT[1] was a European project in which cognitive studies of actual unaltered translator behaviour are carried out based on key-logged and eye-tracking data. More details of this workbench are described in Chap. 3 of this book.

A screenshot of the CAT workbench can be seen in Fig. 4.1.

4.2.2 Post-editing Process

Almost 70 years have elapsed since the publication of the first report on the application of computers to translation by Warren Weaver in 1947. Drawing on that report, translation studies research and general users of MT engines still seem to agree "that the output of the machine must be submitted for review to a *post-editor*" (Kay et al. 1994, p. 39 as cited in Krings 2001, p. 2). The process a post-editor performs on the MT output constitutes what has been called post-editing, which can be defined as "reviewing a pre-translated text generated by an MT engine against an original source text, correcting possible errors, in order to comply with a set quality criteria in as few edits as possible (in general)" (Mesa-Lao 2013, p. 4).

Two further questions arise from this definition. First, to what extent an original source text is necessary in a post-editing task? This type of input is not mandatory in all post-editing tasks, such as monolingual post-editing. In this kind of post-editing,

[1]CASMACAT: Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation was a project (2011–2014) co-funded by the European Union under the Seventh Framework Programme Project 287576 (ICT-2011.4.2).

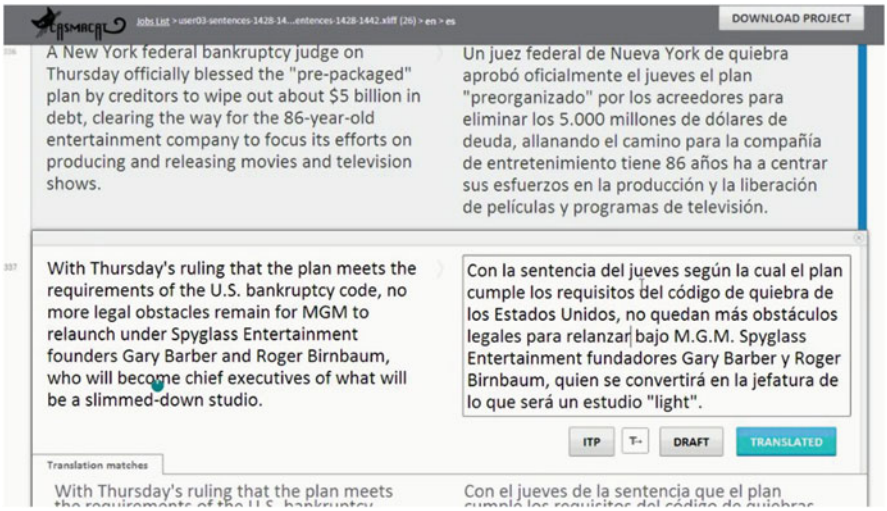


Fig. 4.1 CASMACAT interface with interactive translation prediction mode

a monolingual speaker with no knowledge of the source language can correct the MT output. Secondly, who sets the quality criteria? In general, one could say that a client who needs a post-edited text establishes the quality criteria for the post-editor to follow. These criteria can be based on one of two forms of post-editing: partial (rapid, light or yet fast) post-editing and a complete (conventional or yet full) post-editing. Thus, those forms "are distinguished on the basis of the depth of intervention in the machine translation" (Krings 2001, p. 45).

In partial post-editing, as Krings (2001, p. 54) states, "only its value as a source of raw or 'gist' information remains." Therefore, "only the most egregious machine translation errors are repaired." On the contrary, in full post-editing, "[t]he machine translation serves the complete post-edit apparently only as a source of raw material for an extensive new formulation." (Krings 2001, p. 48). Then, a machine-translated text that is completely post-edited would probably look like a manually translated text, because it would have undergone a deep intervention by the post-editor.

Krings (2001) investigates many of those aspects in the post-editing process. In his study, he assesses machine translation and post-editing objectively and empirically with different groups of translators as participants. He considers the cost and the post-editing effort with access and no access to source text and uses TAPs (Think-Aloud Protocols) as a methodological tool to investigate what is going on in the subjects' minds. Among those aspects, Krings (2001) highlights that "the question of post-editing effort is the key issue in the evaluation of the practicality of machine translation systems (p. 178) and points out the obvious reason for this: "As long as fully automatic high quality machine translation remains an unreachable ideal concept, then the amount of post-editing effort will be the primary determinant

of whether machine translation is worthwhile" (p. 178). But what did Krings (2001) mean by post-editing effort?

4.2.3 Post-editing Effort

When explaining what he meant by post-editing effort, Krings (2001) introduces a fundamental distinction between three types of post-editing effort: temporal, technical and cognitive effort.

The temporal effort is the time spent on the actual post-editing task. Krings (2001) emphasizes the importance of this type of effort in post-editing when he states that "[t]he time savings that occur (or should occur) in comparison to human translation are correspondingly the most important measure for calculating the economic viability of machine translation." (Krings 2001, p. 179). The other two types of post-editing effort are also important and should be considered in the analysis of the post-editing effort. Technical effort, for example, "results from those physical operations that must be performed on a machine translation to correct any errors or defects. This includes all deletions, insertions, and rearrangements" (Krings 2001, p. 54). The third type of effort, cognitive effort, "involves the type and extent of those cognitive processes that must be activated in order to remedy a given deficiency in a machine translation" (Krings 2001, p. 54). Differently from temporal and technical effort, cognitive effort is not directly measurable. Instead, the author uses think-aloud protocol data to determine the cognitive effort required during the post-editing process.

After the publication of Krings' findings, many other empirical pieces of research on post-editing have reported further on some aspects of post-editing, mainly comparing it with human translation in terms of post-editing effort (O'Brien 2004, 2005, 2006, 2007; Carl et al. 2011; Lacruz et al. 2012). This makes sense since the practicability of post-editing needs to be evaluated by comparing it with human translation.

Based on the Krings' (2001) assumption that the post-editing effort should be studied in three dimensions, namely, temporal, technical and cognitive, O'Brien (2007) develops a study that analyses the temporal and technical effort in the post-editing process. In her study, she uses the concept of NTIs (Negative Translatability Indicators), which means "linguistic features that are known to be problematic for MT [Machine Translation]" (O'Brien 2007, n.p.). In this study, O'Brien investigates temporal and technical effort in segments of the source text which present NTIs and compares them to segments in which those indicators have been removed.

The 12 participants of O'Brien's (2007) study were professional translators at IBM. Nine of them performed a post-editing task and three of them performed a manual translation task of an excerpt from a user guide software, which was automatically translated from English into German by using IBM's Websphere MT engine for the post-editing task.

The results of O'Brien (2007) indicate that the post-editing task was performed faster than the manual translation task. Furthermore, the median processing speeds

for segments without NTI, which are measured in words per second, were higher than segments with NTI segments. These differences were statistically significant. However, the author asserts "post-editing effort can sometimes be greater for those sentences [without NTIs] than for those that contain NTIs" (O'Brien 2007, n.p.) Thus, she concludes that removing NTIs does not always lead to less temporal effort in post-editing task.

As regards the technical effort, O'Brien (2007) concludes that the segments without NTIs require less deletions and insertions than those segments with NTIs, indicating that, on average, both technical effort and the time spent on the post-editing task (temporal effort) are reduced when NTIs are absent. Therefore, it is useful to evaluate the allocation of effort in removing those NTIs before submitting to machine translation in the translation market.

Carl et al. (2011) also carried out experiments in which they compared human translation with post-editing. They focused on the time spent on those tasks, on productivity gains and on the quality of target texts rendered by seven translators who produced seven versions of three post-edited and three manually translated texts. The study analyses two kinds of post-editing effort, namely temporal and technical effort. Although the authors do not explicit use the term "technical effort" in their study, we infer it by what they call properties of the manual and the post-edited translations. Those properties include the number of characters in both post-edited translations and manually translated versions, the number of deletions, insertions, navigations keystrokes and mouse clicks. Carl et al. (2011) found out that, on average, there were more deletions, navigation keystrokes and mouse clicks in post-editing than in manual translation while fewer insertions were found in post-editing tasks. Thus, those findings imply that more technical effort is demanded in manual translation than in post-editing. In their study, Carl et al. (2011) also points out that their results indicate that translators spend less time in post-editing than in manual translation and that post-editing presents differences in gaze behaviour, i.e. the number of fixations in ST and TT windows are not as evenly distributed as in manual translation. Furthermore, both the total reading time and the fixation count on the TT proved to be significantly higher in post-editing task compared to manual translation when Carl et al. (2011) applied an unpaired two-sample *t*-test ($p < 0.01$).

As far as cognitive effort is concerned, we have seen that Krings (2001) evaluates it by analysing think-aloud data. The author assumes that this kind of data "permit the development of a number of empirical parameters for determining cognitive post-editing effort" (Krings 2001, p. 179). Therefore, Krings (2001) analyses this data assuming that the higher the verbalization effort, the higher the cognitive effort. He concludes that, among the processes he analyses, verbalization effort was considerably higher in post-editing without source text when compared with post-editing with source text. His results also indicate that *translation* and *post-editing with source text* present almost the same rate of verbalization effort, indicating similar levels of cognitive effort.

Pause analysis has been a common method to analyse cognitive effort in translation process research. When applying this methodology specifically to post-editing process, O'Brien (2006) and Lacruz et al. (2012) found interesting results.

O'Brien (2006) analyses pauses in post-editing and triangulates them with the Choice Network Analysis method and key-logged data generated with the aid of the software Translog. According to O'Brien (2006, p. 16), by applying Choice Network Analysis, it is possible to identify "those parts of a sentence that are most changed during post-editing. It is assumed that for these changes to be implemented, cognitive effort is required". Although her results suggest that analysing pauses is a useful indicator of cognitive effort in post-editing, she asserts that it is very difficult to correlate cognitive effort with pauses, source text difficulty and target text quality. This is probably due to the fact that pause duration and frequency are subject to individual differences, thus justifying the need of supplementary methods to analyse cognitive effort such as Choice Network Analysis and keyboard monitoring using Translog.

Lacruz et al. (2012) complement O'Brien's (2006) study by introducing average pause ratio as a metric to establish a relationship between pauses and cognitive effort in post-editing. Lacruz et al. (2012) assert that the average pause ratio is sensitive to the number of pauses and pause duration. The authors state that they assessed the cognitive effort required to post-edit a segment by using a measure of technical effort that counts the number of complete editing events. According to the authors:

> We classify post-edited segments as having required more or less cognitive effort on the part of the post-editor based on a metric that counts the number of *complete editing events.* In many circumstances, collections of individual editing actions can be considered to naturally form part of the same overall action, which is what we label as complete editing event (Lacruz et al. 2012, n.p.)

In the same paper, Lacruz et al. (2012) report a case study with a professional translator with 25 years of experience as a freelance translator and with no previous experience in post-editing. They classify the post-edited segments rendered by him into more or less cognitively demanding on the basis of more or fewer complete editing events in the segment. Their results indicate that the average pause ratio was higher for cognitively less demanding segments (with two or fewer complete events) than for cognitively more demanding segments (with four or more complete editing events).

In line with recent studies in translation process research that have been using eye-tracking data as part of a methodology designed to investigate cognitive effort, it is indeed possible to include metrics such as fixation duration and fixation count for such purposes.

4.3 Methodology and Procedures

4.3.1 Selection of Source Texts

Excerpts from two drugs leaflets originally published in English were selected for the present study from the EMEA corpus, a parallel corpus made out of documents

from the European Medicines Agency (EMEA).[2] One of the selected texts was about human insulin to treat diabetes (Text 1) whereas the other text was about an anticancer medicine (Text 2). Each of the two texts consisted of 20 segments to be post-edited. Care was taken in attempting to choose texts that had the same level of difficulty translating. We would also argue that, despite being specialized texts, these texts could be edited by post-editors without specific training in medical translation. The experiment is contained in the TPR-DB (Chap. 2) under the study name CEMPT13, and can be downloaded from the TPR-DB website.

4.3.2 Selection of Machine Translation Outputs

A ready-made MT engine was engineered into the CASMACAT workbench using Moses in order to generate MT outputs into Portuguese. The highest-scoring translation hypotheses produced by the system were presented to post-editors as raw MT output. The data available in the EMEA corpus for the English-Portuguese language pair was also used to generate the translation search graph delivered by the MT system under the interactive condition (see Chap. 3 for a more detailed introduction to ITP).

4.3.3 Participants

Twenty-one subjects were recruited overall, of whom sixteen produced usable data. All of the participants had Brazilian Portuguese as their L1 and English as their L2 and had received academic training on post-editing MT outputs as part of their undergraduate degree. The average age of the post-editors was 29.5 years (range 22–39).

4.3.4 Conducting the Post-editing Task

Participants received an information sheet with details regarding the general purpose of the research and the practicalities involved in taking part in the data collection. They could also read on a paper hardcopy the post-editing guidelines that they should follow to perform the task. The guidelines distributed in hardcopy were: (1) Retain as much raw MT as possible; (2) Do not introduce stylistic changes; (3) Make corrections only where absolutely necessary, i.e. correct words and phrases that are clearly wrong, inadequate or ambiguous according to Brazilian Portuguese

[2] http://opus.lingfil.uu.se/EMEA.php

grammar; (4) Make sure there are no mistranslations with regard to the English source text; (5) Publishable quality is expected.[3]

The decision to participate was voluntary and made upon awareness of the details on this sheet, including the fact they would have their eye movements and keyboard activity recorded.

Two different conditions were evaluated: (1) traditional post-editing (PE) and (2) post-editing with interactive translation prediction (ITP). Each of the two texts was shown on both conditions among the different participants. Each editor processed each text once under one of the two conditions experimental conditions at the Laboratory for Experimentation in Translation (LETRA), located at Federal University of Minas Gerais, Brazil.

Before starting to carry out their tasks, participants were introduced to the CASMACAT workbench and to the two different conditions used in the experiment. They were given time to familiarize themselves with the CASMACAT workbench. Before starting to perform both tasks, participants were asked to complete an online questionnaire with information about their translation profile (biodata, working languages and previous experience in post-editing).

4.4 Data Analysis and Results

For the purposes of this chapter, the analysis of temporal, technical and cognitive effort will focus on eye-tracking data in different areas of interest (AOIs), each AOI corresponding to one segment in the source text and its translation provided by the CASMACAT workbench (target text). The software, Tobii Studio 3.2.3, was used to create the AOIs. The following procedures were followed. First, PE and ITP post-editing sessions were edited in the Tobii Studio software, so that a set of scenes related to one segment at a time was created. Secondly, raw data related to fixation duration and fixation count on each AOI for all participants were extracted into an excel spreadsheet. Finally, statistical analysis was performed using the R software package. The cut-off point for significance level was set at 0.05.

Temporal effort was measured by the total time spent by each participant to complete the task whereas technical effort was measured by the number of keystrokes and mouse events. In order to measure cognitive effort, we have analysed both fixation duration and fixation counts.

The results were analysed based on five variables: total task time, keyboard activity, fixation count, fixation duration and TER score. An exploratory data analysis was conducted to determine whether the sample distribution was normally distributed. Results for the Shapiro-Wilk normality test (Shapiro and Wilk 1965) indicated that the distribution of only one out of the five variables, i.e., fixation

[3]Similar guidelines were also used in the other CASMACAT studies, including the EFT14 study (Chap. 3), the LS14 study (Chap. 5) and the CFT14 study (Chaps. 7 and 8).

duration, did not deviate significantly from a normal distribution either in traditional post-editing ($W = 0.9008$, $p < 0.08$) or interactive post-editing ($W = 0.9305$, $p < 0.25$).

4.4.1 Temporal Effort

Temporal effort refers to the amount of time needed to post-edit the MT output. We investigated how much time participants spent in each condition. We hypothesised that if subjects were faster in the ITP condition, then ITP could be recommended as a good practice.

Table 4.1 shows the distribution of the total time spent on traditional post-editing (PE) compared to interactive post-editing (ITP), as well as per ST character.

The overall production time spent in both conditions was recorded and extracted by using the Tobii Studio software. On average, participants spent 1,005,161 ms (16 min 75 s) on traditional post-editing (PE) whereas they spent 1,225,812 ms (20 min 43 s) when using post-editing with interactivity (ITP).

Table 4.1 Total time spent (ms) on PE and ITP and per character

Participants	Duration (ms)		Duration (ms)—normalized by ST characters		Difference between task durations (PE-ITP)	Total duration average (PE + ITP)/2
	PE	ITP	PE	ITP		
P02	1,231,219	1,280,000	976	801	−48,781	1,255,609
P03	1,083,447	1,269,000	678	1006	−185,553	1,176,223
P04	1,438,495	1,566,000	1140	979	−127,505	1,502,247
P05	2,343,441	2,327,000	1466	1845	16,441	2,335,220
P06	818,592	1,179,000	649	737	−360,408	998,796
P09	873,094	762,000	546	604	111,094	817,547
P10	863,664	1,427,000	684	892	−563,336	1,145,332
P11	968,101	897,000	605	711	71,101	932,550
P12	930,180	979,000	737	612	−48,820	954,590
P13	653,330	902,000	408	715	−248,670	777,665
P14	722,673	1,220,000	573	763	−497,327	971,336
P15	707,302	1,014,000	442	804	−306,698	860,651
P16	1,030,441	1,458,000	817	912	−427,559	1,244,220
P18	606,403	1,041,000	480	651	−434,597	823,701
P19	1,074,724	1,231,000	672	976	−156,276	1,152,862
P21	737,477	1,061,000	584	663	−323,523	899,238

Contrary to what was expected, participants did not become faster in the ITP condition. Actually, they spent significantly more time with ITP according to the results of a *Wilcoxon signed-rank Test* ($Z = 10$, $p = 0.001$).[4]

4.4.2 Technical Effort

Technical effort is gauged here by measuring the number of keystrokes and mouse events produced by each participant in both PE and ITP.

Table 4.2 provides an overview of the number of keystrokes and mouse events performed by each participant in the two tasks.

On the one hand, participants working in PE had 809 keystrokes and 76 mouse events on average. On the other hand, ITP and on-line learning techniques increased keyboard activity as participants produced 1019.50 keystrokes and 103.75 mouse events on average.

Contrary to what was expected, if we compare keyboard activity for both conditions, traditional post-editing required significantly fewer keystrokes ($Z = 31$,

Table 4.2 Keystrokes and mouse events performed on PE and ITP

Participants	Keystrokes		Mouse events		Keystrokes—normalized by characters		Mouse events—normalized by characters	
	PE	ITP	PE	ITP	PE	ITP	PE	ITP
P02	412	424	71	81	0.3267	0.2653	0.0563	0.0507
P03	642	963	80	112	0.4018	0.7637	0.0501	0.0888
P04	939	1054	101	175	0.7446	0.6596	0.0801	0.1095
P05	510	799	114	264	0.3191	0.6336	0.0713	0.2094
P06	964	776	71	84	0.7645	0.4856	0.0563	0.0526
P09	1372	839	14	40	0.8586	0.6653	0.0088	0.0317
P10	694	1624	66	53	0.5504	1.0163	0.0523	0.0332
P11	537	547	79	66	0.3360	0.4338	0.0494	0.0523
P12	1273	1065	50	33	1.0095	0.6665	0.0397	0.0207
P13	315	600	72	80	0.1971	0.4758	0.0451	0.0634
P14	952	856	97	85	0.7550	0.5357	0.0769	0.0532
P15	517	1023	65	76	0.3235	0.8113	0.0407	0.0603
P16	746	919	71	140	0.5916	0.5751	0.0563	0.0876
P18	1876	2472	47	50	1.4877	1.5469	0.0373	0.0313
P19	483	594	92	76	0.3023	0.4711	0.0576	0.0603
P21	709	1757	121	245	0.5623	1.0995	0.0960	0.1533

[4]Provided that this non-parametric test is computed with the median, such value is given for each condition (PE and ITP) as follows: PE = 901,637 and ITP = 1,199,500.

$p = 0.05$) than ITP post-editing. As for the mouse events, there was not a significant difference ($Z = 38$, $p = 0.12$) between both conditions.

4.4.3 Cognitive Effort

In order to compare the cognitive effort required for PE and ITP, fixation count (i.e. number of fixations) and fixation duration were calculated. Building on Duchowski (2007) and Jakobsen and Jensen (2008), we assume that longer fixation duration and a higher number of fixations are indicative of more cognitive effort.

Table 4.3 shows the fixation count and mean fixation duration performed by each participant under PE and ITP.

As can be seen in Table 4.3, participants had longer fixation duration during PE. This result can be considered favourable for ITP. A paired *t*-test showed that fixation duration was significantly lower for interactive post-editing: $t(15) = 8.75$, $p < 0.001$). This result could be explained by the new translation suggestions made by the CASMACAT workbench so that during ITP participants retype shorter sequences and read the emerging new translation proposals.

Opposite results were obtained when analysing fixation count. Participants working with PE had significantly fewer fixations ($Z = 0$, $p < 0.001$) than when working with the ITP setting.

Table 4.3 Fixation count and mean fixation duration for PE and ITP

Participants	Fixation duration (ms)		Fixation count		Fixation count—normalized by the number of words in the AOIs	
	PE	ITP	PE	ITP	PE	ITP
P02	413	223	1801	4233	135	288
P03	389	310	1336	2648	91	199
P04	568	257	1331	2038	100	139
P05	641	287	2031	4079	138	306
P06	440	228	977	3396	73	231
P09	429	289	1151	1905	78	143
P10	518	248	710	3533	53	241
P11	763	357	724	1783	49	134
P12	361	290	1107	2454	83	167
P13	431	287	985	1976	67	148
P14	513	273	674	2471	50	168
P15	393	243	872	2024	59	152
P16	708	357	883	3151	66	215
P18	393	263	840	2813	63	192
P19	505	287	1339	2511	91	188
P21	577	297	700	4263	52	290

4.4.4 Comparing Post-editing Effort with Translation Edit Rate

TER scores (Snover et al. 2006) were calculated using the post-edited versions as reference in order to measure the amount of editing performed on each of the segments. TER measures the minimum number of edits (insertions, deletions, substitutions or reordering) that are needed to transform the MT output into the post-edited segment. Therefore, the higher the TER score, the more edits were performed.

The following scatterplots summarize the results (Figs. 4.2 and 4.3) with respect to TER scores and fixation duration in PE and ITP.

As can be seen in Figs. 4.2 and 4.3, the scatterplots suggest there is no correlation between the two variables. This result is confirmed by a Spearman's correlation coefficient run for traditional post-editing ($r_{sp} = -0.391$, $p = 0.13$) and interactive post-editing ($r_{sp} = -0.280$, $p = 0.29$).

Higher fixation values are concentrated in two different areas which might refer to segments that demanded more post-editing effort. However, a qualitative analysis should be conducted in order to find out the reasons for such a result. Unexpectedly,

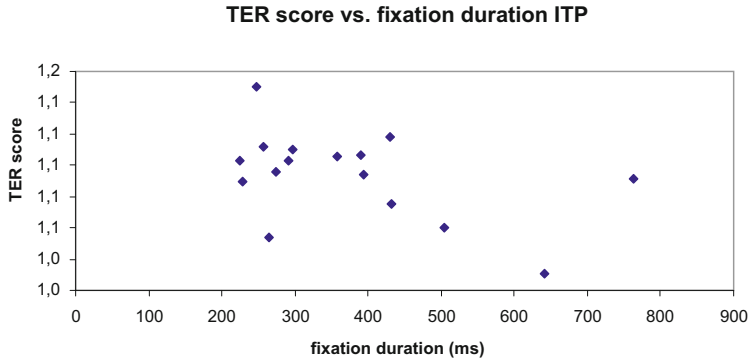


Fig. 4.2 Scatterplot correlating TER score and fixation duration for ITP

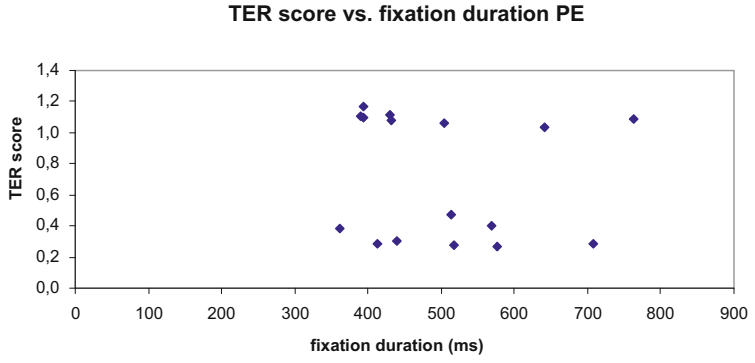


Fig. 4.3 Scatterplot correlating TER score and fixation duration for PE

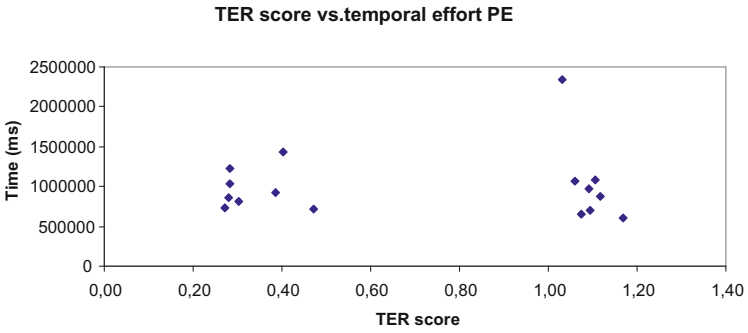


Fig. 4.4 Scatterplot correlating TER score and temporal effort in PE

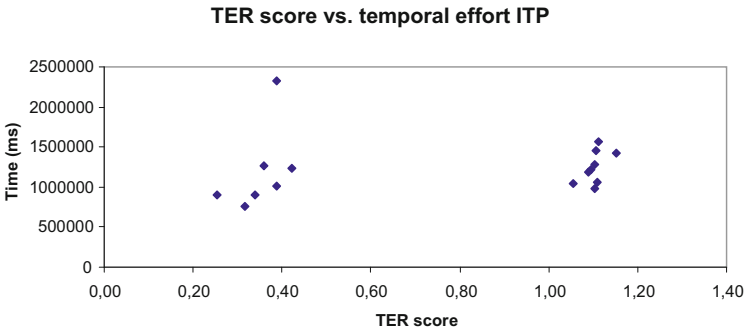


Fig. 4.5 Scatterplot correlating TER score and temporal effort in ITP

segments with TER scores ranging from approximately 0.2 to 0.4 had fixation duration as high as TER scores above 1.0. However, lower TER scores should have required fewer edits and therefore shorter fixation duration.

The following scatterplots summarize the results (Figs. 4.4 and 4.5) with respect to TER scores and temporal effort in PE and ITP.

Figure 4.4 shows that there was no correlation between TER score and temporal effort in traditional post-editing ($r_{sp} = -0.197$, $p = 0.46$). However, Spearman's correlation coefficient indicates that interactive post-editing had a positive correlation between TER scores and temporal effort ($r_{sp} = 0.570$, $p = 0.02$). Increases in time spent to complete the task were associated with higher TER scores.

4.5 Discussion

Throughout the analysis developed in this chapter, we have tested two hypotheses: (1) ITP would contribute to a decrease in post-editing effort, and (2) ITP post-editing effort would positively correlate with TER scores. By analysing temporal, technical

and cognitive effort in ITP post-editing compared to traditional PE, we were able to observe whether the interactivity reduces the effort dedicated to post-editing.

Our results show that, contrary to what was expected, participants neither became faster when post-editing in the ITP condition nor showed a reduction in the number of keystrokes when working in that condition. Interestingly, an opposite and favourable result emerges when fixations are analysed. Participants had significantly shorter fixation duration when working with ITP.

Bearing those results in mind, we could then speculate the reasons for the negative results regarding temporal and technical effort. They may be explained by the lack of familiarity of the participants with the interactive tool or their unwillingness to incorporate the auto-complete changes provided by the CASMACAT workbench. As described in the methodology section, participants had some training on post-editing but not real professional experience as post-editors. This may have had an impact on the time spent and the changes made on each task (See Chap. 5 for an in-depth analysis of the learning effect in ITP).

In addition to that, analysing the amount of editing performed on each of the segments during the task in comparison to the amount of editing predicted by TER scores provides interesting results from the industry perspective since semi-automatic translation edit rate metrics such as TER scores are used to predict quality estimates. Therefore, investigating the correlation between actual and predicted post-editing effort could bring helpful insights for establishing fair and reasonable price rates for post-editors, Language Service Providers (LSPs) and customers.

Our results show that there was no correlation for traditional post-editing but a positive correlation between TER scores and temporal effort in ITP post-editing. As expected, an increase in the time spent to complete the task was associated with higher TER scores. A correlation between the actual and the predicted effort can be considered encouraging regarding the use of semi-automatic translation edit rate metrics for establishing a threshold of "good" and "bad" MT outputs in the context of ITP post-editing.

Some of our results related to post-editing effort suggest that ITP post-editing may offer a successful path; however, since only a few post-editors participated in this data collection, the current study should be considered only as an initial exploration of interactivity on post-editing processes, particularly for English-Portuguese language pair. The small sample may also have interfered with some of the non-significant results, so it would be beneficial to conduct further experiments with a larger set of participants as well as to explore some of the qualitative data related to participants' previous training, experience and willingness to use interactive machine translation.

4.6 Concluding Remarks

As we have stated in the review of existing literature, recent studies (Koehn 2009; Plitt and Masselot 2010; Federico et al. 2012; Flournoy and Duran 2009; Green et al. 2013) suggest that post-editing is, on average, more efficient than translating from

scratch. However, we have seen that such evidence is not at all clear from the outset. Thus, in this chapter we have set about to investigate the effect of interactivity in human-machine post-editing process in English-Portuguese translation.

On the one hand, our results have shown that, contrary to what was expected, participants neither became faster when carrying out ITP-related post-editing tasks nor showed a reduction in the number of keystrokes when working in that condition. These negative results raise questions related to possible implications of these results for the industry with respect to gains related to temporal and technical effort in post-editing. They also provide food for thought concerning future research directions.

On the other hand, however, our results have also indicated that ITP-related tasks have a positive impact on cognitive effort in post-editing as shown by significantly shorter fixation duration when participants worked in the ITP condition. Another positive result relates TER scores with temporal effort in ITP post-editing.

Altogether, evidence from results related to hypotheses 1 and 2 highlights the relevance of combining a quantitative and a qualitative approach when assessing different types of effort in post-editing. This combined approach seems to grow in importance as far as cognitive effort is concerned. After all, gains for both post-editors and the industry will only be meaningful if less effort also leads to qualitatively better output.

Our small-scale results seem to indicate that this is what actually happens when one investigates effort with respect to ITP post-editing. Nevertheless, as we have stated above, one should carry out larger-scale studies to arrive at more robust and concluding evidence. Research in post-editing is somehow still in its infancy, particularly in scarcely related language pairs such as English-Portuguese. With our results, we hope to have provided elements to expand on-going research in post-editing and paved the way for further studies which may confirm some of the exploratory claims that we have made in this chapter.

Acknowledgements The work described in this chapter was carried out within the framework of the EU project CASMACAT: Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation, funded by the European Union 7th Framework Programme Project 287576 (ICT-2011.4.2). Website: http://www.casmacat.eu. Brazilian researchers were funded by CNPq, the Brazilian Research Council (grant 307964/2011-6), and FAPEMIG, the Research Agency of the State of Minas Gerais (grant SHA/PPM-00170-14).

References

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., & Vilar, J. M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics, 35*(1), 3–28.

Carl, M., Dragsted, B., Elming, J., Hardt, D., & Jakobsen, A. L. (2011).The process of post-editing: A pilot study. In B. Sharp, M. Zock, M. Carl, & A. L. Jakobsen (orgs.), *Proceedings of the 8th natural language processing and cognitive science workshop* (Copenhagen studies in language series, Vol. 41, pp. 131–142).

Casacuberta, F., Civera, J., Cubel, E., Lagarda, A. L., Lapalme, G., Macklovitch, E., & Vidal, E. (2009). Human interaction for high quality machine translation. *Communications of the ACM, 52*(10), 135–138.

Duchowski, A. (2007). *Eye tracking methodology: theory and practice*. Clemson: Springer.

Federico, M., Cattelan, A., & Trombetti, M. (2012). Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the tenth conference of the association for machine translation in the americas (AMTA)*. AMTA 2012. Retrieved October 30, 2014.

Flournoy, R., & Duran, C. (2009). Machine translation and document localization at adobe: From pilot to production. In *MT Summit XII: Proceedings of the twelfth machine translation summit*.

Green, S., Heer, J., & Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *SIGCHI conference on human factors in computing systems* (pp. 439–448). ACM.

Isabelle, P., & Church, K. (1998). Special issue on: New tools for human translators. *Machine Translation, 12*(1/2).

Jakobsen, A. L., & Jensen, K. T. H. (2008). Eye movement behaviour across four different types of reading task. *Copenhagen Studies in Language, 36*, 103–124.

Kay, M., Gawron, J. M., & Norvig, P. (1994). *Verbmobil: A translation system for face-to face dialog*. Stanford: Center for the Study of Language and Information.

Koehn, P. (2009). A process study of computer-aided translation. *Machine Translation, 23*(4), 241–263.

Krings, H. (2001). *Repairing texts: Empirical investigations of machine translation port-editing processes* (Trans. G. Koby, G. Shreve, K. Mischericow, & S. Litzar). Ohio: Kent State University Press.

Lacruz, I., Gregory, M. S., & Angelone, E. (2012). Average pause ratio as an indicator of cognitive effort in post-editing: A case study. In S. O'Brien, M. Simard, & L. Specia (Eds), *Proceedings of the AMTA 2012 workshop on post-editing technology and practice (WPTP 2012)*. Retrieved from http://amta2012.amtaweb.org/AMTA2012Files/html/2/2_paper

Langlais, P., & Lapalme, G. (2002). TransType: development-evaluation cycles to boost translator's productivity. *Machine Translation, 17*(2), 77–98.

Mesa-Lao, B. (2013). *Introduction to post-editing – The CasMaCat GUI* . Retrieved from http://bridge.cbs.dk/projects/seecat/material/hand-out_post-editing_bmesa-lao.pdf

O'Brien, S. (2004) Machine translatability and post-editing effort: How do they relate? In *Translating and the computer*. London: Aslib.

O'Brien, S. (2005). Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation, 19*, 37–58.

O'Brien, S. (2007). An empirical investigation of temporal and technical post-editing effort. *Translation and Interpreting Studies, 2*(1), 83–136.

O'Brien, S. (2006). Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Language and Cultures, 7*(1), 1–21.

Plitt, M., & Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. In The Prague bulletin of mathematical linguistics no. 93 (pp. 7–16). ISBN 978-80-904175-4-0. doi:10.2478/v10108-010-0010-x.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA-2006* (pp. 223–231).

Chapter 5
Learning Advanced Post-editing

Vicent Alabau, Michael Carl, Francisco Casacuberta, Mercedes García Martínez, Jesús González-Rubio, Bartolomé Mesa-Lao, Daniel Ortiz-Martínez, Moritz Schaeffer, and Germán Sanchis-Trilles

Abstract This chapter reports the results of a longitudinal study (LS14) in which the CASMACAT post-editing workbench was tested with interactive translation prediction (ITP). Whereas previous studies with the CASMACAT workbench (Sanchis-Trilles et al., *Machine Translation*, 2014) or similar systems (Langlais et al., *Machine Translation, 15*, 77–98, 2004) tested user interaction only for a few days, the aim of this study was primarily to find out whether and how the performance of professional post-editors improved over time when working with the CASMACAT ITP feature. We were also interested in uncovering any specific profiles of translators depending on personal factors such as previous experience in

V. Alabau (✉) • G. Sanchis-Trilles
Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, Camino de Vera s/n, 46021 Valencia, Spain

Sciling S.L., Valencia, Spain
e-mail: valabau@sciling.com; gsanchis@sciling.com

M. Carl • B. Mesa-Lao
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark
e-mail: mc.ibc@cbs.dk

D. Ortiz-Martínez • F. Casacuberta
Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, Camino de Vera s/n, 46021 Valencia, Spain
e-mail: dortiz@prhlt.upv.es; fcn@prhlt.upv.es

M. García-Martínez
Computer Laboratory, University of Maine, Le Mans, France
e-mail: mercedes.garcia_martinez@univ-lemans.fr

J. González-Rubio
Unbabel Lda., 1000-201 Lisboa, Portugal
e-mail: jesus@unbabel.com

M. Schaeffer
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

Institute for Language, Cognition and Computation University of Edinburgh, Edinburgh, UK
e-mail: mschaeff@inf.ed.ac.uk

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_5

post-editing and typing skills. Finally, the aim was also to collect feedback from the post-editors in order to know more about their views regarding this type of technology.

Keywords CASMACAT workbench • Interactive post-editing • Interactive translation prediction • Learning behavior in interactive post-editing • Production time • Typing time

5.1 Introduction

The way texts are produced changes with every technological invention. From paper and pencil to type-writers and computers, each new technology gives rise to new types of texts, new styles of authoring, and new ways of how texts are generated and perceived. Today we are experiencing increased automation of text production, in particular through the Internet and through novel forms of editing, authoring and translating digital content.

Within EU CASMACAT project (see Sanchis-Trilles et al., *Machine Translation*, 2014 and also Chap. 3 in this volume), we have developed an advanced post-editing platform with an interactive translation prediction mode, context dependent completions during the translation process (Langlais et al. 2004). Even though this feature was designed to help translators in their translation production, within a 3-days field study in a professional translation agency[1] Carl et al. (2013) it semed to hamper translators rather than help them to produce faster translations. Investigating some of the screen recordings, we hypothesized that post-editors might need to get more extended exposure to the CASMACAT workbench as its novel editing features might require completely different translation styles and translation procedures, which first would have to be learned (Sanchis-Trilles et al. 2014). This assumption is in line with experiences gained in a similar translation prediction system, TRANSTYPE (Langlais et al. 2004), where it was suggested that "over a longer period [the system] is expected to give a much better picture of its possibilities".

Accordingly, we conducted a longitudinal study (LS14) which involved five post-editors working alternatively with CASMACAT's traditional post-editing mode and the Interactive Translation Prediction (ITP) mode over a period of 6 weeks. The aim was to test whether post-editors become faster when working with ITP as they become more acquainted with this type of assistive technology, and to investigate whether exposure to this workbench over a longer period of time has an effect on editing behaviour.

[1] Field trials of the CASMACAT workbench were carried out at Celer Soluciones SL, Madrid, who were partner in the CASMACAT consortium

The LS14 study took place in May and June 2014. It was followed in July 2014 by the third CASMACAT field trial (CFT14), for which a more detailed description is contained in Chap. 7 of this volume. The CFT14 study was conducted at the same translation agency, aiming at assessing whether post-editors profit from ITP online learning as compared to traditional post-editing.[2] Seven post-editors participated in the CFT14 study from which four had also taken part in the previous longitudinal study (LS14). As a side effect, we can thus investigate what the four post-editors who participated in both studies have learned, compared to those three post-editors who only participated in the CFT13 study.

The CFT14 study differs from the LS14 study with respect to:

- the text type in LS14 was general news, while CFT14 was a specialized text from the medical domain extracted from the EMEA corpus.[3]
- The number of source text words was also quite different in these two studies: LS14 involved 24 source texts of 1000 words each, while CFT14 involved only two source text with 4500 words each (texts were much longer in CFT14, so as to test the online learning effect with tokens that occurred several times within each text).

Both studies combined add up to around 225,000 source text words which were translated into 249,000 target text words. The studies are included in the publicly available TPR-DB.[4]

Results show that LS14 participants became indeed faster over the period of 6 weeks working with the ITP system and, according to the projection of the data collected, they could have been even more productive after 6–7 weeks of regular exposure to this new technology.

A closer look at the way post-editors became acquainted with ITP suggests that learning to work with this interactive technology requires a different way of controlling the typing speed. In order to be able to fully benefit from the ITP suggestions (i.e. the translation autocompletions) provided by the system, post-editors need to check more frequently the proposals of the ITP system. Since all post-editors in the LS14 study were touch typists, they could only fully benefit from the ITP suggestions once they gradually learned to avoid overwriting new suggestions and thus saving typing effort.

Section 5.2 introduces the LS14 study. It gives background on the participants, the experimental design and the results of the study. Section 5.3 compares behavioral patterns of LS14 participants with CFT14, and tries to describe what exactly is being learned over time. Section 5.4 corroborates these findings with the feedback from participants, as acquired on the basis of questionnaires.

[2] See also Chap. 3 for a comparison of online learning and active learning in the CASMACAT tool.

[3] http://opus.lingfil.uu.se/EMEA.php.

[4] The TPR-DB is available online free of charge from: http://sourceforge.net/projects/tprdb/. The TPR-DB website is at: https://sites.google.com/site/centretranslationinnovation/tpr-db.

5.2 A Longitudinal Study with Interactive Translation Prediction (LS14)

5.2.1 Participant Profiles

Five professional translators {P01, P02, P03, P04, P05a} were recruited by Celer Soluciones SL to take part in the study. Participants were 33 years old on average (range 26–42) and all of them were regular users of computer-aided translation tools (mainly SDL Trados and WordBee) in their daily work as professional translators. All participants but one (P04) had previous experience in post-editing MT as a professional service, and all post-editors considered themselves to have excellent typing skills. For three of the four participants with post-editing experience {P01, P02, P05a}, their workload involving post-editing services did not exceeded 10 % of their projects as reported in an introductory questionnaire. The fourth participant (P03) with post-editing experience reported that 75 % of their workload as a professional translator involved post-editing projects. The five post-editors can be grouped in two groups, L_1 and L_2, as follows[5]:

- L_1: {P01, P02, P05a} are the more experienced translators/post-editors
- L_2: {P03, P04} where:
  - P03: has no formal translator training and only 1 year experience
  - P04: has 3 years formal translator training and experience, but no post-editing experience

5.2.2 Text Type

The source texts involved in this longitudinal study were pieces of general news extracted from the WMT 2014 corpus. Each source text contained 1000 words on average distributed over 48 segments on average (range 39–61).

5.2.3 Experimental Design

The experimental design involved 24 different source texts which were post-edited from English into Spanish over a period of 6 weeks (four texts per week). MT was

[5]More specific data on the participants' age, level of experience, professional education, etc., is available in the CRITT TPR Database (metadata folder).

provided by the CASMACAT server and the participants were asked to work under the following conditions:

- *Condition 1*: Traditional post-editing (P), i.e. no interaction is provided during the post-editing process.
- *Condition 2*: Interactive post-editing (PI), i.e. interaction is provided during the post-editing process in the form of ITP.

Every week, all post-editors worked on the same four source texts counterbalancing texts/conditions among participants in order to avoid any possible text/tool-order effect (two texts in condition 1 and two texts in condition 2). During the first and the last week of the study, post-editors worked from Celer Soluciones SL while their eye movements were recorded using an eye-tracker. From week 2 to week 4, post-editors worked from home as they usually do when completing jobs for the company. Meeting the participants at the company the first week was useful to make sure they understood the assignment before starting to post-edit. Post-editing guidelines were given, similar to those discussed in Chap. 3, as well as a hands-on tutorial on how ITP works from the user perspective (condition 2). During the last week of the experiment, participants returned to Celer Soluciones SL so that a second sample of their eye movements could be recorded and so that we could gather their feedback and their comments on the technology they had been using.

Each post-editor post-edited 1154 segments, i.e., in total more than 140,000 source text words (half of them in each condition, as shown in Chap. 2, Appendix A). Presentation of texts and task order were counterbalanced, such that participants post-edited in the PI condition first and post-edited in the P condition afterwards half the time. In addition, texts were grouped in two lists: two participants post-edited list A (during their weekly assignments) in condition P and post-edited list B in condition PI, while the remaining three participants post-edited list A in condition PI and post-edited list B in condition P.

5.2.4 Results

In Sect. 5.2.4.1 we provide an overall comparison of the translation durations, in terms of *FdurN*, *KdurN* and *PdurN*, which show that on average all translators slow down in the PI mode. Section 5.2.4.2 shows individual differences in post-editing behaviour: for some of the post-editors total post-editing time can be predicted by typing durations, while for other types of post-editors typing duration is less indicative of the total post-editing time (see Sect. 5.3).

5.2.4.1 Overall Post-editing Durations

The evaluation of the LS14 data is based on three different parameters computed at the segment level[6]:

1. *FdurN*: production time per segment, excluding pauses >200 s, normalised by the number of characters in the source segment.
2. *KdurN*: duration of coherent keyboard activity per segment excluding keystroke pauses >5 s, normalised by the number of characters in the source segment.
3. *PdurN*: duration of coherent keyboard activity per segment excluding keystroke pauses >1 s, normalised by the number of characters in the source segment.
4. *Mins*: number of average manual insertions per source text character
5. *Mdel*: number of average manual deletions per source text character

Table 5.1 gives an overview of average post-editing durations (in ms) and typing activities per source text character for all five post-editors in the two conditions during the 6 weeks. The data show that post-editors needed more insertions but less deletion keystrokes in the PI condition than in the P condition. On average, there are 0.416 manual insertions per source text character in the P condition and 0.538 per source text character in PI, but there are more manual deletions in P (0.371 per source text character) than in PI (0.254). A Wilcoxon-Mann-Whitney test for categorical data revealed that there were more manual insertions in PI than in P ($W = 3{,}410{,}539, p < 2.2e-16$) and more manual deletions in P than in PI ($W = 5{,}617{,}674, p < 2.2e-16$).

Table 5.1 Overall typing activity (insertions + deletions) and production times in the LS14 data

Participant	Cond	*Mins*	*Mdel*	*FdurN*	*KdurN*	*PdurN*
P01	PI	0.744	0.399	563.64	254.3	113.33
P01	P	0.595	0.545	529.71	215.86	88.51
P02	PI	0.5	0.223	456.53	173.06	68.24
P02	P	0.416	0.346	439.87	157.46	68.51
P03	PI	0.429	0.187	623.81	223.79	85.26
P03	P	0.353	0.319	573.68	167.51	63.77
P04	PI	0.569	0.280	684.30	230.22	130.28
P04	P	0.362	0.329	701.46	161.53	88.32
P05a	PI	0.447	0.181	320.72	158.18	69.99
P05a	P	0.354	0.314	284.43	138.20	54.25
AV.	PI	0.538	0.254	529.80	207.91	93.42
AV.	P	0.416	0.371	505.83	168.11	72.67
AV.	Total	0.477	0.312	517.82	188.01	83.05

[6]Further insights on this metrics are reported in Chap. 2

As shown by *KdurN* values, post-editors needed between 138.2 and 215.86 ms per character for post-editing (P) while it took them on average between 158.18 and 254.30 ms in the PI mode. Duration values for *FdurN* and *PdurN* show a similar pattern.

5.2.4.2 Individual Differences in P and PI

Figure 5.1 plots the relationship between *FdurN* and *KdurN* for the five participants in the LS14 study. Each point in the graph shows the average *KdurN*/*FdurN* ratio per source text and post-editing condition over 1 week of post-editing activity. Each dot represents the average per-character post-editing duration of approximately 2000 source text words per week and per condition in either of the two post-editing modes (P or PI). That is, each post-editor is represented with six dots representing 6 weeks for the P task and six dots for the performance of the 6 weeks in the PI task.

Different post-editors show different *KdurN*/*FdurN* relations: Experienced post-editors form group L_1 show a strong correlation between these two durations ({P01: R = 0.78, P02: R = 0.78, P05a: R = 0.82}), which is not the case for the less experienced translators in L_2: {P03: R = 0.74, P04: R = 0.40}). This suggests that experienced post-editors use their time more efficiently and predicably while they work on a segment. Despite being a professional translator, post-editor P04 showed a much weaker correlation between *KdurN* and *FdurN* (R = 0.40) than the other post-editors, probably related to the fact that he had no previous post-editing experience. P03—the only one without formal training despite working as a freelance translator for Celer Soluciones SL—showed a slightly weaker correlation between these two measures (R = 0.74).

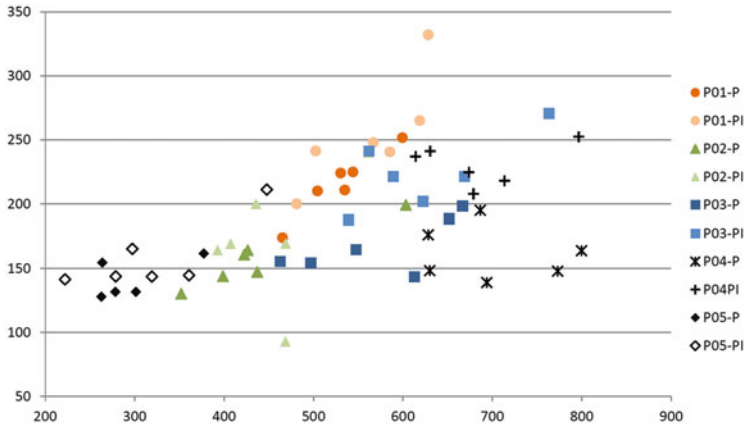


Fig. 5.1 LS14 study—*FdurN* (*horizontal*) vs. *KdurN* (*vertical*) for all five participants

5.2.4.3 Projecting Learning in P and PI

While the five post-editors show different behaviour, they become substantially quicker in the PI condition over time. However, in the baseline condition (P), there was no improvement over time. Figure 5.2 plots the effect of time on post-editing durations measured in terms of *Kdur* per source text character (i.e. *KdurN*) for the two CASMACAT settings. For this analysis, skipped segments with either zero tokens in the final target text and/or with zero total editing duration and segments with more than one revision were excluded. Segments with more than one revision were excluded, because participants complained that often when a segment was revisited, the initial MT output rather than the already corrected text appeared, which meant that translators had to edit text which they had already corrected. In total, 12 % of the data was excluded.

Despite the general downwards trend in PI over time, Fig. 5.2 shows a difference in efficiency in week 1 and 6 as compared to the other weeks. The reason for these peaks in production time for weeks 1 and 6 might be the experimental setup itself, since these 2 weeks involved eye-tracking apparatus and the request to post-edit from the company:

1. Having to work from the company office, rather than from home, seems to have had a negative impact on post-editor's performance. During weeks 2–5, post-editors worked from home, which is what they are used to, since all of them work as freelancers.

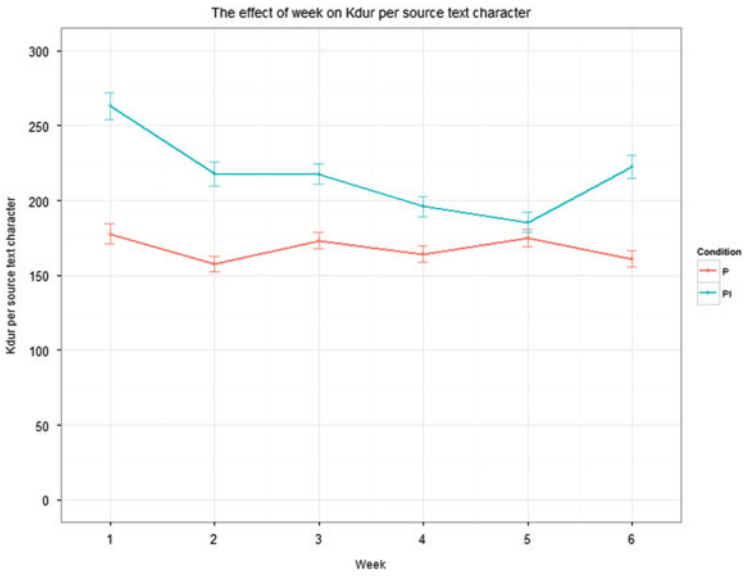


Fig. 5.2 LS14 study—productivity as reflected in *KdurN* taking into account 6 weeks

2. The ITP mode involves a great amount of internet traffic: a new translation prediction is sent over the internet for (almost) every keystroke. This adds to the traffic from the gaze samples (at a rate of approximately 300 Hz), which are also sent over the internet to a remote server, so that a delay in response was frequently observed in the office of the translation agency when using CASMACAT in the PI setting

In addition to this, using an eye-tracker involved limited head movement and sometimes recalibration during the process of post-editing was necessary. Together, these aspects may have had a negative effect on participants' productivity in weeks 1 and 6, or—in other words—the data might show a lab effect.

The productivity drop for week 6 under PI can also be found in the difficulty of the texts themselves: TER values were computed for all the texts in LS14, and values were particularly higher for texts in week 6. We could identify text 20 in week 6 (post-edited under PI by participants P01, P03 and P05) as one of the most difficult texts to post-edit. Text 20 in LS14 was of a more specialized nature of legal text. This different degree in text specialization could be the reason for both lower MT quality and thus requiring more edits from the post-editors, as reflected in the higher number of edits recorded in TER values.

Assuming that working at home and working in the office are two different conditions, we calculated a learning projection based only on the 4 weeks when post-editors worked in the office. Figure 5.3 plots the two conditions in

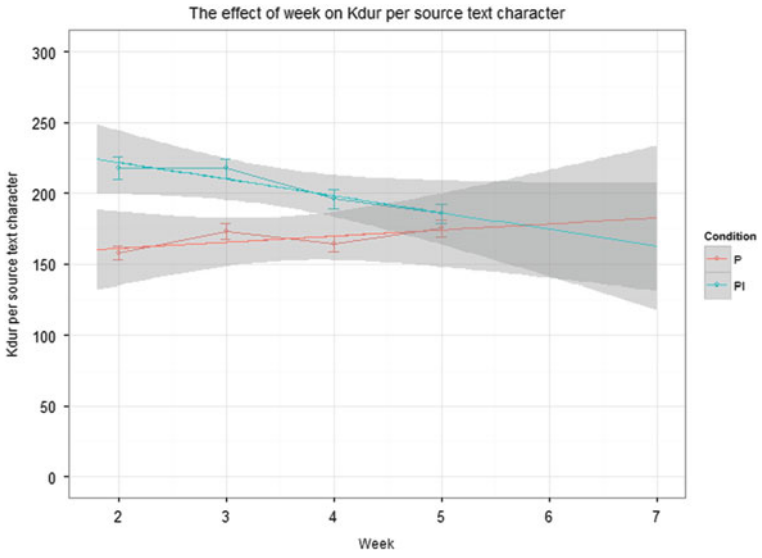


Fig. 5.3 LS14 study—productivity projection as reflected in Kdur based only on the data from weeks 2–5 (working from home)

LS14 showing that post-editing under the PI condition could have become—theoretically—more efficient already after 6 weeks.

The two regression lines in Fig. 5.3 are based on simple linear models and show the projection of the average post-editing time under the PI and the P conditions over a hypothetical timeframe of 7 weeks. The grey areas around the linear regression lines represent the 95 % confidence region for each regression. According to this projection, it is between weeks 6 and 7 that post-editors would become more efficient under the PI condition than under the P condition. While this is a hypothetical assumption, assuming a linear relationship between time spent working on the CASMACAT workbench and *Kdur*, this projection clearly shows a learning effect for the PI condition, which is absent in the P condition.

5.3 What is Learned During ITP

Singla et al. (2013) have shown that post-editor profiles can be detected automatically and single post-editors can be identified with a certain degree of accuracy on the basis of process data. They create n-gram models based on activity microunits,[7] as well as part-of-speech sequences to automatically cluster post-editors. Discriminative classifier models are used to characterize post-editors based on a diverse range of translation process features. Singla et al. (2013) conclude that classification and clustering of participants could be used to develop translation tool features which are customized to best fit an individual's behaviour.

However, as shown above, when working with the ITP system over a longer period of time, post-editors seem to change and adapt their behaviour, which indicates that translator profiles do not only refer to a static set of properties but a translator's profile can tell us also something about how the individual learns and adapts to new situations.

In this section we assess what it is that post-editors have learned in the 6 weeks during which they were working with the CASMACAT ITP mode. We compare the behaviour of the post-editors involved in the LS14 study with that in the subsequent CFT14 field trial.[8] We briefly introduce the participants in the CFT14 field trial, outline the differences of the texts used in LS14 and CFT14, and highlight the learning effects by comparing the two studies.

[7] Activity units are presented and discussed in Chap. 2. For an alternative approach to define activity microunits, see also Chaps. 8 and 14 in this volume.

[8] For more detailed information on the CFT14 data see Chap. 7.

5.3.1 Participants in LS14 and CFT14

Seven post-editors contributed to the CFT14 field trial. These can be separated into two groups: C_1:{P01,...,P04} are the four post-editors which previously participated in the LS14 study. In addition there was a group of three new post-editors C_2:{P05, P06, P07}, which had no experiences with the CASMACAT PE and ITP modes. This makes it interesting to investigate how the behaviour of the four C_1 post-editors who worked in both studies is different from the new C_2 post-editors.

Table 5.2 shows a general overview of the participants' profiles involved in both studies. The most salient factors in the metadata collected for subject profiling are:

1. P04 did not have previous post-editing experience
2. P03 did not have formal translator training and was less experienced, despite being a regular freelance translator for Celer Soluciones SL
3. P05 had much more experience as a professional translator than the rest
4. P05a did not participate in the CFT14 field trial

Note that we make a distinction between P05a and P05 in this table to differentiate between two different post-editors who were not simultaneously in LS14 and CFT14 and had the same participant number.

5.3.2 Texts in LS14 and CFT14

There were a few differences in the LS14 and the CFT14 studies.

1. For LS14, the goal was to compare the CASMACAT ITP and post-editing (P) modes, while CFT14 aimed at comparing post-editing (P) and ITP with online learning (PIO). A detailed description of the differences is contained in Chap. 3.

Table 5.2 Information about participants in the LS14 and CFT14 studies

Participants	P01	P02	P03	P04	P05a	P05	P06	P07
Gender	F	M	F	F	M	F	F	M
Years of translator training	4	4	0	3	14	5	4	4
Years of professional experience	8	8	1	3	14	27	3	11
Post-editing experience	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Took part in LS14 study	Yes	Yes	Yes	Yes	Yes	No	No	No
Took part in the CFT14 study	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes

Table 5.3 Comparing properties of EMEA corpus translations and the news translations

Study	TType	*HTra*	*CrossS*	*CrossT*	*SLen*	*TLen*
LS14	News	0.612	1.60	1.29	25.0	27.85
CFT14	EMEA	0.445	1.44	1.23	21.0	22.93

2. In order to appreciate online learning capacities, texts were much longer in CFT14 than in LS14, but there were only two texts for each translator, one to be post-edited in the P mode and the other in the PIO mode.[9]
3. Whereas the LS14 data is based on an English-to-Spanish news text, the CFT14 study used a medical text extracted from the EMEA corpus.

As shown in Table 5.3, segments on the source side (*SLen*) as well as on the target side (*TLen*) are on average shorter in the medical text than in the news text. The medical text has also a lower translation ambiguity, as indicated by the lower average word translation entropy *HTra*. This is likely due to dense terminology in the medical text, and the reduced choices for medical and chemical term translations, as compared to expressions in the news text. EMEA translations are also syntactically closer to the source text: lower *CrossS* and *CrossT* values indicate greater syntactic similarity between the source and the target language.[10] In summary, translations of the medical text tend to be more literal than news text translation.[11]

Despite the different nature of these texts, it can be expected that the experience with the ITP post-editing mode that C_1 translators obtained during the 6 weeks of the LS14 experiment would also carry over to the CFT14 study, while it is likely that the fresh translators in the group C_2 who do not have this experience thus show different behaviour.

5.3.3 Typing and Translation Times

The aim of ITP is to reduce the relative time spent on mechanical translation production (i.e. typing). Taking into account individual differences in typing speed, orientation and translation times, we measure the desirable learning effects as the ratio of coherent keystroke activities (*Kdur*) and the filtered total production duration (*Fdur*): *KdurN* indicates the amount of coherent typing activity, while *Fdur* is the overall translation time, so that the ratio *KdurN*/*FdurN* indicates the relative proportion used for typing, the amount of which we want to reduce with the ITP mode.

[9] A comparison of the PIO mode and active learning is discussed in Chap. 3, this volume.

[10] See Chap. 2 in this volume for more details on these metrics.

[11] See also Chaps. 9 and 13 in this volume for a discussion on word translation literality, and how the *Cross* and the *HTra* features are indicators for this end.

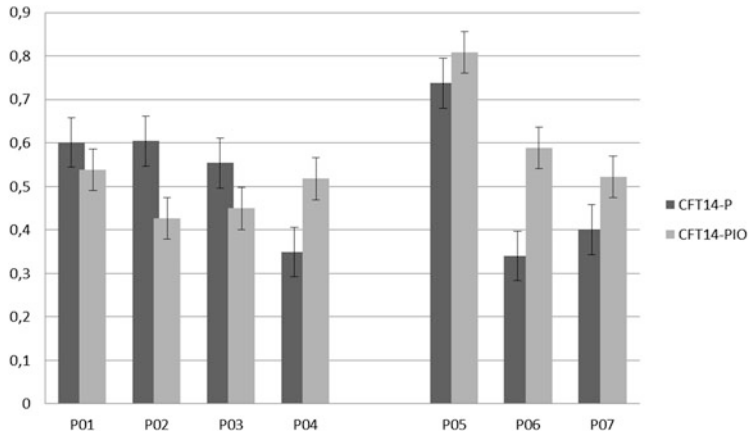


Fig. 5.4 Ratio of typing time (*Kdur*) and production time (*Fdur*) for C_1 and C_2 translators

We take the traditional post-editing mode (P) as a baseline and compare the differences in relative typing effort for the ITP mode across different post-editors, and the two groups C_1 and C_2 with and without extended exposure to CASMACAT. We find that this measure provide another suited indicator for translator profiles, and to capture the learning IPT effects.

Figure 5.4 shows that most of the post-editors in the C_1 group ({P01, P02, P03}) have a lower proportion of coherent keystroke activities (*Kdur/Fdur*) in the PIO mode than in the P mode. That is, in the interactive ITP mode these post-editors seem to have learned to accept interactive suggestions which reduces the amount of their coherent typing time, which is not the case for the translators in the C_2 group. C_1 translators seemed to accept the interactive translation suggestions more often than the new C_2 translators by less frequently overwriting the ITP proposals.

Post-editor P04 is an exception in the C_1 group, which might be explained by the fact that she did not have any prior experience with post-editing MT output and performed already in the most unpredictable way during the LS14 study (see Fig. 5.1). P05 has the highest *KdurN*/*FdurN* ratio, indicating her ability to make use of her time in the most productive way. Comparing the performance patterns in Fig. 5.4 and taking into account that P04 is (one of) the least experienced translators, while P05 is the most experienced one suggest that the *KdurN*/*FdurN* measure captures some important features.

All post-editors self-rated their typing skills as excellent in an introductory questionnaire and, indeed, their typing speed caused many cases of overwriting behaviour as they continued typing even though the right suggestions by the ITP system were already pasted in the target text. Learning to control this overwriting behaviour was also reported by the post-editors themselves when providing user feedback, as reported in the next section.

5.4 Eliciting User Feedback

User's feedback was collected with a questionnaire that post-editors completed at the end of both studies and in which, apart from answering the questions, participants could also make further comments.[12]

The user feedback derived from the longitudinal study was collected in week 6 right after post-editing the last text in the study. Post-editors had to answer the following five questions:

1. If Celer Soluciones SL (or any other LSP) ever gave you the chance to post-edit with or without interactivity, what would you prefer?
2. In your daily work as a professional translator, do you prefer to translate from scratch instead of post-editing machine translation?
3. Would you use CASMACAT as a post-editing tool for your future projects?
4. According to your own personal opinion, what are the advantages of using interactivity while post-editing MT?
5. According to your own personal opinion, what are the disadvantages of using interactivity while post-editing MT?

The aim of the first question was to know if, after having post-edited using interactivity over an extended period of time, participants would choose ITP over a "traditional" form of post-editing. All participants, except P03, stated that they would still prefer to post-edit without interactivity. Interestingly, P03 was the only one without formal translator training and with less than 2 years of translation experience. She suggested that ITP becomes an effective way to retrieve equivalents as you type ("ITP helped me to find equivalents").

When trying to find out more about the resistance to adopt ITP for post-editing purposes, in the open section of the questionnaire both P01 and P02 provided feedback along these lines:

> having to post-edited with interactivity demands a controlled typing speed and this is difficult to achieve when you are an experienced touch typist.

Advanced touch typists need to be aware of the fact that they will only benefit from ITP when they stop overwriting most of the suggestions offered by the system. As was also visible in the collected screen recordings, P01 and P02 are the two participants with more cases of overwriting behaviour due to their fast typing speed.

With respect to the second question, four out of the five post-editors in LS14 answered "It depends (on the text type, quality of the machine translation, etc.)". P02 was the only one who would always prefer to translate instead of post-edit.

The third question in the questionnaire wanted to explore how likely it was that translators would adopt the CASMACAT workbench as a professional tool. P02 and P05a were the only ones who would not use the workbench for further post-editing

[12] The questionnaire used to collect the user feedback presented in this section is available at this *introductory questionnaire*.

projects claiming that existing commercial CAT tools already serve this purpose. P01, P03 and P04 stated that they would adopt this workbench for post-editing purposes in the future.

When asked about the benefits of ITP, the responses collected were diverse: P05a stated that he was not able to mention any advantages and P02 argued that he rarely benefited from the suggestions provided by the system. The rest of the participants offered a more positive view on ITP, acknowledging, for instance, that the idea behind ITP certainly helps to decrease the technical effort (typing). However, they would have to invest more time in order to increase productivity using this novel workbench by learning not to overwrite many of the ITP suggestions. In line with this finding, P01 mentioned "I have to retrain myself on typing for ITP purposes".

With respect to the disadvantages of ITP, all participants (except P03) mentioned that it is difficult to become familiar with the fact that the target text is constantly changing. It is difficult to pay attention to the source text, the target text and, in addition, to all the suggestions triggered by the ITP. In addition, P02 suggested that another area of the screen could be used to show these predictions—similar to how translation memory matches are shown in a separate window.

The feedback collected seemed to offer a clear cut difference between the extremely positive attitude towards ITP shown by P03 (the only one without translator training and less experience) and the negative views offered by P05a (the participant with most years of formal training and many years of experience). These two extremes in terms of experience and formal training certainly played a decisive role for ITP acceptance.

5.5 Discussion

The aim of this study was to explore the benefits of working with interactive machine translation combined with online learning techniques for post-editing purposes. Results from the LS14 study showed how professional translators needed an average of 6 weeks (see Fig. 5.3) to become familiar with interactivity features for post-editing purposes. The crucial factor in order to obtain a successful interaction between the post-editor and the ITP featured in CASMACAT is directly related to their typing behaviour. Only after post-editors stop overwriting most of the suggestions provided by the system can productivity gains be reached by using ITP. Touch typists find this trade-off between typing speed and the suggestions provided by the system somehow difficult to achieve. This study shows that after weeks of interaction, a successful interaction can be achieved. It would be interesting to conduct further studies to explore if non-touch typists or non-professional translators with a slower keyboard activity, become more easily acquainted with this technology within a shorter timespan.

Most of the participants reported that they would prefer to work without interactivity but with online learning, a technique which is described in more detail in Chaps. 3 and 7 in this volume.

Acknowledgements The work described in this chapter was carried out under the auspices of the EU project CASMACAT: Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation, supported by the European Union 7th Framework Programme Project 287576 (ICT-2011.4.2). Website: http://www.casmacat.eu.

References

Carl, M., Martinez, M.G., Mesa-Lao, B., Underwood, N., Keller, F., & Hill, R. (2013). *CASMACAT project tech report: D1.2: Progress report on user interface studies, cognitive and user modeling*. European Commission.

Langlais, P., Lapalme, G., & Loranger, M. (2004). Transtype: Development evaluation cycles to boost translators productivity. *Machine Translation, 15*, 77–98.

Sanchis-Trilles, G., Alabau, V., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., et al. (2014). Interactive translation prediction versus conventional post-editing in practice: A study with the CasMaCat workbench. *Machine Translation, 28*(3–4), 217–235.

Singla, K., Carmona, David O., Gonzales, A., Carl, M., & Bangalore, S. (2013). Predicting post-editor profiles from the translation process. In *Proceedings of the Workshop on Interactive and Adaptive Machine Translation, AMTA Workshop*, Vancouver, Canada (pp. 51–60).

Chapter 6
The Effectiveness of Consulting External Resources During Translation and Post-editing of General Text Types

Joke Daems, Michael Carl, Sonia Vandepitte, Robert Hartsuiker, and Lieve Macken

Abstract Consulting external resources is an important aspect of the translation process. Whereas most previous studies were limited to screen capture software to analyze the usage of external resources, we present a more convenient way to capture this data, by combining the functionalities of CASMACAT with those of Inputlog, two state-of-the-art logging tools. We used this data to compare the types of resources used and the time spent in external resources for 40 from-scratch translation sessions (HT) and 40 post-editing (PE) sessions of 10 master's students of translation (from English into Dutch). We took a closer look at the effect of the usage of external resources on productivity and quality of the final product. The types of resources consulted were comparable for HT and PE, but more time was spent in external resources when translating. Though search strategies seemed to be more successful when translating than when post-editing, the quality of the final product was comparable, and post-editing was faster than regular translation.

Keywords Translation • Post-editing • External resources • Translation process • Translation quality

J. Daems (✉) • S. Vandepitte • L. Macken
Department of Translation, Interpreting and Communication, Ghent University, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
e-mail: joke.daems@ugent.be

M. Carl
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

R. Hartsuiker
Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*, New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_6

6.1 Introduction

With the increasing need for faster and cheaper translations due to the increasing amount of text to be translated, computer-aided translation has become more and more widespread. While correcting machine translation output by means of post-editing is now a relatively common task for translators, professional translators are still reluctant to do it, and it is still not clear exactly how regular translation differs from post-editing.

A better understanding of the differences between human translation and post-editing can improve the field of translation in numerous ways. On the one hand, the knowledge can be used to improve translation tools to better aid translators with their work, by indicating in which cases a translator should be allowed to work from scratch, or in which cases he can benefit from the presence of machine translation output. On the other hand, insight in these differences can help understand the reluctance of professional translators to post-edit and can help colleges and universities to teach translation students the appropriate skill sets required for the increasingly technological translation work. Recent studies indicate that certain types of college students would make decent post-editors (Yamada 2015).

In this chapter, we focus on the usage of external resources by student translators translating and post-editing newspaper articles from English into Dutch. For both types of activity, we compare the number and type of resources consulted. We also investigate whether consulting different types of resources and spending more or less time consulting external resources leads to a decrease or increase in productivity and/or quality of the final product.

6.2 Related Work

The field of translation process research is rapidly evolving. Where, originally, rather intrusive methods such as think aloud protocols (TAP) had to be used in order to study the translation process, new tools such as keystroke logging tools and eye-trackers have helped researchers gather data in more ecologically valid ways. The Translation Process Research Database (TPR-DB), which contains over 1300 translation and post-editing sessions, is one example of advanced data collection in the field (see Chap. 2 in this volume). Originally containing Translog data (Jakobsen and Schou 1999; Carl 2012), the TPR-DB has since been enriched with data from CASMACAT (Alabau et al. 2013, and Chap. 3), a state-of-the-art workbench for translation and post-editing, with added keystroke logging capacities.

Yet some aspects of the translation process remain elusive even with these advanced tools. The usage of external online resources, for example, which can provide insights into translators' problem-solving strategies (Göpferich 2010) or uncertainty management (Angelone 2010), is not so easily analyzed. For regular

translation, search queries can be related to source text meaning, meaning transfer or target text production. For post-editing, however, the machine translation output comes into play as well. Whereas the presence of this MT output is intended to facilitate and speed up the translation process, professional translators seem to benefit less from post-editing than translation trainees (Garcia 2011). This could be caused by insecurity about the quality of the MT output, which leads to a higher number of consulted resources, which could, in turn, negatively affect productivity. A better understanding of the usage of external resources during translation and post-editing is needed to obtain a more profound insight into successful problem-solving strategies with regard to quality and productivity.

External resources are usually registered by means of screen capture software such as Camtasia Studio (Göpferich 2010). The drawback of this software, however, is the fact that the data still needs to be replayed and manually encoded for automatic analysis, which can be quite time-consuming. TAP can provide some idea of the resources consulted, but participants' utterances are often incomplete and researchers still need to look at the screen recordings in parallel to make sense of their data (Ehrensberger-Dow and Perrin 2009). Some previous research has made use of data gathered with the TransSearch tool to get a better insight in translators' queries (Macklovitch et al. 2008), but they are limited to one type of resource (TransSearch) and don't take other types of resources into account. The present study attempts to solve these issues by introducing a new method for the analysis of external resources by means of Inputlog (Leijten and Van Waes 2013), a keystroke logging tool originally intended for writing research, which logs all Windows-based applications. In a recent study, Inputlog has been used to analyze the external resources used by a professional communication designer when creating a proposal (Leijten et al. 2014). To the best of our knowledge, Inputlog's logging of external resources has not been used for translation research before the present study. We've opted for a combination of CASMACAT and Inputlog to be able to fully grasp the translation process with external resources. As described in Chap. 2, Sect. 2.7.1, an extra table for the TPR-DB can be created, which accommodates the Inputlog data and allows for a more thorough analysis of external resources, adding an extra layer to the translation process research options the TPR-DB currently provides.

6.3 Methodology

6.3.1 Participants

Participants were ten master's students of translation, who had passed their English General Translation exam. Eight participants were female, two participants were male, and ages ranged from 21 years old to 25 years old. Two participants wore contact lenses and one participant wore glasses, yet the calibration with

the eyetracker was successful for all three participants. Students had no previous experience in post-editing. To prevent exhaustion effects, each session was spread over two half days on different days. Participants received a gift voucher of 50 euros for each half-day session, amounting to 100 euros per participant.

6.3.2 Text Selection

We tried to control for text difficulty as much as possible, as we are mainly interested in investigating differences between post-editing and human translation, and wanted to exclude other potential influential factors. A number of newspaper articles were selected from Newsela,[1] a website which offers newspaper articles at various reading levels, originally intended for use in the classroom. What makes this site so useful is the fact that texts are not just ranked according to existing readability metrics, but that context and the difficulty of a topic is taken into account as well. We selected articles from different topics with the highest possible Lexile® levels (between 1160 L and 1190 L[2]), and selected 150–160 words from each article as potential texts. Lexile® measures are a scientifically established standard for text complexity and comprehension levels, giving a more accurate representation of how challenging a text is than existing readability measures. The scores are usually used in classrooms to provide students with texts of their appropriate reading levels. Our study is—to the best of our knowledge—the first one to apply these measures for translation research. As additional control measures, we then manually compared the texts for readability, potential translation problems and machine translation quality. Texts with on average less than fifteen or more than twenty words per sentence were discarded, as well as texts that contained too many or too few complex compounds, idiomatic expressions, infrequent words or polysemous words. The machine translation was taken from Google Translate, and annotated with our two-step Translation Quality Assessment approach (Daems et al. 2013). We discarded the texts that would be too problematic, or not problematic enough, for post-editors, based on the number of structural grammatical problems, lexical issues, logical problems and mistranslated polysemous words. The final corpus consisted of eight newspaper articles of 150–160 words long, each consisting of 7–10 sentences.

[1] newsela.com

[2] The authors would like to thank MetaMetrics® for their permission to publish Lexile scores in the present chapter. https://www.metametricsinc.com/lexile-framework-reading

6.3.3 Experimental Setup

Each participant translated four texts and post-edited four different texts. To counter fatigue effects, the tasks were performed in two sessions, with two translation and two post-editing tasks in each session. We used a Latin square design to eliminate task order effects, as can be seen in Table 6.1. Across all participants, each text was translated five times and post-edited five times.

We used a combination of logging tools to be able to analyze the translation and post-editing process in detail. Whereas think-aloud protocols (TAP) are often used to elicit problem-solving strategies and other steps in the translation process (Angelone 2010; Ehrensberger-Dow and Perrin 2009), they have been shown to influence the translation process itself (Jakobsen 2003; Krings 2001). We therefore opted to use keystroke logging tools, which are capable of logging the process without interfering with it. The first tool is CASMACAT (Alabau et al. 2013, Chap. 3, this volume), a translator's workbench which doubles as a keystroke logging tool. Unlike other keystroke logging tools, it has the functionality and interface of an actual translator's workbench, allowing for a more realistic experimental setup. In this study, we used a simplified version of CASMACAT, without interactive translation. Another reason for selecting CASMACAT was the fact that it is compatible with the EyeLink2000 eye-tracker. We collected the gaze data with the EyeLink2000 to add an extra layer of information to our other data. Though we will not report on gaze data in the present chapter, it must be noted that a chinrest was used to gather the gaze data, which limited participants' movements, and which could have some effect on our results. In addition to CASMACAT, we also used the keystroke logging tool Inputlog (Leijten and Van Waes 2013). Though Inputlog was originally intended for writing research within the Microsoft Word environment, its capability to log all applications and browser tab information enables us to extract information on the usage of external resources. As CASMACAT only logs what happens within the CASMACAT interface, we needed to add Inputlog to our tool set to analyze the entire translation process, including the usage of external resources.

Table 6.1 Latin square design, mixed text order and task order

Participant		P1	P3	P5	P7	P9	P2	P4	P6	P8	P10
Session1	task1	PE_1	PE_8	PE_7	PE_6	PE_5	HT_1	HT_8	HT_7	HT_6	HT_5
	task2	PE_2	PE_1	PE_8	PE_7	PE_6	HT_2	HT_1	HT_8	HT_7	HT_6
	task3	HT_3	HT_2	HT_1	HT_8	HT_7	PE_3	PE_2	PE_1	PE_8	PE_7
	task4	HT_4	HT_3	HT_2	HT_1	HT_8	PE_4	PE_3	PE_2	PE_1	PE_8
Session2	task5	HT_5	HT_4	HT_3	HT_2	HT_1	PE_5	PE_4	PE_3	PE_2	PE_1
	task6	HT_6	HT_5	HT_4	HT_3	HT_2	PE_6	PE_5	PE_4	PE_3	PE_2
	task7	PE_7	PE_6	PE_5	PE_4	PE_3	HT_7	HT_6	HT_5	HT_4	HT_3
	task8	PE_8	PE_7	PE_6	PE_5	PE_4	HT_8	HT_7	HT_6	HT_5	HT_4

Columns are labeled with participant codes (ranging from P1 to P10), cells contain codes for the task type (*PE* post-editing, *HT* human translation) and text (ranging from 1 to 8)

The first session consisted of the following steps: first, participants filled out an introductory survey, asking them about their experience with an attitude towards post-editing; second, they performed the LexTALE test (Lemhöfer and Broersma 2012) to be able to measure their English proficiency; third, they copied a text of 150 words, so that they could get used to the keyboard and the chin rest of the eye-tracker; fourth, they translated a text in the CASMACAT interface, consisting of four segments that were post-edited and four segments that were translated manually, to get them acquainted with the tool and task; and finally, participants translated two texts and post-edited two texts. For both types of task, the students were instructed to make sure the final product was of publishable quality. Each segment in the CASMACAT interface contained one sentence.

The second session started with another warm up task within CASMACAT, consisting of four segments to be post-edited and four segments to be translated manually, followed by the actual tasks: two texts to be translated manually and two texts to be post-edited. After these tasks, participants had to look at the texts again and highlight the most problematic passages for one translation task and one post-editing task. They were asked to add comments to these passages in a Word document. At the end of the session, participants had to fill out another survey, asking them about their experience and their attitude towards post-editing.

6.4 Analysis

The final dataset consisted of CASMACAT and Inputlog data (xml-files) for all 80 sessions. Using the scripts provided with the TPR-DB, the CASMACAT xml-files were prepared for word alignment. A first, automatic, alignment was done with Giza++ (Och and Ney 2003), which we then manually corrected with the YAWAT tool (Germann 2008). Data from the aligned files was extracted and converted to more manageable table formats with another TPR-DB script (see Chap. 2). From the Inputlog data, we extracted the focus events with the provided software (focus events contain information on the opened application or screen, time spent in the application, and keystrokes). We then manually grouped the different events into categories: dictionary, web search, concordancer, forum, news website, encyclopedia, etc. Figure 6.1 shows an overview of the most common categories for human translation and post-editing. As can be seen, most types of external resources are only sporadically used, with the exception of search engines, concordancers, dictionaries, and encyclopedias. We therefore limit ourselves to these four categories for further analysis, and group the other external resources together in a generic category ‘other’.

A next step was to combine the CASMACAT and Inputlog data for subsequent analysis. Since this is the first study where data from both tools are combined, the TPR-DB had to be updated to accommodate for the new data. An InjectIDFX-script was developed to merge Inputlog data with the CASMACAT xml-files. CASMACAT only logs the keystrokes and events within the CASMACAT interface.

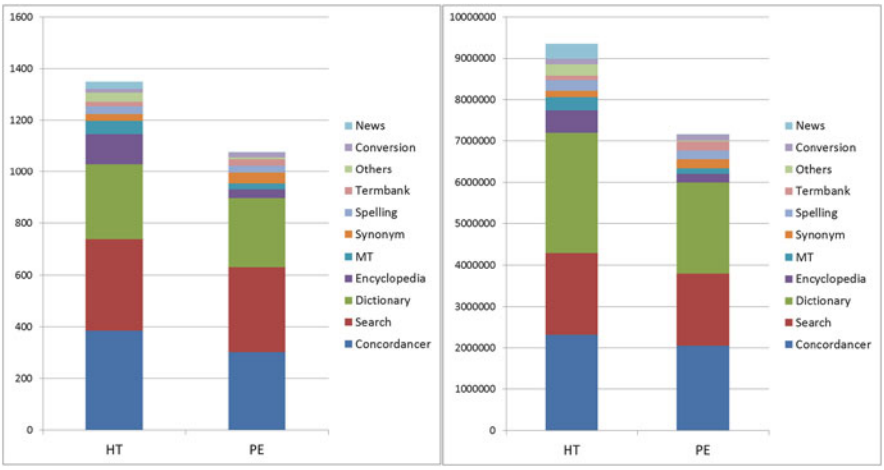


Fig. 6.1 General overview of resource types used in and human translation (HT) and post-editing (PE), expressed in total number of resource hits (*left*) and total duration (*right*) over all 80 sessions

The xml-files themselves contain a 'blur'-event whenever a person leaves the CASMACAT interface and a 'focus'-event whenever they return to the CASMACAT interface, but whatever happens between the blur and the focus-event is unknown. By adding the Inputlog data to the xml-files, we can analyze what happens when a person leaves the CASMACAT interface as well. We added an extra table: the EX-table, containing information on external resources consulted, the time spent in the resource, and keystrokes made within the external resource. We added an extra column to the EX-file where we added the categories we had assigned to the various Inputlog events. An extract from an EX-file can be seen in Table 6.2 below.

Looking at the 'Focus' column and corresponding category label in Table 6.2, we see the participant moving from the main document (CASMACAT, EXid 3) to a new tab in Google Chrome (EXid 4), where he types 'woorden...' (see 'edit'), leading him to the Dutch spelling website 'Woordenlijst' (EXid 5). He then types 'groot-bri' to look up the Dutch spelling of Britain (Groot-Brittannië). After this search, he returns to the CASMACAT interface (EXid 6) for 2 min, after which he again opens a new tab in Google Chrome (EXid 7) for the next search: 'linguee', allowing him to go to the Linguee concordancer (EXid 8), where he looks up the translation of 'in fact' (EXid 9) before returning to the CASMACAT document once more (EXid 10).

It is currently impossible to automatically map external resources to the correct segment. In the data file, there is a column for the last segment that was open before the CASMACAT interface was left, and the first segment to be opened after returning to the CASMACAT interface, but the search itself could be related to either one, or even an entirely different segment. For example, a person can look up a word in a dictionary while translating the first segment of a text. If the person goes back to the CASMACAT interface without closing the screen with the search

Table 6.2 Excerpt from EX-file

EX id	Focus	Time	Dur	ST segN	ST segL	STidN	STidL	KD idN	KD idL	Edit	Category
...											
3	Translate—T1_T5_PE_P9.xlf—204—Google Chrome	−53,975	0	9629	−1	5	−1	0	−1	–	MAIN
4	Nieuw tabblad—Google Chrome	81,778	3360	9630	9629	15 + 16	12 + 13 + 14	122	121	**woorden[..]nlij**	NAVIGATION
5	Woordenlijst Nederlandse Taal—Officiële Spelling—Google Chrome	85,138	3937	9630	9629	15 + 16	12 + 13 + 14	122	121	**groot-bri**	SPELLING
6	Translate—T1_T5_PE_P9.xlf—204—Google Chrome	89,075	123,512	9630	9629	15 + 16	12 + 13 + 14	122	121		MAIN
7	Nieuw tabblad—Google Chrome	212,587	3548	9633	9632	75	70	193	192	**linguee**	NAVIGATION
8	Linguee | Nederlands-Engels woordenboek (en andere talen)—Google Chrome	216,135	2718	9633	9632	75	70	193	192	**n fact**	CONCORDANCER
9	in fact—Nederlandse vertaling—Linguee woordenboek—Google Chrome	218,853	4765	9633	9632	75	70	193	192		CONCORDANCER
10	Translate—T1_T5_PE_P9.xlf—204—Google Chrome	223,618	264,006	9633	9632	75	70	193	192	eed	MAIN
...											

Each time the participant switches to another screen or application, a focus event is recorded, with code EXid and a label found in column 'Focus'. Time is time in ms since the beginning of the session, Dur is the time in ms spent in a particular focus event. STsegL represents the last segment opened in CASMACAT before leaving the tool, STsegN is the next segment opened after returning to the CASMACAT tool. STidL and STidN represent the last source token before leaving CASMACAT and the next token after returning to CASMACAT. KDidL and KDidN contain the ID of the last keystroke before leaving CASMACAT and the next keystroke after returning to CASMACAT. The actual characters typed within a focus event are shown in the column 'edit'. Each focus event is given a corresponding category

query on it, the next time that person opens the search query, this will show up exactly like the search made during the first segment in the data. It would require a lot of extra manual work to label each external resource with the correct segment. In the future, we will try to better map the CASMACAT and Inputlog data by looking at keystrokes or by filtering on the time spent on certain pages. At the moment, however, we grouped the information from the EX-files per session, and not per segment so as to not incorrectly link certain resources to segments. This information was added to the more general SS-file, a table containing an overview of the different sessions. For the different categories (Dictionary, Concordancer, Encyclopedia, Search, and Other) we added a column containing the number of times that resource was consulted in that particular session, and a column containing the time spent in that resource during the session. To be able to better compare the data across all sessions, we normalized the counts and durations by dividing them by the number of source text tokens.

6.4.1 Differences in Usage of External Resources Between HT and PE

Before assessing the impact of the usage of external resources, we wanted to check whether or not there is a difference in the external resources used in regular translation (HT) or post-editing (PE). We used the R statistical software (R Core Team 2014), the lme4 package (Bates et al. 2014) and the lmerTest package (Kuznetsova et al. 2014) to perform a linear mixed effects analysis of the relationship between the total time spent in external resources normalized by dividing by the number of source text tokens, and the type of task (post-editing and human translation). As fixed effect, we entered task. To account for between participant and between text variation, we added intercepts for participants and text as random effects, without random slope. We did test a model with random slope for task, but the slope did not significantly improve the model, so we left it out in the final model. The model with fixed effect was significantly different from the null model without fixed effect ($p = 0.006$), reducing the Akaike's Information Criterion (AIC) value from 1256.8 to 1251.3. AIC (Akaike 1974) is a method designed for model selection, based on a comparison between models. It is shown to have a sound theoretical foundation (Burnham and Anderson 2004). Burnham and Anderson provide the following strategy and rules of thumb when assessing plausible models: the best model is considered the one with the lowest AIC value—in the above case 1251.3—and the plausibility of the model that you compare with it is determined by the difference between both AIC values—in this case the difference between 1256.8 and 1251.3, i.e. 5.5. According to Burnham and Anderson, if the difference is less than 2, there is still substantial support for the model, if the difference is between 4 and 7, there is considerably less support, and models that differ from the best model by more than ten points have basically no support. For the present models, we can

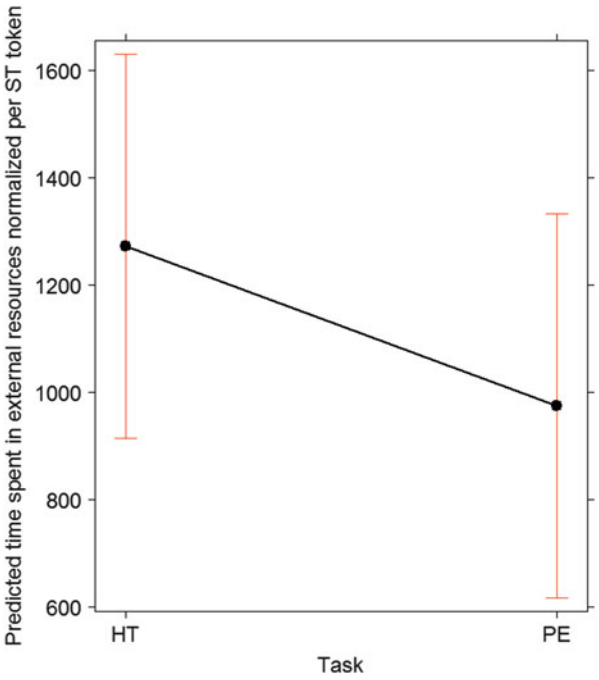


Fig. 6.2 Effect plot of relationship between task (HT = human translation, PE = post-editing) and predicted time (in ms) spent in external resources normalized per ST token. Error bars represent 95 % confidence intervals

conclude that the null model without fixed effects (and AIC value of 1256.8) is not supported enough, so we drop it in favour of the model with fixed effect (and AIC value of 1251.3). The model summary further showed that significantly more time is spent in external resources in human translation, compared to post-editing: about 297 ms $\pm$ 105 (standard errors). The effect plot obtained with the effects package (Fox 2003) is depicted in Fig. 6.2 below. This plot indeed confirms that less time is spent in external resources when post-editing than when translating. Though the confidence intervals in Fig. 6.2 overlap to some extent, this does not affect the statistical significance found (Goldstein and Healey 1995). Visual inspection of normal Q-Q plots indicated right skewed data, which is presumably due to the natural boundary at zero, which is an integral part of the data: It is impossible to spend less than 0 s in external resources, fifty per cent of data points are below 1000 ms, with very few observations above 2000 ms.

In addition to the overall comparison of time spent in external resources, we wanted to check whether the time spent in each type of external resource differed between both methods of translation. We restructured our data of the session summary table (cf. Chap. 2, Sect. 2.3) to be able to perform the appropriate analysis. An excerpt of the new data file can be seen in Table 6.3 below.

Table 6.3 Restructured data for comparative analysis of usage of external resources between human translation and post-editing

Session	Participant	Text	Task	ExternalSource	CountSource	DurSource
P01_P01	P01	T1	P	Dictionary	0.033898305	228,3,785,311
P01_P01	P01	T1	P	Concordancer	0.084745763	369,7,909,605
P01_P01	P01	T1	P	Encyclopedia	0	0
P01_P01	P01	T1	P	Search	0.096045198	417,0,225,989
P01_P01	P01	T1	P	Other	0	0

The column CountSource contains the number of times each resource was consulted during a particular session, normalized per ST token, and the column DurSource contains the time spent in each external resource during a particular session, also normalized per ST token

We again performed a linear mixed effects analysis of the relationship between the time spent in external resources normalized per ST token and the type of task, but this time also in relation to the type of external resource (dictionary, concordancer, encyclopedia, search, other). As fixed effects, we entered task and external resource with interaction term (as we are interested in the combined effect of task and external resource type). Again, we had intercepts for participants and texts, without random slope as random effects (both models were tested, but the model without random slope performed better). The model with fixed effects and interaction was significantly different from the null model without fixed effects ($p < 0.001$), reducing AIC from 5693.7 to 5650.7, but—contrary to our expectations—not significantly different from the model without interaction between task and type of external resource ($p = 0.896$; AIC = 5643.8). The drop1 test showed that none of the predictors (with or without interaction) were significant. We therefore conclude that type of external resource and task are not significantly inter-dependent on each other with regards to the time spent in external resources, even though the overall time spent in external resources was significantly different between human translation and post-editing. The model summary only showed significance for the time spent in encyclopedias and 'other resources'. Both are used significantly less than dictionaries, concordancers and search queries: encyclopedias lowered the duration in the resource per token by about 250 ms (±60 ms), and 'other resources' lowered it by about 150 ms (±60 ms). The effect plot of the model with interaction can be seen in Fig. 6.3. As we can see, there seems to be some trend to spend more time in each resource when translating than when post-editing, but these differences were not found to be significant within the current model.

From these two analyses, we can conclude that overall, the ten translation students spend more time in external resources when translating than when post-editing, though the time spent in each specific resource is not significantly different between the two conditions. In the following sections, we take a closer look at possible effects of the usage of external resources, namely the impact of external resources on overall productivity, and the impact of external resources on the final quality.

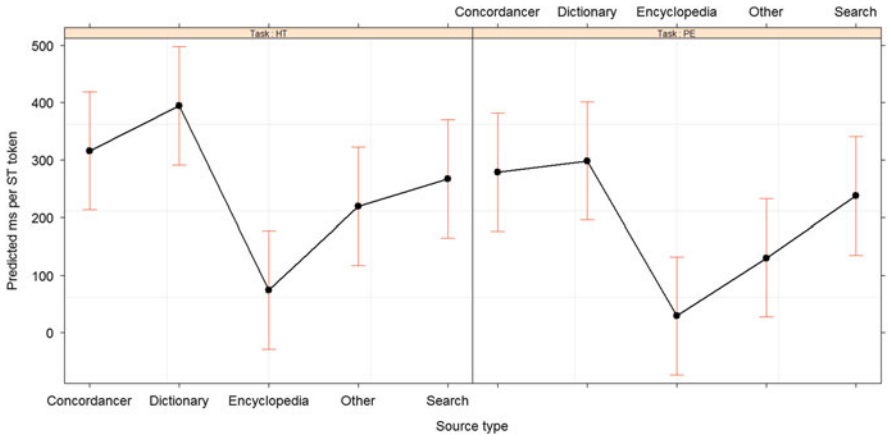


Fig. 6.3 Effect plot of predicted time (in ms) spent in each type of external resource, normalized per ST word, for both task types (*left*: HT = human translation, *right*: PE = post-editing)

6.4.2 Impact of External Resources on Productivity

There are two conceivable ways in which the usage of external resources affects productivity. On the one hand, we can expect total translation time to increase when a person spends more time in external resources, on the other hand, it is possible that the time spent in external resources decreases the overall time needed to translate a text, as a translator looks up external resources to solve problems.

We first take a closer look at the overall difference in time between human translation and post-editing by performing a linear mixed effects analysis. Total time normalized per ST token was taken as the dependent variable, and task as the predictor variable. Intercepts for text and participant were added as random effects. The model with predictor variable performed significantly better than the null model ($p = 0.0116$), reducing the AIC value from 1370.6 to 1366.2. Significantly more time per token was needed for the regular translation task compared to the post-editing task: 523.43 ms (± 202.14; $p = 0.0119$). This effect is visualized in Fig. 6.4 below.

In a next step, we added the time spent in external resources as a predictor, plus the interaction with task, so as to assess the combined effect of task and time spent in external resources on overall time. This model performed significantly better than the model with only task as predictor ($p < 0.001$), reducing the AIC value from 1366.2 even further to 1321.9. However, when we tested the model with interaction against a model without interaction, there was no significant difference, and the model without interaction reduced the AIC value to 1319.9. In addition, the drop1 function showed that only the time spent in external resources was a significant predictor. The AIC value for the final model, which included only the time spent in external resources as predictor, was 1318.6. We can conclude that, even though the

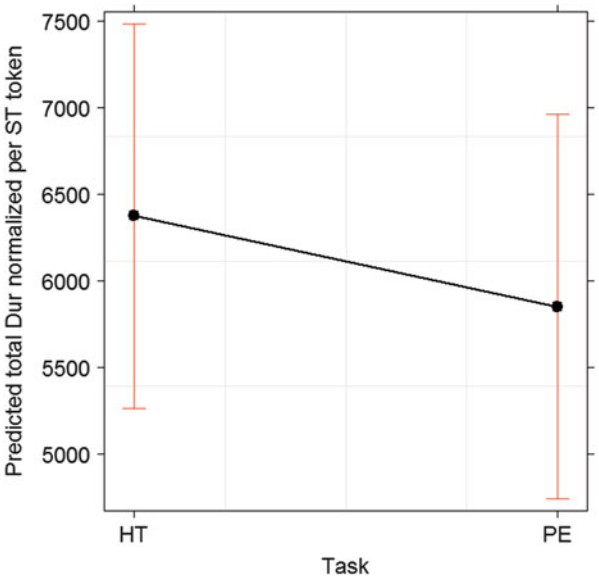


Fig. 6.4 Effect plot of predicted total time (in ms) normalized per ST token for both task types (HT = human translation; PE = post-editing). Error bars represent 95 % confidence intervals

total time, and the time spent in external resources is significantly higher for human translation than for post-editing, the time spent in external resources is a much better predictor of overall time than the task type. The model summary shows that every millisecond spent in external resources per ST token corresponds to a total time per token to increase by 1.348 ms (± 0.145; $p < 0.001$), thus causing us to reject the hypothesis that the time spent in external resources reduces the overall time needed. The effect plot can be seen in Fig. 6.5 below. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality.

6.4.3 Impact of External Resources on Quality

Another crucial aspect to take into account is a text's final quality. Spending more time in external resources (and thus increasing the overall time needed) can be justified if this extra time also brings about an increase in quality. While quality assessment is not always straightforward, we have developed a translation quality assessment approach which allows us to look at the most important problems after translation. It is beyond the scope of this chapter to expand on our methodology, but it has been discussed in more detail in Daems et al. (2013, 2014). The main difference between our approach and other approaches is that we look at acceptability and adequacy as two aspects of quality: quality with regards to the

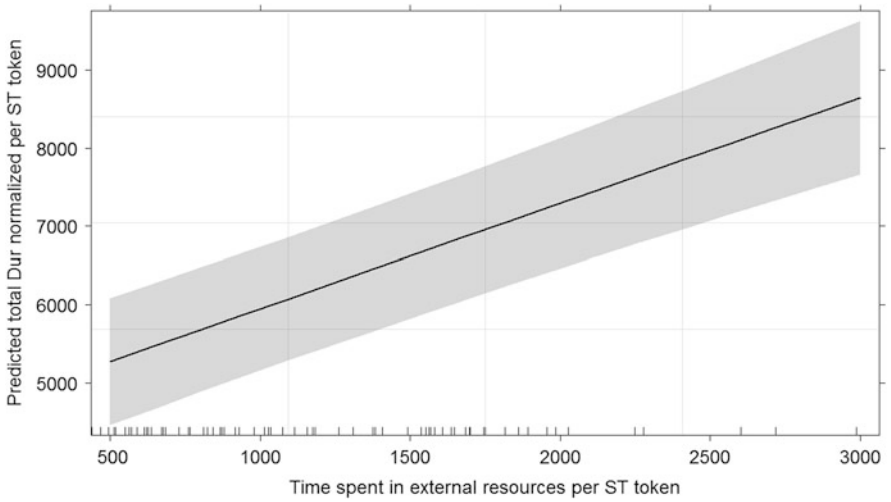


Fig. 6.5 Effect plot of relationship between time spent in external resources normalized per ST token and total time normalized per ST token (both in ms)

final text as a good text in the target language and culture, and quality with regards to the correspondence between source and target text. Acceptability and adequacy each contain various subcategories (such as, for example, grammar, spelling, style and lexicon for acceptability; and word sense, deletions and contradictions for adequacy), allowing for a fine-grained error analysis. Each error category also receives an error weight from zero to four, indicating the severity of the error for the specific text type (for example, a contradiction error receives a weight of four, whereas a capitalization error receives a weight of one). We do also provide an overall quality score. The overall score is calculated by summing up the error scores for acceptability and adequacy and subtracting those acceptability items which were caused by adequacy errors, so as to not penalize the same problem more than once. For example, a word sense error (adequacy) can also lead to a logical problem (acceptability), as is the case in the following situation: The source text contains the verb 'to spend', meaning 'to spend money' (e.g. 'families continue to spend cautiously'), but this is translated as 'doorbrengen' in Dutch, meaning 'to spend time'. The word 'doorbrengen' in this sentence is both a word sense error and a logical problem in the target text. Rather than summing up both error scores in these situations, we only count the error score for the word sense error. Two of the authors highlighted and labeled all errors in the translations, after which we held a consolidation phase where problematic cases were discussed and resolved. Our analyses were conducted on data containing only those errors both annotators agreed on. As with the information on external resources, the error count and score for each category was added to the session file (SS) and normalized by dividing through the number of words in the source text.

6.4.3.1 Overall Quality

Before looking at the effect the usage of external resources has on quality, we looked at the effect of the task on quality. We fit a linear model with normalized total error score as dependent variable and task as predictor variable. In this model, task was not a significant predictor of total error score in itself ($p = 0.669$). We can therefore conclude that there is no significant difference in overall quality between both types of translation (post-editing and human translation).

We then fit a linear mixed effects model to analyze the relationship between overall error score normalized per ST token and the normalized total time spent in external resources. Normalized total error score was the dependent variable, task and time spent in external resources with interaction were added as predictor variables and text and participants were added as random effects, both with random slope for task. This model performed better than the null model without predictors, though only just so ($p = 0.09$), reducing AIC from −306.57 to −306.97, which—according to Burnham and Anderson (2004)—is a negligible reduction. Backward elimination of non-significant effects with the step function showed a significant effect for all variables, with the exception of the slope added to the variable text. In the final model, this slope was left out, leading to a further reduction of the AIC value to −309.59. The main effects of task (post-editing vs. translation) are positive and significant ($p = 0.05$), increasing the average total error score per ST token in the translation condition with 0.035 units (±0.0174). Taking the interaction effect of the total number of external resources into account, however, we see something else entirely. The slope for the time spent in external resources is set at 0.000015 for the post-editing condition (±0.000008587; $p = 0.079$), which is reduced with 0.0000286 points (±0.00001: $p = 0.0118$) in the translation condition. This interaction effect can be seen in Fig. 6.6 below. Inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality.

The differences in slope seem to indicate a difference in the effect of consulting external resources for both types of task. In the case of post-editing, spending a longer time in external resources does not lead to an increase in quality, but rather a decrease, indicating that the resource consulting strategies are not successful. In the case of translation, however, the extra consulted resources do seem to pay off, leading to a decrease in overall error score. This is perhaps not such a surprising result, given that our participants are students with experience in translation, but not in post-editing. It can be assumed that they have developed successful resource consulting strategies when translating throughout their studies, whereas post-editing is a new type of translation, giving rise to different problems, questions, and strategies, which are not always as successful as when translating. We speculate that a possible explanation for these findings can be found in the machine translation (MT) quality. On the one hand, students might be too trusting of MT quality (as evidenced by the fact that less time is spent in external resources when post-editing), on the other hand, they encounter very different problems when post-editing than when translating from scratch, making it hard to find the exact cause of a problem, and—in extension—to decide on the most appropriate external resources

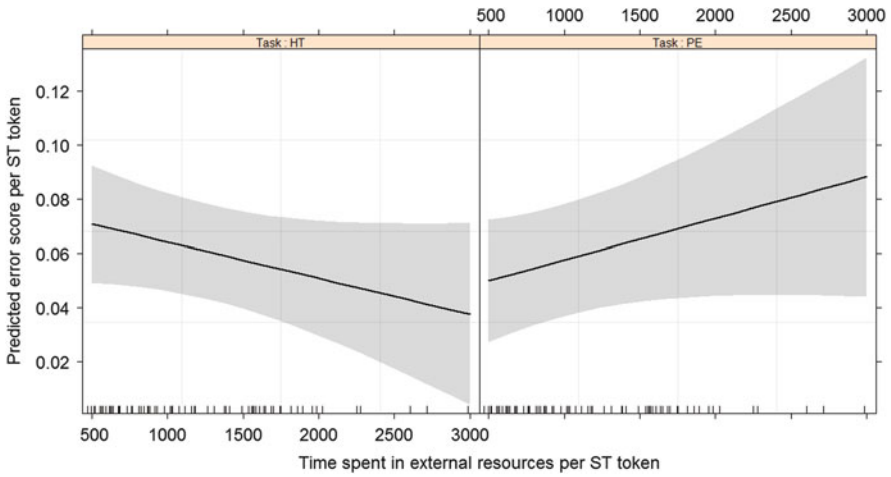


Fig. 6.6 Effect plot of the predicted relationship between time spent in external resources normalized per ST token and overall error score normalized per ST token, for both types of task (*left*: HT = human translation, *right*: PE = post-editing)

to consult. Perhaps the machine translation output primes certain—misguided—search strategies, leading to the students being unable to solve problems even when consulting external resources. Another explanation could be that, when translating from scratch, students look up external resources in sentences that are not so difficult to begin with, which would be reflected in extra time spent in external resources for sentences that already have low error scores.

In addition to this global analysis, we wanted to look at the effect of time spent in the various external resources normalized per ST token on overall quality. We performed a linear mixed effects analysis to assess the relationship between the total error score per ST token and the time spent in the various external resources per ST token. The full model contained the duration of all external resources as possible predictor variables (dictionary, encyclopedia, search, other, concordancer). Text and participant were added as random factors, with added random slope for task. The model with predictor variables did, however, not perform better than the null model ($p = 0.243$), increasing AIC from −309.49 to −306.2. We used the step function from the lmerTest package to assess the necessity of each variable through automatic backward elimination of effects. Only the random effects were significant according to this function. This might indicate that quality is influenced more by differences between texts and differences between participants than the types of external resources consulted. Additional correlation analyses showed no significant correlation between the students' LexTALE proficiency scores and the total error score. We did find a low but significant correlation ($r = 0.296$, $p < 0.01$) between the total error scores and how tiring students perceive post-editing to be. What is remarkable, however, is that the students who perceive post-editing as being less tiring than human translation have higher error scores. This could indicate that

those students are not critical enough: the fact that they perceive human translation as being more tiring could indicate that they struggle with human translation—potentially leading to high error scores—and the fact that they perceive post-editing as less tiring could indicate that they trust the machine translation output too much—again leading to higher error scores. These assumptions warrant further investigation in future research.

6.4.3.2 Acceptability

After looking at quality in general, we took a closer look at our two aspects of quality: acceptability and adequacy, beginning with the first. Inspection of exploratory box plots showed no obvious difference between the acceptability score normalized per ST token for both tasks, which was confirmed by fitting a simple linear model with acceptability error score as dependent, and task as predictor variable. In this model, task was not a significant predictor of the acceptability error score ($p = 0.35$), which is in line with the findings from the overall error score.

We then set out to statistically assess the relationship between time spent in external resources and acceptability error score. We performed a linear mixed effects analysis with normalized acceptability error score as dependent variable and task and normalized time spent in external resources with interaction as predictor variables. Participant was added as a random effect, with added random slope for task. This model, however, did not significantly perform better than the null model ($p = 0.57$). Backward elimination of non-significant effects with the step function showed that none of the predictor variables significantly added to the model. Only participant as random effect with random slope for task was retained, leading us to conclude that neither the overall time spent in external resources nor task type has a significant effect on the acceptability error score, but acceptability error score is most likely influenced by between participant differences. In their 2010 paper, Carl and Buch-Kromann also found no significant relationship between longer translation times and the fluency—which corresponds to our notion of acceptability—of student translators.

The following step was to see whether time spent in specific external resource types had an effect on acceptability error score. We performed a linear mixed effects analysis to assess the relationship between the total acceptability error score per ST token and the time spent in the various external resources per ST token. The full model contained the duration of all external resources as possible predictor variables (dictionary, encyclopedia, search, other, concordancer). Text and participant were added as random factors, with added random slope for task. We used the step function from the lmerTest package to assess the necessity of each variable through automatic backward elimination of effects.

On the basis of this analysis, we again only retained participant as a random effect, with random slope for task, and the duration for dictionary as a predictor variable. This was the only predictor variable found to have an impact on overall acceptability quality. The final model was tested against a null model without

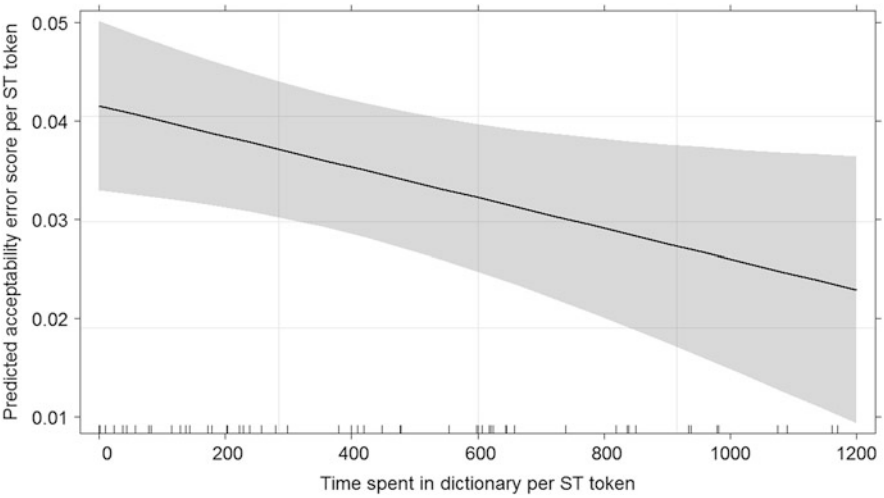


Fig. 6.7 Effect plot of the predicted relationship between time spent in dictionaries normalized per ST token and acceptability error score normalized per ST token

predictor variable, and was found to provide a significantly better fit ($p = 0.01762$), reducing AIC from −384.9 to −388.53.

The effect plot can be seen in Fig. 6.7 below. Residual plots did not reveal any obvious deviations from homoscedasticity or normality. Each millisecond spent in dictionaries affects the acceptability error score per ST token with −0.000016 points (±0.000006). So each second spent to look something up in a dictionary can reduce the acceptability error score for that word with approximately 0.016 units. We can conclude that dictionaries seem to be the only external resource that significantly reduces the acceptability errors made, making it perhaps the most useful resource with regards to acceptability issues.

6.4.3.3 Adequacy

A second aspect of quality is adequacy. We again fit a linear model, this time with normalized adequacy error score as dependent variable and task as predictor variable. As was the case for acceptability, no significant effect was found ($p = 0.527$).

We then performed a linear mixed effects analysis with normalized adequacy error score as dependent variable and normalized time spent in external resources as predictor variable to assess the relationship between time spent in external resources and adequacy quality. Participant and text were added as a random effects, with added slope for task. This model, however, did not perform better than a model without fixed effects ($p = 0.7$), increasing the AIC value from −346.67 to −344.82. Backward elimination of non-significant effects with the step function from the lmerTest package showed only text to be a significant random effect, without

slope. We can conclude that the overall time spent in external resources does not significantly influence the obtained adequacy error score. This finding is in line with the findings by Carl and Buch-Kromann (2010) that there is no notable correlation between accuracy—which corresponds to our notion of adequacy—and translation time.

The next step was to look at the influence of the different types of resources. We applied the same methodology to assess the relationship between the total adequacy error score normalized per ST token and the time spent in the various external resources normalized per ST token. Again, the full model contained the duration of all external resources as possible predictor variables (dictionary, encyclopedia, search, concordancer, other), as well as the task predictor variable. Text and participant were added as random factors, with added random slope for task. We used the step function from the lmerTest package to assess the necessity of each variable.

On the basis of this analysis, we only retained task as a random effect, without random slope. This time, the only predictor that came out of the analysis as having a significant effect on overall adequacy error score, was the time spent in encyclopedias. The final model was tested against a null model without predictor variable, and was found to provide a significantly better fit ($p = 0.04182$), reducing AIC from -352.39 to -354.53.

The effect plot can be seen in Fig. 6.8 below. Residual plots did not reveal any obvious deviations from homoscedasticity or normality. Each millisecond spent in encyclopedia affects the adequacy error score per ST token with 0.000056 points (± 0.000027). So each second spent to look something up in an encyclopedia can increase the adequacy error score for that word with approximately 0.056 units.

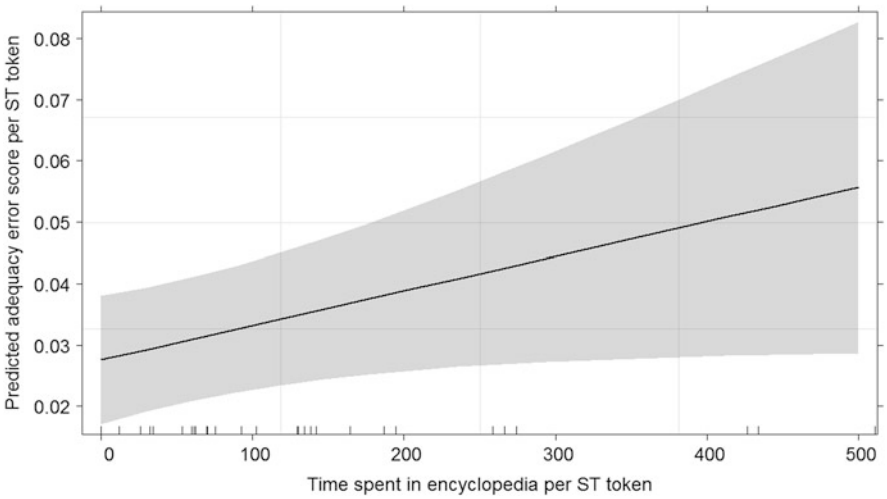


Fig. 6.8 Effect plot of the predicted relationship between time spent in encyclopedias normalized per ST token and adequacy error score normalized per ST token

Of course we do not claim this relationship to be causative. It is presumably not the consulting of the encyclopedia which increases the error score, but the need to consult more encyclopedias can be an indication of the difficulty of the translation. The fact that the effect on adequacy error score is positive might mean that consulting encyclopedias is not always a successful strategy. A possible explanation could lie in the nature of encyclopedias: they provide additional information on a topic, but they do not always provide clues on how to translate terms. Closer inspection of the data shows that sometimes, participants try to look up concepts that are not typical encyclopedia entries, such as 'officially enforced anger'. Additionally, an encyclopedia such as Wikipedia sometimes provides corresponding pages in other languages, but these pages do not always exist or are not always informative. One participant, for example, looked up 'Federal Bureau of Investigation' in Wikipedia, of which the corresponding Dutch page also uses the English term. While the participant spent almost half a minute looking at the Wikipedia pages for 'Federal Bureau of Investigation', this did not help him find an adequate translation. Another participant looked up 'law enforcement agency' and unsuccessfully opened the German page because there was no corresponding Dutch page. The above findings need to be considered with caution, as the overall time spent in encyclopedias is negligible compared to the time spent in other types of external resources (see Fig. 6.1).

6.5 Conclusion

We have conducted a balanced experiment comparing the usage of external resources in human translation and post-editing for general text types, and the effects on time and quality of a text, using a unique combination of state-of-the-art keystroke logging tools. We discussed the addition of Inputlog data to the TPR-DB by means of EX-files (see Chap. 2), containing information on the usage of external resources in a format that is easy to use with the existing TPR-DB tools. This study moves beyond the limitations of previous studies, that either had to make do with manual observation of external resources (Göpferich 2010) or looked at data from within one type of external resource only (Macklovitch et al. 2008).

We found a significant difference in time spent in external resources for both task types (with translation requiring more time). In contrast with our expectations, we found no statistical evidence for the hypothesis that translators use different types of resources, and in different quantities when translating or post-editing, though there seems to be a trend to spend more time in each resource when translating than when post-editing. Significantly less time is spent in encyclopedias and other types of resources compared to dictionaries, concordancers and search engines, for both types of translation.

The overall time needed to translate a text was significantly higher for translation than for post-editing, which is in line with previous findings (Plitt and Masselot 2010). We further found that the time spent in external resources significantly

increases the total time needed to translate a word, indicating that even though the resources might help translators solve translation problems, this goes at the cost of overall productivity. While participants needed significantly more time to translate than to post-edit a word, the effect of time spent in external resources was greater than the effect of the task type.

In a final analysis, we looked at the effect of external resources on the quality of a text. The overall quality of a translation did not seem to be significantly influenced by one specific type of resource, but rather by the overall time spent in external resources, as well as by the task type. When looking at post-editing, longer consultation of external resources was accompanied by higher overall error scores, whereas the opposite was true for human translation, where longer consultation of external resources was accompanied by lower overall error scores. This leads us to believe that participants are more successful in problem solving by consulting different resources when translating than when post-editing. This finding is in line with the suggestion by Yamada that post-editing requires different skills from human translation (2015). With regards to the acceptability aspect of quality, we found no significant difference between human translation and post-editing. When looking at the effect of each type of external resource on acceptability quality, we found that extra time spent consulting dictionaries does bring about an increase in acceptability quality, perhaps making it worth the loss in productivity. With regards to the adequacy aspect of quality, we again found no significant difference between human translation and post-editing. When looking at the effect of each type of external resource on adequacy quality, we found that spending more time in encyclopedias does not bring about a decrease in error score, but rather an increase. This indicates that longer searches do not necessarily lead to better translations with regards to adequacy.

In sum, we can conclude that, whereas search strategies during the translation process are more effective than those used when post-editing, post-editing is still faster than human translation without negatively affecting the final quality of the product.

6.6 Future Work

While the analyses in this chapter have given us a general idea of the effects of external resources and the differences between human translation and post-editing, it might be interesting to look at the texts more closely as well. Due to practical constraints, we performed our analyses on the text level, whereas a more fine-grained approach might give us more practical insights. In the future, we want to better map the resource events to the relevant segments, so that we can perform analyses on the segment level rather than the text level. Taking a closer look at search queries might also provide useful insights in the type of things translators look up in both conditions. Perhaps the external resources used are comparable, but the types of queries are not, or the time spent on each type of query is not.

In addition, we want to take a closer look at the problematic passages as highlighted by the participants and the machine translation quality for the post-editing task. As between participant differences seemed to have a great effect on the results, it can be interesting to perform more in-depth analyses of individual problem solving strategies.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., Martínez, M., et al. (2013). CASMACAT: An open source workbench for advanced computer aided translation. *The Prague Bulletin of Mathematical Linguistics, 100*, 101–112. doi:10.2478/pralin-2013-0016.

Angelone, E. (2010). Uncertainty, uncertainty management and metacognitive problem solving in the translation task. In G. Shreve & E. Angelone (Eds.), *Translation and cognition* (pp. 17–40). Amsterdam; Philadelphia: Benjamins.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7.*http://CRAN.R-project.org/package=lme4

Burnham, K., & Anderson, D. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*, 261–304.

Carl, M. (2012). The CRITT TPR-DB 1.0: A database for empirical human translation process research. In S. O'Brien, M. Simard, & L. Specia (Eds.), *Proceedings of the AMTA 2012 workshop on post-editing technology and practice (WPTP 2012)* (pp. 9–18). Stroudsburg, PA: Association for Machine Translation in the Americas (AMTA).

Carl, M., & Buch-Kromann, M. (2010). Correlating translation product and translation process data of professional and student translators. In *Proceedings of EAMT*, Saint-Raphaël, France.

Daems, J., Macken, L., & Vandepitte, S. (2013). Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE. In *Proceedings of the MT summit XIV workshop on post-editing technology and practice* (pp. 63–71).

Daems, J., Macken, L., & Vandepitte, S. (2014). On the origin of errors: A fine-grained analysis of MT and PE errors and their relationship. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 62–66). Reykjavik, Iceland: European Language Resources Association (ELRA).

Ehrensberger-Dow, M., & Perrin, D. (2009). Capturing translation processes to access metalinguistic awareness. *Across Languages and Cultures, 20*(2), 275–288.

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software, 8*(15), 1–27. http://www.jstatsoft.org/v08/i15/

Garcia, I. (2011). Translating by post-editing: Is it the way forward? *Machine Translation, 25*, 217–237.

Germann, U. (2008). Yawat: Yet another word alignment tool. In *46th annual meeting of the association for computational linguistics: Human language technologies; demo session*, 20–23. Columbus, OH.

Goldstein, H., & Healey, M. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, 158*, 175–177.

Göpferich, S. (2010). The translation of instructive texts from a cognitive perspective. In F. Alves, S. Göpferich, & I. Mees (Eds.), *New approaches in translation process research* (pp. 5–65). Frederiksberg: Samfundslitteratur.

Jakobsen, A. (2003). Effects of think aloud on translation speed, revision and segmentation. In F. Alves (Ed.), *Triangulating translation: Perspectives in process oriented research* (pp. 69–95). Amsterdam: Benjamins.

Jakobsen, A., & Schou, L. (1999). Translog documentation. In G. Hansen (Ed.), *Probing the process in translation: Methods and results* (pp. 1–36). Frederiksberg: Samfundslitteratur.

Krings, H. (2001). *Repairing texts. Empirical investigations of machine translation post-editing processes*. Kent, OH: Kent State University Press.

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2014). *lmerTest: Tests in linear mixed effects models. R package version 2.0-20*. http://CRAN.R-project.org/package=lmerTest

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392. doi:10.1177/0741088313491692.

Leijten, M., Van Waes, L., Schriver, K., & Hayes, J. (2014). Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research, 5*(3), 285–337.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods, 44*, 325–343.

Macklovitch, E., Lapalme, G., & Gotti, F. (2008). TransSearch: What are translators looking for? In *AMTA-2008: MT at work: Proceedings of the eighth conference of the association for machine translation in the Americas* (pp. 412–419), Waikiki, Hawai'i, St. Honolulu.

Och, F., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29*(1), 19–51.

Plitt, M., & Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localization context. *Prague Bulletin of Mathematical Linguistics, 93*, 7–16.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Yamada, M. (2015). Can college students be post-editors? An investigation into employing language learners in machine translation plus post-editing settings. *Machine Translation, 29*, 49–67.

Chapter 7
Investigating Translator-Information Interaction: A Case Study on the Use of the Prototype Biconcordancer Tool Integrated in CASMACAT

Julián Zapata

> *[R]egardless of how our universe got to be the way it is, we can start our story with a world based on information.*
> — Ray Kurzweil, How to Create a Mind (2013)

Abstract This chapter introduces translator-information interaction (TII) as the field of study that investigates translators' interaction with (digital) information and information tools. In particular, the current chapter examines translators' interaction with a prototype biconcordancer (BiConc) tool integrated in the CASMACAT workbench. The BiConc was introduced in the third CASMACAT field trial (The data of the third CASMACAT field trial is stored in the TPR-DB under the study name CFT14, cf. Chap. 2, this volume.) a post-editing experiment involving seven English-to-Spanish professional translators. In addition to external online tools, the BiConc was one of the informational resources that participants could use while post-editing two machine-translated texts under two different conditions: (1) traditional post-editing and (2) interactive post-editing with online learning (A description of the CASMACAT online-learning mode is provided in Chap. 3 in this volume). In the case study reported in this chapter, only the segments in which participants used the CASMACAT BiConc tool were examined. On the basis of screen recordings, the present study analyses the way translators interacted with the BiConc and other informational resources in order to solve a particular problem while post-editing. Overall, the chapter argues that human-centered research is essential not only in the understanding of the cognitive processes involved in translation activity, but also in the development and the improvement of tools intended to better address the professional needs of translators. Thus, this case study and subsequent TII investigations can be used to inform the efficient integration of the BiConc tool and other informational resources to CASMACAT and other future-generation (web-based) translation environments.

J. Zapata (✉)
School of Translation and Interpretation, University of Ottawa, Ottawa, ON, Canada
e-mail: jzapa026@uottawa.ca

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_7

Keywords Human-information interaction • Information behaviour • Information retrieval • Information tools • Usability

7.1 Introduction

In our day, translation is essentially both a computer-interaction task and an information-interaction task. Indeed, throughout history, human translators have used an array of tools not only to write their translations but also to search and store information. In the digital age, information and communication technologies (ICTs),[1] and in particular language technologies (LTs),[2] are integral parts of the translation field, and have decidedly had a significant impact on translation research, practice and teaching.

The current chapter introduces the notion of translator-information interaction (TII) as the field of study that investigates translators' interaction with (digital) information and information tools. This new notion complements that of translator-computer interaction (TCI), coined by Sharon O'Brien in 2012. TII and TCI represent logical extensions of the fields of human-information interaction (HII) and human-computer interaction (HCI) respectively. Now, although TII and TCI are emerging fields of research, the interaction of translators with computers and digital information is not a recent phenomenon, as O'Brien (2012, pp. 103–104) explains:

> Already with the introduction of the electronic typewriter, with only two lines of memory, and the use of dictaphones, translation became a computer-interactive task. This was followed by the introduction of word-processing software [. . .][,] a development that would have required some translators to interact with a computer for the first time. Not long after the mass embracing of word processing, came the introduction of Translation Memory tools [and] terminology management programs, which are [. . .] not restricted to the [parallel] storage of terms [in two languages], but also store phrases and sometimes even sentences or larger chunks of text [. . .].

In sum, in the age of ICTs, translators have adopted different types of computer tools in an effort to facilitate their work and carry out their tasks effectively (Austermühl 2001; Bowker 2002). For instance, parallel bilingual resources, as

[1] ICTs are defined as the bulk of technological applications based on computing, microelectronics, telecommunications and multimedia, the combination and interconnection of which allow people to search, capture, process and transmit data of different nature (text, audio, image, video, etc.); to interact with each other and with machines; to access information; and to spread and share information (Touré et al. 2009, p. 35).

[2] LTs are defined in this chapter as the bulk of natural language processing (NLP) applications that facilitate the active or passive use of a natural language. Certain LTs are developed for the general public, while others are developed for language professionals (e.g., writers, translators, terminologists, etc.). LTs may be divided in two categories: spoken-language-based and written-language-based. Each one of these categories may be divided into two types: passive applications (e.g., unchangeable information on the web or electronic/online dictionaries and term banks) and active applications (e.g. text processing software, spellcheckers and speech recognition systems).

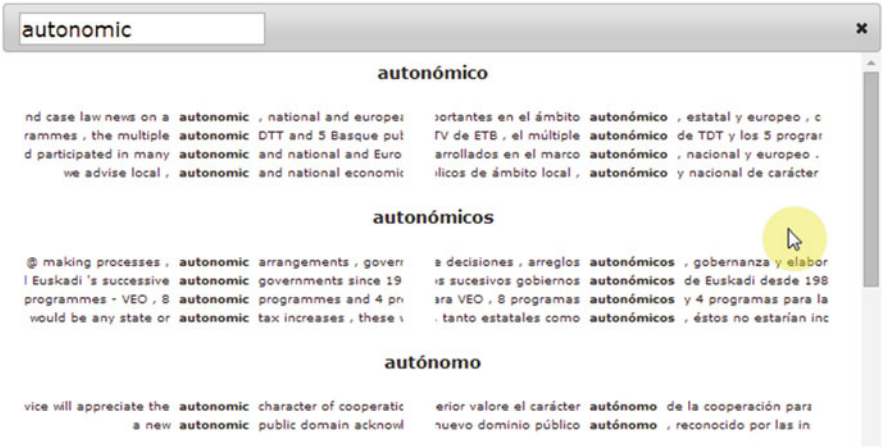


Fig. 7.1 Example of a search result in the prototype biconcordancer tool integrated in the CASMACAT translator's workbench

described above by O'Brien, have been built and used for over two decades (Langlois 1996). The present chapter deals in particular with an increasingly popular type of parallel bilingual resource: bilingual concordancers, or *biconcordancers*. This type of informational tool allows the user to search for a word, or any character string, within a previously-aligned bilingual parallel corpus. Hence, the search result consists of a list of segments in a language A containing the searched character string, and their corresponding segments in a language B, as illustrated in Fig. 7.1.

By way of a literature review and a case study, a new approach in empirical translation process research (TPR) is proposed in this chapter, that is, the investigation of translator's interaction with (digital) information and informational tools, or TII. More specifically, the chapter reports on translators' interaction with the first prototype of a biconcordancer (BiConc) integrated in the CASMACAT workbench. The BiConc tool was introduced in the third CASMACAT field trial (CFT14), a post-editing experiment conducted with seven English-to-Spanish professional translators in a Madrid-based translation company.

Several questions motivate TII research: how well do human translators work with the information and informational tools they currently have at their disposal? How accurate, rich and relevant is the information they find? How user-friendly are informational tools for translators? How can the information and the tools be improved to maximize translators' performance and well-being, and the quality of their work? As far as the CASMACAT workbench is concerned, how can the performance of the built-in BiConc tool be assessed and improved? What are the advantages and drawbacks of integrating an informational tool to a translation environment, as compared to having an array of external web-based resources? These are some of the questions that motivated the present investigation, which

remains exploratory given the scope and limitations of a pilot study and of the Translation Data Analytics (TDA)[3] project, and are partly dealt with in this chapter.

Overall, the chapter argues that human-centered research is essential not only in the understanding of the cognitive processes involved in translation activity, which is TPR's ultimate goal (Balling and Carl 2014; Jakobsen and Jensen 2008; Jakobsen 2003, 2011; O'Brien 2009), but also in the development and the improvement of tools intended to better address translators' professional needs (Carl et al. 2011). Thus, this case study and subsequent investigations in the same vein can be used to inform the efficient integration of the BiConc tool and other informational resources to the CASMACAT workbench and other future-generation translation environments.

7.2 Theoretical Framework: From Translator-Computer Interaction to Translator-Information Interaction

It is difficult to think about translation today without thinking about computer tools and technologies. In recent years, there has been an increased awareness of the significance of technologies and the role they play in translation research, teaching and practice. More than ever before, translation researchers, trainers and professionals are aware of the importance of improving existing tools and creating new tools to cope with the evolution of technology and the ever-changing professional needs of translators. There is a tangible need to design and develop ergonomic and flexible interfaces that take the human factor into consideration and that are adapted to the translator's workflow and needs (Carl et al. 2011; LeBlanc 2013; O'Brien 2012; Taravella and Villeneuve 2013), since any application that is too rigid impedes the work that it is meant to support (Karamanis et al. 2011, p. 49). Given the current state of affairs, research that takes the human factor into account is bound to play a more prominent role in translation tool design and implementation in the years to come.

Simply put, HCI research focuses on designing computer applications that are *useful*, *usable* and *universal* (Shneiderman 2008). A typical HCI research project seeks to design or redesign a particular computing technology in order

[3]This pilot study was carried out within the framework of the TDA project held in July-August 2014 at the Centre for research and innovation in translation and translation technology (CRITT), located at the Copenhagen Business School, in Denmark. The aim of the TDA project was to explore and analyse translator-computer interaction data available in the CRITT TPR-DB in an effort to assess and elaborate methods to produce data-driven user profiles, to investigate differences in communication styles, and to identify patterns of user behavior for more and less successful man–machine communication. The TDA project was supported by the European Union's 7th Framework Program (FP7/2007-2013) under grant agreement 287576 (CASMACAT).

to (1) improve upon or enhance a given experience or (2) create a quiet different experience than before (Harper et al. 2008, p. 58):

> In both situations, initial research is conducted by learning more about people's current experiences [...]. Ethnographic studies, logging of user interaction and surveys are commonly deployed. Based on the findings gathered, we begin to think about why, what, and how to design something better. To aid the process, usability and user experience goals are identified and conceptual models developed. Prototypes are built, evaluated, and iterated, demonstrating whether the user goals have been met or whether the new user experience is judged to be enjoyable, pleasurable or valuable by the target group.

Thus, usability studies combining tool use and translation processes are therefore more than necessary in translation research, as O'Brien (2012, pp. 116–117) argues:

> [TCI] would likely benefit from an increased focus on ethnographic-style, cognitive ergonomic studies of both translation tools and the translation process itself [...]. More experimental studies of translator-tool interaction could be carried out using formal usability research methods such as screen recording, eye tracking, and observation, the results of which could then be used by translation technology developers to improve the specifications of tools for the benefit of translators and, ultimately, the end users of those translations.

As stated in the introductory section, the work described in this chapter aims at proposing a new approach in empirical TPR, that is, the investigation of the way translators interact with (digital) information and informational tools. Thus, TII would complement O'Brien's notion of TCI. Furthermore, the idea that TII is a larger discipline that encompasses TCI is put forward. Indeed, some HCI and HII researchers argue that HII constitutes a larger discipline, since it looks beyond computers. It focuses on the interaction between humans and the information in the environment, in all its complexity, regardless of the tools used to facilitate such interaction (Fidel 2012; Gershon 1995; Marchionini 2008); the computer just happens to be one of the mediums that facilitate or mediate the interaction with the information we need and produce. Humans have always been in constant interaction with information, be it via machines or not. Our world is based on information (Kurzweil 2013, pp. 2–3).

The study of the interaction between humans and information is not new. However, with the advent of ICTs and, in particular, of the Internet, the field of HII has become particularly popular within the research communities in computer science and an array of other disciplines (Fidel 2012, pp. 17–21). The massive influx of mobile, Internet-connected devices has led humans to new ways of accessing enormous quantities of information and services at any time and from practically anywhere, making it necessary to investigate HII from every angle and every field, and to strengthen HII as a multidiscipline.

Two research areas related to HII are particularly well grounded today, and offer a great potential in empirical TPR: information retrieval (IR) and information behavior (IB). The former investigates the models and mechanisms of (computer) systems that allow or facilitate the retrieval of information. The latter examines information research strategies, information evaluation criteria, and the modalities and contexts of information use. In other words, while IR focuses on developing and improving informational tools, IB investigates the ways of browsing the different

sources of information, and of evaluating if the information found is adequate for solving a given problem in order to use it according to the constraints set by the context (Fidel 2012, pp. 35–37).[4] Thus, IB informs IR research: In the search of informational tools that are more efficient, it is necessary to meticulously investigate translators' interaction with the different informational resources that are currently made available to them. It is also essential to include real users working in real-life situations when assessing the usability and the performance of new tools and tool prototypes, which in return helps designers and developers in making key decisions about particular aspects and features of a user interface.

In sum, TII offers a great potential in empirical TPR and translation studies in general since, in the search for translation tools that are efficient, ergonomic and well-adapted to translator's needs, it is necessary to thoroughly study translators' interaction with information and with the different informational resources they use to carry out their tasks. Let us now illustrate TII research by presenting, after a brief overview of the CFT14 experiment, the methodology and the results of this case study looking into professional translators' interaction with the CASMACAT built-in BiConc tool prototype and other external informational tools, and with the information retrieved in those resources.

7.2.1 The CFT14 Experiment: An Overview

This pilot TII study was performed based on data collected during the third CASMACAT field trial (CFT14), carried out in June 2014 by researchers from the CRITT (Alabau et al. 2014).[5] The CFT14 consisted in delivering the CASMACAT workbench to professional English-to-Spanish translators and having them post-edit two 4500-word medical specialized texts (package leaflets for schizophrenic patients) under two different conditions: (1) traditional post-editing with no assistance during the process (P), and (2) post-editing through interactive translation prediction featuring online learning (PIO). The 2 texts consisted of 131 and 141 segments respectively. They were pre-translated using a statistical machine translation (SMT) engine and then loaded into the CASMACAT environment for participants to post-edit them. An eye-tracker was used to record participants' gaze behavior. Lastly, a questionnaire followed the experiments.

[4] According to Fidel (2012, p. 85) context is important because, even before carrying out any search, it is context that shapes the informational needs, since the motivation to search for information is not only cognitive, but also contextual.

[5] The team of researchers listed below are to be acknowledged for their work on the CASMACAT workbench and, in particular, for running the CFT14 experiment and providing us with the data presented in this section: Vicent Alabau, Michael Carl, Francisco Casacuberta, Mercedes García Martínez, Jesús González-Rubio, Bartolomé Mesa-Lao, Philipp Koehn, Daniel Ortiz-Martínez, and Moritz Schaeffer.

The principal goals of this field trial were: (1) to assess the benefits in terms of productivity derived from introducing online-learning techniques; (2) to investigate how post-editors use informational tools during the post-editing process, in particular the built-in BiConc tool; (3) to assess how professional reviewers use the newly-introduced CASMACAT electronic pen functionalities while reviewing post-editors' output; and (4) to collect feedback from reviewers using the electronic pen as an additional input method for revision (*ibid*.).

All post-editors were freelance translators recruited by *Celer Soluciones* SL, a Madrid, Spain-based translation company. Participants were 35 years old on average. They were all regular users of language technologies in their day-to-day work. All participants but one had experience post-editing machine-translated texts as a language service.[6] More detailed data on the participants' age, expertise, education, etc., is available in the CRITT TPR database[7] (metadata folder; see also Hvelplund and Carl (2012) for a description).

Participants were all given the time to familiarize themselves with the CASMACAT workbench; some of them were using it for the first time. Likewise, in order to ensure an equal distribution of texts and conditions across participants, variables were counterbalanced from participant to participant.

To measure whether participants become faster when post-editing with interactive translation prediction and online learning techniques (goal 1 of this field trial), task completion times and keystroke activity were measured and analyzed. Time was measured using *FDur*, *KDur* and *PDur* values (see Chap. 2, Sect. 2.4.6, for a definition of these values). In order to measure the productivity benefits derived from introducing online-learning techniques during the post-editing process, the amount of technical effort (i.e. the number of insertions and deletions needed to correct the raw SMT output) was calculated for the two conditions. Keystroke activity was measured by using *Mdel* values (i.e., number of manually generated deletions) and *Mins* values (i.e., number of manually generated insertions). It is important to make the distinction between manual and automatic insertions and deletions since the interactive translation prediction functionality triggers a number of automatic insertions and deletions that do not require any technical effort (i.e. typing activity) from the post-editor (*ibid*,). Table 7.1 compiles the keyboard activity and production time measures across participants.

Now, usability studies such as the CFT14 should take into account the translation/post-editing process as a whole in order to control for any possible confounding variables that may have an impact on the data. Results of the CFT14 (see Alabau et al. (2014)) (also reported in this volume; see Chap. 4) show in particular that post-editors did not seem to be faster under the PIO condition. However, a more in-depth qualitative analysis of the process data collected shows

[6]Only participant 4 (P4) reported that she did not have any experience in post-editing. As it will be seen in the Methodology section below, this does not have an impact on the results of the pilot experiment reported in this chapter.

[7]Available at: https://sites.google.com/site/centretranslationinnovation/tpr-db

Table 7.1 Overall typing activity measures and production times

Participant	Cond	Ins/ST char	Del/ST char	Fdur	Kdur	Pdur
P1	P	0.88	0.79	469	290	138
P1	PIO	0.73	0.38	467	245	117
P2	P	0.85	0.70	418	265	129
P2	PIO	0.66	0.25	572	234	105
P3	P	0.45	0.41	420	227	71
P3	PIO	0.47	0.32	579	257	95
P4	P	0.54	0.46	657	217	112
P4	PIO	0.67	0.21	517	261	142
P5	P	0.63	0.53	331	262	132
P5	PIO	0.45	0.31	325	253	120
P6	P	0.51	0.45	704	230	84
P6	PIO	0.40	0.14	433	230	88
P7	P	0.68	0.63	530	197	63
P7	PIO	0.41	0.32	444	217	75
Average	PIO	0.54	0.27	476	242	106
Average	P	0.65	0.57	504	241	104
Average	P + PIO	0.60	0.42	490.43	241.79	105.07

that an explanation for this can be found in the participants' information behaviour. Actually, working with online-learning techniques was observed to have a positive impact in terms of efficiency gains, but only when the time used by post-editors to search information is not taken into account (*ibid.*). Thus, it is evident that overall task completion times might not be a good indicator of performance when the post-editor needs to conduct informational searches to verify the quality of and improve the SMT system output. Now, even though participants did not become faster in terms of task times, their keyboard activity, as reflected in *Mins* and in particular in *Mdel* values, shows that post-editors had to type less when post-editing with interactivity and online learning techniques (condition PIO) as opposed to doing traditional post-editing (condition P). This means that online-learning techniques may help post-editors to save some effort during their work: Participants working under the P condition deleted 0.65 keystrokes and inserted 0.57 keystrokes on average per source text (ST) character. However, in the PIO condition, they inserted 0.54 keystrokes and deleted 0.27 keystrokes per ST character on average. Thus, a comparison of keyboard activity in both conditions shows that there was a decrease in the number of insertions and deletions in the PIO condition. Since both texts were comparable in size and translation difficulty, this decrease in technical effort (i.e., typing activity) must be attributed to the expected benefits of online-learning techniques during the post-editing process. See also Chap. 3 for similar findings.

This being said, based on this data alone, one cannot explain the fact that there were no significant benefits in terms of efficiency gains when overall task times are considered. Preliminary observations of screen recordings of all post-editing sessions pointed to the fact that participants often double-checked, in various informational resspurces, solutions proposed by the SMT system, even when

those solutions had been populated throughout segments by the machine-learning technique implemented for the PIO condition.

The present chapter deals primarily with the second main goal of the CFT14 experiment: the investigation of post-editors' interaction with informational tools and, in particular, with CASMACAT's built-in BiConc tool. The following sections describe the methodology and the results of this pilot investigation.

7.2.2 Methodology

For the purposes of this pilot TII study, only the segments in which CFT14 participants used the CASMACAT BiConc tool were examined. By using the BiConc, post-editors were able to retrieve information such as term equivalents and collocations (see Fig. 7.1 in the introduction Sect. 7.1), which would guide them in making an informed decision while solving a particular translation problem. The BiConc's search results are sorted by their relative frequencies (i.e., the most probable translations are shown first) based on the training data available in CASMACAT.

Using the CFT14 log files (i.e., the "event.xml" files), a script using the Cygwin[8] terminal was run to extract data about the post-editing segments where the BiConc tool was used at least once. A total of 55 instances of BiConc use were found. For each one, the script provided us with the following data: *Event ID* (i.e., information on the participant's identity, the text number and the post-editing condition; e.g., "P01_P2" (meaning "Participant 1, post-editing condition P, text 2"); *segment ID* (e.g., "10804"); and *token(s) searched* (e.g., "autonomic"). With the segment ID in hand, it was then possible to extract, from the CFT14 log files, the source segment (i.e., the original segment in English), the raw SMT output, and the participant's final target (i.e., the final segment in Spanish after the entire project was saved). An MS Excel spreadsheet was created to store and analyse these data. For each one of the 55 instances found, the data was stored in columns as follows: Event ID, segment ID, token(s) searched, source segment, raw SMT output, and final target segment.

The core of this pilot investigation was the examination of screen-capture videos.[9] Thanks to these videos, it was possible to observe and analyze the way translators interacted with the BiConc tool (and other external informational resources) in order to solve a given problem while post-editing those segments. Additional columns were then added to the Excel spreadsheet to store data such as information relevance (see *Experimental results and analysis* below); the

[8]The Cygwin package is available at: www.cygwin.com

[9]The videos are available in .fbr format in the following address: http://bridge.cbs.dk/field_trial3/VIDEO/. While playing the files, it is necessary to forward the video to the specific segment being analyzed. The segment ID can be seen on the left hand side of the CASMACAT user interface.

external informational resources used, if any; and notes (i.e., other observations and hypotheses, some of which are reported in Table 7.2 in the following section).

7.3 Experimental Results and Analysis

A first glance at the dataset allowed us to notice that only three out of the seven participants in the CFT14 study made use of the BiConc; event IDs only showed activity for participants P1, P3 and P7 in both post-editing conditions (i.e., P and PIO).[10] P7 carried out the most searches in the BiConc (24 in total); P3 carried out 20, and P1 carried out 11 searches. Figure 7.2 illustrates the use count of the BiConc per participant and per condition.

It is worth noting that participants who did not use the BiConc were also the ones who reported using fewer external resources overall. Also, among the reasons for not using the BiConc, participants P2, P4, P5 and P6 reported in the questionnaire that they forgot that they had this possibility and only used those informational tools with which they were already familiar (see Chap. 5).

It can also be observed that participants who did use the BiConc made it in both P and PIO conditions, but with a considerable difference between them. P1 and P7 used the BiConc fewer times in PIO. Generally, this could be attributed to the fact that successful searches[11] followed by edits in the text resulted in improved SMT outputs, since solutions approved by the user are populated throughout segments thanks to the online-learning technique implemented. However, P3 shows the opposite search pattern, with many more searches in the PIO condition. To find an

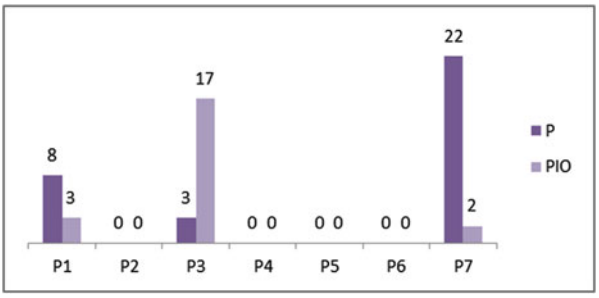


Fig. 7.2 Use count of BiConc tool per participant and per condition

[10]By examining the videos, it was possible to notice that the BiConc tool was not accessible to P4 in neither condition (i.e. that the BiConc tool button did not appear on the CASMACAT interface). The reason for this issue is unknown. Thus, only half of participants who had access to the tool actually made use of it.

[11]The notion of information relevance will be discussed below.

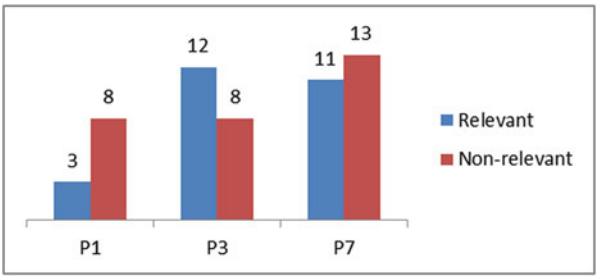


Fig. 7.3 Information relevance measurements per participant

explanation for this difference, it is necessary not only to look into the experimental design (see Sect. 7.2.1) but also to closely examine the screen recordings for P3's post-editing process. P3 post-edited text 1 in PIO making use of the BiConc but with few cases of successful information retrieval (see Fig. 7.3), which seems to have affected her confidence in the BiConc when post-editing the second text (under the P condition), where she still made a fair number of searches during the post-editing task, but preferred external resources over the CASMACAT built-in BiConc tool.

In addition to the number of times post-editors actually used the BiConc, it was also important to investigate the number of times such searches led to successful cases of information retrieval. This can be associated with the concept of *relevance*, extensively discussed in the HII literature.[12] As pointed out by Fidel (2012, p. 26), the evaluation process is almost always necessary when retrieving information (from digital information systems). Indeed, once information is acquired, a person examines and evaluates that information to discern what is relevant (and what is not) to the particular problem they are trying to solve.

Determining information relevance has been considered a monumental, complex endeavour, primarily because the judgement of relevance can be both subjective and dynamic (*ibid*., pp. 27–32). As this challenge is being acknowledged, it is argued that, as far as this chapter is concerned, the assessment of information relevance is based merely on whether or not the information found in the BiConc tool by the post-editor was the information used[13] to solve the problem at hand (in other words, if the information found was the information kept in the final target text, when the entire project was being approved and saved).

[12]While relevance has been mainly associated with the performance evaluation of information systems, it has also been associated with the human processes that take place when people determine how relevant a piece of information is, and the elements that shape these processes (Fidel 2012, p. 27).

[13]As it can be observed in the screen videos, post-editors may "use" the information found in different ways: they can copy/paste it, or they could type it into the post-editing interface, for instance.

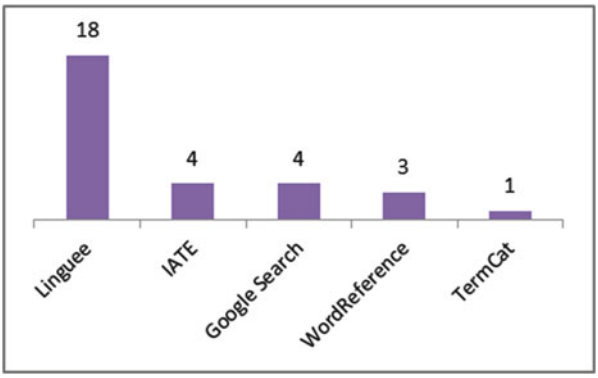


Fig. 7.4 Use count for external informational resources

A close examination of the search queries and results (and final target segments) reveals that the relevance rate varies among participants, with participants P1 and P7 having more relevant than non-relevant cases (see Fig. 7.3). On average, only 47 % of the BiConc searches (26 out of 55) provided participants with satisfying results (i.e., post-editors kept this information in the final target segment).

Furthermore, having a close look at the screen recordings, it can also be observed that participants who used the BiConc also used other Internet resources, such as term banks, dictionaries and corpora, to complement their information retrieval efforts. In addition to the CASMACAT BiConc, P1, P3 and P7 also searched information using Google (a search engine); Linguee (a biconcordancer); WordReference (a bilingual dictionary); and IATE and TermCat (terminology databases).

Remarkably enough, as shown in Fig. 7.4, for cases where the BiConc was used, the most frequently used external resource was another biconcordancer tool (i.e., Linguee), which was also observed to be extensively used by all CFT14 participants throughout the entire experiment.[14] These results reveal the extent to which today's professional translators favour this type of parallel bilingual resource over any other type of tool (e.g., dictionaries or term banks), complementing thus the results of some recent studies in the same vein (cf. Simard (2013); Daems et al., Chap. 6 in this volume). Hence, it can be concluded from this pilot investigation alone that online biconcordancers need to be constantly enhanced and improved, and integrated more tightly into translation environments.

It is worthwhile noting at this point that quantitative data (e.g., use counts and relevance measurements) may not be sufficient to investigate the TII process and

[14]This observation is based on screen recording examinations (i.e., by looking at the videos it was possible to observe that this particular resource was extensively used by all post-editors throughout the experiment). However, no exact figures on the total use of external resources in the CFT14 are available. Logging software such as Inputlog (Leijten and Van Waes 2013) will be included in future investigations.

Table 7.2 Some information behaviour noted during the BiConc tool use analysis using screen recordings

1. Having found a useful solution in the BiConc, the translator uses external resources to double-check information; results turn out to be the same
2. Although a good solution was proposed by the BiConc, the translator opts for a solution coming arguably from their cognitive background
3. Having searched for a term in both the BiConc and an external resource, with no results, the translator opts to leave notion implicit
4. The BiConc was used only after a query in an online bilingual dictionary yielded no results
5. When typing a good solution provided by the BiConc, the interactive post-editing system automatically inserted another good solution. The translator opts to keep the latter
6. A good solution was found in the BiConc, but translator made an adaptation, based on the information found
7. The translator copied/pasted a solution from BiConc. The font format from the BiConc (type, size and color) was kept in the text field; then, the translator opened a text editor to convert text into plain text, and copied/pasted it again into the CASMACAT environment
8. The translator searched both in the BiConc and in external resources, but solution provided by the BiConc was preferred
9. The BiConc took several seconds to display results; the translator could not wait and searched in an external resource; when returning to the CASMACAT workbench, the results of the query in the BiConc was being shown and turned out to be the same as found in the external tool

evaluate the quality of information and information systems. Thus, it becomes necessary to further examine, through different data-collection methods such as input logging, screen recording, eye-tracking, active observations with video recordings, and interviews, different forms of behavior; and to formulate more-detailed hypotheses about the TII processes and the usability of information tools. For example, other translator-information behaviour observed in the CFT14 screen recordings is compiled in Table 7.2[15].

The list in Table 7.2 is neither exhaustive nor objective. The ability to describe this behavior may depend, for instance, on the researcher's own perception and valuation of the post-editing process and quality, and on their particular research goals. Likewise, to add to this list of observations, other-data collection methods, as mentioned above, would need to be combined with screen-recordings in order to triangulate the data and provide more-detailed analyses of post-editors' information behaviour. For instance, can eye-movement recordings provide an insight into the cognitive processes that take place when a translator opts to leave a certain notion implicit (see observation 3 in Table 7.2) or when she chooses a solution that does not come from any of the sources consulted (observation 2)? Can external video recordings provide information on the participants facial or physical reactions when interacting with the system (i.e. on the physiological usability of the system; see

[15]This behaviour can be attributed to one, two or all three participants who made use of the BiConc tool.

Hornbæk (2006)) (see observations 5, 7 and 9, for instance) and browsing the different sources of information (observation 1), and on why a certain piece of information appears to be relevant or not (observation 8)? Can we learn, through interview questions, why does the post-editor prefer some tools and resources over others (observation 4 in Table 7.2, and Fig. 7.4), or why would they prefer making an adaptation, or why is a piece of information inappropriate to solve a given problem (observation 6)? Hornbæk (2006) describes how different sources of data and an analysis of the relations between the different aspects of usability (efficiency, effectiveness and user satisfaction), and between subjective and objective measures, could provide a wider picture of the usability and the quality-in-use of a system or a system's feature.

For the purposes of this TII study, data triangulation would have been ideal, but was not possible given the scope and limitations of a pilot experiment and of the TDA project, as stated in the introductory section. In future experiments, these combined observations and further analyses will inform researchers, for instance, about certain preferences of individuals or about the cognitive processes involved in translation and information-retrieval tasks, or about technical problems with the workbench's user interface (see observation 7 in Table 7.2) or with the system as a whole (see observation 9).

Lastly, it would be very appropriate, from a usability point of view, to design and carry out longitudinal studies where the learning effect over a period of time could be observed. Indeed, a longitudinal study with the CASMACAT workbench was carried out before the built-in BiConc tool was introduced (see Chap. 5 in this volume) and showed that over time post-editors become faster when using ITP. It would be interesting to conduct further studies of this kind to investigate how the interaction with the BiConc and other information tools can change over time, how long it takes for a user to get fully acquainted with a given tool or with a given feature of a tool, or if there is a possible trade-off between different features of a system (e.g., it would be interesting to observe if a tight integration of information tools into a translation environment and an acquaintance with the tools by the user after a certain period of time can increase the benefits of the ITP feature in terms of efficiency gains).

Having discussed the results of this pilot investigation and formulated a few areas for future work, let us now point towards new directions in TII research.

7.4 Discussion: Towards Web-Based Translation Environments

This pilot study and other CASMACAT-related experiments point towards a major area of research in TPR and translation technology: The need for a tighter integration of Internet-based informational tools and translation environments. Empirical TPR needs to pay greater attention to the study of translators' interaction

with (digital/Internet-based) information and to the optimal integration of such information into translation tools and the translation workflow.

ICTs, and particularly the Internet, have dramatically evolved over the past decades, and have led to major changes affecting not only individuals and organizations but society in general. They have made information accessibility constant, transparent and increasingly comprehensive. Indeed, the challenge is no longer to access information, but to be able to filter relevant information (Aubert et al. 2010, pp. 8–9) according to the context of use (Fidel 2012, p. 85).

The Internet is considered the informational resource par excellence, the "*El Dorado*" of knowledge (Duval 2012, p. 50). Now, the fact that it is becoming accessible practically anywhere and anytime leads humans to develop new behavior and new ways of interacting with information; of understanding, using and producing information. The Internet is arguably becoming translator's primary source for information retrieval (Borja 2008; Simard 2013). Few translators still take the time to open, even to carry along their (huge) paper dictionaries, paper term records and language books, to name only a few "traditional" informational resources. On the web, translators can find hundreds of monolingual and bilingual dictionaries, concordancers and biconcordancers, terminology databases, grammar and conjugation guides, encyclopaedia and other documentation; in sum, practically all the information that may be useful when producing a translation. Therefore, a tighter integration of these tools is necessary: further studies need to be conducted to make informational resources easily accessible, flexible, user friendly and adapted to translator's preferences and to the changing conditions of HCI. Likewise, further cognitive studies are needed to examine the impact of ICTs, particularly the Internet, on the translator's behaviour[16] and cognitive abilities, and on the translation process as a whole.

In sum, the Internet will play an increasingly important role in TII research, both for understanding translator-information behavior and for improving the quality of the information and informational tools used by translation professionals. As web-based translation environments such as the CASMACAT workbench become more and more popular and efficient, it becomes essential to conceive new, and possibly better, ways of making these environments work together well with the information translators need to carry out their tasks efficiently and effectively.

[16]Cognitive psychology studies have shown that some cognitive functions such as reading, learning and memorizing are affected by the (intensive) use of the Internet. In fact, people will turn to a search engine to search answers to even the simplest question. Just knowing that a piece of information is readily available anywhere and anytime leads humans not to memorize it (Duval 2012).

7.5 Conclusion

In this chapter, the notion of translator-information interaction (TII) was introduced as the field of study that investigates translators' interaction with information, complementing thus Sharon O'Brien's notion of translator-computer interaction (2012). To illustrate TII research, the chapter reported on a pilot study examining translators' interaction with a prototype biconcordancer (BiConc) tool integrated in the CASMACAT workbench during the third CASMACAT field trial (CFT14). A systematic analysis of such interaction was possible through screen recording observations, which allowed to look well beyond the data provided by the CFT14 log files alone. This investigation was nonetheless of exploratory nature given the scope and limitations of a pilot study and, even more importantly, the complexity inherent to TII research. The study of the interaction between humans and information is complex because it implies considering every element and every aspect of the informational work: the interaction process and the changes that result from that interaction at the level of the individuals searching for information and the tools or systems used to retrieve the information (Marchionini 2008, p. 171). It is also worth considering a possible interplay between the information provided by the various tools and the translator's cognitive background (i.e., their knowledge). The translator looks for a given piece of information they do not know or they are uncertain about. Now, when judging the quality of a suggestion by the system, trust (i.e., trust in oneself) may also play a significant role. In other words, as observed in the behaviour described in Table 7.2 in Sect. 7.3, the interaction between the post-editor's cognitive background and the information provided by the tools is potentially an interaction of trust.

With this chapter, several research questions for future TII research were raised: how well do human translators work with the information and informational tools they currently have at their disposal? How accurate, rich and relevant is the information they find? How user-friendly are informational tools for translators? How can the information and the tools be improved to maximize translators' performance and well-being, and the quality of their work? How can the performance of an existing tool be assessed and improved? What are the advantages and drawbacks of integrating an informational tool to a translation environment, as compared to having an array of external resources? These questions can only be partly dealt with in a pilot investigation like the one described here. Only a larger-scale study with a larger sample size and combining different sources of data can provide a wider, and potentially better, picture of the TII processes and the usability of information systems and tools.

From this exposition, it may be concluded that TII studies, however complex they are, will be essential in the development and the improvement of tools intended to better address the needs of translators at the digital age. In the words of, Carl et al. (2011),

> [d]evelopment of translation tools could benefit from incorporating knowledge of human translation behavior and translator styles [...]. As Knight et al. (2007)[17] point out, "the combination of [...] usability studies and cognitive modeling [may help to] make an informed decision about critical aspects of a user interface."

In the age of translation technology, mobile computing and ubiquitous information, research on TII will become increasingly important in empirical TPR. Behavioural studies that explore information interaction will play a crucial role in the design and development of new tools that are user-friendly and adapted to translators' informational needs and to the changing reality of the translation industry.

References

Alabau, V., Carl, M., García-Martínez, M., González-Rubio, J., Mesa-Lao, B., Ortiz-Martínez, D., et al. (2014). *D6.3: Analysis of the third field trial*. Technical report, CasMaCat project.

Aubert, B., Cohendet, P., & Montreuil, B. (2010). *L'innovation et les technologies de l'information et des communications*. Québec: CEFRIO.

Austermühl, F. (2001). *Electronic tools for translators*. Manchester: St. Jerome.

Balling, L. W., & Carl, M. (2014). Production time across languages and tasks: A large-scale analysis using the CRITT translation process database. In J. W. Schwieter & A. Ferreira (Eds.), *The development of translation competence: Theories and methodologies from psycholinguistics and cognitive science* (pp. 239–268). Newcastle Upon Tyne: Cambridge Scholars.

Borja, A. (2008). Corpora for translators in Spain. The CDJ-GITRAD corpus and the GENTT project. In M. Rogers & G. Anderman (Eds.), *Incorporating corpora: The linguist and the translator* (pp. 243–265). Clevedon: Multilingual Matters.

Bowker, L. (2002). *Computer-aided translation technology: A practical introduction*. Ottawa: University of Ottawa Press.

Carl, M., Dragsted, B., & Jakobsen, A. L. (2011). On the systematicity of human translation processes. In *Actes de la conférence de Tralogy*. Retrieved from http://lodel.irevues.inist.fr/tralogy/index.php?id=103

Duval, C. (2012). L'impact du Web en 4 questions. *La Recherche, 467*(September), 46–50.

Fidel, R. (2012). *Human information interaction. An ecological approach to information behaviour*. Cambridge, MA: MIT Press.

Gershon, N. (1995). Human Information Interaction. In *4th international world wide web conference.*

Harper, R., Rodden, T., Rogers, Y., & Sellen, A. (2008). *Being human: Human-computer interaction in the year 2020*. Cambridge: Microsoft Research.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human Computer Studies, 64*(2), 79–102.

Hvelplund, K. T., & Carl, M. (2012). User activity metadata for reading, writing and translation research. In V. Arranz, D. Broeder, B. Gaiffe, M. Gavrilidou, M. Monachini, & T. Trippel (Eds.), *Proceedings of the eighth international conference on language resources and evaluation. LREC 2012: Workshop: describing LRs with metadata: Towards flexibility and interoperability in the documentation of LR* (pp. 55–59). Paris: ELRA.

[17] Knight, A., Pyzark, G. & Green, C. (2007). When two methods are better than one: combining user study with cognitive modeling. In *CHI'07 proceedings* (pp. 1783–1788).

Jakobsen, A. L. (2003). Effects of think aloud on translation speed, revision and segmentation. In F. Alves (Ed.), *Triangulating translation. Perspectives in process oriented research* (pp. 69–95). Amsterdam: Benjamins.

Jakobsen, A. L. (2011). Tracking translators' keystrokes and eye movements with Translog. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research. Integrative approaches in translation studies* (pp. 37–55). Amsterdam: Benjamins.

Jakobsen, A. L., & Jensen, K. T. H. (2008). Eye movement behaviour across four different types of reading task. *Copenhagen Studies in Language, 36*, 103–124.

Karamanis, N., Luz, S., & Doherty, G. (2011). Translation practice in the workplace: A contextual analysis and implications for machine translation. *Machine Translation, 25*(1), 35–52.

Kurzweil, R. (2013). *How to create a mind. The secret of human thought revealed* (p. 336). New York: Penguin.

Langlois, L. (1996). Bilingual concordancers: A new tool for bilingual lexicographers. In *Proceedings of the 2nd international conference of the american machine translation association.* Retrieved from http://mt-archive.info/AMTA-1996-Langlois.pdf

LeBlanc, M. (2013). Translators on translation memory (TM). Results of an ethnographic study in three translation services and agencies. *The International Journal for Translation and Interpreting Research, 5*(2), 1–13.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392.

Marchionini, G. (2008). Human-information interaction research and development. *Library and Information Science Research, 30*(3), 165–174. Retrieved from http://www.ils.unc.edu/~march/Marchionini_Inf_interact_LISR_2008.pdf.

O'Brien, S. (2009). Eye tracking in translation-process research: Methodological challenges and solutions. *Copenhagen Studies in Language, 38*, 251–266.

O'Brien, S. (2012). Translation as human–computer interaction. *Translation Spaces, 1*(1), 101–122. doi:10.1075/ts.1.05obr.

Shneiderman, B. (2008). Foreword. In A. Sears & J. A. Jacko (Eds.), *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications* (2nd ed., pp. xix–xx). New York: Lawrence Erlbaum Associates.

Simard, T. (2013). *Analyse comparative de la couverture et de l'acceptabilité des solutions d'un dictionnaire bilingue spécialisé, d'une banque de données terminologiques et d'un concordancier en ligne: application au domaine de la traduction médicale*. University of Ottawa. Retrieved from http://www.ruor.uottawa.ca/fr/handle/10393/24929

Taravella, A. M., & Villeneuve, A. O. (2013). Acknowledging the needs of computer-assisted translation tools users: The human perspective in human-machine translation. *The Journal of Specialised Translation, 19*(January), 62–74. Retrieved from http://www.jostrans.org/issue19/art_taravella.pdf.

Touré, M. A., Mbangwana, M., & Sène, P. A. (2009). Que sont les TIC : Typologie des outils et des systèmes. In T. Karsenti (Ed.), *Intégration pedagogique des TIC en Afrique. Stratégies d'action et pistes de réflexion* (pp. 33–56). Ottawa: CRDI.

Part III
Modelling Translation Behaviour

Chapter 8
Statistical Modelling and Automatic Tagging of Human Translation Processes

Samuel Läubli and Ulrich Germann

Abstract Advanced translation workbenches with detailed logging and eye-tracking capabilities greatly facilitate the recording of key strokes, mouse activity, or eye movement of translators and post-editors. The large-scale analysis of the resulting data logs, however, is still an open problem. In this chapter, we present and evaluate a statistical method to segment raw keylogging and eye-tracking data into distinct *Human Translation Processes* (HTPs), i.e., phases of specific human translation behavior, such as *orientation*, *revision*, or *pause*. We evaluate the performance of this automatic method against manual annotation by human experts with a background in Translation Process Research.

Keywords Computer-aided translation • Computer-assisted translation • Post-editing • Quantitative data analysis • Translation processes • Unsupervised sequence modelling

8.1 Introduction

8.1.1 Background

Krings (2001, p. 24) once described "the construction of a machine translation system capable of translating as well as a human being" as being "more difficult to achieve than man's conquest of the moon". Nevertheless, state-of-the-art machine translation (MT) systems have nowadays reached a level of quality where their incorporation into human translation workflows significantly increases the productivity of professional translators in a post-editing (PE) set-up, where bilingual experts revise MT output rather than translate from scratch (Green et al. 2013). The

S. Läubli (✉)
School of Informatics, The University of Edinburgh, Edinburgh, UK

Autodesk Development Sàrl, Neuchatel, Switzerland
e-mail: samuel.laubli@autodesk.com

U. Germann
School of Informatics, The University of Edinburgh, Edinburgh, UK
e-mail: ugermann@inf.ed.ac.uk

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_8

increasing popularity of PE in the translation industry has enticed researchers in MT to try to find new ways of using MT to make human translation faster and less cognitively demanding. Alabau et al. (2014), for example, use interactive MT to provide translators with automatic sentence completion, and they use automatic MT quality estimation to identify and highlight parts in the MT output that are likely to be of poor quality and therefore probably need revision (see also Chaps. 3, 4, 5, and 10).

Unfortunately, the ultimate effectiveness of such approaches is difficult to evaluate. Empirical investigations have so far focused primarily on measuring the impact of MT-based productivity tools on temporal translation effort, e.g., by comparing how long it takes translators to post-edit similar texts with and without automatic sentence completion. Apart from observing changes in the *average* translator efficiency under various conditions, most of the studies, however, also note vast variance in how much *individual* translators benefited from MT support features (e.g. Plitt and Masselot 2010; Underwood et al. 2014). Moreover, the focus on temporal effort often results in neglecting the impact of such translation aids on cognitive effort and user satisfaction. For this reason, we can often tell *whether or not* particular forms of machine assistance affect human translation performance, but we know surprisingly little about *how* they affect the underlying human translation processes.

In Translation Process Research (TPR), this question is commonly approached by analysing translator activity data (TAD) from recorded translation sessions. Recordings typically include keystrokes, mouse clicks, and eye movements of translators performing a given translation task. By analysing translation logs (Elming et al. 2014), for example, found that translators spend less time on reading the source text ("orientation") when post-editing than when translating from scratch. (This, by the way, could be a partial explanation for the time savings associated with post-editing vs. translation from scratch.)

Modern, research-oriented translators' workbenches with advanced user observation and logging capabilities, such as Translog II (Carl 2012) and CASMACAT (Alabau et al. 2014) greatly facilitate the collection of TAD for TPR. The complexity and sheer abundance of information contained in recordings of translation sessions, however, make it impossible to analyse these data manually. As a consequence, TPR studies often base their analysis on heuristic aggregations and visualisations, and tend to consider only a small subset of all available data. Well aware of this limitation, Jakobsen (2011) concedes that "with the present state of technological development, it still seems relevant, perhaps even necessary, to examine small volumes of eye movement and keystroke data manually and selectively". However, he adds that "this should not prevent us from pursuing a larger goal. The potential for large-scale computational analysis of translator activity data is there, or will very soon be there, and the prospect of creating a computational model of how expert human translators typically execute their skill seems within reach."

Jakobsen's vision is what we aim at in this work: to allow large-scale analysis of translator activity data through statistical models. The goal is to (1) automatically identify human translation processes in unlabelled data from recorded translation

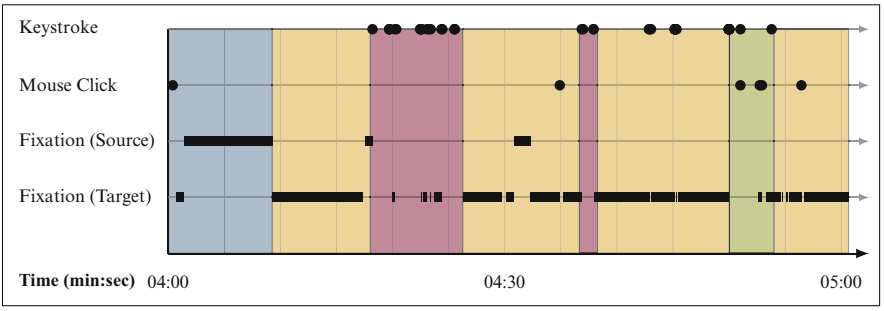


Fig. 8.1 Illustration of the basic aim of our work: learning sequence classification models based on keylogging and eye tracking data to automatically segment recorded translation sessions into sequences of human translation processes (*shaded areas*; this illustration is not based on actual data)

sessions, and (2) segment recorded translation sessions into sequences of the identified processes, thus revealing when and how often a translator executed each of them, as illustrated in Fig. 8.1. This information will enable Translation Process researchers to identify new activity patterns and regularities in massive amounts of data.

8.1.2 Approach

In line with current TPR, we assume that when translating or post-editing, human translators go through a sequence of distinct, latent, high-level *Human Translation Processes* (HTPs), such as reading the source text, reading a draft translation, revising the draft translation, etc. Translators jump back and forth between HTPs, but execute only one at any given time. Which one that is, cannot be observe directly. However, each process manifests itself in a characteristic pattern of observable behaviour: key strokes, mouse activity, eye movement, etc. (Carl and Jakobsen 2009; Carl 2010).

For modeling purposes, we assume that the probability of executing a particular HTP next is fully determined by the current HTP, so that we can model HTPs as states in a first-order Hidden Markov Model (HMM), and their characteristic patterns of observable behaviour as their "emissions".

Once a Hidden Markov Model (HMM) of the overall translation or post-editing process has been trained (Sect. 8.4.2), we use the Viterbi algorithm (Viterbi 1967) to segment and annotate raw translation activity logs with HTP labels.

Our claim is that the method proposed here makes it possible to infer meaningful translation processes from unlabelled keylogging and eye tracking data automatically.

8.1.3 Terminological Clarifications

Three terms used frequently in this chapter require explanation: *translation*, *human translation process*, and *translation log*.

First, within the context of this work, translation subsumes, and often specifically means post-editing.[1] Although this chapter is titled "Statistical Modelling and Automatic Tagging of Human Translation Processes", it largely focuses on modelling the behaviour of post-editors and thus, strictly speaking, on a specific form of translation. The statistical modelling approach proposed in Sect. 8.4, however, is not limited to post-editing.

Second, we refer to the high-level processes involved in both translation from scratch and post-editing as *human translation processes* (HTPs). Specifically, this means processes such as orientation (i.e., getting acquainted with the source text or MT output) or revision (i.e., adapting the target text). We clearly distinguish these from *translation actions*, which result from executing these processes, such as pressing a key or looking at a word. The distinction between (latent) *human translation processes* and (observable) *translation activities* is grounded in recent TPR literature (e.g., Carl 2010). In the research literature, these concepts are also often referred to as *cognitive* vs. *technical translation processes* (Krings 2001), or *translation events* vs. *translation acts* (Toury 1995).

Finally, by *translation log* we mean the record of all translation actions observed during a translation session.

8.2 Foundations

8.2.1 MT Research and Post-editing

While post-editing of automatically produced draft translations is becoming more and more wide-spread as a standard mode of operation in professional translation, post-editing does not work equally well for everyone. Plitt and Masselot (2010), for example, found that post-editing increased the translation throughput by 131 % for their fastest, but by only 20 % for their slowest translator. Similar studies (Guerberof 2009; Green et al. 2013) also report considerable variance in the benefits of post-editing over translation from scratch (with a positive effect on average).

A number of factors have been identified that might contribute to this variance: translation direction (Plitt and Masselot 2010), translators' professional working experience (Plitt and Masselot 2010; Green et al. 2013), text types (Läubli et al. 2013), text difficulty (Green et al. 2013), and MT quality (Koehn and Germann

[1] We acknowledge that this view is not shared by everyone. For example, Moritz Schaeffer (personal communication) argues that post-editing should be considered a task/skill distinct from translation per se. Not every effective translator is an effective post-editor, and vice versa.

2014), and others. But as both MT and industrial research are primarily interested in *whether* and *how much* rather than *why* and *how* post-editing accelerates the translation process with a certain MT system or in a particular working environment, the impact of the aforementioned factors on the actual *performance* of post-editing is hardly ever investigated in detail. As a result, surprisingly little is known about why post-editing is faster than translation from scratch, and why certain translators benefit more from it than others.

With regard to exploring the effects of post-editing in more detail, the aforementioned studies suffer from their focus on post-editing time alone. In a study that investigated the cognitive effort involved in post-editing, Koponen et al. (2012) recorded not only the time, but also the keystrokes of eight post-editors. They conclude that, while post-editing time in relation to segment length can be a good indicator of cognitive effort, recorded keystrokes are "very useful as a way to understand how translators work", and that "studying the strategies of different post-editors can be potentially very useful for post-editing practice." Here, MT research comes into contact with translation process research.

8.2.2 Translation Process Research and Post-editing

Translation Process Research (TPR) is a branch of descriptive translation studies that investigates the underlying cognitive and mental processes rather than the products resulting from human translation. In contrast to theoretical (or normative) translation studies, TPR is grounded in "(observable and reconstructable) facts of real life rather than merely speculative entities resulting from preconceived hypotheses and theoretical models" (Toury 1995, p. 1). In a nutshell, TPR aims at understanding translation through observation.

The first large-scale TPR study on post-editing was conducted by Krings (1995, 2001). Observing translation students and a number of professional translators adapt machine-translated product manuals in a paper-and-pencil setting, he developed a fine-grained taxonomy of the processes involved in post-editing. Krings was particularly interested in assessing the effort involved in post-editing processes, which he factored into technical and cognitive effort (cf. Fig. 8.2). While the author found technical effort to be directly observable—Krings used two video cameras

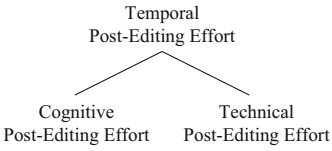


Fig. 8.2 Differentiation of fundamental post-editing effort factors as suggested by Krings (2001)

to record the post-editors at work—he relied on think-aloud protocols[2] (TAP) to determine cognitive post-editing effort. This enabled Krings not only to characterise but also to quantify the post-editing processes he identified. For example, he found that "fifteen percent of the processes observed [in post-editing] were physical writing processes, 12 percent were target text evaluation processes", etc. (Krings 2001, p. 529).

The use of TAPs is controversial in TPR (Toury 1995, p. 235; Jakobsen 2003). As an alternative, Jakobsen (1999, 2003) proposes to record translators' keystrokes while they are working. Unlike thinking aloud, he argues, keystroke logging is unobtrusive to translators and, albeit "no substitute for the information that can be elicited through think[ing] aloud" (Jakobsen 2003), it enables meaningful characterisations of translation or post-editing processes. For example, pauses in typing activity might be indicative of cognitive processing (O'Brien 2006; Koponen et al. 2012; cf. also Sect. 8.2.1). Furthermore, O'Brien (2007) proposes to record translators' eye movements and pupil dilation by means of eye trackers, based on Just and Carpenter's (1980) eye–mind hypothesis: the fundamental assumption that what the eyes are focussing on is what the mind is attending to. Research tools such as Translog-II and CASMACAT nowadays greatly facilitate the recording of translator activity data (TAD) such as keystrokes and eye movements in TPR studies, making TAD-based experiments the predominant paradigm in current translation process research (Krings 2005; Carl et al. 2014).

While the collection of TAD from translation or post-editing sessions is now fairly straightforward with the aforementioned tools, analysing the raw data resulting from such recordings is still tedious and difficult. The TAD from a recorded translation session normally consists of thousands of keystrokes and eye fixations with many specific attributes (cf. Sects. 8.3.2 and 8.5.1.1). This makes it impossible to fully analyse and compare all raw data resulting from a TPR study manually. TAD-based TPR studies thus often focus on analysing small subsets of the available data by means of aggregation and visualisation.

As for aggregation, Carl (2010) suggests two concepts for representing basic translation actions: fixation units (FU) and production units (PU). A FU is defined as a sequence of two or more eye fixations on the source text such that the time interval between any two fixations does not exceed a given threshold. Similarly, a PU is a sequence of two or more keystrokes with no time interval between any two successive keystrokes exceeding a given threshold (cf. also Chap. 2, Sects. 5.3 and 5.4). The motivation for working with FUs and PUs is the assumption that "a lapse of time of more than [a given threshold] indicates a shift in the translators [sic] mind to another textual unit to be translated" (Carl 2010). Carl and Kay (2011) show that the number of PUs strongly correlates with translation time when using a threshold of 800 ms. However, at the same time, they also show that the choice of threshold values has a considerable impact on how a translation session is segmented. This exposes a weakness associated with aggregating TAD into FUs

[2] Also referred to as thinking-aloud protocols (Toury 1995) and talk-aloud protocols (Gerloff 1986).

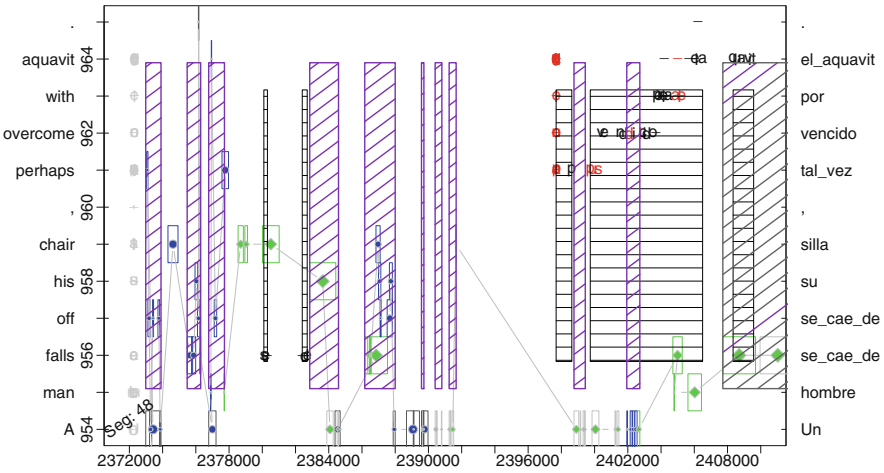


Fig. 8.3 Translation progression graph of an excerpt from a recorded post-editing session

and PUs: although using a certain threshold can be motivated empirically (Carl and Jakobsen 2009; Carl and Kay 2011), any fixed threshold does not take variance between individual translators into account. The PU threshold to identify coherent sequences of writing, for example, would need to be higher for slower than for faster typists.

Apart from analysing FUs and PUs quantitatively, many TPR studies are based on qualitative analyses of visualised TAD. In particular, translation progression graphs (Perrin 2003) are used to visualise text insertions and deletions, eye fixations, and also FUs and PUs in recorded translation sessions over time. Aggregations and visualisations enable TPR experts to identify and characterise human translation processes (HTPs) in TAD, but the use of translation progression graphs in particular limits TPR studies to small subsets of the available data. The graph shown in Fig. 8.3, for example, visualises 36 s from a post-editing session with a total duration of 45 min. Furthermore, one could argue that the visual analysis of translation progression graphs by an expert involves a considerable amount of data interpretation and is thus by no means objective. Nevertheless, TPR studies based on analyses of FUs and PUs and/or progression graphs have yielded valuable findings about the processes involved in human translation and post-editing (see for example Hvelplund 2011; Dragsted and Carl 2013; Massey and Ehrensberger-Dow 2014; Schaeffer and Carl 2014, to name but a few).

Martínez-Gómez et al. (2014) proposed the first unsupervised approach to translator modelling, showing that automatically identified activity patterns in the TAD of recorded translation sessions can be used to predict indicators of a translator's expertise.

In the remainder of this chapter, we will show that activity patterns inferred through unsupervised learning can not only be used to make predictions, but also

to inform the descriptive analysis of recorded translation sessions in what we will refer to as data-driven TPR. The modelling technique for HTPs that we propose in Sect. 8.4 is based on the same foundation as Martínez-Gómez et al.'s (2014) method (i.e., k-means clustering), but extends it through incorporating a notion of sequentiality, namely that the classification of observable translation actions under a specific HTP—such as orientation or revision—does not only depend on these observations, but also on the preceding HTP. In Sect. 8.5, we show that models trained in this way are capable of identifying meaningful translation processes in TAD without prior assumptions or human intervention.

8.3 Data Collection

8.3.1 Recording Translation Actions with the CASMACAT *Workbench*

The translation logs used in our experiments were collected with the CASMACAT workbench, an interactive translation interface jointly developed by experts in MT and TPR in a 3-year project within the European Union's Seventh Framework Programme (for details, see Chap. 3).

One of the main advantages of the CASMACAT workbench over conventional TPR tools such as Translog II is that it provides a more realistic translation environment. As Läubli et al. (2013) point out, experiments aimed at assessing post-editing efficiency often isolate participating translators from essential productivity support such as translation memories, thus compromising their ecological validity. The same holds for TPR studies conducted with research-oriented software such as Translog II. Dragsted and Carl (2013), for example, note that the participants of their study "found themselves in an unusual situation in a lab working with programmes which were unfamiliar to them and without their usual aids and tools." The CASMACAT platform, in contrast, is a state-of-the-art translation workbench[3] and as such arguably provides a more ecologically valid means of gathering TAD from working translators and post-editors.

CASMACAT can record keyboard and mouse activity, as well as, in combination with an eye tracker, eye movements and gaze fixations. The outcome of recording a translation session is TAD that links the source and target texts with activities performed to analyse, create, or modify them (Carl and Jakobsen 2009).

[3]CASMACAT is based on the web-based MATECAT workbench, which is deployed and actively used in production at several translation and IT companies (Federico et al. 2014).

8.3.2 *The* CASMACAT *Field Trial 2014*

In our experiments, we used translation logs from the 2014 CASMACAT "field trial[4] (CFT14; see also Chap. 10). Seven professional post-editors carried out two tasks each, under two experimental conditions: traditional post-editing (PE) and post-editing with interactive translation prediction based on MT with online learning (PIO; see Chinea-Rios et al. 2014; Sanchis-Trilles et al. 2014). Each task consisted of post-editing a text of roughly 4500 words from the medical domain (patient information leaflets), which had been or was being automatically translated from English into Spanish. All of the tasks were carried out under experimental conditions involving an eye tracker, resulting in 14 recorded translation sessions with logged keyboard, mouse, and gaze actions.

From these, we used the sessions produced under the traditional PE condition for our experiments. They are between 96 and 204 min long (mean: 144 min; standard deviation: 37 min) and contain, on average, 18,515 keystrokes, 696 mouse clicks, and 9794 eye fixations on the source and target texts.

8.3.3 *Creating Gold Standard Annotations*

The central hypothesis of our work is that HTPs automatically inferred from data are in correspondence with HTPs that are well-known in TPR. The statistical models described in Sect. 8.4 allow segmentation of recorded translation sessions into HTP phases. To validate our hypothesis, we wanted to compare the output of the automatic segmentation process with manual segmentations of TAD into HTPs by human experts (cf. Sect. 8.5.2). To this end, we compiled a collection of manually annotated excerpts from recorded PE sessions.[5] The annotation was based on video replays of the CFT14 PE sessions: we used the CASMACAT workbench's replay function to recreate and record as video post-editing sessions from existing translation logs. The resulting videos show what the translators were typing and what they were looking at. The annotation task comprised seven excerpts of 5 min each (A–G), taken from five CFT14 PE sessions. All excerpts were split into short video sequences of 3 s (henceforth: "snippets").

We asked 12 annotators (A1–A12) to identify which HTP a translator was executing in each of the snippets by choosing the most appropriate label from a TPR taxonomy of post-editing processes.

[4]The data are available from http://sourceforge.net/projects/tprdb/.

[5]The data are available from http://www.casmacat.eu/?n=Main.Downloads.

8.3.3.1 Annotators

We engaged 12 participants (six male, six female) in the annotation task. Participants were aged between 22 and 38 years (mean: 27.8) and pursued a master's (7) or Ph.D. (5) degree. All of the participants were familiar with at least the foundations of translation and post-editing processes. Eleven annotators stated that their previous (8) and/or current (11) degree programme was related to translation process research or translation studies, and six had experience working as professional translators. Participants were reimbursed EUR 40.00 for classifying 825 segments and completing two short surveys, which took them roughly 3.5 h in total.

8.3.3.2 Annotation Procedure

Each participant carried out nine tasks using a purpose-built browser-based annotation interface:

- a test task of 25 snippets, meant to get annotators acquainted with the classification process. The labels assigned in this task were not stored and are thus not evaluated;
- eight tasks of 100 snippets each, that is, each of the seven excerpts A–G, in random order. Session B appeared twice in order to measure each participant's intra-annotator agreement.

8.3.3.3 Tagset

All annotators were asked to classify each snippet as one of the following six HTPs:

- **Orientation: source text** (Os)
  The translator is reading without inserting or deleting text, mainly focussing on the source text.
- **Orientation: target text** (Ot)
  The translator is reading without inserting or deleting text, mainly focussing on the target text.
- **Orientation: source and target text** (Ost)
  The translator is reading without inserting or deleting text, focussing on both the source and the target texts.
- **Revision: linear** (Rl)
  The translator is editing the target text. Every word is edited only once, in linear order.
- **Revision: scattered** (Rs)
  The translator is editing the target text. Some source words are edited several times.

– **Pause** (P)
  The translator is idle, e.g., because he or she is waiting for the interface to respond.

This tagset was designed to characterise different phases of the post-editing process and is described in detail in Chap. 14. We chose it mainly due to its clarity and simplicity. Krings (2001) and others (cf. Sect. 8.2.2) provide more fine-grained taxonomies to characterise post-editing processes, but we considered them to be less suitable for the classification task at hand since the differences between class definitions are often very subtle. For example, Krings distinguishes between reading an entire text or sentence ("SOURCE/READ") and giving direct attention to an element (usually a word; "SOURCE/FOCUS") of the source text (Krings 2001, p. 514f). Using the six classes defined by Oster, we were hoping to ensure that annotators easily understand the classification task as well as reproduce their own annotations, which was confirmed by a post-experimental survey and (mostly) high intra-annotator agreements (cf. Sect. 8.3.3.4), respectively.

8.3.3.4 Intra- and Inter-Annotator Agreement in Human Annotation

Once all annotators had completed their work, we assessed both intra- and inter-annotator agreement using Fleiss' Kappa (Fleiss 1971), which measures the extent to which annotators' agreement on the classification of items exceeds the agreement that would be expected by chance:

$$\kappa = \frac{\text{observed agreement} - \text{chance agreement}}{1 - \text{chance agreement}} \tag{8.1}$$

A κ-value of 0 means that there is no agreement beyond chance level at all, while 1 indicates perfect agreement between all annotators at all times. κ values are meaningful for relative comparisons such as which classes annotators agree on more or less often in a classification task. However, absolute κ values are generally not comparable across different tasks and datasets, particularly if different numbers of annotators are involved (Sim and Wright 2005).

All of the 12 annotators classified the same seven sessions A–G. Consequently, there are 12 classifications for each of the 100 snippets per session. We chose as the gold standard annotation for each snippet the label that the majority of annotators assigned to it; in case of a draw, we used the label that most recently appeared in the session's gold label history. The intra- and inter-annotator agreement values are shown in Tables 8.1 and 8.2 and are calculated for three and six classes. The six classes mode reflects agreement using the full tagset described in Sect. 8.3.3.3. In the three classes mode, we merged the subclasses for orientation and revision, respectively, i.e., $\{\text{Os}, \text{Ot}, \text{Ost}\} \rightarrow \text{O}$ and $\{\text{Rl}, \text{Rs}\} \rightarrow \text{R}$. The agreement scores for this mode thus reflect how well annotators could distinguish between the basic HTPs

Table 8.1 Intra-annotator agreement (κ) in the CFT14-Gold dataset over three and six HTP classes

Ann.	3 Cl.	6 Cl.
A7	0.94	0.89
A3	0.95	0.87
A1	0.83	0.86
A11	0.80	0.68
A4	0.86	0.67
A5	0.84	0.65
A10	0.79	0.64
A9	0.85	0.59
A12	0.76	0.59
A2	0.73	0.56
A8	0.64	0.42
A6	0.54	0.31

Records are sorted in descending order by intra-annotator agreement over six classes

in screen recordings of post-editing sessions: orientation (O), revision (R), and pause (P).

Table 8.1 lists the intra-annotator agreements for each participant based on 100 doubly-classified snippets. The scores indicate that (1) most annotators were rather consistent in labelling the snippets using the full tagset, with A6 and A8 being the exceptions, and (2) that—not surprisingly—nearly all annotators performed better when only three basic classes were distinguished. Apparently the differences between linear (Rl) and scattered (Rs) revision, and between the three subclasses of orientation (Os, Ot, and Ost) are too subtle for human annotators to consistently tell them apart.

This conjecture is supported by the inter-annotator agreement scores shown in Table 8.2. Annotators showed high agreement in distinguishing between the three basic classes O, R, and P. Among these, they agreed least on identifying pauses (P), but even here, nearly nine out of twelve annotators agreed on average ($\bar{\kappa}$: 0.73). In contrast, the fine-grained distinctions between the subclasses of orientation and revision were a lot more controversial. This holds for scattered revision (Rs; $\kappa = 0.15$), but also for linear revision (Rl; $\kappa = 0.43$) and, to a lesser extent, source text orientation (Os; $\kappa = 0.49$). The scores indicate that it is particularly difficult to differentiate between scattered and linear revision. This was confirmed by the annotators in a post-experimental survey.

Table 8.2 Number of snippets, inter-annotator agreement (Fleiss' κ), and mean agreement with gold standard label ($\overline{\kappa}$) per class with three and six classes

Three classes	#	κ	$\overline{\kappa}$	Six classes	#	κ	$\overline{\kappa}$
O	202	0.65	0.87	Os	16	0.49	0.75
				Ot	81	0.54	0.79
				Ost	104	0.70	0.81
R	452	0.76	0.93	Rl	406	0.43	0.72
				Rs	41	0.15	0.62
P	46	0.51	0.73	P	51	0.51	0.71
All	700	0.68	0.90	All	700	0.45	0.73

8.4 Statistical Modelling of Translator Behaviour

As mentioned earlier, our model assumes that translators go through a sequence of latent high-level HTPs to solve the translation task: orientation, revision, pauses, etc. They switch back and forth between them, but they execute only one at any given time. While there is no way to determine unobstrusively[6] which state they are in at any time during the process, but the underlying states manifest themselves in specific patterns of observable physical behaviour such as keystrokes, eye movements, etc. (Just and Carpenter 1980).

8.4.1 Observations

The HTP's characteristic patterns can be described in terms of the (1) type, (2) frequency, and (3) combinations of observable actions that they trigger. For example, orientation phases typically comprise multiple eye fixation and few if any keystrokes (the latter e.g. for navigation); revision involves fewer or shorter eye fixations and more keyboard activity. Moreover, the co-occurrence of actions also provides hints at what a particular observation was triggered by. Pressing the backspace key within a sequence of many alphanumeric keystrokes, for example, suggests that the translator is currently probably drafting. If the backspace key is however pressed once or twice after a mouse click, this suggests that the translator is currently revising.

In order to capture such patterns of co-occurrence, we slice the translation log evenly into short time windows and count for each event type the number of occurrences in each time interval. Each time window can thus be represented as

[6]Of course we could ask them, but that would interrupt precisely those mental processes that we want to eavesdrop on and force the translator to reflect on what might otherwise be a subconscious, automatic process. This is one of the main arguments against think-aloud experimental protocols.

a vector of counts, where each dimension of the observation vector corresponds to translation actions of a particular type.

The identification of the underlying states as 'orientation', 'revision', etc. is, by the way, post-hoc and not part of the model itself—our method is entirely data-driven and does not require a-priori knowledge, information or assumptions about the different HTP types (e.g., that there is a revision process where we would expect deletions and mouse clicks). Only a few parameters (e.g., the number of hidden states) are chosen a-priori and not determined during the training process itself, but even their settings can be evaluated empirically, as discussed in Sect. 8.5.1.

8.4.2 Building a Model

Model building works as follows. Given the complete set E of all user actions recorded during a translation session, we first define a set of *relevant* actions $E' \subseteq E$. In the work presented here, we consider three different sets of relevant actions:

E'_1: keystrokes (`<keyDown>`);
E'_2: keystrokes and mouse clicks (`<keyDown>`, `<mouseDown>`); and
E'_3: keystrokes, mouse clicks, and eye fixations (`<keyDown>`, `<mouseDown>`, and `<fixation>`).

For keystrokes, we furthermore distinguish between

- deletion: DEL and BACKSPACE keys (`<keyDownDel>`);
- control: CTRL, ALT, and WIN keys (`<keyDownCtrl>`);
- navigation: ARROW UP, DOWN, LEFT, and RIGHT keys (`<keyDownNav>`); and
- all other (mostly alphanumeric) keys (`<keyDownNormal>`).

For eye fixations, we differentiate

- fixations on the source text (`<fixationSource>`); and
- fixations on the target text (`<fixationTarget>`).

Once the subset of relevant actions is defined, we segment the recorded translation sessions into consecutive time windows of equal length w, as shown in Fig. 8.4b. For each window, we count the number of occurrences per action. Each window forms an observation (i.e., a feature vector), with the action types being its dimensions and the number of occurrences per action being the respective values, as illustrated in Fig. 8.4c. For example, during the first 5000 ms, we observe 0 keystrokes, 2 mouse clicks, 11 source text fixations, and 3 target text fixations, producing the feature vector $\langle 0, 2, 11, 3 \rangle$. The output of the feature extraction process for an entire recorded translation session of n consecutive time windows is thus a sequence of n feature vectors $O = o_1, o_2, \ldots, o_n$.

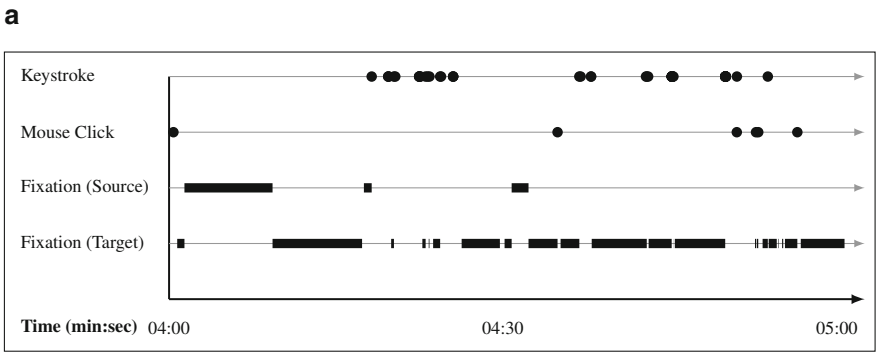


Fig. 8.4 The feature extraction process. Note that the keystroke event types (deletion, control, navigation, and alphanumeric) have been folded into a single event type in this illustration. (**a**) Define a subset of relevant action types. (**b**) Parametrise the recorded translation sessions into time windows of equal length. (**c**) Count the number of occurrences per action type in each window. The counts in each window form an observation, i.e., each column corresponds to a feature vector

8.4.3 Unsupervised Sequence Modelling

We assume that each feature vector is randomly "produced" or "emitted" by the underlying HTP, and that there is a certain amount of variance in the feature values for each HTP class. For ease of modeling, we approximate the emission probability by a mixture of m Gaussians (GMM)[7] per state and action type. To obtain initial translation action emission probabilities, we pool all observations contained in the training data together—regardless of their origin and order—and use the `k-means++` algorithm (Arthur and Vassilvitskii 2007) to cluster them into k classes, where k is the number of underlying HTPs (hidden Markov states). This parameter will be empirically optimised (see Sect. 8.5.1).

Our model is thus a fully connected GMM-HMM as illustrated in Fig. 8.5. Assuming that each of the hidden states will correspond to a HTP, we initialise the observation probability densities with the means and covariances of the observations assigned to the equivalence classes obtained in the k-means clustering step, such that each equivalence class initialises the probability density of a HMM state. The transition probabilities are initialised uniformly, such that the model assigns the same likelihood to the transition between any two HTPs at this point.

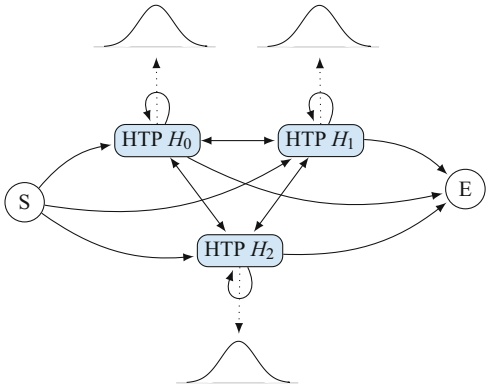


Fig. 8.5 A fully-connected hidden Markov model (HMM) with three hidden states H_0, H_1, and H_2. Transition probabilities are omitted. Each hidden state is assumed to correspond to a distinct human translation process (HTP)

[7] In fact, a mixture of Poisson distributions would have been the appropriate choice here, as the action counts are not continuous but discrete data. The mixture model approach allows us to better fit the asymmetrically distributed data with the symmetric distributions such as the Poisson distribution, because of the skewedness of the actual data. An even better option would be to use more general two-parameter models such as the Conway-Maxwell-Poisson distribution, which allows a better fit to heavy or thin tails in the distribution (see Shmueli et al. 2005 for details on the Conway-Maxwell-Poisson distribution).

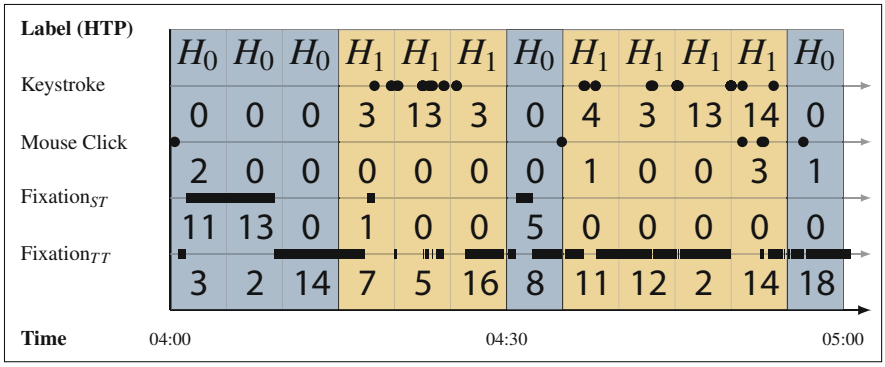


Fig. 8.6 Segmenting translator activity data (TAD) into human translation processes (HTPs) based on a GMM-HMM model learnt from unlabelled data

The transition probabilities and observation probability densities are then trained with the EM algorithm (Baum 1972; Dempster et al. 1977).[8] Good settings were determined by grid search over a range of reasonable parameter settings (cf. Sect. 8.5.1).

8.4.3.1 Segmentation and Labelling

Once the HMM has been trained, we use the Viterbi algorithm (Viterbi 1967) to segment recorded translation sessions into sequences of HTPs by annotating each observation, within its local context, with the most likely HTP underlying it. Figure 8.6 examplifies the outcome for a 1 min excerpt from a recorded translation session (`P07_P21` in CFT14), which has been parametrised into feature vectors covering $w = 5000$ ms each. Each observation has been labelled as either an instance of HTP H_0 or H_1, segmenting the excerpt into three H_0 phases (4:05–4:15, 4:30–4:35, 4:55–5:00) and two H_1 phases (4:15–4:30, 4:35–4:55).

8.5 Experiments

The automatic segmentation and labelling procedure described here was evaluated against the performance of human experts in an annotation task whose objective it was to annotate excerpts of translation logs with conjectures about the underlying HTP (Sect. 8.5.2). In order to do so, we first had to determine good settings for the

[8] We have implemented the modelling approach described in this chapter in `segcats`, available at http://github.com/laeubli/segcats. The clustering and EM algorithms are based on `scikit-learn` version 14.1 (see Pedregosa et al. 2011, and http://scikit-learn.org/stable/).

GMM-HMM's meta-parameters: the set of e observed actions, the window length w, the number of HTPs k, and the number of GMM mixture components m.

8.5.1 Finding Optimal Model Parameters

8.5.1.1 Training Data

We optimized the meta-parameters via grid search on the seven recorded post-editing (PE) sessions contained in the CFT14 dataset. Each of these sessions was converted into a sequence of feature vectors as described in Sect. 8.4.1. As the feature extraction process involves two of the meta-parameters, namely the set of event types e and the window length w, we prepared 12 different parametrisations ($\{E'_1, E'_2, E'_3\} \times \{500\,\text{ms}, 1000\,\text{ms}, 3000\,\text{ms}, 5000\,\text{ms}\}$) of the training data.

8.5.1.2 Experimental Procedure

The goodness of model fit of each model was measured by computing the total log-likelihood of the training data. We used the following parameter values (see also Sect. 8.4.1):

- monitored actions e:
  - E'_1 keystrokes in four categories: alphanumeric, deletion, control, navigation;
  - E'_2 keystrokes as in E'_1, plus mouse clicks;
  - E'_3 keystrokes and mouse clicks as in E'_2, plus eye fixations in two categories: fixations on the source text, fixations on the target text;
- window length w: 500 ms, 1000 ms, 3000 ms, 5000 ms;
- number of HMM states k: 2, 3, ..., 10;
- number of GMM components per HMM state m: 1, 2, ..., 10.

The cross product of these sets of parameter values yields ($3 \times 4 \times 9 \times 10 =$) 1080 distinct parameter settings, all of which we investigated. As the semi-random centroid initialisation in the `k-means++` algorithm has been shown to have a considerable impact on the grouping of translation actions in the HMM states, we trained ten models per parameter configuration. This allowed us to derive the mean ($\overline{LL}$) and standard deviation ($std(LL)$) of the total log likelihood of the training data for each of the 1080 parameter configurations.

8.5.1.3 Findings

Our analysis of the impact of the e, w, k, and m parameters on the likelihood of the training data ($\overline{LL}$) can be summarised[9] as follows:

- Shorter window lengths w generally increase $\overline{LL}$. The best models are learnt from recorded translation sessions that have been parametrised into $w = 500$ ms segments.
- There is no clear impact of the set of monitored actions e on $\overline{LL}$. An interesting finding is that while using eye tracking data (E'_3 vs. E'_1 or E'_2) increases $\overline{LL}$ with short window lengths, it has the opposite effect with longer window lengths.
- In general, $\overline{LL}$ increases sharply with the number of HMM states n up to $n = 4$ or 5, then increases moderately up to $n = 7$, and remains relatively stable with $n >= 7$. The best models have nine or ten HMM states.
- Models with fewer than four GMM components m per HMM state perform considerably worse than models with four or more components. This finding is very consistent across models with different parameter values for e, w, and n. The best models in terms of $\overline{LL}$ have between seven and ten GMM components.

The model that scored best in terms of $\overline{LL}$ overall was trained on TAD segmented into time windows of length $w = 500$ ms, considering keystrokes, mouse clicks, and eye fixations ($e = E'_3$). It has $k = 10$ HMM states with $m = 8$ GMM components each, but similar results were achieved with $k = 8$ or 9 and $m = 7$–10.

8.5.2 *Validation of the HTPs Models Against Human Performance*

To determine whether the annotations produced by the automatic procedure are meaningful to human experts, we tested how well automatic annotations match those by human experts. We used the gold data set described in Sect. 8.3.3 as a benchmark. For the purpose of comparison, the automatically inferred HTP classes were manually mapped to the set of labels $\{O, R, P\}$ used by the human expert annotators. The mapping procedure is described in more detail later in this section.

8.5.2.1 Method of Comparison

Unfortunately, CASMACAT's replay mode is not faithful to the lapse of real time. As a consequence, the video sequences shown to annotators do not match the timing information in the underlying TAD. This makes it impossible to compare the label sequences assigned by the automatic procedure and the human experts directly.

[9] Details can be found in Appendix B of Läubli (2014).

We therefore aggregated the assignments and compared how often experts and models assigned the O, R, and P labels in each of the seven excerpts in CFT14-Gold. In other words, we did not assess how well experts and models classified what a post-editor was doing at a certain time, but rather how exactly they estimated the total time that the post-editor spent on the orientation (O), revision (R), and pausing (P) HTPs in the whole excerpt (5 min). We compared the number of 3 s long segments per HTP in the gold standard to the labels assigned by individual experts and models, measuring the root mean square error (RMSE) for each of the HTPs (classes), i.e.

$$RMSE_x = \sqrt{\frac{1}{n}\sum_{t}^{n}(\hat{x}_t - x_t)^2}, \tag{8.2}$$

where n is the number of classified excerpts, $\hat{x}_t$ is the number of segments in excerpt t classified as HTP $x \in \{\mathrm{O}, \mathrm{R}, \mathrm{P}\}$ in the gold standard, and x_t is the number of segments classified as HTP $x \in \{\mathrm{O}, \mathrm{R}, \mathrm{P}\}$ by an individual model or annotator in the same excerpt t.

In addition to the root mean square errors per class, we give two total RMSE values: $RMSE_{\text{sum}}$ and $RMSE_{\text{w.avg.}}$. $RMSE_{\text{sum}}$ is the sum of the RMSE values for all classes O, R, and P:

$$RMSE_{\text{sum}} = RMSE_O + RMSE_R + RMSE_P \tag{8.3}$$

$RMSE_{\text{w.avg.}}$ is the sum of all RMSE per class weighted by their relative class frequency f in the gold standard:

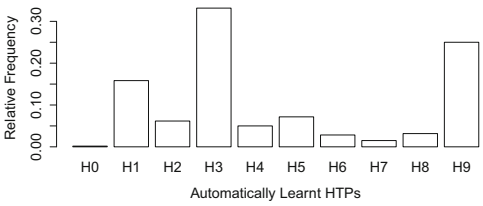
$$RMSE_{\text{w.avg.}} = (f_O \times RMSE_O) + (f_R \times RMSE_R) + (f_P \times RMSE_P) \tag{8.4}$$

8.5.2.2 Data

The experiment was based on the logs of the aforementioned seven post-editing sessions in the CFT14 dataset (Sect. 8.3.2). The model was trained on the complete raw TAD from these seven sessions. The test data consisted of seven excerpts (5 min each) from the same seven post-editing sessions that were annotated manually, as described in Sect. 8.3.3.

8.5.2.3 Experimental Procedure

For the evaluation, we chose the model that showed the best fit of the data in terms of log-likelihood ($e = E'_3$, $w = 500\,\text{ms}$, $k = 10$, $m = 8$; cf. Sect. 8.5.1). Next,



Session time (s)	...	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0	12.5	13.0	13.5	14.0	14.5	...
Model output	...	H2	H2	H5	H7	H1	H1	H3	H3	H3	H3	H1	H3	H3	H6	H3	H4	H1	H9	...
Mapping	...	R	R	O	O	O	O	R	R	R	R	O	R	R	O	R	R	O	P	...
Aggregation	...	O						R						R						...

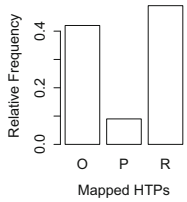


Fig. 8.7 Mapping of automatically learnt HTPs (H_0–H_9) to orientation (O), revision (R), and pause (P) in `Session_G.csv` from the CFT14-Gold dataset. The session was automatically labelled with H_0–H_9. The mapping of these states to O, R, and P was defined manually (cf. Fig. 8.8)

we manually mapped each of the model's ten HTPs[10] to either the orientation (O), revision (R), or pause (P) HTP from the TPR tagset described in Sect. 8.3.3.3. This process is illustrated in Fig. 8.7.

The mapping was based on visual analysis of the probability densities for the observable translation actions in each HTP. If the probability mass for alphanumeric and/or deletion keystrokes in an automatically learnt HTP was centered clearly above zero, we tagged it as an instance of revision (R). Otherwise, we tagged it as orientation (O), unless the probability mass for all observable actions was centered around zero, in which case we tagged it as pause (P). Figure 8.8 shows the probability densities for all actions in the first HTP of the model, which we mapped to the revision (R) HTP according to the aforementioned guidelines.

With this mapping of automatically learnt HTPs to the HTPs used to describe the post-editing process in TPR, we were able to compare the number of segments classified as O, R, and P by the model to the number of segments classified as O,

[10]Recall that each HMM state corresponds to an automatically learnt HTP in the models learnt through the unsupervised sequence modelling approach proposed in Sect. 8.4.

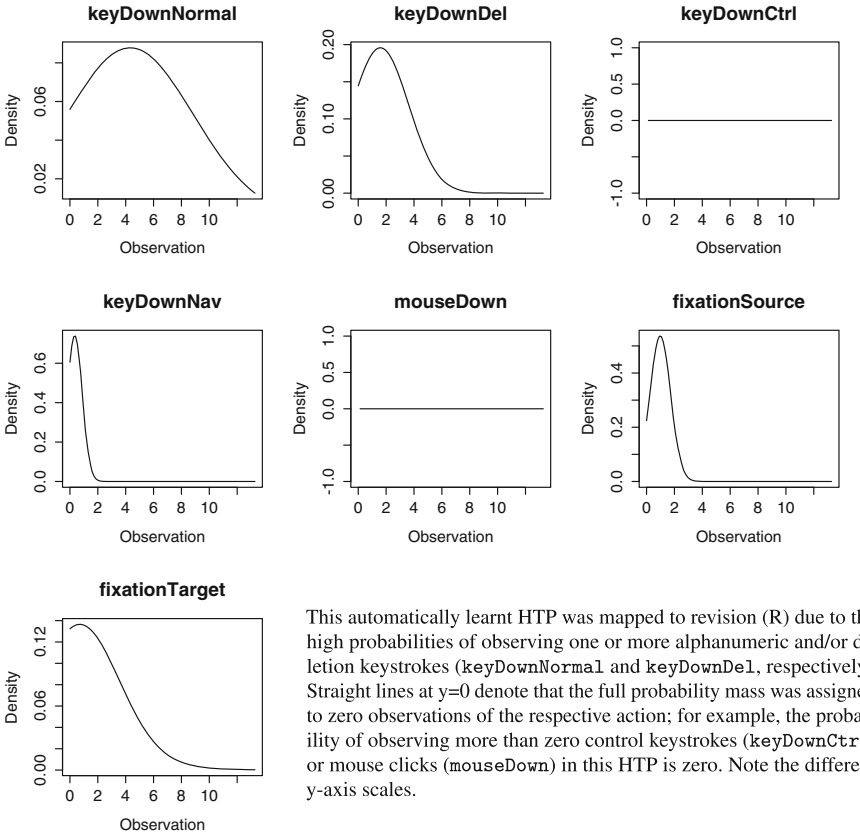


Fig. 8.8 Probability densities for observable translation actions in the first HTP (H_0) of the best-performing model trained on the CFT14 post-editing sessions

R, and P segments in the gold standard. However, as the best-performing model operated on TAD segments of 500 ms while the gold standard contained labels for segments of 3000 ms, we aggregated every subsequent group of six labels in the model output into one label based on the majority of labels in that group; in case of a draw, we used the label that was closest to the end of the group. The only exception was pausing: if a group contained any non-pause label (O or R), the respective segment was labelled as O or R, again based on which of the two occurred more frequently.

This resulted in a sequence of O, R, and P labels—each corresponding to 3000 ms of translator activity data—for each of the seven CFT14 excerpts and each of the 12 annotators, as well as for the best-performing statistical model. For each of the annotators and the model, we counted the number of O, R, and P segments assigned to each of the excerpts, and calculated the RMSE with respect to the counts in the gold standard.

8.5.2.4 Results

Table 8.3 shows the RMSE values for all human annotators (A1–12) and the best-performing statistical model (M) against the gold standard, ordered by total RMSE (summed deviation for O, R, and P; lower is better). The best-scoring annotator's classifications (A3) deviate, on average, by 2.13 segments per class from the gold standard. As each segment corresponds to 3 s of a post-editing session, this value can be interpreted as follows: if asked how long a translator spends on orientation (O), revision (R), and pause (P) in a given excerpt of 5 min (i.e., 300 s), annotator A2's predictions will deviate by 6.39 s (2.13 segments × 3 s) per class from the gold standard on average; more precisely, his or her predictions for the time spent on orientation will deviate by 10.02 s (3.34 segments × 3 s), by 4.8 s for revision (1.6 segments × 3 s), and by 6.21 s (2.07 × 3 s) for pause.

In direct comparison with the 12 human annotators, the statistical model (M) ranks 11th out of 13. It is remarkable that two annotators performed worse than the statistical model (M) in this evaluation. Even when we concede that there may have been external reasons for their poor performance (distraction, lack of time or lack of commitment), we have to keep in mind that all annotators had prior experience in TPR, and that the type of annotation they were asked to produce is closely related to core skills in TPR: interpreting translation logs. Moreover, as they were shown video replays of the translation process as monitored (cf. Sect. 8.3.3.2), they had a lot more information at their disposal than the statistical models: annotators saw the actual source and target texts, the directions and durations of successive eye fixations, etc. The statistical model, in contrast, bases its classifications solely on the *number* of keystrokes, mouse clicks, and eye fixations in isolated segments of 500 ms and the immediately preceding HTP (by virtue of the first-order Markov Model), and is very limited in the scope if its model. Unlike humans, it has a very

Table 8.3 Root mean square error (RMSE) per class and annotator (A1–12) or model (M) in number of 3 s long segments

	RMSE per HTP			RMSE total	
	O	R	P	Sum	W. Avg.
A3	3.34	1.60	2.07	7.01	2.13
A4	3.12	2.51	3.74	9.37	2.77
A9	3.40	2.04	4.00	9.44	2.56
A1	3.91	3.51	2.10	9.52	3.53
A7	5.96	1.85	4.74	12.55	3.23
A5	4.34	2.90	5.74	12.98	3.50
A12	7.16	7.05	1.36	15.57	6.71
A11	6.27	3.25	7.06	16.58	4.37
A2	2.93	7.89	6.14	16.96	6.34
A6	5.67	5.63	5.84	17.14	5.66
M	9.91	6.55	6.22	22.68	7.50
A8	11.10	9.65	4.33	25.08	9.72
A10	12.08	15.59	6.55	34.22	13.98

limited notion of context and cannot remember or aggregate over time what the post-editors were doing.

The inter-annotator agreement (see Table 8.2) indicated that human expert annotators by and large can agree on the classification of segments. The fact that the model under evaluation is within the range of performance of human annotators is strong evidence that the HTPs it automatically inferred from data are meaningful within the context of TPR.

8.6 Conclusion and Outlook

We have presented a statistical model of human translation and post-editing processes that allows automatic annotation of HTP logs with information about the sequence of translation processes executed by the translator at the time. Not only does the annotation show good agreement with annotation by human experts, the states discovered automatically in an unsupervised fashion also display good face validity: visual inspection of the distribution of event probabilities of the various translation action types "makes sense" from the perspective of current Translation Process Research—we can easily map unlabeled HMM states to different HTP types known from the literature (cf. Chap. 14).

With this level of annotation quality, automatic annotation of translation logs promises to be a valuable addition to the Translation Process researcher's toolbox. Unlike manual annotation by human experts, our approach is able to annotate large volumes of TAD at low cost, which might enable us to mediate the effects of the inevitably high levels of noise in the data by analysing very large data sets. Unlike the use of aggregation heuristics, classification is data-driven and can be adapted to the data at hand. By building translator-specific models, we may be able to differentiate individual translation styles. Knowing whether a translator was reading (orientation) or editing (revision) before and after pauses may help us understand what triggered the pause. Was the translator unfamiliar with the term in the source language (a pause during source-side orientation might suggest so), or thinking about a better translation while fixating a source word during revision? Answering these questions may help discover and quantify different translation styles, empirically establish translator profiles, as well as identify the bottleneck in human translation where translators and post-editors might benefit most from additional assistance. It may also prevent us from investing resources into solving issues that, in fact, do little to slow down the overall translation and post-editing process.

8.7 Availability

The tools developed during the course of this work have been released as free, open-source software under the GNU General Public License v3.0. They are available at https://github.com/laeubli/segcats and https://github.com/laeubli/viscats. The manually annotated translation sessions used for evaluation (cf. Sect. 8.3.3) are available at http://www.casmacat.eu/?n=Main.Downloads.

Acknowledgements This work was supported in part by the European Union Seventh Framework Programme for Research, Technological Development and Demonstration (FP7/2007–2013) under grant agreement no. 287576 (CASMACAT).

References

Alabau, V., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., Germann, U., et al. (2014). Casmacat: A computer-assisted translation workbench. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden (pp. 25–28).

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, LA (pp. 1027–1035).

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In *Proceedings of the 3rd Symposium on Inequalities*, Los Angeles, CA (pp. 1–8).

Carl, M. (2010). A computational framework for a cognitive model of human translation processes. In *Proceedings of ASLIB Translating and the Computer* (Vol. 32), London, UK.

Carl, M. (2012). Translog-II: A program for recording user activity data for empirical reading and writing research. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey (pp. 4108–4112).

Carl, M., & Jakobsen, A. L. (2009). Towards statistical modelling of translators' activity data. *International Journal of Speech Technology, 12*(4), 125–138.

Carl, M., & Kay, M. (2011). Gazing and typing activities during translation: A comparative study of translation units of professional and student translators. *Meta, 56*(4), 952–975.

Carl, M., García, M. M., & Mesa-Lao, B. (2014). CFT13: A resource for research into the post-editing process. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland (pp. 1757–1764).

Chinea-Rios, M., Sanchis-Trilles, G., Ortiz-Martínez, D., & Casacuberta, F. (2014). Online optimisation of log-linear weights in interactive machine translation. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland (pp. 3556–3559).

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*(1), 1–38.

Dragsted, B., & Carl, M. (2013). Towards a classification of translation styles based on eye-tracking and keylogging data. *Journal of Writing Research, 5*(1), 133–158.

Elming, J., Winther Balling, L., & Carl, M. (2014). Investigating user behaviour in post-editing and translation using the CASMACAT workbench. In S. O'Brien, L. Winther Balling, M. Carl, M. Simard, & L. Specia (Eds.), *Post-editing of machine translation* (pp. 147–169). Newcastle upon Tyne: Cambridge Scholars Publishing.

Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., et al. (2014). The MateCat tool. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, Dublin, Ireland (pp. 129–132).

Fleiss, J. L. (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378.

Gerloff, P. (1986). Second language learners' reports on the interpretive process: Talk-aloud protocols of translation. In J. House & S. Blum-Kulka (Eds.), *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies* (pp. 243–262). Tübingen: Narr.

Green, S., Heer, J., & Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *Proceedings of the 2013 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, Paris, France

Guerberof, A. (2009). Productivity and quality in the post-editing of outputs from translation memories and machine translation. *International Journal of Localisation, 7*(1), 11–21.

Hvelplund, K. (2011). *Allocation of cognitive resources in translation: An eye-tracking and key-logging study*. Ph.D. thesis, Copenhagen Business School, Copenhagen, Denmark.

Jakobsen, A. L. (1999). Logging target text production with Translog. *Copenhagen Studies in Language, 24*, 9–20.

Jakobsen, A. L. (2003). Effects of think aloud on translation speed, revision and segmentation. In F. Alves (Ed.), *Triangulating translation. Benjamins translation library* (Vol. 45, pp. 69–95). Amsterdam, Netherlands: John Benjamins.

Jakobsen, A. L. (2011). Tracking translators' keystrokes and eye movements with Translog. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research: Integrative approaches in translation studies. Benjamins translation library* (Vol. 94, pp. 37–56). Amsterdam, Netherlands: John Benjamins.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354.

Koehn, P., & Germann, U. (2014). The impact of machine translation quality on human post-editing. In *Proceedings of the Workshop on Humans and Computer-assisted Translation (HaCaT)*, Gothenburg, Sweden (pp 38–46).

Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). Post-editing time as a measure of cognitive effort. In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP)*, San Diego, CA (pp. 11–20).

Krings, H. P. (1995). *Texte reparieren. Empirische Untersuchungen zum Prozeß der Nachredaktion von Maschinenübersetzungen*. Habilitation thesis, University of Hildesheim, Hildesheim, Germany.

Krings, H. P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes*. Kent, OH: Kent State University Press.

Krings, H. P. (2005). Wege ins Labyrinth – Fragestellungen und Methoden der Übersetzungsprozessforschung im Überblick. *Meta, 50*(2), 342–358.

Läubli, S. (2014). *Statistical modelling of human translation processes*. Master's thesis, University of Edinburgh, Edinburgh, UK.

Läubli, S., Fishel, M., Massey, G., Ehrensberger-Dow, M., & Volk, M. (2013). Assessing post-editing efficiency in a realistic translation environment. In *Proceedings of the 2nd Workshop on Post-Editing Technology and Practice (WPTP)*, Nice, France (pp. 83–91).

Martínez-Gómez, P., Minocha, A., Huang, J., Carl, M., Bangalore, S., & Aizawa, A. (2014). Recognition of translator expertise using sequences of fixations and keystrokes. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA)*, Safety Harbor, FL (pp. 229–302).

Massey, G., & Ehrensberger-Dow, M. (2014). Looking beyond the text: The usefulness of translation process data. In D. Knorr, C. Heine, & J. Engberg (Eds.), *Methods in writing process research*. Frankfurt am Main, Germany: Peter Lang (pp. 81–89).

O'Brien, S. (2006). Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Languages and Cultures, 7*(1), 1–21.

O'Brien, S. (2007). Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology, 14*(3), 185–205.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., & Grisel, O. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Perrin, D. (2003). Progression analysis (PA): Investigating writing strategies at the workplace. *Journal of Pragmatics, 35*(6), 907–921.

Plitt, M., & Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague Bulletin of Mathematical Linguistics, 93*, 7–16.

Sanchis-Trilles, G., Ortiz-Martínez, D., & Casacuberta, F. (2014). Efficient wordgraph pruning for interactive translation prediction. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT)*, Dubrovnik, Croatia (pp. 27–34).

Schaeffer, M., & Carl, M. (2014). Measuring the cognitive effort of literal translation processes. In *Proceedings of the Workshop on Humans and Computer-assisted Translation (HaCaT)*, Gothenburg, Sweden (pp. 29–37).

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 54*(1), 127–142.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257–268.

Toury, G. (1995). *Descriptive translation studies and beyond. Benjamins translation library* (Vol. 4). Amsterdam, Netherlands: John Benjamins.

Underwood, N., Mesa-Lao, B., Martinez, M. G., Carl, M., Alabau, V., Gonzalez-Rubio, J., et al. (2014). Evaluating the effects of interactivity in a post-editing workbench. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland (pp. 553–559).

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory, 13*(2), 260–269.

Chapter 9
Word Translation Entropy: Evidence of Early Target Language Activation During Reading for Translation

Moritz Schaeffer, Barbara Dragsted, Kristian Tangsgaard Hvelplund, Laura Winther Balling, and Michael Carl

Abstract This study reports on an investigation into the relationship between the number of translation alternatives for a single word and eye movements on the source text. In addition, the effect of word order differences between source and target text on eye movements on the source text is studied. In particular, the current study investigates the effect of these variables on early and late eye movement measures. Early eye movement measures are indicative of processes that are more automatic while late measures are more indicative of conscious processing. Most studies that found evidence of target language activation during source text reading in translation, i.e. co-activation of the two linguistic systems, employed late eye movement measures or reaction times. The current study therefore aims to investigate if and to what extent earlier eye movement measures in reading for translation show evidence of co-activation. Results show that the number of translation alternatives for a single word and differences between source and target text in terms of word order have an effect on very early and late eye movement measures. Results are interpreted in terms of semantic and structural cross-linguistic priming: items which have a similar word order in source and target texts are likely

M. Schaeffer (✉)
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

Institute for Language, Cognition and Computation, University of Edinburgh, Edinburgh, UK
e-mail: moritzschaeffer@gmail.com

B. Dragsted • L.W. Balling • M. Carl
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

K.T. Hvelplund
Department of English, Germanic and Romance Studies, University of Copenhagen, Copenhagen, Denmark

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_9

to have similar syntactic structures. These items are therefore more likely to prime structurally. Source items which have few translation alternatives are more likely to share a semantic representation and are hence more likely to prime semantically than items with more translation alternatives. Findings support the literal translation hypothesis.

Keywords Co-activation • Priming • Translation • Entropy • Eye movements

9.1 Introduction

It has been a subject of debate in translation process research (TPR) whether translation is a sequential process or whether and to what extent comprehension and production activities may occur in parallel (Carl and Dragsted 2012; Balling et al. 2014). In the sequential, or vertical perspective, human translation is described (Gile 1995) as a process in which the translator first reads a source-language (SL) segment, then formulates a "meaning hypothesis", i.e., assigns a meaning to the translation segment by drawing on SL and general world knowledge, and possibly external information sources, and then checks the meaning hypothesis for plausibility. Having finished the processes involved in understanding the source text (ST), the translator moves on to reformulating the meaning hypothesis in the target language (TL), drawing again on general world knowledge and on knowledge of the TL, and checks for fidelity and general acceptability, continuously revising the target text (TT) until a satisfactory version has been arrived at. In the same vein, according to the Interpretive Model (Lederer 1994) translation is a process in which the translator understands the text, deverbalizes its language and re-expresses its sense in the TL.

Common to these models is that they view ST reading as a phase distinct from the reformulation phase and characterised largely by the same processes as reading for monolingual comprehension. In contrast to this, the horizontal/parallel view holds that TL reformulation commences during ST comprehension, and that the process involved in reading for translation is different from reading for monolingual comprehension (see e.g. Jakobsen and Jensen 2008; Schaeffer et al. forthcoming). In line with this view, Carl and Dragsted (2012) propose that the ST is understood and meaning hypotheses are generated only to the extent required to keep on producing target text. Deep ST understanding is prompted by problems occurring in the TT. If TT production is interrupted, for instance because the translator is not able to retrieve an appropriate TL equivalent or is considering which translation to choose out of several alternatives (see below), the missing information needs to be retrieved. This may lead to increased eye movement activity and gaze time on a ST word or passage with a view to verification or reinterpretation (ibid.: 143–144).

Schaeffer and Carl (2013: 185) propose a different kind of model in which "... both horizontal and vertical processes are always active at the same time." Schaeffer and Carl (ibid) argue that "... that the horizontal process is an early

process while the vertical processes depend on context which becomes available later, as processing advances in the chunk or text . . . ".

This study assumes that translators read the ST with TT production in mind; hence, different processes are involved in reading for translation than in reading for monolingual comprehension. Previous studies which found evidence of co-activation of the two linguistic systems during ST reading, i.e., studies which found support for the hypothesis that translation is a parallel/horizontal process, employed late eye movement or other late behavioural measures. This study tests the hypothesis that target-language-specific aspects have an impact during very early stages of ST processing. If target language specific aspects have an impact on early eye movement measures, this would allow for a much stronger claim regarding the horizontal/parallel view, because early eye movement measures are more indicative of automatic processes than late measures, and any effect is more likely to allow for conclusions regarding bilingual lexicon.

This study analyses a subset of the CRITT TPR-DB, described in Chap. 2, in order to test whether target-language-related aspects have an effect on early and late eye movement measures on the source text. The study only considers eye movements on the ST, given the object of interest is whether or not the two linguistic systems are co-activated during reading for translation: it is therefore not of interest to us whether target-language-related aspects have an effect on eye movements during TT reading.

9.2 Theoretical Background

Section 9.2 is split up into nine subsections that describe dependent variables and predictors used in the final linear mixed models described in Sect. 9.4. Section 9.2.1 introduces the dependent variables. Section 9.2.2 discusses how translators often read the ST and type the TT concurrently. This is taken as a coarse indicator of co-activation during translation. Section 9.2.3 introduces the literal translation hypothesis and the two predictor variables of central interest, *Cross* and *HTra* (described in Chap. 2, and below). This section also explains how these two features of the CRITT TPR-DB relate to the literal translation hypothesis. Section 9.2.4 describes previous studies on the effect of translation alternatives on behavioural measures. This includes studies which employ single words as stimuli and more naturalistic studies which employ longer texts and translation production. Section 9.2.5 introduces cross-linguistic priming. The core of our argument and evidence is that shared semantic and structural representations prime and that they form the basis for a literal (interim) translation of the source text. Section 9.2.6 presents several models of the bilingual lexicon and how these relate to our findings. Section 9.2.7 argues that it is paramount to employ earlier eye movement measures if findings are to be related to how the bilingual brain represents language during translation. Relevant previous studies are discussed in this section. Word frequency is a further predictor variable in our models. Section 9.2.8 therefore describes studies which have investigated the effect of word frequency on monolingual reading and

translation. Section 9.2.9 explains why *STseg* is included as a predictor in our models. Previous research suggests that translators become faster as they progress in the text. *STseg* numbers segments in each source sequentially and can therefore be an indicator for this facilitation effect.

9.2.1 Eye Movements in Reading and Translation

There is a long tradition for analysing eye-movements in reading (see for instance Rayner 2009 for a comprehensive overview). A basic assumption (the so-called eye-mind assumption) in eye movement research is that "the eye remains fixated on a word as long as the word is being processed" (Just and Carpenter 1980: 330). Hence, for text comprehension, according to Clifton et al. "[. . .] how long readers look at a word is influenced by the ease or difficulty associated with accessing the meaning of the word" (Clifton et al. 2007: 248). Gaze duration is thus taken to signal the cognitive effort associated with processing a particular item, and fixations in reading tend to be longer on items requiring effortful processing, for instance words which are less frequent, words containing spelling errors, ambiguous words, words which are inappropriate in a given context, etc. (e.g. McConkie and Yang 2003: 413).

In recent years, eye-tracking has also been used increasingly in translation research to investigate cognitive processes in translation. Studies have examined a broad range of aspects of translation, including cognitive load and translation memory matches (O'Brien 2006), reading for translation (Jakobsen et al. 2008; Schaeffer et al. forthcoming; Hvelplund 2015), translator styles (Dragsted and Carl 2013), etc. Eye-tracking has also been used extensively in combination with key logging (e.g. Jakobsen 2011; Carl and Dragsted 2012).

The following eye movements measures from the CRITT TPR-DB were used in the current study: first fixation duration (*FFDur*), gaze duration on ST words (*FPDurS*), regression probability (*Reg*), regression path duration (*RPDur*), probability of a fixation or skipping probability is calculated on the basis of the existing data, and total reading time on the ST (*TrtS*). Most of these measures exist for both the ST and the TT. First fixations are considered to be indicative of early (lexical) processing (Rayner 1998). Gaze duration (*FPDurS*) is the sum of the fixations on $word_n$ before the eyes move on to a different word either to the left or right of $word_n$. Gaze duration therefore also describes the processing of $word_n$ in terms of lexical processing, although gaze duration is a later stage in lexical processing than first fixation duration. Regression path duration (*RPDur*) includes all the fixations that are summed under the name gaze durations, but regression path duration also includes fixations on words that are situated to the left of $word_n$, i.e. regression path duration includes regressions to earlier words that had already been read. Regression path duration therefore represents processes which integrate aspects of a word with prior words. Probability of a regression (*Reg*) describes the probability that the eyes move to the left from $word_n$. It is normally assumed that a regression occurs because $word_n$ is difficult to integrate. Total reading time (*TrtS*) is the sum of all fixations

on word$_n$, regardless of when these took place. In this sense, total reading time is a very late measure of word processing and includes post-lexical integration processes and gaze during translation revisions. The measure skipping rate or probability of a fixation describes the number of times or the likelihood that a word$_n$ is not fixated at all. A word$_{n+1}$ to the right of the fixation on word$_{\mathbf{n}}$ can be pre-processed during fixation on word$_{\mathbf{n}}$ and may be guessed (e.g. Ehrlich and Rayner 1981), and it may therefore not need to be fixated at all. Skipping rate is normally computed on the basis of a first run, i.e., on the basis of a forward movement through the text. The first run normally ends if either the end of the sentence is reached or a regression is made. The probability of a fixation reported here is different. The probability of a fixation reported here describes whether a word has been fixated at all—irrespective of whether this occurred in the first or any subsequent runs through the text. Thus, a probability of a fixation of zero reported here refers to a situation in which a word received no fixation at all during the whole of the session.

Whether an observed effect occurs during early or late eye movement measures may be an indicator of whether it is cognitively determined or evidence of willed behaviour. A very early effect may give an indication regarding the automatic cognitive mechanism underlying the effect while late eye movement measures are more likely to reflect rather conscious behaviour.

Popular measures of gaze activity in TPR have been average fixation duration, total reading time on the word, segment or whole text and pupil size. One of the few studies that have applied first fixation duration (*FFDur*) to examine aspects of cognitive processing during translation is Rydning and Lachaud (2010). Comparing *FFDur* of professional translators and bilingual non-translators, the authors found that professional translators were able to recognise the meaning of polysemous words outside of context more quickly than bilinguals (2010: 99) as indicated by significantly shorter *FFDur*. This finding was further supported by the same effect on TRT. In the introduction we argue that during translation, source and target-related processes are tightly intertwined. The following section presents preliminary and coarse evidence to support this view.

9.2.2 Concurrent ST Reading and TT Writing in Translation

As detailed in Chap. 10, Grosjean (1997) hypothesised that a bilingual's two languages are always active to a certain extent and that this is best described by a continuum of co-activation of the two languages. Grosjean argues that it is the context of the language use, which determines where on the continuum the bilingual is currently situated. It is highly likely that during translation, both languages are active at the same time. A range of studies (Macizo and Bajo 2006; Ruiz et al. 2008; Schaeffer et al. forthcoming; Wu and Thierry 2012; Balling et al. 2014) (described in more detail in Chap. 10 and below) support this hypothesis. One indicator of co-activation may be the time translators spend reading the ST while typing the TT. During concurrent ST reading and TT typing, both languages must

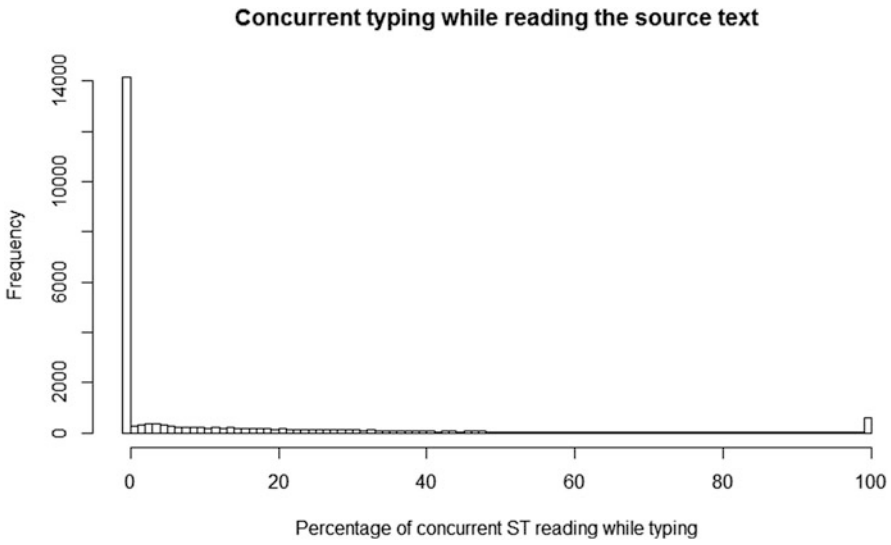


Fig. 9.1 Distribution of percentage of concurrent ST reading and TT typing of total production time (per production unit)

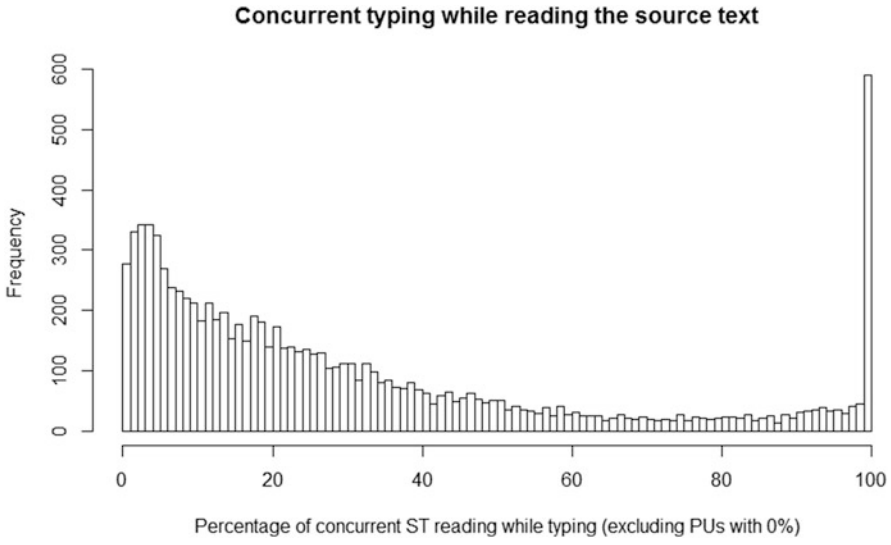


Fig. 9.2 Distribution of the percentage of the production time translators concurrently read the ST while typing the TT (only for those production units where some concurrent ST reading while typing occurred)

be activated simultaneously. The *ParalS* feature in the PU files of the TPR-DB (see Chap. 2) captures this kind of manifest concurrency. Figures 9.1 and 9.2 visualise the distribution of the percentage of the production time translators read the ST

while typing the TT (see Sect. 9.3 for a brief outline of the data sets included in this study).

Figures 9.1 and 9.2 illustrate that, while it is relatively rare for concurrent ST reading and TT typing to occur for a complete production unit (Fig. 9.1), there are many shorter stretches of time when translators read the source text while typing the target text (Fig. 9.2). Approximately 40 % of the PUs in the data (9148 out of a total of 23,294) has some concurrent typing while reading the ST. These findings mirror earlier research which has found that the mean duration of instances of this kind of manifest concurrency is around 429 ms, and considerably longer for ST processing units (846 ms) and TT processing units (1141 ms) (Hvelplund 2011: 143).

The next section introduces the literal translation hypothesis in relation to the two features of the CRITT TPR-DB which will be used to predict early and late eye movement measures during translation.

9.2.3 The Literal Translation Hypothesis

Malmkjaer (2005) argues that many phenomena, which have been claimed to be translation universals are actually socially constrained norms and hence subject to diachronic and meandering changes. Malmkjaer, instead, highlights only one possible hypothesis regarding the translation process which could be a cognitively determined universal, namely the literal translation hypothesis, which "... has been implied or explicitly studied by many scholars, and does not seem to have a single source..." (Chesterman 2011: 23).

There are several formulations of this hypothesis, but on the most general level, it can be summarised as follows: a literal translation is the first or default solution a translator applies to the source text, often only as an interim solution before a less literal translation is considered or produced. Carl and Schaeffer (forthcoming) propose a definition of literality that allows for quantification of the phenomenon. According to their definition, a translation is literal if the three following literality criteria are fulfilled:

1. Word order is identical in the ST and TT.
2. ST and TT items correspond one-to-one.
3. Each ST word has only one possible translated form in a given context.

A word, sentence or text will rarely fulfil *all* criteria at the same time (see Figs. 9.3 and 9.4). As such, the criteria should jointly be seen as a prototype of a literal translation. In this view, an interim representation serves as a reference in the process of creating a text, which is more or less acceptable according to target norms while remaining more or less adequate with respect to the source. The *Cross* feature in the TPR-DB (see Chap. 2) indicates the similarity in word order between source and target texts (literality criteria 1 and 2): words with a *Cross* value of 1 have a similar (relative) position in both the source and the target sentences. Words with a higher *Cross* value (either positive or negative) represent different word orders in the

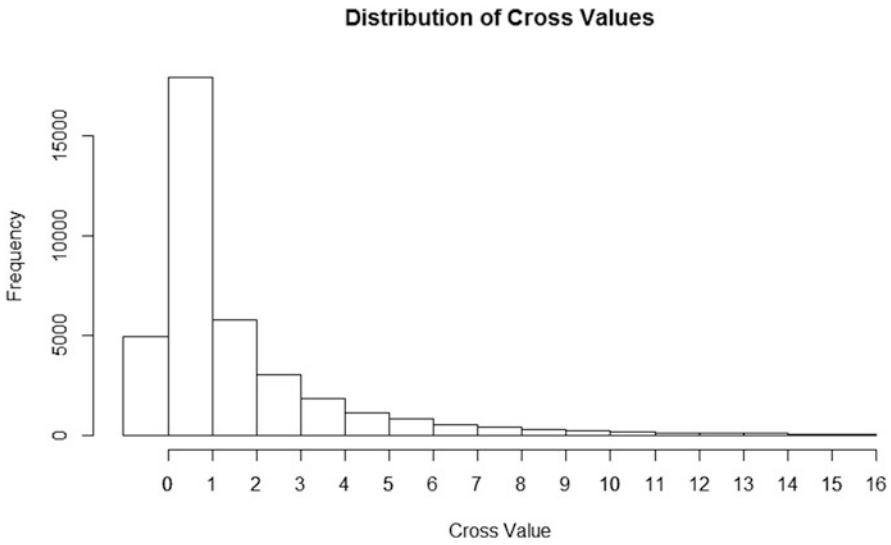


Fig. 9.3 Distribution of absolute *Cross* values (values over 16 have been excluded for ease of presentation and because there are very few of these)

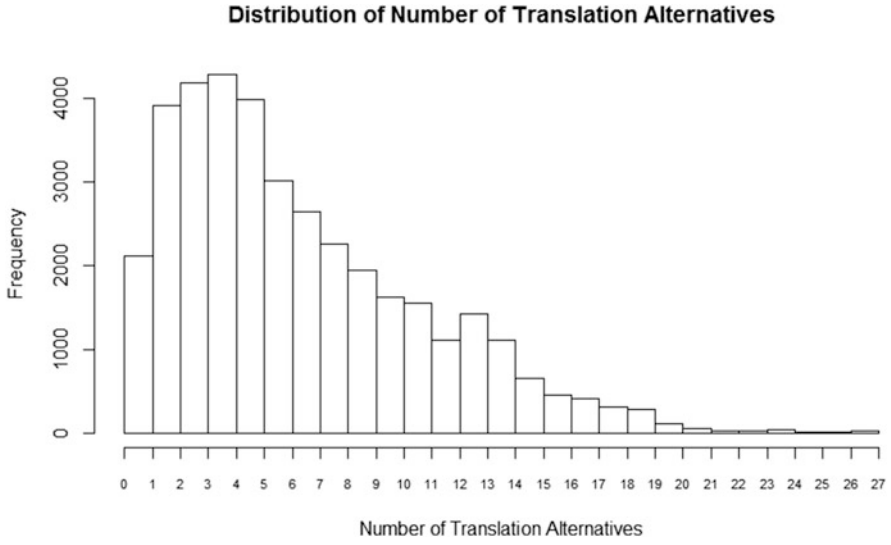


Fig. 9.4 Distribution of number of translation alternatives

source and the target. Figure 9.3 shows that a *Cross* value of 1 is very frequent—the distribution has a large peak at *Cross* 1 (Fig. 9.3). However, about 40 % of the words in the data has a *Cross* value above one (14,653 out of a total of 37,519). The vast majority (over 90 %) of words have more than one alternative translation (Fig. 9.4).

The fact that most words have more than one translation alternative (see below) highlights that the literal translation hypothesis describes an ideal or prototype that often remains an interim representation. It is more likely that translators maintain the same (relative) word order in the TT as in the ST (*Cross* value of 1). What Fig. 9.3 suggests is that translators tend to maintain the (relative) source word order in individual translation units. In the case of relative word order, the literal default often goes beyond an interim representation and finds its way into the final TT.

The degree to which two items are literal, according to the definition above, may predict how easy it is to process this item. If the literality of an item has an effect on early eye movement measures, it is likely that this has to do with how the brain represents language during translation, that horizontal processes occur early, and that it is the co-activation of the two linguistic systems which makes reading from translation fundamentally different from monolingual reading. If, however, literality only has an effect on late eye movement measures, it is more likely that translation is a more sequential process. If there is an effect on both early and late measures, this would lend support to the model proposed by Schaeffer and Carl (2013) who argue that horizontal processes take place early and vertical monitor processes take place late.

9.2.4 Translation Alternatives

In the TPR-DB, the number of alternative translations for a single source word is described by the word translation entropy, as measure *HTra*, described in Chap. 2. Word translation entropy describes the degree of uncertainty regarding which lexical TT item(s) are chosen given the sample of alternative translations for a single ST word: if the probabilities are distributed equally over a large number of items, the word translation entropy is high and there is a large degree of uncertainty regarding the outcome of the translation process. If, however, the probability distribution falls unto just one or a few items, entropy is low and the certainty of the TT item(s) to be chosen is high.

The above presentation and discussion of previous research has concerned the matter of concurrent ST processing and TT processing in translation and at what point during the translation process the translator's TL is activated. Our overall hypothesis is that TL-related aspects have an impact very early on during the translation process, and that reading during translation, by derivation, is fundamentally different from monolingual reading for comprehension. The early eye movement measures on ST items are examined in order to establish to what extent they vary as a function of word translation entropy. Thus, the following research question will be examined:

How early do target-language-related aspects have an effect on eye movements?

An effect of target-language-specific aspects on early eye movement measures is indicative of co-activation of the two linguistic systems during reading for

translation and of priming (see below), if items, which are likely to share a representation or which are likely to be closely linked, result in facilitation. The next sections will briefly introduce relevant priming studies.

9.2.5 Cross-Linguistic Semantic and Structural Priming and Translation

Co-activation of linguistic items may shorten reading (or reaction) times when the linguistic items in question share a semantic representation. As described in Chap. 10, a number of cross-linguistic priming studies suggest that semantic and/or syntactic representations may be shared between two languages (e.g. Hartsuiker et al. 2004; Duñabeitia et al. 2010; Bernolet et al. 2013). The cross-linguistic semantic priming effect in single word studies is generally the observation that a very short presentation of a word in one language followed by the presentation of its translation equivalent facilitates word recognition of the translation equivalent as compared to when an unrelated word in the other language is presented first. Cross-linguistic structural priming studies typically present a prime sentence in one language with a particular syntactic structure and participants are then asked to, e.g., describe a picture in a different language. The priming effect is the likelihood that participants use the same syntactic structure in the description of the picture which they processed during reading of the prime sentence (in the other language). While there is some evidence that the mechanism behind cross-linguistic structural priming is a shared abstract representation of syntax (e.g. Hartsuiker et al. 2004), there is also evidence that overlap in word order is necessary for (cross-linguistic) structural priming to occur (Hartsuiker et al. 1999; Bernolet et al. 2007; Loebell and Bock 2003; Kidd et al. 2014). However, other studies show the opposite (Desmet and Declercq 2006; Shin and Christianson 2009; Chen et al. 2013). While similar word order is likely to also represent a similarity in syntactic structure, two languages may produce a similar syntactic structure without necessarily producing a similar word order.

9.2.6 The Bilingual Lexicon and Translation

Priming from L1 to L2 has been reported in many studies, but priming from L2 to L1 often fails to be observed. This is referred to as translation asymmetry. The BIA+ (Dijkstra and van Heuven 2002) model argues that L1 words prime L2 word recognition, because L1 words are used more often than L2 words which are used less often and which thus require more effort to activate.

The Distributed Conceptual Features Model (DCFM) (De Groot 1992), unlike other models, suggests that a word's meanings are distributed over a number of different senses. Finkbeiner et al. (2004: 16) propose the Sense Model, which is very similar to the DCFM. According to the Sense Model, the semantic representations

associated with a lexical item are bundled. The Sense Model argues that the observed translation asymmetry is due to the fact that L1 items have more associated senses than L2 items. According to the Sense Model, L2 words share fewer senses with L1 translation equivalents, because they have a smaller number of associated senses, while L1 primes have a higher number of associated senses and therefore the priming effect from L1 to L2 is also stronger. Finkbeiner et al.'s (2004) Sense Model argues that the degree of semantic overlap between two words predicts how strong the priming effect is. A large semantic overlap between two words results in a strong priming effect and if the overlap is small or the L2 senses are not known to the bilingual, the priming effect is weaker. The number of alternative translations for a single ST word may also partly represent the (lack of) semantic overlap. There are a number of single word studies, which have shown that words with more than one possible translation are recognised and produced more slowly (e.g. Tokowicz and Kroll 2007; Laxén and Lavaur 2010; Boada et al. 2012; Eddington and Tokowicz 2013).

In sum, the strength of the priming effect may depend on the degree to which a linguistic item fulfils the literality criteria described above: the more literal (according to the above definition) an item is, the stronger the priming effect. If the syntactic structures in ST and TT are similar it is likely that the word order is also similar and if the overlap in semantic representations for a lexical item is similar, it is likely that a word will tend to be translated in the same way by different translators. In other words, the degree to which two items share structural and semantic representations may predict the strength of the priming effect. If *Cross* and *HTra* have an effect on early eye movement measures, this would lend support to models of the bilingual lexicon which posit non-selectivity. Such an early effect would further lend support to the DCFM and the Sense Model, given that *HTra* is a continuous variable which may describe the graded overlap in terms of semantics between a source word and its translation.

9.2.7 Time Course of Co-activation During Translation

Previous studies (Dragsted 2012; Carl and Schaeffer forthcoming) have shown that the number of alternative translations for a source word has an effect on total reading time on the ST (*TrtS*, the sum total of all fixation durations on a particular word) during translation, such that total reading times increase as a function of the number of translation alternatives. However, it is not clear at what point in time cognitive processing of the various alternative translation options commences.

Various studies suggest that the TL is activated during source text reading. Macizo and Bajo (2006) found that interlingual homographs, which are ambiguous only in the target language but not in the source language, resulted in longer reaction times, but only in the reading for translation condition. These authors further found that cognates resulted in shorter reaction times, but again only during the reading for translation condition. Ruiz et al. (2008) found that ST words with low frequency TT equivalents resulted in longer reaction times than ST words with high frequency

TT equivalents. The authors of both these studies argue that this is evidence for co-activation.

Balling et al. (2014) found an effect of congruence on total reading time of a group of words, such that ST segments which required re-ordering in the TL were read for longer, and interpreted these findings as evidence for co-activation.

While the studies previously discussed worked with various measures of total reading time, the first study, to our knowledge, which tested the effect of the target language on early eye movement measures is Schaeffer et al. (forthcoming). This study manipulated the number of target words required to translate a single source word. Two kinds of items were embedded in the same sentence frames: one-to-one (the ST word was likely to be translated using just one TT word) and one-to-many (the ST word was likely to be translated using more than one TT word). Participants read these sentences in two conditions: reading for comprehension and reading for translation. Schaeffer et al. found a 20 ms effect on average fixation durations: during reading for translation, the average fixation duration across the whole sentence was 20 ms longer than during reading for comprehension. This increase in average fixation duration cannot be explained in terms of motor aspects of target text production, because participants were asked to first read the source sentence and were told to only start typing once they knew how they would translate the sentence. Schaeffer et al. further found that participants made on average 16 fixations more per sentence during reading for translation and the number of regressions also doubled, as did total reading time. The significant increase in all relevant eye-movement measures suggests that reading for translation is fundamentally different from reading for comprehension. For first fixation duration, the effect of the number of required TT words was only found when reading for translation, not during monolingual comprehension. However, the effect was relatively large (23 ms). This study suggests that target language-specific aspects are activated already very early during ST reading for translation.

In sum, most studies which found evidence of co-activation during translation employed late eye movement measures or reaction times. The current study therefore aims to investigate if and to what extent earlier eye movement measures in reading for translation show evidence of co-activation. More specifically, the aim is to study whether early eye movement measures are affected by the number of alternative translations for a single source text word and by word order differences. Drawing on the CRITT-TPR database (Carl 2012), we evaluate the hypothesis that the number of TT alternatives (*HTra*) and *Cross* value have an effect on early eye movement measures.

9.2.8 The Effect of Word Frequency in Reading and Translation

Given that word frequency has a large effect on eye movements during monolingual reading and on different behavioural measures during translation, we included monolingual frequency as a predictor in our statistical model. It is well

established that word frequency is an important variable in cognitive processing. High-frequency words are perceived and produced more quickly and more efficiently than low-frequency words (e.g. Brysbaert and New 2009), and according to Rayner (1998), there is "abundant evidence that the frequency of a fixated word influences how long readers look at the word" (Rayner 1998: 376). The relationship between word frequency and reading/production time has also been studied in translation. Balling and Carl (2014) found that source word frequency has a significant effect on production time, such that higher average frequencies of ST words are associated with shorter production times, especially in student translators.

9.2.9 Discourse and Translation

A facilitation effect has been identified in translation (Englund Dimitrova 2005) where translators become faster as they progress in the text. Englund Dimitrova (2005: 30) argues that "global decisions regarding the TT and the task are made at the beginning of the task ... and the growing text representation (mental and in the form of a TL version), will facilitate certain aspects...". The effect was quite coarse and the behavioural measures employed in her study were mainly based on keylogging. In addition, the data set was relatively small (nine participants) in comparison to the current data set. We therefore decided to investigate the current dataset with the more sensitive eye movement measures in order to find out whether there is a facilitation effect similar to the tentative effect Englund Dimitrova found. The predictor here is *STseg*, which is the sequential numbering of sentences in a given ST. A facilitation effect would suggest that the discourse model is built as translators progress in the task making their task easier as they advance in the text.

9.3 Research Design and Methods

The present study investigates the effect of cross-linguistic syntactic re-ordering and word translation entropy on early eye movement measures in translation. It is based on a subset from the CRITT Translation Process Research (TPR) Database which consists of 295 recordings from nine different studies (ACS08, BD08, BD13, BML12, HLR13, KTHJ08, MS12, NJ12, SG12).

There are 42,211 items (ST tokens) in the nine studies. We decided to exclude punctuation (commas, full stops, hyphens, brackets etc.) and numbers, as well as currency signs (e.g. £), given that these carry little information relevant for the present purpose. This resulted in the exclusion of 11.12 % of the data. Misspellings of words, for instance *government* and *govenment*, will affect the entropy value since they are treated as two distinct translation options. In reality, they represent only one translation option, viz. the properly spelled *government*. Nevertheless, no correction of misspellings has been made, partly for practical reasons and partly under the

assumption that misspellings are approximately evenly distributed over the data set. In addition, datapoints which were more than 2.5 standard deviations above or below a participant's mean for a particular dependent variable were excluded. This resulted in the exclusion of less than 3 % for each dependent variable (apart from "skipping probability" where less than 5 % were excluded).

9.4 Data Analyses

For the analyses, we used R (R Development Core Team 2014) and the lme4 (Bates et al. 2014) and languageR (Baayen 2013) packages to perform linear mixed-effects models (LMEMs). To test for significance, we used the R package lmerTest (Kuznetsova et al. 2014), which implements ANOVA for mixed-effects models using the Satterthwaite approximation to estimate degrees of freedom. The models consisted of the *Cross*, and *HTra* features, as discussed above, the ST frequency (*Prob1*), segment identifier (*STseg*) and length of the ST words in number of characters (*LenS*) as predictors. Since *Cross* can be both positive and negative, we only used absolute values. Given that word length (*LenS*) and frequency (*Prob1*) have a strong effect on eye movements during reading, it was important to control for these. As random variables, we included participant (*Part*), item, a unique text (*Text*) identifier and *Study*.

The dependent variables were first fixation duration (*FFDur*), gaze duration (*FPDurS*), regression probability (*Reg*), regression path duration (*RPDur*), probability of a fixation or skipping probability (ProbFix), and total reading time on the source (*TrtS*). All of these measures only relate to the source text.

Continuous dependent variables (*FFDur*, *FPDur*, *RPDur*, and *TrtS*) were transformed with the natural logarithm because they were not normally distributed. Collinearity was assessed by inspecting variance inflation factors for the predictors; all values were low, indicating that collinearity between predictors was not a problem (Table 9.1).

Table 9.1 Dependent variables, predictors and random factors for the LMEMs described in more detail below

Dependent variables	Predictors	Random variables
FFDur	~*Prob1* + *LenS* + *STseg* + *HTra* + abs(*Cross*)	+ (1|ParticipantUnique) + (1|Item) + (1|TextUnique) + (1|Study)
FFDurS		
RPDur		
TrtS		
ProbFix		
Reg		

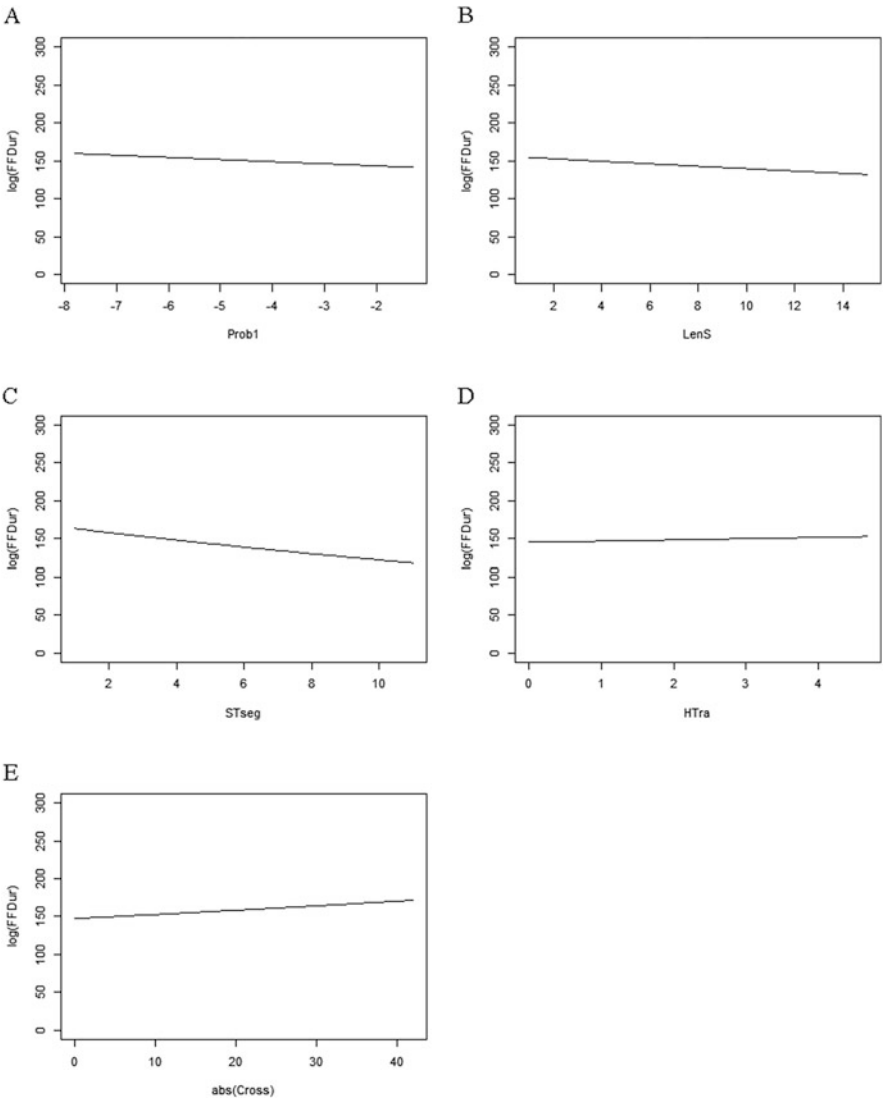


Fig. 9.5 (**a**–**e**) The effects of frequency (*Prob1*), word length (*LenS*), segment identifier (*STseg*), word translation entropy (*HTra*), and absolute values of crossing word re-ordering (*Cross*) on first fixation durations (*FFDur*)

Section 9.4.1 summarises the results in a table. Each dependent variable is then discussed separately. The plots in Figs. 9.5, 9.6, 9.7, 9.8, 9.9 and 9.10 are partial effects plot and they illustrate the effect of the given predictor when all other predictors are held constant. For ease of comprehension, the logarithmically transformed dependent variables were back transformed to their actual values.

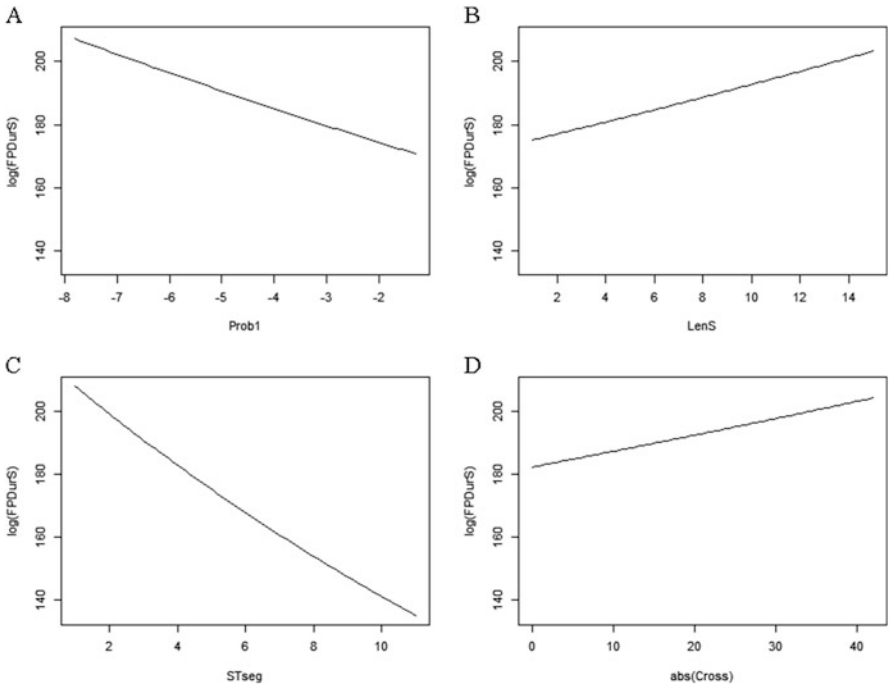


Fig. 9.6 (**a–d**) The effect of frequency (*Prob1*), word length (*LenS*), segment identifier (*STseg*), and absolute *Cross* on (log) gaze durations

9.4.1 Results

9.4.2 First Fixation Durations

First fixation durations represent the first contact with a $word_n$, before the eyes either re-fixate $word_n$ or move on to word to the left or right. All the low-level aspects of word recognition such as integration of visual features of letters occur during a first fixation duration in addition to processing of morphological and phonological aspects all of which result in lexical access. In addition to the processing of $word_n$, $word_{n+1}$ is pre-processed in terms of visual features such as word length.

The effect of frequency on *FFDur* was significant and in the expected direction. The effect of *LenS* was significant, but in the opposite direction of what would intuitively be expected; recall, however, that *FFDur* is the duration of a single fixation, the first on the word, which does not automatically become longer for longer words (e.g. Hyönä and Olson 1995). As indicated in Table 9.2 and Fig. 9.6b, the effect of *LenS* on *FPDurS* was in the expected direction, suggesting that the longer reading times for longer words are due to re-fixations. The significant effect

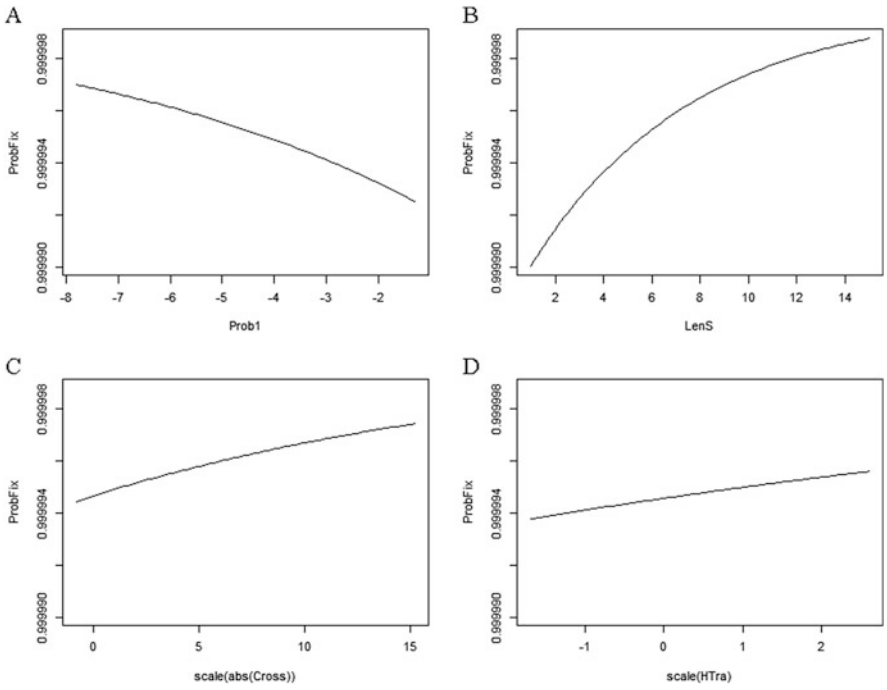


Fig. 9.7 (**a–d**) The effect of frequency (*Prob1*), word length (*LenS*), scaled absolute *Cross*, and scaled word translation entropy on the probability of a fixation

of *STseg* on first fixation durations suggests that translators become faster as they progress in the translation—even for such an early measure.

Both *Cross* and *HTra* were positively significantly correlated with *FFDur*. This suggests that target-language-specific aspects play a role at the earliest stages of reading, i.e. SL and TL are co-activated from the very first visual encounter with an ST word. In addition, words with fewer alternative translations and lower *Cross* values require less effort to process than words with a higher number of alternative translations and higher *Cross*. This may indicate that these ST words are more likely to prime and facilitate their TT equivalents than words with a higher word translation entropy and higher *Cross* value. The *Cross* effect was relatively large, suggesting that re-ordering and structural priming play a large role during the early stages of reading during translation. This seems to confirm the marginally significant effect found in Chap. 10. Together, this further lends support to the literal translation hypothesis, as defined above, in that the default rendering procedure during ST reading in translation is to generate an interim representation in which ST word order and TT word order are identical, where ST and TT items correspond one-to-one and in which each ST word has only one possible translated form. When this is not possible, because of context, target norms or for any other reason, cognitive effort increases.

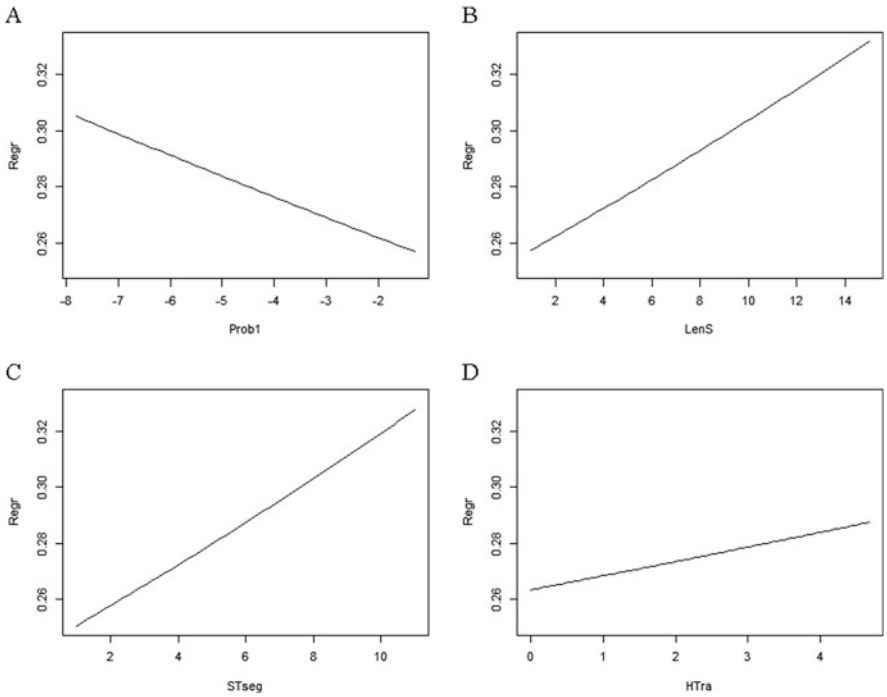


Fig. 9.8 (**a–d**) The effect of frequency (*Prob1*), word length (*LenS*), segment identifier (*STseg*), and word translation entropy (*HTra*) on the probability of a regression

Ideally, linear mixed models should have normally distributed residuals; visual examination of the residual distribution showed that this was not the case for this analysis and the following analysis of gaze duration; instead, the residuals showed a somewhat bimodal distribution. This suggests that, although this model explains a number of aspects of the translation process, there may be important variation that our predictors do not capture. This is not surprising given the large number of variables that may affect translation. It may be noted that mean first fixation durations, gaze durations and regression path durations are relatively short compared to monolingual reading. It may also be noted that the effects of word length and frequency are rather small in comparison to monolingual reading: the effect of e.g. frequency on *FFDur* is typically in the region of 20–30 ms and on gaze duration normally around 50–60 ms while here, it is around 6 ms and 20 ms, respectively. However, it is unlikely that this is task related. It is more likely that this is because of the way fixations are calculated in the different studies.

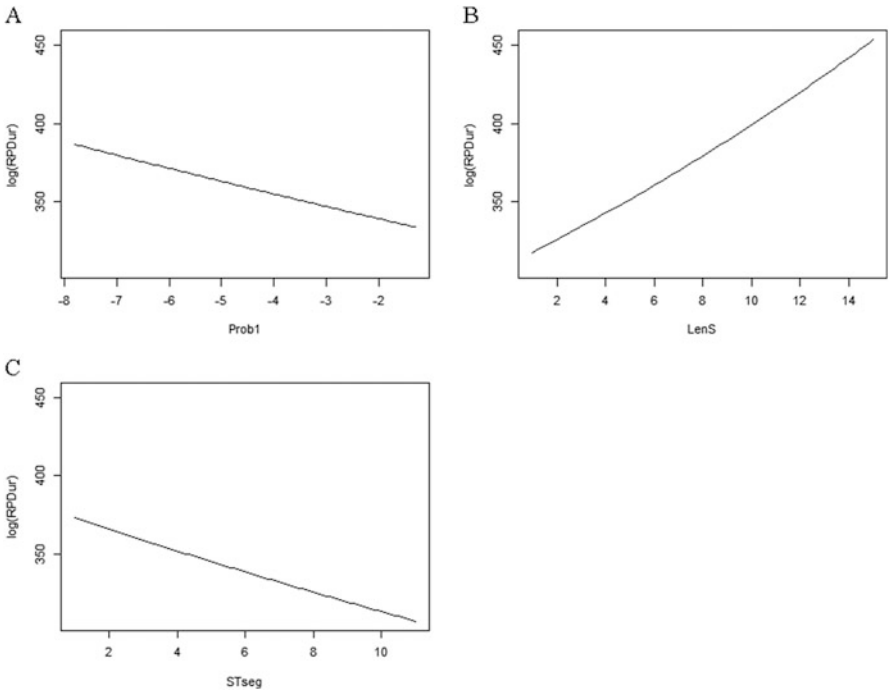


Fig. 9.9 (**a–c**) The effect of frequency (*Prob1*), word length (*LenS*), and segment identifier (*Stseg*) on regression path durations.

9.4.3 First Pass Gaze Duration

First pass gaze duration (*FPDurS*) is the sum of all fixation durations on a word before the eyes move to a different word and these, hence, represent a later stage in lexical processing. A reader might re-fixate a word either because it is long or because it is difficult to understand or integrate, or because it is ambiguous in some way.

As expected, word frequency (*Prob1*) negatively significantly correlated with first pass gaze duration; as mentioned above, word length (*LenS*) also had a significant positive effect, also as expected. The segment identifier (*STseg*) had a negative slope. Somewhat surprisingly, word translation entropy did not have a significant effect on gaze duration. However, *Cross* had a positive slope and was marginally significant. While the effect of *Cross* lingers on into gaze durations, the effect of word translation entropy appears very early on (in first fixation durations and probability of a fixation), and it only surfaces again in total reading time. This suggests that initial activation of shared representations is relatively automatic and that these automatically activated shared representation serve as a reference in the production of the target text—as evidenced by the effect of *HTra* and *Cross* on *TrtS*.

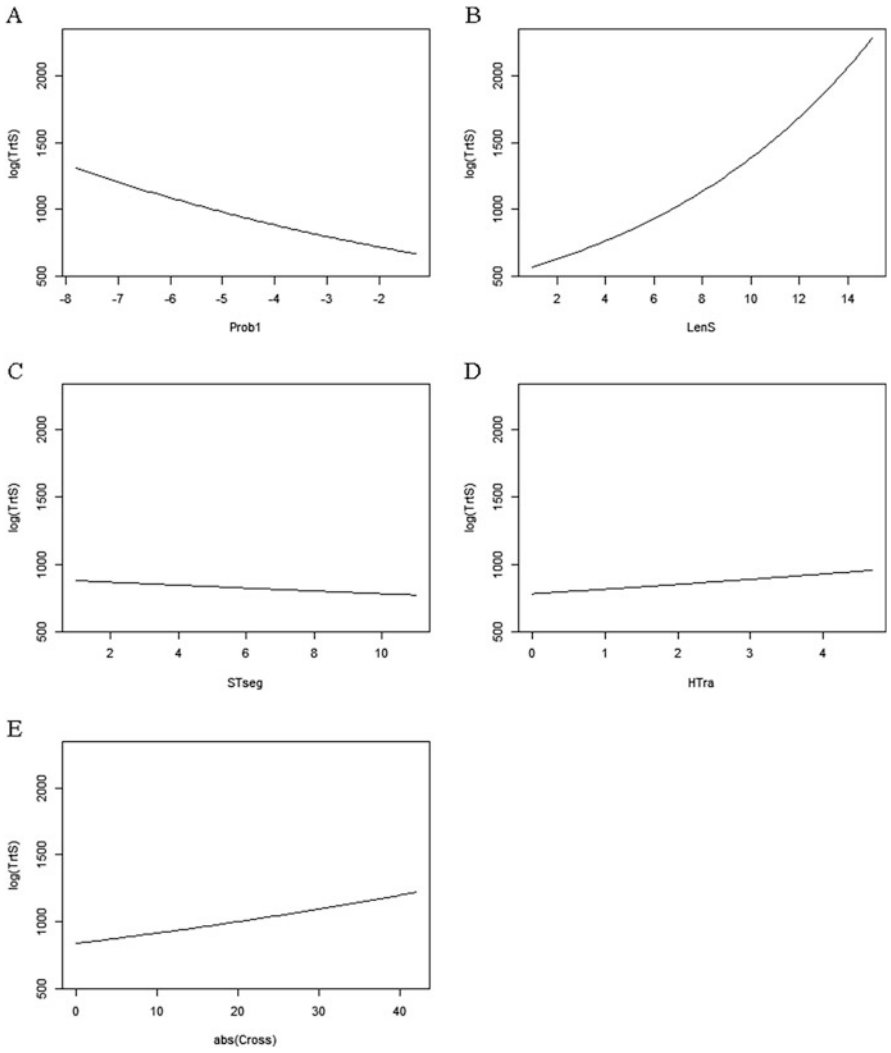


Fig. 9.10 (**a**–**e**) The effect of frequency (*Prob1*), word length (*LenS*), segment identifier (*STseg*), word translation entropy (*HTra*), and absolute *Cross* on total reading times

9.4.4 Probability of a Fixation

Probability of a fixation describes the likelihood that a word is fixated. Some words are never fixated, because they can be guessed from the context and/or they can be pre-processed during the fixation(s) on the prior word. Probability of a fixation can thus give an indication of how predictable a word is or how well it can be pre-processed.

Table 9.2 LMEM results for all predictors and dependent variables

	Prob1					*LenS*					*STseg*				
	β	SE	*t*	*p*		*β*	SE	*t*	*p*		*β*	SE	*t*	*p*	
FFDur	−1.86	4.49	−4.15	3.45E-05	***	−1.12	2.25	−4.99	6.50E-07	***	−3.24	2.04	−15.90	<2E-16	***
FPDurS	−2.98	4.96	−5.96	1.91E-09	***	1.07	2.50	4.27	2.05E-05	***	−4.31	2.29	−18.82	<2E-16	***
Reg	−0.04	0.02	−2.36	0.0182	*	0.03	0.01	3.26	1.10E-03	**	0.04	0.01	5.16	2.41E-07	***
RPDur	−2.27	7.22	−3.14	1.69E-03	**	2.54	3.63	7.00	3.05E-12	***	−1.94	3.14	−6.17	2.18E-09	***
ProbFix	−0.14	0.03	−4.76	1.97E-06	***	0.15	0.02	9.71	<2E-16	***					
TrtS	−1.03	8.21	−12.59	<2E-16	***	9.91	4.12	24.04	<2E-16	***	−1.31	3.79	−3.46	5.45E-04	***
	HTra					*Cross*									
	β	SE	*t*	*p*		*β*	SE	*t*	*p*						
FFDur	1.10	4.24	2.59	9.70E-03	**	3.62	1.49	2.43	0.0151	*					
FPDurS						2.70	1.60	1.69	0.0915	†					
Reg	0.03	0.01	1.77	0.0769	†										
RPDur															
ProbFix	0.08	0.03	2.81	4.99E-03	**	0.05	0.03	1.75	0.08082	†					
TrtS	4.34	7.88	5.51	3.90E-08	***	8.96	2.13	4.21	2.51E-05	***					

†p < .1, *p < .05, **p < .01, ***p < .001

The model for probability of a fixation did not converge when all the random effects were included. We therefore excluded one random effect that showed the least variation in the other models, namely *Text. STseg* was also excluded, because the model did not converge when this predictor was included. In addition, *Cross* and *HTra* were scaled (the variables were scaled by subtracting the mean and dividing by the standard deviation). As expected, higher frequency words were less likely to be fixated than lower frequency words. Equally expected was the effect of word length on the probability of a fixation (positively associated, such that longer words were more likely to be fixated). Very surprising are the effects of *Cross* and *HTra* on the probability of a fixation. Although these effects were modest in size and, in the case of *Cross* only a marginally significant effect was identified, the results suggest that translators anticipate target-language-specific aspects of upcoming words and skip these if they are easy to process. These effects underline the fact that activation of target-related aspects occurs very early.

9.4.5 Probability of a Regression

Probability of a regression describes the likelihood that the eyes move to $word_{n\text{-}m}$ from $word_n$. A regression is normally indicative of integration problems.

For regression probability (*Reg*), the effects of frequency and word length were again in the expected direction and significant. The effect on *Cross* did not reach significance, which is surprising, given that re-ordering might require integration of previously read words. However, it seems that re-ordering by visual inspection of previous words occurs much later as captured by the effect of *Cross* on total reading time (see below). The effect of word translation entropy on regressions was only marginally significant. The fact that *Cross* had no effect on regressions while word translation entropy had a weak effect might suggest that semantic integration of words with higher word translation entropy needs to be resolved contextually by regressing to earlier words. However, neither *Cross* nor *HTra* had an effect on regression path duration (see below), confirming the findings from earlier measures, i.e. ambiguities in terms of semantics (*HTra*) and structural ambiguities or difficulties (*Cross*) have an early effect (*FFDur*) and are resolved late (during *TRT*). Together, this, once again, strengthens the view that horizontal processes occur early and vertical processes occur late. The fact that *STseg* had a relatively large significant and positive effect on *Reg* suggests that translation is an iterative process, i.e. it seems to be common that already translated text is re-read—presumably and especially during the revision phase.

9.4.6 Regression Path Duration

Regression path duration (*RPDur*) refers to the sum of all fixations on a $word_n$ in addition to fixations on prior words before the eyes move on to words situated to

the right of word$_n$. *RPDur* is a relatively late measure and indicative of integration problems.

The effects of frequency and word length on regression path duration were in the expected direction and significant. The effect of *STseg* on regression path duration was more modest than frequency and word length, but all were highly significant. Again, the negative slope suggests that integration is less costly towards the end of the text, given that translators have a relatively good discourse model of both the ST and the TT towards the end of the text, making it easier to integrate difficult words. Visual inspection of the residuals showed a relatively normal distribution. The fact that neither *Cross* nor *HTra* had an effect on regression path durations confirms the findings from earlier measures: shared semantic and structural representations are activated automatically and early, and serve as a basis for production and monitoring during much later processes.

9.4.7 Total Reading Time

TrtS is a very late measure which includes all fixations on a word$_n$—irrespective of when these have taken place.

For total reading time, all effects were highly significant and mirrored those on first fixation durations (apart from word length). Both the effect of *Cross* and the effect of *HTra* on total reading times were relatively strong, positive and highly significant. Again, these findings suggest that the initially and automatically activated shared structural and semantic representations serve as a basis for later regeneration of the ST in the TL and for later monitoring processes.

9.5 General Discussion

The picture that emerges from our findings is that reading for translation is fundamentally different from reading for monolingual comprehension. Monolingual reading in L1 is the most well-researched type of reading, but no target-language-specific aspects play a role in this kind of reading. This is the first study, to the authors' knowledge, which employs earlier eye movement measures and such a broad range of target languages and such a large corpus of eye movements. Early eye movement measures are crucial if the time-course of the cognitive model is to be investigated and they are also important if conclusions regarding the organisation of the mental representations are to be drawn from the findings: late eye movement measures are likely to be indicative of willed behaviour while early eye movement measures are likely to be indicative of more automatic processing. It is not very surprising that target-language-specific aspects play a role during the later processes in reading for translation where TT production is involved, unlike in monolingual reading, which does not involve text production.

However, it is not likely that TT production, i.e. the actual typing, is responsible for the observed effects on early eye movement measures: in the study by Schaeffer et al. (forthcoming), participants were instructed to only start writing once they had a translation in mind, and eye movements were only recorded during the reading phase and not after the TT production was started. The studies by Macizo and Bajo (2006) and Ruiz et al. (2008) also separated the reading phase from the (oral) text production stage and also found target-language-specific effects, but only when the reading purpose was translation, not when the reading purpose was repetition. In other words, even when reading and writing are kept experimentally separated, target-language-specific effects on ST reading are observable.

We found an effect of word translation entropy and syntactic source-target language reordering on first fixation durations and the probability of a fixation. This supports the integrated nature of the bilingual lexicon and cross-linguistic priming: relative word order and semantic overlap between lexical items of two different languages can be quantified and has an observable effect on eye movements during translation. We observed an early and a late effect of word translation entropy and word order, which further confirms what has long been suggested in translation studies (e.g. Englund Dimitrova 2005; Krings 1986), i.e. that translation is subliminal and automatic to a certain extent (see also Wu and Thierry 2012) and partly conscious and willed behaviour. Schaeffer and Carl (2013: 173) argue that "... identification of shared aspects is automatic and there is no conscious control over how source and target are aligned cognitively ..." The evidence provided in the current chapter supports this view and further supports more generally the model proposed by Schaeffer and Carl (2013: 185) which posits "... Early during source text reading, shared representations are activated which then serve as a basis for regeneration in the target language." The early effect of *Cross* and *HTra* is evidence of horizontal, automatic processes while the late effect of these target-language-related aspects on *TrtS* is evidence of vertical monitor processes.

Our results show that words which have been translated in the same way by different translators are more likely to prime and facilitate processing, while words which are "translation ambiguous" (Eddington and Tokowicz 2013), i.e. words with more than one possible translation, are less likely to prime and more likely to inhibit processing already at a very early stage (during first fixation duration)—most likely because of the way the bilingual lexicon represents these items. Our results show that he degree to which two items share structural and semantic representations predicts the strength of the priming effect. That *Cross* and *HTra* have an effect on early eye movement measures lends support to models of the bilingual lexicon which posit non-selectivity, such as the BIA+ model (Dijkstra and van Heuven 2002). This effect further lends support to the DCFM (De Groot 1992) and the Sense Model (Finkbeiner et al. 2004), given that *HTra* is a continuous variable which describes the graded overlap in terms of semantics between a source word and its translation.

Most cross-linguistic structural priming studies have employed comprehension-to-production paradigms: typically, the prime sentence is read and the target sentence is produced and the priming effect describes the influence of the read

sentence on the produced sentence. The study by Kidd et al. (2014) is the only study, to the authors' knowledge, to report structural cross-linguistic priming during comprehension. These authors found a priming effect from a comprehended English sentence to how a German sentence was comprehended. Given that for the present study, we only considered eye movements on the source text and given that we interpret the effects of *HTra* and *Cross* as priming effects, they are situated at the border between comprehension and production—especially in the case of the early effects. The results show that the *Cross* value has an effect on first fixation durations such that words with higher *Cross* values elicited significantly longer first fixation durations. This finding is in line with Kidd et al. (2014), in that low *Cross* values prime and facilitate processing, while words with a higher *Cross* value do not prime and inhibit. In other words, what these findings suggest is that SL representations prime TL-related processes during source text comprehension. *Cross* describes the degree of overlap between source and target in terms of word order. When this overlap is high, co-activated or shared structural representations facilitate source text reading during translation, because, during this kind of reading, production-related representations are already active at a very early stage. E.g. Schoonbaert et al. (2007) found that cross-linguistic syntactic priming can be boosted if the verb in the prime and target sentence is a translation equivalent. During translation, most target words are of course translation equivalents of source words and the relatively modest boost observed in priming studies can be assumed to be much stronger during translation.

It is interesting that *STseg* has a relatively large effect on first fixation durations and that *STseg* has a significant and often relatively large effect on all relevant eye movement measures (apart from probability of a fixation). The consistency and size of this effect suggest that creating a discourse model is of great importance during translation, making a faster processing possible.

Finally, it seems obvious, on the basis of the evidence presented here, that early processes are horizontal and that the output from the early processes serves as a basis for late, vertical processes. It is highly likely that production-related processes and source language reading processes cannot be separated. This conclusion becomes even stronger considering the naturalistic data used in the present investigation, in comparison to some of the other studies mentioned. At the same time, those less naturalistic studies show the interesting fact that separation of SL and TL processes does not even occur when these processes are separated experimentally.

According to these findings, Malmkjaer is right when she argues that the literal translation hypothesis is one of the very few phenomena, which qualifies "... for the status of cognitively determined universals ..." (2005: 17): it should be highlighted that the data for the present study consisted of one source language (English) and six rather distinct target languages (Danish, Spanish, Estonian, Chinese, Hindi, and German). In other words, had we found these effects in one language only, but not in others, it would be possible that the effects are specific to a particular language combination, or a specific target language, rather than a phenomenon which holds across language combinations. One other aspect of the current study may lend further weight to Malmkjaer's claim: while the processes during first fixation durations are of course not completely automatic, an individual has far

less willed control over the processes which are at play during the first 250 ms of the processing of a word than is the case for total reading time. In other words, it is likely that the role of primed representations highlights cognitively determined constraints rather than willed behaviour.

While the present study includes a relatively broad sample of target languages compared to the literature, it is of course limited considering the vast number of different languages across the globe and the findings will require further corroboration. It remains beyond dispute, however, that a multitude of concurrent processes are at play during (reading for) translation, which suggest that reading for translation is fundamentally different from monolingual reading.

References

Baayen, R. H. (2013). *languageR: Data sets and functions with 'Analyzing linguistic data: A practical introduction to statistics'*. Available at: http://cran.r-project.org/package=languageR.

Balling, L. W., & Carl, M. (2014). Production time across languages and tasks: A large-scale analysis using the CRITT translation process database. In J. W. Schwieter & A. Ferreira (Eds.), *The development of translation competence: Theories and methodologies from psycholinguistics and cognitive science* (pp. 239–268). Newcastle upon Tyne: Cambridge Scholars Publishing.

Balling, L. W., Hvelplund, K. T., & Sjørup, A. C. (2014). Evidence of parallel processing during translation. *Meta, 59*(2), 234–259.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). {*lme4*}*: Linear mixed-effects models using Eigen and S4*. Available at http://cran.r-project.org/package=lme4.

Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2007). Shared syntactic representations in bilinguals: Evidence for the role of word-order repetition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(5), 931–949.

Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2013). From language-specific to shared syntactic representations: The influence of second language proficiency on syntactic sharing in bilinguals. *Cognition, 127*(3), 287–306.

Boada, R., Sánchez-Casas, R., Gavilán, J. M., García-Albea, J. E., & Tokowicz, N. (2012). Effect of multiple translations and cognate status on translation recognition performance of balanced bilinguals. *Bilingualism: Language and Cognition, 16*(01), 183–197.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Carl, M. (2012). The CRITT TPR-DB 1.0: A database for empirical human translation process research. In *AMTA 2012 workshop on post-editing technology and practice*.

Carl, M., & Dragsted, B. (2012). Inside the monitor model: Processes of default and challenged translation production. In *Translation: Corpora, computation, cognition. Special issue on the crossroads between contrastive linguistics, translation studies and machine translation, 2*(1), 127–145.

Carl, M., & Schaeffer, M. (forthcoming). Literal translation and processes of post-editing. In: *Translation in transition: Between cognition, computing and technology*. Amsterdam: Benjamins.

Chen, B., Jia, Y., Wang, Z., Dunlap, S., & Shin, J.-A. (2013). Is word-order similarity necessary for cross-linguistic structural priming? *Second Language Research, 29*(4), 375–389. doi:10.1177/0267658313491962.

Chesterman, A. (2011). Reflections on the literal translation hypothesis. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research: integrative approaches in translation studies* (pp. 23–35). Amsterdam: John Benjamins.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 341–371). Amsterdam: Elsevier.

De Groot, A. M. B. (1992). Determinants of word translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(5), 1001–1018.

Desmet, T., & Declercq, M. (2006). Cross-linguistic priming of syntactic hierarchical configuration information. *Journal of Memory and Language, 54*(4), 610–632.

Dijkstra, T., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition, 5*(3), 175–197.

Dragsted, B. (2012). Indicators of difficulty in translation – correlating product and process data. *Across Languages and Cultures, 13*(1), 81–98.

Dragsted, B., & Carl, M. (2013). Towards a classification of translation styles based on eye-tracking and keylogging data. *Journal of Writing Research, 5*(1), 133–158.

Duñabeitia, J. A., Perea, M., & Carreiras, M. (2010). Masked translation priming effects with highly proficient simultaneous bilinguals. *Experimental Psychology, 57*(2), 98–107.

Eddington, C. M., & Tokowicz, N. (2013). Examining English–German translation ambiguity using primed translation recognition. *Bilingualism: Language and Cognition, 16*(02), 442–457.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*(6), 641–655.

Englund Dimitrova, B. (2005). *Expertise and explicitation in the translation process*. Amsterdam: John Benjamins.

Finkbeiner, M., Forster, K., Nicol, J., & Nakamura, K. (2004). The role of polysemy in masked semantic and translation priming. *Journal of Memory and Language, 51*(1), 1–22.

Gile, D. (1995). *Basic concepts and models for interpreter and translator training*. Amsterdam: John Benjamins.

Grosjean, F. (1997). The bilingual individual. *Interpreting – International Journal of Research and Practice in Interpreting, 2*, 163–187.

Hartsuiker, R. J., Kolk, H. H. J., & Huiskamp, P. (1999). Priming word order in sentence production. *The Quarterly Journal of Experimental Psychology, 52A*(1), 129–147.

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science, 15*(6), 409–414.

Hvelplund, K. T. (2011). *Allocation of cognitive resources in translation: An eye-tracking and key-logging study*. PhD thesis, Copenhagen Business School.

Hvelplund, K. T. (2015). Four fundamental types of reading during translation. In A. L. Jakobsen & B. Mesa-Lao (Eds.), *Translation in Transition*. Amsterdam: John Benjamins.

Hyönä, J., & Olson, R. K. (1995). Eye fixation patterns among dyslexic and normal readers: effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(6), 1430–1440. doi:10.1037/0278-7393.21.6.1430.

Jakobsen, A. L. (2011). Tracking translators' keystrokes and eye movements with Translog. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research. Integrative approaches in translation studies* (pp. 37–55). Amsterdam: John Benjamins.

Jakobsen, A. L., & Jensen, K. T. H. (2008). Eye movement behaviour across four different types of reading task. In S. Göpferich, A. L. Jakobsen, & I. M. Mees (Eds.), *Looking at eyes. Eye-tracking studies of reading and translation processing* (Vol. 36, pp. 103–124). Copenhagen: Samfundslitteratur.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354.

Kidd, E., Tennant, E., & Nitschke, S. (2014). Shared abstract representation of linguistic structure in bilingual sentence comprehension. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-014-0775-2.

Krings, H. P. (1986). *Was in den Köpfen von Übersetzern vorgeht: eine empirische Untersuchung zur Struktur des Übersetzungsprozesses an fortgeschrittenen Französischlernern*. Tübingen: Günter Narr Verlag.

Kuznetsova, A., Christensen, R. H. B., & Brockhoff, P. B. (2014). *lmertest: Tests for random and fixed effects for linear mixed effect models (lmer Objects of lme4 Package). R package version 2.0-6*. Available at http://www.cran.rproject.org/package=lmerTest/.

Laxén, J., & Lavaur, J.-M. (2010). The role of semantics in translation recognition: Effects of number of translations, dominance of translations and semantic relatedness of multiple translations. *Bilingualism: Language and Cognition, 13*(02), 157.

Lederer, M. (1994). *La traduction aujourd'hui. Le modèle interprétatif*. Paris: Hachette.

Loebell, H., & Bock, K. (2003). Structural priming across languages. *Linguistics, 41*(5), 791–824.

Macizo, P., & Bajo, M. (2006). Reading for understanding and reading for translation: Do they involve the same processes? *Cognition, 99*, 1–34.

Malmkjaer, K. (2005). Norms and nature in translation studies. *Synaps, 16*, 13–19.

McConkie, G. W., & Yang, S.-N. (2003). How cognition affects eye movements during reading. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 413–427). Oxford: Elsevier.

O'Brien, S. (2006). Eye-tracking and translation memory matches. *Perspectives, 14*(3), 185–205.

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.r-project.org/.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*(8), 1457–1506. doi:10.1080/17470210902816461.

Ruiz, C., Paredes, P., Macizo, P., & Bajo, M. T. (2008). Activation of lexical and syntactic target language properties in translation. *Acta Psychologica, 128*, 490–500.

Rydning, A. F., & Lachaud, C. (2010). The reformulation challenge in translation. In G. M. Shreve & E. Angelone (Eds.), Translation and cognition (pp. 85–108), vi, 381 pp.

Schaeffer, M. J., Paterson, K., McGowan, V. A., White, S. J., & Malmkjær K. (forthcoming). Reading for translation. In A. L. Jakobsen & B. Mesa-Lao (Eds.), *Translation in Transition*. Amsterdam: John Benjamins.

Schaeffer, M. J., & Carl, M. (2013). Shared representations and the translation process: A recursive model. *Translation and Interpreting Studies, 8*(2), 169–190.

Schoonbaert, S., Hartsuiker, R. J., & Pickering, M. J. (2007). The representation of lexical and syntactic information in bilinguals: Evidence from syntactic priming. *Journal of Memory and Language, 56*(2), 153–171.

Shin, J.-A., & Christianson, K. (2009). Syntactic processing in Korean-English bilingual production: Evidence from cross-linguistic structural priming. *Cognition, 112*(1), 175–180. doi:10.1016/j.cognition.2009.03.011.

Tokowicz, N., & Kroll, J. F. (2007). Number of meanings and concreteness: Consequences of ambiguity within and across languages. *Language and Cognitive Processes, 22*(5), 727–779.

Wu, Y. J., & Thierry, G. (2012). Unconscious translation during incidental foreign language processing. *NeuroImage, 59*(4), 3468–3473.

Chapter 10
Syntactic Variance and Priming Effects in Translation

Srinivas Bangalore, Bergljot Behrens, Michael Carl, Maheshwar Ghankot, Arndt Heilmann, Jean Nitzke, Moritz Schaeffer, and Annegret Sturm

Abstract The present work investigates the relationship between syntactic variation and priming in translation. It is based on the claim that languages share a common cognitive network of neural activity. When the source and target languages are solicited in a translation context, this shared network can lead to facilitation effects, so-called priming effects. We suggest that priming is a default setting in translation, a special case of language use where source and target languages are constantly co-activated. Such priming effects are not restricted to lexical elements, but do also occur on the syntactic level. We tested these hypotheses with translation data from the TPR database, more specifically for three language pairs (English-German, English-Danish, and English-Spanish). Our results show that response

S. Bangalore
Interactions Corporation, New Providence, NJ, USA

B. Behrens
Department of Literature, Area studies and European Languages, University of Oslo, Oslo, Norway

M. Carl
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

M. Ghankot
Indian Space Research Organisation, Hassan, Andhra Pradesh, India

A. Heilmann (✉)
English Linguistics Department, RWTH Aachen, Aachen, Germany
e-mail: arndt.heilmann@rwth-aachen.de

J. Nitzke
Department for Language, Culture and Translation Studies in Germersheim (FTSK), University of Mainz, Mainz, Germany

M. Schaeffer
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

Institute for Language, Cognition and Computation University of Edinburgh, Edinburgh, UK

A. Sturm
University of Geneva, 36, rue Prévost-Martin, 1205 Geneva

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_10

times are shorter when syntactic structures are shared. The model explains this through strongly co-activated network activity, which triggers a priming effect.

Keywords Cognitive Effort • Priming • Eye Tracking • Translog • Literal translation • Co-activation • Keystroke Logging

10.1 Introduction

A range of single word studies have investigated the effect of translation ambiguity on behaviour during translation recognition and translation production. Tokowicz and Kroll (2007) noted that when their English-Spanish bilingual participants translated single word stimuli with more than one possible translation, their response times were slower in comparison to words with only a single translation. Tokowicz and Kroll attributed this effect to active competition between translation alternatives. A selection between all possible alternatives is cognitively effortful as all items have to be compared with regards to their appropriateness, and once the appropriate item has been singled out, others have to be suppressed.

Up to now, the effect of translation alternatives on behaviour during translation has been dominated by single word studies (e.g. Laxén and Lavaur 2010; Boada et al. 2012; Eddington and Tokowicz 2013). To the best of our knowledge, the first study to investigate behavioural measures for translations of whole texts is the one by Dragsted (2012). In her study, she found increased total reading time (on source text words), number of fixations (on source text words) and pauses in the production for words with high lexical variance. It therefore seems likely that processing of source text (ST) lexical items involves the activation of target language (TL) competitors. However, a raw count of the number of competitors does not directly reflect their influence on behavioural times, since some options may be more dominant than others, i.e. receive more neural activation. In the following, we will explain how such influence can be accounted for.

As a transfer process including the reproduction of an initial source message in another context, every translation is a selection of a final target formulation out of many possible target formulations (Neubert and Shreve 1992). However, the details of this selection process and the factors influencing it are largely unknown. Whenever an ST is translated by n translators producing TT_n translations, each single translation TT_i is selected out of many possible target texts. Each selection of the actual elements of TT_i is determined by the characteristics of the target language, its morphology, syntax, pragmatics and stylistics, the translation brief and target audience etc., but also by the individual translator, her background and experience. Each final target text TT_i is thus a selection from possible options in the target language which were available to one particular translator at one particular point in time. It is highly unlikely that any two translators will produce exactly the same translation of the same source text. In cases where every translator produces a different translation, one would assume the selection process to be cognitively

demanding, as all possible realizations of TT elements are assumed to have been potentially available to all translators. In cases where all translations of a given source text unit are identical, this can be taken as a sign of lacking choice, as there might have only been a single correspondence in the target language. Consequently, the translation was comparatively easy as the translator did not have to make any choice. Translation competence can be defined in terms of selection and selection effort, namely as "the ability to generate a series of more than one viable target text (TT_1 , TT_2 ... TT_n) for a pertinent source text" and "the ability to select only one viable TT from this series, quickly and with justified confidence." (Pym 2003, 489). However, this does not imply that the selection process and its outcome are the same for all competent translators.

Based on the assumption that different translations created by different translators reflect the options which were available to all translators, Carl and Schaeffer (forthcoming; see also Chap. 9) describe this concept with the term of word translation entropy. Word translation entropy is also a feature in the TPR-DB which is described in Sect. 2.4.7.

The idea behind word translation entropy is that the distributions of the translation probability for each word should be a better predictor than the raw count of translation options. As some translation options can be chosen by more than one translator, such choice behaviour can inform about selection processes in translation. To account for the selection variance, translation entropy measures are higher when each translator produces a different translation and entropy values are low when only a limited number of translation alternatives have been realized. Entropy is a measure of uncertainty in choices.

We propose to use entropy as a measure of a translator's cognitive effort in making choices during translation. Carl and Schaeffer show that when the translation of a word resulted in a high translation entropy i.e. high variation, these words were also more effortful to process than words with low word translation entropy. This affected total reading times of the words on the source and of the target text. Schaeffer et al. (Chap. 9) also found an effect on first fixation durations and skipping probability (on source text words).

10.2 Entropy as a Measure of Variation

The notion of entropy in the sense it is discussed here is borrowed from information entropy and was introduced by Claude E. Shannon. He used the term as a description of the unpredictability or uncertainty of the content of messages. A high information entropy value indicates much uncertainty, which, when used to describe the translation process, represents a set of co-activated translation possibilities that are equally good choices for the translation of a source text item. Claude E. Shannon (1951) used the term information entropy as a measure of the amount of information that is transmitted in a communication process. "Variance" and "information" are interchangeable in this context (Miller 1956). Entropy increases when variation

Table 10.1 Example probability distributions of hypothetical translations. TT_1–TT_6 exemplify the effect of probability distributions on entropy (***H***)

	TT_1	TT_2	TT_3	TT_4	TT_5	TT_6	H
$p(s \rightarrow t_i)$	1.00						0.00
	0.50	0.50					1.00
	0.25	0.25	0.25	0.25			2.00
	0.50	0.16	0.17	0.17			1.79
	0.18	0.18	0.16	0.16	0.16	0.16	2.58
	0.30	0.14	0.14	0.14	0.14	0.14	2.51

increases. The concept of Entropy is denoted by the symbol ***H*** and represents the average amount of non-redundant information provided by each item entering a system. Entropy ***H*** is computed based on the probability ***p*** of an item entering the system and its information. The probabilities $\boldsymbol{p(s \rightarrow t_i)}$ of an ST item ***s*** and its possible translation $\boldsymbol{t_{i\ldots n}}$ are computed as the ratio of the number of alignments $\boldsymbol{s \rightarrow t_i}$ counted in TTs over the total number of observed TT segments, as in Eq. (10.1). The information of a probability p is defined as $\boldsymbol{I(p) = -log_2(p)}$, and entropy ***H*** is the expectation of that information as defined in Eq. (10.2):

$$p(s \rightarrow t_i) = \frac{counts(s \rightarrow t_i)}{\#translations} \tag{10.1}$$

$$H = \sum_{i=1}^{n} p_i I(p_i) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{10.2}$$

Table 10.1 describes the effect of probability distributions on entropy (***H***): if all six translators choose the same translation realization for a given word, the probability of this translation is at its maximum (6/6 = 1) and entropy is at its minimum (0), but as soon as translators opt for different target realizations, entropy increases: if one option has a probability of 0.30 and five other options have each a probability of 0.14, then entropy is relatively high (2.51). If there are four different options, but all four options have the same probability (0.25), entropy is higher than when one of the four options has a higher probability (0.50) than the other three (0.16, 0.17, 0.17). For example, the entropy value 2.51 calculated in the following way:

$$2.51 = -1^{*}\left(0.30^{*}\mathrm{LOG}_2 0.30 + 4^{*}\left(0.14^{*}\mathrm{LOG}_2 0.14\right)\right) \tag{10.3}$$

Instead of counting all possible translation alternatives for a given source item, entropy captures the weight of each of these alternatives and may hence be a better reflection of the cognitive environment of translators working on a given text. In other words, it captures the distribution of probabilities for each translation option, so that more likely choices and less likely options are weighted accordingly. The following section examines possible factors which might have an influence on entropy.

10.2.1 Co-activation and Translation

The first question to be addressed is the onset of the selection process and the effect of the selection process on eye movements during reading for translation. At what point during the translation process does the translator start with the mental production of the target text, and to what extent does this mental production process interfere with source text comprehension?

Studies suggest that both languages of a bilingual are always active. Grosjean (1997) argued that activation of the bilingual's two languages is situated on a continuum which has a relatively monolingual state at one extreme and highly co-activated bilingual state at its other extreme. Grosjean argued that it is the context of the language use which determines where on the continuum the bilingual is currently situated: if both interlocutors speak the same two languages, it is more likely that both languages are active, while when only one interlocutor speaks two languages or two interlocutors do not speak the same two languages, it is more likely that the bilingual(s) are situated closer to the monolingual mode. Translation would situate the bilingual firmly towards the very extreme of the bilingual state. A range of studies supports this hypothesis.

Macizo and Bajo (2006) presented professional translators and naïve bilinguals (Spanish/English) with single sentences containing interlingual homographs. In a masked self-paced reading paradigm, participants were instructed to either read the sentence for oral repetition or for oral translation. According to the condition, participants had to read the Spanish sentences and either translate them into English, or repeat them. The homographs made the sentence ambiguous when their meaning in the other language became activated: e.g. *presente* in Spanish is very similar to the English *present*. While the Spanish word is not ambiguous in the sentence, when translating it into the English word *present*, it could either refer to the present moment or to a gift. Macizo and Bajo found that the ambiguous homograph slowed down reaction times, but only when the purpose of reading was to translate. This effect was more pronounced for naïve bilinguals than for professional translators.

Ruiz et al. (2008) used essentially the same experimental design, but manipulated the frequency of the equivalent target word. They kept the monolingual frequency of critical words in the Spanish source sentence constant while the equivalent target words had either a high or a low frequency. Ruiz et al. found that reaction times were slowed down when the equivalent English target word had a low frequency, but again, this was only the case in the translation condition. We interpret their findings in terms of online parallel activation of source and target items during translation; i.e. both languages are active to a high degree during translation.

Schaeffer et al. (forthcoming) used a similar experimental design: this study compared reading for comprehension with reading for translation, but instead of self-paced reading, they used an eye-tracking paradigm. Furthermore, the authors manipulated the number of target words required to translate a single source word embedded in the same sentence frame. For example, *worry* and *laugh* were embedded in the same sentence frame (Many of the fishermen will *worry/laugh*).

Whereas the translation of *worry* into German requires three words *sich Sorgen machen*, *laugh* can be translated by a single word *lachen*. Schaeffer et al. found that the first fixation duration was 23 ms longer when more than one target word was needed for the translation. Again, this effect occurred only in the reading for translation condition. This study further supports the idea that translation occurs online and that target items are activated early during source text reading.

Wu and Thierry (2012) lend further support to the automatic co-activation of the two languages which they observed even though the experimental design discouraged it. In their ERP study, participants were asked to press a button in response to the presentation of circles or squares. Participants were told that sometimes words would appear on the screen, but were instructed to ignore these. 15 % of these words were interlingual homophones, i.e. their Chinese translation would sound similar to either of the words *circle* or *square*. Wu and Thierry found an N200 effect for these homophones, suggesting that participants had to inhibit their spontaneous reaction of pressing the button any time the English word activated the Chinese words for either *square* or *circle*. Thus, co-activation could be detected in an environment where it was explicitly discouraged and even irrelevant to the task. We therefore assume that both the ST and TT language are simultaneously activated during the entire translation process. That means that the translator becomes engaged in exploring and selecting potential target text elements as soon as she starts reading the source sentence. As both languages may be activated to the same degree, it is likely that they influence one another during this selection process. One form of this mutual influence is priming (see Sect. 10.2.3 below). In addition, given that in the studies by Bajo and colleagues, the effect was more pronounced for bilinguals than for translators, it is possible that translators are better able to control co-activation, due to their training and constant exposure to both languages simultaneously.

The question remains however, whether it is not more beneficial for translators to retain a specific source text construction if this is possible in the target structure. Such a strategy would be cognitively less demanding than the search of an alternative formulation. This question is addressed in the following section.

10.2.2 The Literal Translation Hypothesis

Like many other concepts in Translation Studies, the concept of literal translation is the object of various definitions (Chesterman 2011, 24). However, it is important to be able to quantify literality if the aim is to show that whatever effect is observed is not language specific, if the aim is to produce a model of translation which is language independent. Carl and Schaeffer (forthcoming) propose a definition of literality which allows for quantification of the phenomenon. According to their

definition, a translation is literal when the following three literality criteria are fulfilled:

1. Word order is identical in the ST and TT.
2. ST and TT items correspond one-to-one.
3. Each ST word has only one possible translated form in a given context.

Literality criterion 3 is of particular interest as it refers to translation entropy. Expanding this criterion to syntactic features, we stipulate that the translations are structurally literal if an ST sentence is translated into the target language with a single syntactic structure by all translators in a given sample. Syntactic entropy measures the uncertainty that different translators will produce the same TT structure for a ST sentence. Syntactic entropy is an indicator for the literality of translations on a syntactic level, and we introduce syntactic literality to the three literality criteria above:

4. All translations of a given source sentence are translated into the target language with the same syntactic structure.

Thus syntactically literal translation would be one with syntactic entropy of 0. Using entropy measures, literality can be studied using a quantitative approach. In line with Ivir's (1981) notion of formal correspondence, literality has been associated with less cognitive effort than non-literal translations. Ivir (1981, 58) describes the translation process as follows:

> The translator begins his search for translation equivalence from formal correspondence, and it is only when the identical-meaning formal correspondent is either not available or not able to ensure equivalence that he resorts to formal correspondents with not-quite-identical meanings or to structural and semantic shifts which destroy formal correspondence altogether. But even in the latter case he makes use of formal correspondence.

Equally related to this notion of formal correspondence as employed by Ivir is Toury's (1995, 275) "law of interference" which postulates that "(. . .)in translation, phenomena pertaining to the make-up of the source text tend to be transferred to the target text." Similar to Ivir, Toury used this law of interference to posit that less cognitive effort is involved in the production of literal translations as they are a kind of "default setting" in the translating mind. In sum, we argue that the default option for a translator is to consider a literal translation which is more likely to be activated first due to a priming effect (see below) and we further argue that if the default is not acceptable or if other, less literal options are activated, this leads to more cognitive effort.

10.2.3 Priming and Variation in Translation

Priming is a psychological effect that affects language in response to stimuli so that the prior encountered element is repeated or processed faster. This effect has been

observed in studies involving one language for semantic representations, but more relevant for the present purpose is that this has also been observed for structural representations in tasks involving one language (cf. Pickering and Ferreira 2008). In addition, there is some evidence that structural priming has also been observed in studies involving two languages, i.e., in cross-linguistic structural priming studies (e.g. Hartsuiker et al. 2004). These studies suggest that semantic and structural representations are shared between languages when these are similar in the two languages (e.g. Duñabeitia et al. 2010; Bernolet et al. 2013). It is likely though, that the mechanism underlying cross-linguistic structural priming requires a similar construction i.e. congruent word order in both languages (Hartsuiker et al. 1999; Hartsuiker and Westenberg 2000; Bernolet et al. 2007; Loebell and Bock 2003; Kidd et al. 2014). If the word order of the source text can be transferred to the translation, this can result in lower total reading times as has been shown by Jensen et al. due to a possible "automatic transfer of L1 syntax to all types of L2 processing" (Jensen et al. 2009, 333). However, there is also evidence that syntactic structures can be primed across languages if the word order in both languages is different. Desmet and Declercq (2006) tested a sentence completion task that showed syntactic priming effects for relative clause attachment from Dutch to English, even though word order restrictions such as verb final position of Dutch sentence is different from the word order in English.

Shin and Christianson (2009) investigated priming effects of functionally equivalent dative-constructions in Korean and English with the help of a sentence recall task. The English target sentence was presented via audio and was either a double object or prepositional object construction. These sentences were followed by a Korean prime either with a prepositional dative construction, post-positional dative construction or double object construction. In their analysis, they found evidence for an argument-order independent priming effect of post-positional dative constructions, primed by functional correspondences, as this construction is the functional equivalent of the canonical English prepositional dative. Similarly, Chen et al. (2013) observed priming effects of English passive structures on Chinese passive structures and vice versa, when participants were asked to describe a picture after being exposed to a passive or active priming sentence in the other language. Priming occurred despite different word orders. It is therefore possible that formal correspondences between languages are a strong but not a necessary factor for cross-linguistic syntactic priming.

Cross-linguistic semantic priming has been associated with a facilitation effect and structural cross-linguistic priming can thus be argued to also facilitate translation. Schoonbaert et al. (2007) found that cross-linguistic syntactic priming can be boosted if the verb is a translation equivalent in prime and target sentence. During translation, most words are of course translation equivalents and the relatively modest boost observed in priming studies can be assumed to be much stronger during translation.

Due to the nature of priming as a general psychological effect, it is to be expected that translators are affected by a structure in a source text to a similar degree. Translators that are thus primed by a syntactic structure, are likely to produce

translations with the same syntactic structure in the target language. For the measure of syntactic variation, the logical consequence would be that lower entropy measures are related to priming since a single translation choice with a high translation probability can lower entropy drastically. Syntactic priming effects may depend on several characteristics of the input, for example, a cognate verb with the same argument frame.

10.3 Research Questions and Hypotheses

According to the theoretical framework presented above, we assume that the two languages are co-activated during translation. Furthermore, we hypothesise that priming works as a kind of default setting, i.e. shared syntactic nodes of the cognitive network are activated across the source and the target language providing a facilitation effect for the translator. Such facilitation effects should be reflected in lower cognitive effort, and hence in lower behavioural measures than in cases where translators tend not to use the same ST structure for the TT. For the latter case, we predict comparatively higher behavioural measures.

To measure priming effects in translation, we apply the concept of syntactic variance as measured by entropy. In particular, we address the following research questions:

(RQ1) • Can priming effects account for syntactic entropy in translation?
(RQ2) • What influences priming effects in translation?
(RQ3) • Does syntactic variation in translation (as measured by entropy) have an effect on cognitive effort?
(RQ4) • Do priming effects modulate the cognitive effort related to syntactic entropy?

These four research questions will be answered by testing the following hypotheses:

(H1) • Segments with low entropy values reflect priming effects and are highly correlated with lower behavioural measures as compared to segments with high entropy.
(H2) • We predict that priming probability has a negative effect on behavioural measures such that items which are highly likely to have the same syntactic structure as the source sentence receive less attention than those sentences which are highly unlikely to have the same syntactic structure as the source sentence. It is expected that priming probability interacts with syntactic entropy.

In the following, we will test the above hypotheses on the basis of datasets from two tasks—translation between one source language and three different target languages and monolingual copying of the same texts that were also used in the translation task. The copying task is similar to the translation task in that both tasks require source text reading and typing. However, copying does not involve transfer

between two linguistic systems. In this sense, the copying data serves as a control condition: if the syntactic entropy effects we observe in the translation condition are also found in the copying task, it is likely that they represent monolingual source-language-related processes. If, however, syntactic entropy has no effect on behavioural measures during copying, it is likely that these effects are driven by task and target-language-related processes.

10.4 Translation Condition

10.4.1 Participants

The German data was produced by 24 translators (13 translation students, 11 professional translators), the Danish dataset contains translations from 24 translators (12 translation students, 12 professional translators). The Spanish data collection had the most translators with 32 translators but only five professionals (27 students, 5 professionals).

Sixteen subjects participated in the monolingual copying task. All of them had learned English at school and/or university for 4–18 years. Twelve of them were students currently enrolled in a translation programme, two have a degree in translation, and one was never engaged in translation studies. Due to calibration problems, one participant was excluded. Eye-tracking and keylogging data were thus collected and analysed for 15 participants. Twelve participants in the copying task were native speakers of German, one of Turkish, one of French, and one had German and Dutch as his/her first language.

Participants in the baseline condition had to fill out a questionnaire before the experiment. They were instructed to copy the English text and were informed that comprehension questions would follow the task. Three questions for comprehension followed the task. Keystrokes and gaze data were recorded with a Tobii T120 eye-tracker and processed with Translog II (Carl 2012).

10.4.2 Material

The translation data were extracted from the CRITT-TPR database (see Chap. 2): (SG12 for German, KTHJ08 for Danish, BML12 for Spanish) The datasets contain translations of the same six English source texts with the exception that the Danish Study contains only the first three source texts. The datasets contain eye tracking data from a Tobii T120 eyetracker, and keylogging information recorded with Translog (Jakobsen and Schou 1999) and the resulting data was processed with Translog II (Carl 2012) before analysis.

Table 10.2 Properties of the target texts of the translation and the copying condition respectively into the four target languages: Session (number of target texts), Fdur (in hours), Kdur (in hours), Tlen (number of target tokens)

Study	Session	TL	Task	Texts	Part	Fdur (in hours)	Kdur (in hours)	Tlen (in tokens)
TDA14	48	en	C	1–6	11	6.1	5.8	6792
KTHJ08	69	da	T	1–3	24	6.4	5.5	10,571
SG12	47	de	T	1–6	24	9.4	4.6	6632
BML12	63	es	T	1–6	32	8.2	5.8	8936
Total	227	4	2	6	91	30.1	21.7	32,931

Table 10.2 contains a detailed overview of the produced target texts: it indicates the translation (*Task)*, text copying (*C*), translation from-scratch (*T*) and participants (*Part*) involved, the number of translation sessions (i.e. target texts produced), as well as the duration and the total number of target language tokens for each translation mode. Translation (and copying) duration is measured in two different metrics:

- *Fdur*: total production time for all segments, excluding pauses >200 s.
- *Kdur*: total duration of coherent keyboard activity excluding keystroke pauses >5 s. (in the following, we will use refer to Kdur as coherent typing activity for ease of comprehension)

The BML12 study, for instance, contains 63 from-scratch translations which were produced by 32 translators (participants). Each participant had to edit, post-edit and to translate two texts in each mode, and texts were distributed in a randomized order. As shown in Table 10.2, the translated texts together amount to 32,931 target text words which were produced in the 227 translation sessions. Gaps of keystroke activity for more the 200 s (almost 2.3 min) are excluded, under the assumption that translation activities are interrupted in such instances. However, no such pauses were observed in these studies (Fdur is a standard measure in the database and other datasets do have pauses over 200 s).

Table 10.2 also contains information concerning the monolingual copying condition (TDA14) which will be used to contrast the results from the data acquired from the translation condition. A monolingual task that does not involve code-switching of any kind should not reflect entropy measures. Note that during the copying task 95 % of the text production time has been spent on coherent typing (*Kdur*).

10.5 Analysis

10.5.1 Annotation

A detailed description of the annotation used in this study, together with a discussion of possible alternative annotations is available in Chap. 12. In this section, we shortly summarize the main features.

The ST and the TT were parsed according to clause type, voice and type of argument structure. Clause type was annotated as either an independent or dependent clause. Simple sentences as well as main clauses were tagged as independent (I) while subclauses, were annotated as dependent (D). Voice was either annotated as passive (P) or active (A). The third dimension captured verb-argument structures. When the verb of the clause was subcategorized for a direct object or a complement, it was referred to as transitive (T). When it subcategorized for a prepositional object or no object it was labelled intransitive (I). Other argument structures considered were ditransitive structure (D) but also clauses with empty subjects or extraposed subjects (e.g. *Es comprensible que* . . . .[It is understandable that . . .]). These cases were tagged with (M) as in impersonal. A clause characterized as Transitive, Active and Independent thus receives the tag TAI. Segments with multiple clauses and thus multiple tags are merged to longer tags such as TAI_TAD representing a transitive active main clause with a transitive and active subclause. The probabilities of the different translations were computed on the basis of number of occurrences for each tag.

To assess the first and the second research questions, syntactic structures in the annotated translation data have been classified into two categories: PRIME and DIFFERENT. We consider as PRIME every TT segment that preserves the structure of the corresponding ST segment. The category DIFFERENT contains all segments which show a structural change in the TT segment as compared to its corresponding ST segment.

In addition to the original annotation, two new tags were assigned in a category which describes the relationship between the syntactic structures in corresponding source and target text segments. The tag PRIMED was attributed whenever ST and TT structure were identical. The tag DIFFERENT was used whenever different structures were used in the TT as compared to the corresponding ST segment.

The complete dataset was split up into language specific datasets. To identify cases of priming, the target text segments were annotated in the same way as the source text segments with the same annotation scheme as the source text. Source and target structures were compared and categorized as either a prime if they were the same or as different when their structures did not match (see Table 10.3). Title segments were excluded from the analysis due to unusual grammatical properties. This removed 10 % of the data so that 1156 observations remained for analysis.

To answer the research question on the relationship between entropy and priming effects and cognitive effort (RQ2), behavioural translation data from the three language pairs (English-Danish, English-German, English-Spanish) were annotated

Table 10.3 Example of a priming annotation

Source	Target	Count	Comparison	Priming probability
DAI	DAI	4	PRIMED	0.5
	DAI TAD	2	DIFFERENT	

for their syntactic structure and later jointly assessed in mixed linear models. A monolingual copying task served as a baseline. The baseline measures, in contrast to the translation condition, should not be affected by syntactic entropy since syntactic entropy is driven by the TL and not the SL. This control condition will confirm that syntactic entropy actually measures variation in translation and that it is not due to processes related to monolingual ST comprehension. Further, controls were integrated for the analyses of syntactic entropy by means of multivariate statistics controlling for potential confounding factors.

10.5.2 Statistical Analyses

For the analyses, the program R (R Development Core Team 2014) and the lme4 (Bates et al. 2014) and languageR (Baayen 2013) packages were used to perform linear mixed-effects models (LMMs). Since lme4 does not compute p-values, the R package lmerTest (Kuznetsova et al. 2014) was applied. It uses ANOVA for mixed-effects models using the Satterthwaite approximation to estimate degrees of freedom. The behavioural measures here are reported per source text segment as provided by the .sg files from the CRITT-TPR database.

10.5.3 Linear Regression Modelling

Behavioural measures that were chosen as dependent variable were coherent typing activity per word, total reading times for target text and source text per word as well as the average first fixation duration for each segment. Coherent Typing activity is defined as the duration of coherent keyboard activity excluding pauses that are longer than 5000 ms, measured for the production of each segment (see Chap. 2). Total reading time represents the sum of all fixations on a particular segment. The total reading times and coherent typing activity were normalized by dividing the segment measures by the number of tokens constituting the respective segment.

To assess whether priming is an effect that modulates the effect of syntactic entropy on behavioural measures, the ratio of primed to non-primed structures in each segment was assessed in addition to the prior measures of entropy. This ratio can be conceived of as priming probability p_{syn}. It is computed by dividing the number of primed syntactic structures i.e. translations of segments whose structures

are the same as in the source text segment by the total number of translations of this segment (see Eq. 10.4).

$$p_{syn}\left(ST_{syn} == TT_{syn}\right) = \frac{primed}{\#translations} \tag{10.4}$$

A p_{syn} of one means that all translators chose the same structure and a p_{syn} of zero that no translator chose the same source text structure. p_{syn} enters the model as an interaction effect that is modelled as the product of syntactic entropy and p_{syn}.

The interaction effect of syntactic entropy and priming probability should be negatively correlated with measures of total reading time for example, because higher degrees of priming would facilitate processing and weaken the effect of variation.

10.5.3.1 Control Variables

In order to isolate the effect of entropy, a number of control variables were introduced:

Expertise is a strong determinant of translation behaviour. Experts are usually faster than non-experts. It is conceivable that experts may have developed selection strategies so that the effect from co-activation of translation equivalents is reduced, which would lower the effect of entropy on behavioural measures. Expertise is introduced to the model as a categorical variable: 1 represents professional translators, 0 students.

10.5.3.2 Clause Number Within a Segment

The more clauses a segment contains, the more complex it can be believed to be. It is thus possible that syntactic entropy does not capture variation but complexity since the annotation reflects the number of clauses. In production and in comprehension, the complexity of a segment can be thought to influence behavioural data either because reading speed increases because of the expectation of less important information in subclauses or due to difficulty in tracing coherence structure in complex sentences.

10.5.3.3 Word Length

Normalization by the number of words has the disadvantage that different word lengths cannot be accounted for. By chance, syntactic variation could be high in segments containing multiple long words and thus would be associated with higher reading times since longer words are more prone to multiple regressive eye movements and re-fixations. To control for this, average word length per segment

has been computed for source text segments and the translated target text segment by dividing the total number of characters per segment by the number of words per segment.

10.5.3.4 Word Order Changes

Furthermore, the cognitive effort to align word orders between source and target language may be the reason for increased reading times. For this reason, the *Cross* feature has been introduced in the model given that the degree of reordering that was necessary to produce a translation may have an effect on behavioural measures (for an explanation of the *Cross* feature, see Chap. 2).

10.5.3.5 Inefficiency

Some segments may have been more prone to typographical errors and may have undergone major restructuring efforts. Therefore, inefficiency was introduced as a control variable. It is calculated by the number of characters produced during a translation divided by the final amount of characters in the final translation (see also Chap. 2).

10.5.3.6 Random Variables

The last two confounding factors that are controlled for are idiosyncratic differences of the different languages, different participants and items (i.e. the unique segments that were translated) accountable for some variation in the data. They are modelled as random effects, so that the model considers individual intercepts of each participant of each study, and each item.

10.5.3.7 Control over Confounding Variables for First Fixation Durations

Unfortunately for first fixation duration, no appropriate control for confounding factors apart from transformation demands, the random effects, and expertise exist in the dataset. Word length has a very low or even no impact on first fixation durations which is why it was not controlled for (Pollatsek et al. 2008). The contextual environment, i.e. predictability has a large effect on eye movements (e.g. Ehrlich and Rayner 1981). However, it is not possible to infer how constraining the context is from the data directly. The closest indicator that might capture context is priming probability: it is possible that high contextual constraints lead to a more condensed translation probability distribution (Prior et al. 2011, 107).

10.6 Data Trimming

For reading related measures the data was further trimmed so that average total reading times per word below 200 ms per word were excluded, as were segments that were headlines due to their unusual grammatical properties. Also, participants with an extremely bad data quality were excluded completely from reading time analyses when half of their normalized total reading duration fell below an average reading time of 200 ms. The behavioural measures provided by the three studies and three languages in the TPR database (BML12-Spanish, SG12-German and KTHJ08-Danish) were each logarithmically transformed to reduce skew so that the data assumed a shape more similar to a normal distribution. The behavioural measures for each study were then standardized by centering and scaling and assessed as a single dataset. Data points exceeding ± 2.5 standard units were removed before analysis from all measures. 11.4 % of the data was removed from the coherent typing activity analysis (1121 observations left). 33.6 % of the data for total reading time of the source and average first fixation duration were removed (831 observations left). For the total reading time of the target text 34.2 % have been excluded (833 observations left).

10.7 RQ1: Syntactic Entropy and Priming

To test if priming effects can account for syntactic entropy measures, the segment groups termed *Different* and *Primed* were tested to find out if primed and non primed structures affect syntactic entropy differently. A Wilcoxon-Mann-Whitney significance test for categorical data was conducted. The test revealed significant differences in the distribution of primed and non-primed structures with respect to entropy. The difference between both structure types was highly significant for German ($W = 16{,}250$, p-value <0.001), Spanish ($W = 24{,}476$, p-value <0.001) and Danish ($W = 36{,}572.5$, p-value <0.001).

The groups i.e. primed structures and non-primed structures show a difference of almost one unit of entropy when assessing the median (Fig. 10.1). Priming effects by the source text are a very likely explanation for the very low entropy values of zero to one which indicates that priming can streamline translation and reduce syntactic variation.

10.7.1 RQ2: Restructuring and Priming

If priming effects do not result in adherence to the original structure, there may be a systematic reason. As has been indicated before, priming may be promoted by congruent word order for the same syntactic representation. In order to test this

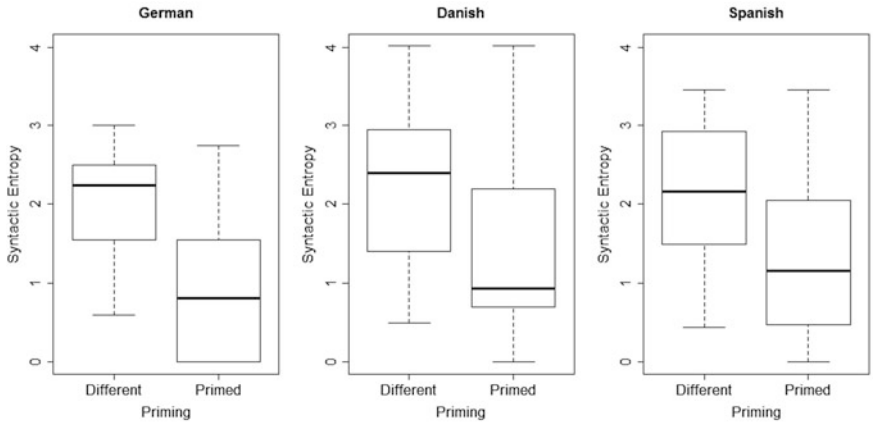


Fig. 10.1 Distribution of primed and non-primed structures in relation to entropy per language-pair

hypothesis, *CrossS* (see Chap. 2), i.e. the relative distortion from source text words to target text words was modelled as a predictor of priming probability with the help of a univariate linear regression. The model was significant with $F(1, 1154) = 40.83$, $p < 0.001$, $R2 = 0.03334$.

Figure 10.2 shows that structuring effort correlates negatively with priming probability. However, primed structures occur also in cases when the average *CrossS* value exceeds the value of 1, which is the literal translation default. Structural priming effects that occur despite congruence also corroborate studies by Chen et al. (2013), Desmet and Declercq (2006) and Shin and Christianson (2009), who provide evidence suggesting that word order similarity is not necessary for priming effects to occur. However, the results clearly indicate higher chances of priming for segments with lower to no restructuring effort. The lower entropy values for primed structures may indicate that increased restructuring effort is eventually a source of deviation from the syntactic representation of the source, since priming is inhibited.

10.8 RQ3: Syntactic Entropy and Behavioural Measures

10.8.1 Total Reading Time (Source)

This section provides the results of the multivariate linear regression analyses, beginning with source text specific eye-tracking measures, followed by production measures. Total reading time of the source was assessed to measure the impact of syntactic entropy (Table 10.4).

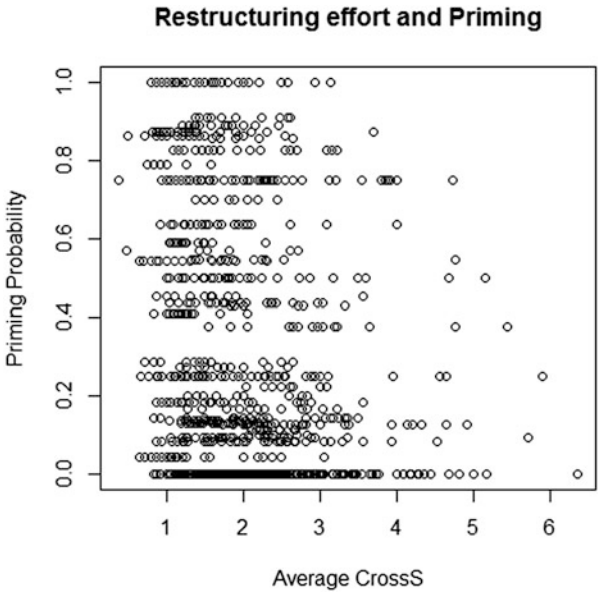


Fig. 10.2 Average cross predicting priming probability

Table 10.4 Linear mixed model (LMM) for the effect of Syntactic entropy on total reading time (source)

Formula	Total reading time (source) ~ syntactic entropy × priming probability + average word length + expertise + CrossS + (1|Participant) + (1|Unique_Segment-ID)			
Variable	β	Standard error	t-value	Significance level
Syntactic entropy	0.16	0.06	2.59	*
Priming probability	0.07	0.07	0.93	
CrossS	0.07	0.03	2.19	*
Number of clauses	0.07	0.03	1.99	*
Expertise	−0.31	0.18	−1.73	†
Average word length	0.17	0.07	2.50	*
Interaction effect	−0.04	0.05	−0.82	

The significance rates reflect participant and item variability
†p < .1, *p < .05

Syntactic Entropy turned out to be positively associated with total reading time of the source text (Fig. 10.3). Similarly, the restructuring effort (*CrossS*), clause complexity (number of clauses) and average word length of the segment displayed positive and significant slopes. The effect for expertise was marginally significant such that professional translators read the source faster than non-professional translators.

The control variables displayed no unexpected behaviour, so that the model seems to measure these dimensions quite well. The fact that no significant

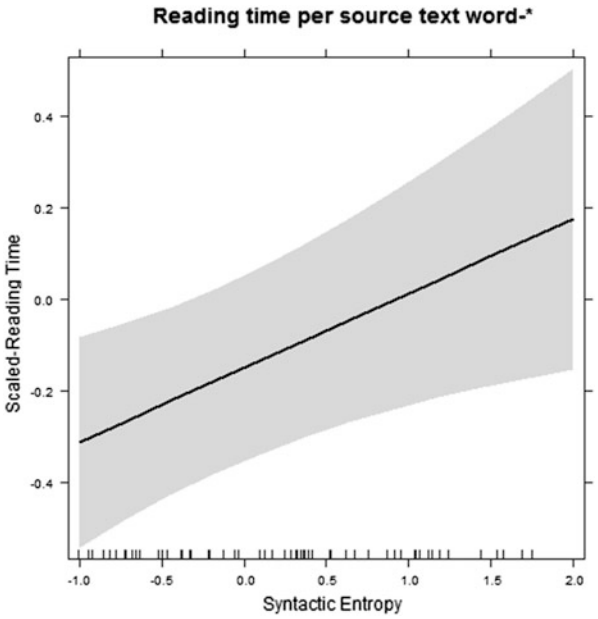


Fig. 10.3 The effect of syntactic entropy on total reading time (source)

interaction effect could be observed for Syntactic Entropy and priming probability nor for priming probability alone contradicts the hypothesis that the effect of entropy on behavioural measures is further modulated by priming effects.

10.8.2 First Fixation Duration

The model for the average first fixation duration of source text words, consisted only of Syntactic Entropy and Priming probability. In this model, entropy displayed a small but marginally significant positive effect on average first fixation duration (Table 10.5).

Further research is warranted to assess the effect of entropy on measures of first fixation duration with more control over confounding factors. The effect of syntactic entropy on a measure of first fixation duration for words may suggest that phenomena of syntactic choice are influencing lexical recognition processes from very early on. A possible reason for higher first fixation durations may be that each of the already activated syntactic choices may compete for integration with the new input, which would of course presuppose that several possible alternative structures are entertained in parallel, rather than serially. However, given the fact that the effect is very weak and given that we only took average first fixation durations into consideration, more research is needed to draw more resilient conclusions.

Table 10.5 LMM for the effect of syntactic entropy on first fixation duration (source)

Formula	Average first fixation duration ~ syntactic entropy × priming probability + (1|Participant) + (1|Unique_Segment-ID)			
Variable	β	Standard error	t-value	Significance level
Syntactic entropy	0.08	0.05	1.77	†
Priming probability	0.02	0.06	0.37	
Interaction effect	−0.04	0.04	−0.96	

The significance rates reflect participant and item variability
†p < .1

Table 10.6 LMM for the effect of syntactic entropy on coherent typing activity (target)

Formula	Coherent typing activity (Target) ~ syntactic entropy × priming probability + average word length (T) + clause complexity + expertise + CrossS + (1|Participant) + (1|Unique_Segment-ID)			
Variable	β	Standard error	t-value	Significance level
Syntactic entropy	0.13	0.04	3.22	**
Priming probability	0.05	0.04	1.36	
Typing inefficiency	0.52	0.02	27.56	***
CrossS	0.01	0.02	0.22	
Number of clauses	0.00	0.02	−0.28	
Average word length	0.41	0.03	14.36	***
Expertise	−0.29	0.15	−1.96	†
Interaction effect	−0.03	0.05	−0.49	

The significance rates reflect participant and item variability
†p < .1, **p < .01, ***p < .001

10.8.3 Coherent Typing Activity

During production, syntactic entropy showed a highly significant and positive association with coherent typing activity. Typing inefficiency and average word length were significant predictors of coherent typing activity. The difference between professional translators and translation trainees was marginally significant and suggests that professional translators tend to be faster writers than non-professionals.

The results obtained here indicate that translation choice in terms of syntactic structure not only slows down reading processes of the source and target text (see below), but also coherent typing activity. Slower typing activity may be caused by higher cognitive load due to selection pressure but also due to revisions. The strong influence of typing inefficiency may be an indication for this (Table 10.6; Fig. 10.4).

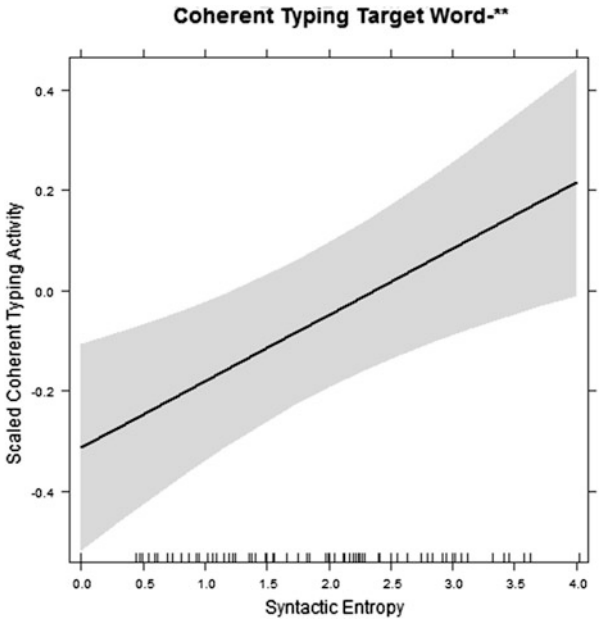


Fig. 10.4 The effect of syntactic entropy on coherent typing activity (target)

10.8.4 Total Reading Time (Target)

For the total reading times of the target text syntactic entropy was again associated with a marginally significant increase of reading duration. Compared with total reading time of the source text, the effect of syntactic entropy on total reading time of the target is slightly less pronounced. This is plausible since in the revision phase of a translation task, gaze accumulates on different stretches of text for detection and correction of typos and mistakes. The effect of syntactic entropy is thus more measurable in the drafting phase of translations. Typing inefficiency was the strongest significant predictor of reading time in the model, followed by average word length. The number of clauses showed a slight but significant decrease in total reading time. Skipping less important stretches of text, e.g. dependent clauses during revision may be a likely explanation (Table 10.7).

Table 10.7 LMM for the effect of syntactic entropy on total reading time (target)

Formula	Total reading time (target) ~ syntactic entropy × priming probability + average word length (T) + clause complexity + inefficiency + expertise + CrossS + (1|Participant) + (1|Unique_Segment-ID)			
Variable	β	Standard error	t-value	Significance level
Syntactic entropy	0.11	0.06	1.83	†
Priming probability	0.00	0.07	−0.05	
CrossS	−0.01	0.03	−0.17	
Expertise	−0.02	0.14	−0.12	
Typing inefficiency	0.24	0.03	8.73	***
Number of clauses	−0.07	0.03	−2.61	**
Average word length	0.23	0.04	5.84	***
Interaction effect	−0.03	0.05	−0.49	

The significance rates reflect participant and item variability
†p < .1, **p < .01, ***p < .001

10.9 Analysis of the Control Group

In order to verify that syntactic entropy is indeed driven by target-language-related and task specific aspects and not by SL processes, the copying data was tested against the entropy values of each language:

behavioural Measures of the Copying condition ~ Syntactic Entropy (German) ...
behavioural Measures of the Copying condition ~ Syntactic Entropy (Danish) ...
behavioural Measures of the Copying condition ~ Syntactic Entropy (Spanish) ...

Cleaning of the copying data was conducted in the same fashion as the translation condition. For coherent typing activity, 14 % of the data was excluded (528 observations left), for total reading time on the source 19 % was excluded (500 observations left), for first fixation durations 19 % of the data was discarded (496 observations left) and total reading time on the target text 42 % of the data had to be excluded (375 observations left).

The same random effects as before entered the equation. The results for the copying condition were controlled for average word length per segment, typing inefficiency and number of clauses per segment. Translation expertise, priming probability, and restructuring are unlikely to play a role in a copying task, which is why they have not been included to avoid unwarranted over-fitting of the model.

10.9.1 Total Reading Time of the Source

The model for total reading time of the source displayed, as expected, only a significant effect for average word length, such that average word length was positively associated with longer average total reading times of the source. Syntactic Entropy was not significant (Table 10.8).

10.9.2 First Fixation Duration (Copying)

Maybe not surprisingly, none of the syntactic entropy values from either of the languages had an effect on first fixation durations during copying (Table 10.9).

Table 10.8 LMMs total reading time (source) (copying)

Formula	Total reading time (source) ~ syntactic entropy(language) + average word length + number of clauses + (1|Participant) + (1|Unique_Segment-ID)				
	Variable	β	Standard error	t-value	Significance level
Spanish	Syntactic entropy	0.02	0.09	0.23	
	Number of clauses	−0.03	0.04	−0.91	
	Average word length	0.34	0.12	2.93	**
German	Syntactic entropy	0.07	0.10	0.76	
	Number of clauses	−0.04	0.04	−1.04	
	Average word length	0.32	0.12	2.64	*
Danish	Syntactic entropy	0.07	0.10	0.63	
	Number of clauses	−0.03	0.05	−0.70	
	Average word length	0.44	0.19	2.37	*

The significance rates reflect participant and item variability
*p < .05, **p < .01

Table 10.9 LMMs for first fixation duration (source) (copying)

Formula	First fixation duration ~ syntactic entropy + (1|Participant) + (1|Unique_Segment-ID)				
	Variable	β	Standard error	t-value	Significance level
Spanish	Syntactic entropy	−0.02	0.05	−0.46	
German	Syntactic entropy	0.05	0.05	0.98	
Danish	Syntactic entropy	−0.01	0.06	−0.25	

Table 10.10 LMMs for coherent typing activity (copying)

Formula	Coherent typing activity ~ syntactic entropy + average word length (T) + number of clauses + inefficiency + (1|Participant) + (1|Unique_Segment-ID)				
	Variable	β	Standard error	t-value	Significance level
Spanish	Syntactic entropy	0.03	0.05	0.61	
	Number of clauses	−0.01	0.02	−0.69	
	Typing inefficiency	7.08	0.44	16.17	***
	Average word length	0.57	0.07	8.10	***
German	Syntactic entropy	0.05	0.06	0.78	
	Number of clauses	−0.01	0.02	−0.77	
	Typing inefficiency	7.08	0.44	16.16	***
	Average word length	0.56	0.07	7.67	***
Danish	Syntactic entropy	0.08	0.06	1.37	
	Number of clauses	−0.01	0.03	−0.32	
	Typing inefficiency	7.91	0.67	11.73	***
	Average word length	0.47	0.10	4.78	***

The significance rates reflect participant and item variability
***$p < .001$

10.9.3 Coherent Typing Activity

As expected, in the control condition, no significant effect for Syntactic Entropy nor clause length could be observed. Typing inefficiency and average word length were highly significant (Table 10.10).

10.9.4 Total Reading Time (Target)

No effect of syntactic entropy on total reading time of the target segment could be found for Danish, Spanish or German entropy values. Only Average word length and typing inefficiency were significant contributors to total reading time in Spanish and German (Table 10.11).

No significant effects of syntactic entropy on any of the behavioural measures during copying were observed. This suggests that syntactic entropy measures an effect that is driven by the target language, i.e. it supports the view that, during translation, both languages are co-activated. It further suggests that translators entertain more than one possible target structure.

Table 10.11 LMMs for total reading time (target) (copying)

Formula	Total reading time (target) ~ syntactic entropy + average word length (T) + number of clauses + inefficiency + (1|Participant) + (1|Unique_Segment-ID)				
	Variable	β	Standard error	t-value	Significance level
Spanish	Syntactic entropy	−0.13	0.09	−1.49	
	Number of clauses	0.00	0.03	−0.14	
	Average word length	0.35	0.12	2.84	**
	Typing inefficiency	2.94	0.73	4.04	***
German	Syntactic entropy	−0.05	0.11	−0.43	
	Number of clauses	−0.01	0.03	−0.34	
	Average word length	0.37	0.13	2.88	**
	Typing inefficiency	2.91	0.73	4.01	***
Danish	Syntactic entropy	0.05	0.14	0.34	
	Number of clauses	−0.02	0.06	−0.36	
	Average word length	0.15	0.24	0.63	
	Typing inefficiency	2.28	1.25	1.82	†

The significance rates reflect participant and item variability
†p < .1, **p < .01, ***p < .001

10.10 General Discussion

Syntactic entropy was a significant predictor of increased total reading time of the source text segments and a marginally significant predictor for average first fixation durations on the reception side of the translation. On the production side total reading time of the target text and coherent typing behaviour were associated with performance decreases (marginally significant in the case of total reading time of the target and significant in the case of coherent typing). Higher behavioural measures may thus be taken as an indication of competition between multiple syntactic translation equivalents and the selection pressure generated from a set of co-activated syntactic realizations increasing cognitive load. This observation corroborates accounts that claim co-activation of linguistic systems during translation (e.g. Macizo and Bajo 2006; Ruiz et al. 2008). Results showed that these effects are driven by the target language and the translation task, since syntactic entropy was not a significant predictor of behavioural measures when participants copied the texts.

Although the syntactic annotation of the data was very shallow, it was possible to measure variation and priming effects. They manifested in structural repetition of syntax found in the source text segment and occurred mainly in the vicinity of low syntactic variation, indicating that many translators were structurally primed by the source. Low syntactic variation is thus likely a result of syntactic priming, influencing translators to reproduce the syntactic structure they read in the source text.

It was surprising that the interaction between priming probability and syntactic entropy was not significant. A deeper level of analysis might lead to different results when, for example, levels of embedding and a finer analysis of the clause type are assessed (see Chap. 12). But since the argument structure is captured in the first dimension of the annotation scheme and subject variation is accounted for to some degree by voice in the second dimension, priming effects due to functional correspondences are probably reliable.

Furthermore, priming effects were hypothesized to be strongly modulated by word order and less so by mere functional correspondence. The results for the linear regression with *CrossS* confirmed the hypothesis, that congruent word order is a strong but not a necessary condition for syntactic priming, since even higher priming probabilities were possible when the literal translation threshold of a *CrossS* value of 1 was exceeded.

While the hypothesis that priming effects are a major factor for decreased syntactic variation could be confirmed, no significant facilitation effect could be observed for the interaction effect of syntactic entropy and priming probability. This may indicate that priming did not have the expected degree of influence on syntactic variation. Other processes that regulate variation may have been underestimated. For example, when a primed structure is incompatible with target norms, translators may choose a different structure in order to produce a target sentence that is compatible with target norms. If many translaters choose the same structure, the resulting entropy value would by lowered in a similar fashion to a priming effect, but it may not display a facilitation effect. This could be why the interaction effect of priming probability and syntactic entropy was not behaving as expected. Another possible explanation for this observation is that options that are primes are monitored carefully to avoid such target language norm violations. This would in turn lead to longer reading times. Decreased translation performance for non-norm conforming structures has been noted by Vandepitte and Hartsuiker (2011). In their study Dutch translators displayed difficulties when translating English SVO structures containing inanimate subjects to Dutch when adhering to this structure. Inanimate subjects tend to not take subject position in Dutch. A monitoring effect may thus cancel the effect of priming in text production.

10.11 Conclusion

The results presented here corroborate the view of shared linguistic representational structures. This chapter shows that the scope of shared linguistic representational structures is not restricted to lexical items but extends to syntax since syntactic co-activation of multiple possible structures is reflected in longer behavioural measures similarly to words with multiple translation alternatives. The results presented here expand and lend further support to the literal translation hypothesis.

Acknowledgements This work was supported by EU's 7th Framework Program (FP7/2007-2013) under grant agreement 287576 (CASMACAT).

References

Baayen, R. H. (2013). *languageR: Data sets and functions with 'Analyzing linguistic data: A practical introduction to statistics'*. Available at: http://cran.r-project.org/package=languageR

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). {*lme4*}*: Linear mixed-effects models using Eigen and S4*. Available at http://cran.r-project.org/package=lme4

Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2007). Shared syntactic representations in bilinguals: Evidence for the role of word-order repetition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(5), 931–949.

Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2013). From language-specific to shared syntactic representations: The influence of second language proficiency on syntactic sharing in bilinguals. *Cognition, 127*(3), 287–306.

Boada, R., Sánchez-Casas, R., Gavilán, J. M., García-Albea, J. E., & Tokowicz, N. (2012). Effect of multiple translations and cognate status on translation recognition performance of balanced bilinguals. *Bilingualism: Language and Cognition, 16*(01), 183–197.

Carl, M. (2012). Translog-II: A program for recording user activity data for empirical reading and writing research. In *LREC*.

Carl, M., & Schaeffer, M. (forthcoming). Literal translation and processes of post-editing. In *Translation in transition: Between cognition, computing and technology*. Amsterdam: Benjamins.

Chen, B., Jia, Y., Wang, Z., Dunlap, S., & Shin, J.-A. (2013). Is word-order similarity necessary for cross-linguistic structural priming? *Second Language Research, 29*(4), 375–389.

Chesterman, A. (2011). Reflections on the literal translation hypothesis. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research: Integrative approaches in translation studies* (pp. 23–35). Amsterdam: John Benjamins.

Desmet, T., & Declercq, M. (2006). Cross-linguistic priming of syntactic hierarchical configuration information. *Journal of Memory and Language, 54*(4), 610–632.

Dragsted, B. (2012). Indicators of difficulty in translation — correlating product and process data. *Across Languages and Cultures, 13*(1), 81–98.

Duñabeitia, J. A., Perea, M., & Carreiras, M. (2010). Masked translation priming effects with highly proficient simultaneous bilinguals. *Experimental Psychology, 57*(2), 98–107.

Eddington, C. M., & Tokowicz, N. (2013). Examining English–German translation ambiguity using primed translation recognition. *Bilingualism: Language and Cognition, 16*(02), 442–457.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal behaviour, 20*(6), 641–655.

Grosjean, F. (1997). The bilingual individual. *Interpreting – International Journal of Research and Practice in Interpreting, 2*, 163–187.

Hartsuiker, R. J., Kolk, H. H. J., & Huiskamp, P. (1999). Priming word order in sentence production. *The Quarterly Journal Of Experimental Psychology, 52A*(1), 129–147.

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science, 15*(6), 409–414.

Hartsuiker, R. J., & Westenberg, C. (2000). Word order priming in written and spoken sentence production. *Cognition, 75*(2), 27–39.

Ivir, V. (1981). Formal correspondence vs. translation equivalence revisited. *Poetics Today, 2*(4), 51–59.

Jakobsen, A. L., & Schou, L. (1999). Translog documentation. In G. Hansen (Ed.), *Probing the process in translation methods and results* (pp. 1–36). Copenhagen: Samfundslitteratur.

Jensen, K. T. H., Sjørup, A. C., & Balling, L. W. (2009). Effects of L1 syntax on L2 translation. In F. Alves, S. Göpferich, & I. M. Mees (Eds.), *Methodology, technology and innovation in translation process research: A tribute to Arnt Lykke Jakobsen* (pp. 319–336). Copenhagen: Samfundslitteratur.

Kidd, E., Tennant, E., & Nitschke, S. (2014). Shared abstract representation of linguistic structure in bilingual sentence comprehension. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-014-0775-2.

Kuznetsova, A., Christensen, R. H. B., & Brockhoff, P. B. (2014). *lmertest: Tests for random and fixed effects for linear mixed effect models (lmer Objects of lme4 Package). R package version 2.0-6*. Available at http://www.cran.rproject.org/package=lmerTest/

Laxén, J., & Lavaur, J.-M. (2010). The role of semantics in translation recognition: Effects of number of translations, dominance of translations and semantic relatedness of multiple translations. *Bilingualism: Language and Cognition, 13*(02), 157.

Loebell, H., & Bock, K. (2003). Structural priming across languages. *Linguistics, 41*(5), 791–824.

Macizo, P., & Bajo, M. T. (2006). Reading for repetition and reading for translation: Do they involve the same processes? *Cognition, 99*(1), 1–34.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Neubert, A., & Shreve, G. (1992). *Translation as text*. Kent, OH: Kent State University Press.

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin, 134*(3), 427–459.

Pollatsek, A., Reichle, E. D., Juhasz, B. J., Machacek, D., & Rayner, K. (2008). Immediate and delayed effects of word frequency and word length on eye movements in reading: A reversed delayed effect of word length. *Journal of Experimental Psychology: Human Perception and Performance, 34*(3), 726–750.

Prior, A., Wintner, S., Macwhinney, B., & Lavie, A. (2011). Translation ambiguity in and out of context. *Applied Psycholinguistics, 32*(01), 93–111.

Pym, A. (2003). Redefining translation competence in an electronic age. Defence of a minimalist approach. *Meta: Translators' Journal, 48*(4), 481–497.

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.r-project.org/.

Ruiz, C., Paredes, N., Macizo, P., & Bajo, M. T. (2008). Activation of lexical and syntactic target language properties in translation. *Acta Psychologica, 128*(3), 490–500.

Schaeffer, M. J., Paterson, K., McGowan, V. A., White, S. J., & Malmkjær K. (forthcoming). *Reading for translation*.

Schoonbaert, S., Hartsuiker, R. J., & Pickering, M. J. (2007). The representation of lexical and syntactic information in bilinguals: Evidence from syntactic priming. *Journal of Memory and Language, 56*(2), 153–171.

Shannon, C. E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal, 30*(1), 50–64.

Shin, J.-A., & Christianson, K. (2009). Syntactic processing in Korean-English bilingual production: Evidence from cross-linguistic structural priming. *Cognition, 112*(1), 175–180.

Tokowicz, N., & Kroll, J. F. (2007). Number of meanings and concreteness: Consequences of ambiguity within and across languages. *Language and Cognitive Processes, 22*(5), 727–779.

Toury, G. (1995). *Descriptive translation studies and beyond* (Vol. 75). Amsterdam: John Benjamins. Benjamins translation library v4.

Vandepitte, S., & Hartsuiker, R. J. (2011). Metonymic language use as a student translation problem: Towards a controlled psycholinguistic investigation. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research: Integrative approaches in translation studies* (pp. 67–92). Amsterdam: John Benjamins.

Wu, Y. J., & Thierry, G. (2012). Unconscious translation during incidental foreign language processing. *NeuroImage, 59*(4), 3468–3473.

Chapter 11
Cohesive Relations in Text Comprehension and Production: An Exploratory Study Comparing Translation and Post-Editing

Márcia Schmaltz, Igor A.L. da Silva, Adriana Pagano, Fabio Alves, Ana LuísaV. Leal, Derek F. Wong, Lidia S. Chao, and Paulo Quaresma

Abstract Few studies using Translog-II in conjunction with eye-tracking data in translation studies have focused on languages which use logographic scripts. This chapter reports on an exploratory study of data related to one text included in MS13, contained in CRITT Translation Process Research Database, with a view to investigating the impact of type of cohesive chain on cognitive effort in Portuguese-Chinese translation and post-editing tasks. Eye-tracking and key logging data were assessed by means of a linear mixed-effects regression model. The results point to no impact of task on the dependent variables, but to an impact of the type of cohesive relations on target text reading and production. The chapter also contributes to developing a methodology for processing of Translog-II data involving Chinese.

Keywords Translog-II • Cohesion • Portuguese-Chinese Translation • Post-editing

M. Schmaltz (✉) • A.L.V. Leal
Department of Portuguese, University of Macau (UM), Macau, China
e-mail: schmaltz.marcia@gmail.com

I.A.L. da Silva
Federal University of Uberlandia (UFU), Uberlandia, Brazil

A. Pagano • F. Alves
Laboratory for Experimentation in Translation, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

D.F. Wong • L.S. Chao
Department of Computer and Information Science, University of Macau, Macau, China

P. Quaresma
Department of Informatics, University of Evora (UE), Évora, Portugal

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_11

11.1 Introduction

Research using Translog-II (Carl 2012) in conjunction with eye-tracking data in translation studies (Carl and Jakobsen 2009; Jakobsen 2011; Hvelplund 2011; Carl and Dragsted 2012; Sjørup 2013; Balling and Carl 2014; Mesa-Lao 2014; among others) has focused on tasks involving Western European languages and consequently alphabetical scripts. However, studies focusing on languages which use logographic scripts are still incipient. This chapter reports on a study of from-scratch translation and post-editing tasks carried out from Portuguese into Chinese by Chinese translators of Portuguese (L2). Drawing mostly on the methodology used by Sjørup (2013), we carried out a study to examine gaze and key logging data from six participants while translating and six other participants while post-editing a 79-word news report. These data are available in CRITT Translation Process Research Database as MS13 (translation session 16, post-editing session 18).

Building on Halliday and Hasan (1976) and Hasan (1984), referents pertaining to the main cohesive chain of the source text (labelled *chain A*) were defined as our focus of enquiry and contrasted with items in a secondary cohesive chain traceable in the same text (labelled *chain B*). Our assumption was that tracking participants (referents) in chain A would be critical for the 12 translators to build a coherent interpretation of the source text (ST) and would require them to retrieve the identity of what was being talked about by referring to another expression either in the co-text or the context of the situation and culture. A higher number of fixations in eye and keyboard activities were thus expected during reading and production of chain A. A secondary chain, in contrast, would have a lesser contribution to the ST and TT (target text) coherence and would thus demand less attention, as well as fewer keyboard and eye activities.

11.2 Review of the Literature

The role of *cohesion* in the establishment of a coherent interpretation of text is one of the many core questions of reading comprehension in translation tasks (Bell 1991; Hatim and Mason 1990). Since it has to do with translators' active participation in understanding an ST unfolding and in building a TT patterned on it, this is an issue particularly well suited to be approached from a translation process perspective.

Among the different resources playing a part in *texture*, i.e., that which makes a text a text and makes it function "as a unity with respect to its environment" (Halliday and Hasan 1976, 2), cohesive devices are responsible for non-structural relations between items in a text. Such relations are established through the creation of semantic bonds, so that one item is interpreted with reference to the other.

One cohesive relation in particular is especially relevant to discourse coherence in text unfolding: this is *participant tracking*, i.e., the mapping of referents pertaining to the main cohesive chains running along a text. Both grammatical

cohesion (more precisely, reference) and lexical cohesion are recruited in participant tracking. By reference is meant a relationship in meaning construed through the use of a personal reference item (personal pronoun or possessive determiner) that enters into a semantic relationship with an item mentioned either before in the text (anaphora) or afterwards (cataphora). Occasionally, reference is made to entities that cannot be retrieved from the text and need to be established situationally; this is referred to as exophoric reference. When two items share the identity of a referent, this is termed *co-referentiality.* Items sharing identity can also be linked through lexical cohesion, be that repetition, synonymy or hyponymy.

Cohesive ties, i.e. semantically bonded items, are particularly important when they form so-called *cohesive chains*, responsible for strong integration of cohesive ties and a more coherent text. A cohesive chain built on participant tracking may be realized through co-reference or lexical cohesion categories that are valid for language in general but that ultimately need to be interpreted in a particular text. Thus, this type of cohesive chain is crucial to text organization and comprehension. Conversely, secondary chains are not essential to participant tracking and are built upon lexical relations that are not text-specific, but general to the lexicon of the language.

When text is processed in translation tasks, Hatim and Mason (1990) argue, translators rely both on contextual and co-textual cues in order to identify cohesive items deemed relevant to a coherent construction of the TT. These cues can be sought in the immediate co-text or demand integration of items that are more distant in the text.

Cohesion has not been extensively examined in translation process research. Denver (2009) investigated adversative-concessive logical-semantic relations in translations from Spanish into Danish. The author found different right and wrong choices among translators and students, but no trace of mental activity in processing relations realized through conjunctions in Spanish, i.e. no verbalization or keystroke, pause or revision signalling that the relations constituted translation problems for the participants.

Angelone (2010) studied uncertainty management and metacognitive problem solving of a professional translator, two students and a bilingual. He classified the textual level at which the participants' metacognitive activity was employed into lexis, term, collocation, phrasal, syntax, sentential, macro level, and unclassified. The macro level category refers, according to the author, to beyond sentence considerations, such as cohesion, coherence, and gender. Only a small part of the elicited verbalizations fell into the sentential and macro level categories.

Both authors relied on think-aloud protocol data; Denver also used key logging data, and Angelone also used screen recordings. To the best of our knowledge, no other translation process research using eye-tracking has addressed translators' or post-editors' processing of cohesion in STs and TTs.

Staub and Rayner (2007) claim that many eye-tracking studies have focused on syntactic parsing, but few have looked into how discourse processing (including cohesion) affects eye movements in reading. Staub and Rayner (2007, 335) argue that recognizing individual words and analyzing grammatical structures of each

sentence does not suffice to understand a text; the reader "must also maintain a representation of the entities and events that have been mentioned, and relate the information that is currently being processed to this stored representation."

Basically, eye-tracking studies focusing on cohesion have so far shown an increase in fixation times due to: long distance between an anaphor and its antecedent (O'Brien et al. 1997); antecedent being a low-frequency word (van Gompel and Majid 2004); and reading a target word and drawing conclusions that have not been explicitly stated in the text (O'Brien et al. 1988).

11.3 Methodology

The results described in this chapter are part of a larger empirical-experimental project carried out by the AuTema-PostEd Group, which aims at tapping into translation and post-editing processes as a source of insight into the role of translators' understanding in task problem solving. In this chapter, we report the results regarding the Portuguese(L2)-Chinese(L1) translation and post-editing of a text about the China Gold Research Institute.

11.3.1 Equipment and Analysis Tools

Data from gaze and keyboard activity were collected and analyzed using Translog-II (Carl 2012, 2013), version 0.1.0189, connected to a Tobii T120 remote eye tracker. The eye-tracking software application Tobii Studio 3.2.1 was also used as a recorder for the participants' verbalizations and gestures. Calibration was performed in both Translog-II and Tobii Studio, the latter running in the background while the participants worked in Translog-II.

Figure 11.1 shows the screen setting in the post-editing task: the ST appears in the top half of the application window, and the TT in the bottom half, which is empty in the translation task. The ST font was Tahoma, and the TT font was SimSun with font size 17. Both texts were double-spaced.

11.3.2 Participants

Originally, 23 professional translators performed two translation tasks (L1 into L2, and L2 into L1) and two post-editing tasks (one in their L1 and another one in their L2) using machine-translated (MT) input provided by the software

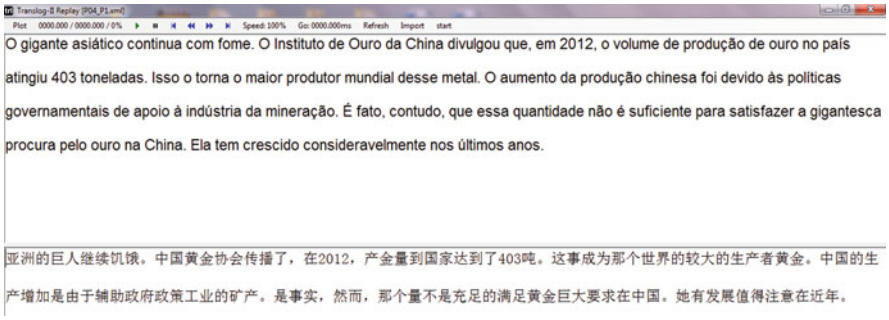


Fig. 11.1 Screenshot of Translog-II user post-editing window

Table 11.1 Results for the three quality assurance criteria

Task	Subject	Mean fixation duration (in ms)	Gaze time on screen (in %)	% of valid win gaze data
Post-editing	P03	317	78.07	70.03
	P09	366	91.98	83.15
	P11	434	93.41	91.80
	P15	439	82.85	77.56
	P19	417	82.28	79.76
	P23	365	86.81	79.37
Translation	P04	421	89.24	76.75
	P06	299	82.98	78.00
	P07	222	63.81	61.95
	P12	309	82.67	88.24
	P18	301	49.59	43.56
	P21	465	83.69	67.16
Post-editing	Mean	389.67	85.90	80.28
Translation		336.17	75.33	69.28

PCT (Portuguese-Chinese Translator[1]) (Wong and Chao 2010). In this chapter, we report the analysis of Text 1 (see Sect. 11.3.4), for which we have data from 17 participants[2] contained in the MS13 study within TPR-DB (translation session 16, post-editing session 18). Only 12 of the participants with more than 1 year of experience and an age range of 23–32 years remained in our sample (see Table 11.1). The others were excluded because of failure to comply with data quality criteria as described in Sect. 11.3.3.

[1]Quadrilingual MT. Available at http://nlp2ct.cis.umac.mo/MT/.

[2]Six participants' Text 1 data were excluded because of technical issues related to tokenization or alignment problems (see Sect. 11.3.6).

All participants provided informed consent. They were Chinese nationals and had an undergraduate degree in Portuguese Studies or a Master's degree in Chinese-Portuguese Translation Studies. All had been granted a fellowship to live in a Portuguese speaking country, and all of them used glasses or contact lenses. None of them had experience in post-editing. Each was paid MOP 90.00 to take part in the experiments.

The participants were instructed to sit approximately 55 cm away from the eye-tracker monitor. They were also told that they could move freely, but were told to keep their eyes on the monitor as much as possible.

11.3.3 Data Quality

For data collection, we tried to cope with the numerous factors that may have had an effect on the quality of the gaze data, especially lighting, glasses, and distance from the monitor. However, in order to ensure consistency in the sample, three data quality assurance criteria were observed. Data should comply with at least two of the three criteria to ensure that the results were not skewed by flawed data.

The first criterion was mean fixation duration: following Sjørup (2013) and Hvelplund (2011), our threshold was established at a minimum of 180 ms. The second criterion was gaze time on screen (GTS), that is, the percentage of time spent gazing on the text in relation to the total time of translation production: once again, following Sjørup (2013) and Hvelplund (2011), our threshold was 30 %. The third criterion, called "% of valid win gaze data," was calculated in terms of the percentage of valid gaze data on the ST and TT token considering the attribute "win" in the XML files produced after each Translog session. More specifically, we divided the number of occurrences of win = 1 (gaze on ST) plus win = 2 (gaze on TT) by the total number of "wins", which included both win = 1 and win = 2 and also win = 0 (gaze not ascribed to either the ST or the TT). As, to the best of our knowledge, no study has reported such a measure before, we arbitrarily established our threshold at ≥40 %.

Table 11.1 shows the figures of the remaining data considering the three criteria mentioned above.

11.3.4 The Experimental Text

The ST, Text 1, is a short news report written in Brazilian Portuguese on the increase of gold extraction and consumer market in China (see Appendix 2). Chain A is the main chain, where the participant being tracked is production volume. Chain B is the secondary chain and refers to the country (i.e. China). Table 11.2 shows both chains and their cohesive devices.

Table 11.2 Main chain and cohesive devices of ST (selected tokens in italics)

Type of chain	Tokens in ST	Explicitation of referents as	Co-reference established through
A	*o volume de produção* atingiu 403 toneladas "*the production volume* increased reaching 403 tons"	–	Not applicable (first item)
	Isso "This"	The fact that the production volume increased	Demonstrative pronoun
	O aumento "The increase"	The production increase	Definite article + lexical noun (synonym)
	essa quantidade "such amount"	The amount of production increase	Demonstrative pronoun + lexical noun (superordinate)
B	*O gigante asiático* "The Asian giant"	–	Not applicable (first item)
	n*o país* "in *the country*"	In the country which is the Asian giant	Definite article + lexical noun (synonym)
	o torna "turns *it* into"	Turns the Asian giant into	Personal pronoun
	n*a China* "in China"	In China, which is the Asian giant	Lexical noun (synonym)

The relative position and length of the selected tokens (words and noun groups) in chain A and chain B were accounted for in our statistical analysis (see Sect. 11.4.2).

11.3.5 Task

After a brief warm-up session, which consisted of a copy test before the experiment, each participant was asked to perform four tasks, randomly assigned to participants: two translations (one into their L1 and another one into their L2), and two post-editing tasks (one in their L1 and another one in their L2) using MT output.

Table 11.3 provides the tasks performed by each participant. The analyses in this chapter refer to T1 and P1 highlighted in Table 11.3.

As a brief, the participants were informed that they should render texts aimed at a target audience analogous to that of the ST. They were told to feel free to produce the human TT or post-edit the MT text without any time constraint, but they could not use any kind of translation aids. As they had little to no experience in post-editing, we provided them with guidelines reported in Mesa-Lao (2014, 225), see also chapters 11.3.5, 11.7, 11.8, and 13 in this volume.

After each task, the participants were requested to provide a retrospective protocol, whereby they could explain whatever they felt like concerning their

Table 11.3 Task distribution across participants

Participant	From scratch		Post-editing	
03	**T1**	T3	P4	P2
04	T2	T4	P3	**P1**
06	T3	T2	**P1**	P4
07	T2	T3	P4	**P1**
09	T3	**T1**	P2	P4
11	**T1**	T3	P2	P4
12	T2	T4	**P1**	P3
14	T3	T2	P4	**P1**
15	**T1**	T4	P3	P2
17	T3	**T1**	P2	P4
18	T4	T2	**P1**	P3
23	**T1**	T3	P2	P4

Note: T = translation; P = post-editing; #1–2 = Portuguese ST; #3–4 = Chinese ST

translation or post-editing, such as difficulties, challenges, strategies, doubts. The retrospective protocols were carried out by means of the Translog-II Supervisor replay function (Jakobsen 2011, 39).

11.3.6 Processing of Chinese Data

In the CRITT TPR-DB are four study folders containing data of translation and post-editing involving both Chinese language and a Latin alphabetic script (see Appendix 1, Chap. 2). As Chinese is a logographic language which does not require blank spaces between the characters (Zang et al. 2011), the processing of the Chinese language data involved additional procedures so that they could be automatically analyzed using the Study Analysis script.[3] These procedures are described in Sects. 11.3.6.1 and 11.3.6.2.

11.3.6.1 Chinese Input System

A logographic language like Chinese requires an input method[4] through a graphic user interface (GUI), which converts sequences of alphabetic letters into Chinese characters. The participants used Sogou[5] as their Chinese input method.

[3] Available at https://sites.google.com/site/centretranslationinnovation/translog-ii.

[4] Basically, there are two categories of Chinese input method, i.e. phonetic readings or root shapes. Most of these input methods can be selected directly from the control panel of MS Windows.

[5] Available at http://pinyin.sogou.com/.

Fig. 11.2 Snapshot of a post-editing session showing Sogou's dialog box. Note: *Circle* indicates fixation

Figure 11.2 shows a snapshot of a post-editing session. In the bottom half of the window, Sogou's dialog box pops up below the line where the participant wants to introduce new characters. A zoom-in shows that while the participant types in *pinyin*[6] a series of alphabetic letters, a number of options are shown out of which one may be the desired corresponding character(s). To select the desired characters and insert them in Translog-II, the participant presses the space bar or the corresponding number key.

As shown in Fig. 11.2, the Chinese input system is prone to word gaze error, since the place where pinyin is typed is not the same place where the Chinese character is inserted. Assuming that the Sogou's dialog box pops up right below the space where the character is supposed to be inserted, we manually[7] attribute the fixation

[6] *Pinyin* is the official phonetic system for transcribing the *Putonghua* (Mandarin) pronunciations of Chinese characters into the Latin alphabet.

[7] The "StudyAnalysis" scripts include a function to refixate gaze mapping to word. However, several translation drifts remain and need to be corrected manually.

to a specific word through Translog-II Supervisor[8] with support of Tobii Studio replay function.

11.3.6.2 Chinese Tokenization and Alignment

The procedures to analyze Translog-II data can be retrieved from the CRITT website.[9] However, some additional steps were required to tokenize and align the data because of Chinese language specificities. As Chinese texts are written as a stream of characters without blank spaces, there is no explicit delimiter to identify word boundaries and automatically tokenize the data using the "StudyAnalysis.pl tokenize" script. To tackle these problems, we came up with an alternative workflow and developed applications to (semi-)automate the process.

Figure 11.3 shows the conventional and the alternative workflows, which contains additional steps 0, 2.5, and 3 as in Fig. 11.3b. Firstly, we fix the incorrect gaze data (see Sect. 11.3.6.1) and save each log file as a new *.xml file. Secondly, a Chinese tokenization step (Step 2.5) is added after the extraction of text data in Step 2: We use the in-house developed tool, *ChiSegmentor* (Leong et al. 2006; Zeng et al. 2013), to automatically identify the word boundaries, and then we manually revise the output drawing on the *Modern Standard Chinese Dictionary* (Li 2010). This information is recorded in the corresponding log file—*.src or *.tgt.

Another change to the workflow is in the alignment step, for which we use *LexAligner* (Tian et al. 2011) to automatically estimate possible word alignments. To check the alignments, we draw on criteria provided by the *Guidelines for Chinese-English Word Alignment* (Li et al. 2009). Because the translation renditions are a result from the processing of an ST (Mossop 2003), we align all ST and TT tokens.

Finally, we run the "StudyAnalysis.pl tables" script to extract several kinds of simple and compound process and product units, which are represented in tables (see Chap. 2 for details). From these units it is possible to generate Translation Progression Graph (TPG) using the R environment for statistical computing.

It is worth noticing that sometimes Chinese and Portuguese tokens have encoding conflict that prevents us from generating TPGs. To overcome this problem, which is identified while running R to generate TPGs, we replace the problematic character with *pinyin* for Chinese and with "a" for Portuguese.[10]

[8] Fix Map device, available at https://sites.google.com/site/centretranslationinnovation/translog-ii.

[9] Available at https://sites.google.com/site/centretranslationinnovation/translog-ii.

[10] For further details, please contact the authors.

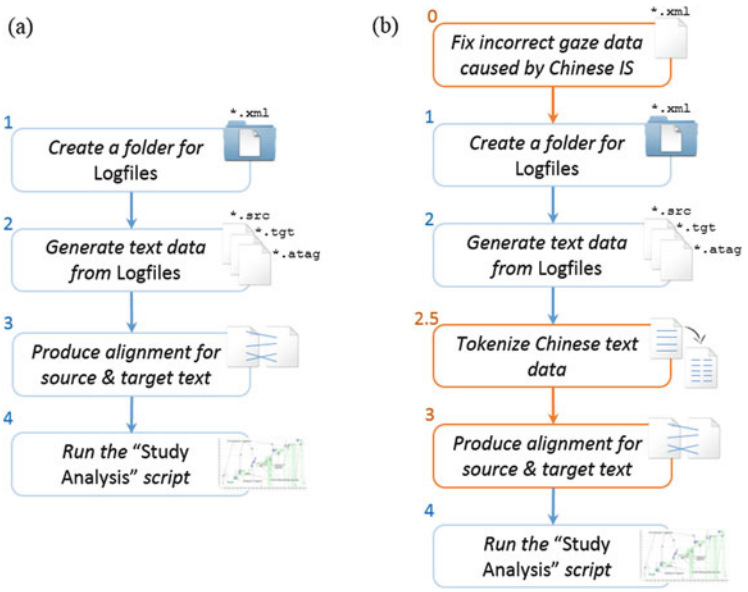


Fig. 11.3 Procedures to prepare Translog-II data for "Study Analysis": (**a**) Conventional process for Roman-alphabet languages; (**b**) Adapted process for the Chinese language

11.3.6.3 Chinese Production Data Extraction

For technical reasons related to the Chinese input method,[11] which is external to Translog-II, the system logs only the text modifications (delete and space keys), but not the actual keystrokes (see Sect. 11.3.6.1). Table 11.4 shows examples of incorrect number of character insertions in the target tokens (TToken) (see Sect. 2.4): in the column Ins, the number "2" refers to the Chinese characters shown on the screen, whereas the column should provide the number of keystrokes actually typed (five insertions for $TTid_{50}$ and nine insertions for $TTid_{51}$, see Fig. 11.2).

A problem also occurs with the duration of the production time of Chinese characters (as reported in Balling and Carl 2014, 260). As shown in Table 11.5, it is common to find long pauses and short production times (most of them of 1 ms), when in fact the participant had virtually no pause and took longer to produce a given token (TToken).

To solve the aforementioned problems, we use Tobii Studio replay function to identify when exactly the participants started and stopped typing keystrokes corresponding to each word logged in Translog-II. While doing this, we also count

[11]There is no immediate connection between the keystrokes and the characters that appear in the text.

Table 11.4 Example of wrong log of keystroke insertions and deletions

TTid	TToken	SToken	Ins	Del	Edit
50	这个	É_fato_,_que_essa	2	0	这个
51	数量	quantidade	2	0	数量

Table 11.5 Example of wrong log of pause and duration (Dur)

TTid	TToken	SToken	Time	Dur	Pause	Edit
50	这个	É_fato_,_que_essa	153911	1	4353	这个
51	数量	quantidade	153913	1	1	数量

Table 11.6 Example of intermediate production alignment problem (Edit1)

TTid	TToken	SToken	Edit1	Time1	Dur1	Pause1	Edit
49	,	,	[那个量]"that amount"	149153	405	10780	那个量
50	这个	É_fato_,_que_essa" It is a fact, that this"		153911	1	4353	这个
51	数量	quantidade "quantitity"		153913		1	数量

Note: Deleted words are in brackets

the actual keystrokes that the participants pressed to produce the characters logged in Translog-II. Then, we manually correct the TT tables.

We also observed that the edited units are coherently aligned to the respective STid and TTid in most of the cases. However, because Translog-II aligns only the initial ST and the final TT, when the characters of the MT text are deleted and/or immediately edited (such as Edit 1 [那个量], which represents the deletion of [that amount] in $TTid_{49}$, Table 11.6), the system does not identify which words they were originally part of (a part of Edit 1 in $TTid_{49}$ should belong to TT_{50} 那个 [that], and another part 量 [amount] should belong to TT_{51}). To account for this, we check all Edit1 and Edit2 actual operations to identify the actual ST and TT tokens (STokens and TTokens).

In sum, a substantial part of our production analysis was built on manually processed data. To ensure quality, all manually extracted data were double-checked. These spreadsheets are available in TPR-DB.

11.4 Data Analysis

11.4.1 Research Question

Our main objective in this study is to compare the cognitive effort demanded for translating and post-editing the selected tokens in the main cohesive chain (chain A) in the ST, which is built on participant tracking, with the cognitive effort

demanded for processing selected tokens in a secondary cohesive chain (chain B) (see Sect. 11.3.4).

Our first research question is "Is it cognitively more demanding to understand and produce a cohesive chain that is built on participant tracking than a secondary cohesive chain?" This question is based on the assumption that participant tracking is crucial to construing a coherent representation of a text (Halliday and Hasan 1976).

Our second research question is "In dealing with cohesive chains, is it cognitively more demanding to translate than to post-edit?". Given that translation takes longer than post-editing (Balling and Carl 2014; Mesa-Lao 2014), it is possible that processing cohesive chains during post-editing is also faster or that cohesive chains are processed differently in the two tasks.

11.4.2 Statistical Analysis and Variables

Our investigation is divided into three statistical analyses: (1) eye movements on chain A and chain B in the ST, (2) eye movements on chain A and chain B in the TT, and (3) keyboard movements relating to chain A and chain B.

In the following, we describe the variables of analysis in the order they were included in the statistical model (for details on the model, see Sect. 11.4.3).

The dependent variables[12] for analyses (1) and (2) were:

- Total reading time on ST and TT token (TrtS and TrtT);
- Number of fixations on ST and TT token (FixS and FixT); and
- First pass duration on ST and TT token (FPDurS and FPDurT).

The dependent variable for analysis (3) was TT token total production time. Time was measured considering any pauses preceding a TT token plus duration (Dur) (see Sjørup 2013, 126–127 for further details).

Our analysis investigates how the dependent variables vary as a function of several explanatory variables, as described below. For further details, see Baayen (2008) and Balling (2008).

The first group of explanatory variables consists of random effects. Random factors are not repeatable and are assumed to have been selected randomly from any given population (Baayen 2008, 241). As such, we included the participants and item, i.e., the selected ST and TT tokens of chains A and B (see Sect. 11.3.4).

The second group of explanatory variables consists of fixed effects, which refer to factors with repeatable levels (Baayen 2008, 241). They were used to account for previous studies that have reported their effect on the results or their importance for TT cohesion. Due to space restrictions, we report only on the ones that were significant in our model (see Table 11.10 in Appendix 1).

[12]Descriptions for these dependent variables are available in Sect. 2.4.6.

Four fixed effects were used in the analysis of both eye and keyboard movements as proxies for processing of chains A and B in both ST and TT:

- Token Length: The length in characters of the ST and TT tokens of chains A and B was expected to affect the dependent variable, as longer words generally receive longer fixations than shorter words (e.g. Rayner 1998; Hyönä et al. 2003; Staub and Rayner 2007);
- Token Position: The position of the ST and TT tokens of chains A and B in the text was expected to have an effect on the participants' gaze behaviour. It may be due to fatigue (e.g. Rayner 1998; Balling 2008, 2013) and/or to a priming effect (Rayner 1998: 390; Staub and Rayner 2007: 331), which, based on Halliday and Hasan (1976), may imply that the beginning of a text deserves more attention because it will determine the understanding of the remaining of the text and it will have items that will serve as referents for items further in the text (see also Chap. 9);
- Token Unigram Frequency: Readers are expected to fixate longer on low-frequency words than on high-frequency words (e.g. Rayner et al. 2005; Rayner 1998). The Corpus of Portuguese[13] and the Corpus of the Peking University Center for Chinese Linguistics[14] were used to measure frequency;
- Token Trigram Probability: High predictability of word association was expected to have an impact on processing effort (Frisson et al. 2005). The variable was computed following McDonald and Shillcock (2003, 650) and considering the selected token and the two preceding tokens as they occur in the text (Balling 2013).

Two variables were added to the analysis of TT Token production, namely:

- Token Character Count: sum of insertions and deletions of TT Token (see Sect. 11.3.6.3); and
- Correctness of Token in the Chain: TT Tokens were assessed as right or wrong in lieu of the ST Tokens; right TT Tokens were assumed to be as instance in which "patterns of lexical cohesion in texts are maintained, subject to the constraints of particular text norms in particular languages" (Hatim and Mason 1990, 200).

Two other fixed effects were included to directly answer our research questions:

- Task: to investigate the effect of translating or post-editing on ST and TT comprehension and on TT production; and
- Type of Chain: to investigate the effect of chain A and chain B on ST and TT comprehension and on TT production.

All continuous variables were naturally logarithmically transformed.

[13] Available at http://www2.lael.pucsp.br/corpora/bp/.

[14] Available at http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp.

11.4.3 Data Analysis: Statistical Models

Following the methods used in Balling and Carl (2014, 250ff.) and Sjørup (2013), we applied a linear mixed-effect regression model (LMER) as implemented in the lme4 package (Bates et al. 2014) in the R environment for statistical computing (version 3.1.2, R development Core Team 2014). The final model, containing only the significant variables, is summarized in the appendix. Table 11.10 in Appendix 1 provides the fixed effects, with variable names in the first column, estimated effect size in the second column, the standard error of this estimate (indicating the amount of variation in the data) in the third column, the t-value in the fourth column, and the associated p-value in the fifth column. We set the significance level at $p \leq 0.05$.

Table 11.11 in Appendix 1 shows the random effects part of the model. Random effects are not associated with p-values, but are included in the model in order to estimate individual effects and dependencies between observations. The standard error of the effects indicates how much variation the different levels capture.

11.5 Results and Discussion

In this session, we report the fixed effects that had a significant impact on the dependent variables.

11.5.1 Comprehension: Eye Movements along Chain A and Chain B in the ST

Table 11.7 summarizes the effects (√) that had a significant impact on our dependent variables related to ST comprehension. Type of chain and type of task were non-significant for all dependent variables relating to source text processing.

ST Token length had a significant effect on total reading time on ST Token (Fig. 11.4) and number of fixations on ST Token (Fig. 11.5). This confirms claims in the literature (e.g. Rayner 1998, 387; Sjørup 2013, 140) that the longer the word or words, the longer the gaze time and the higher the number of fixations.

Position of ST Token had a significant effect on total reading time (Fig. 11.6) and number of fixations on ST Token (Fig. 11.7). The ST Tokens in initial positions were gazed longer and more often than tokens towards the final positions. Since the

Table 11.7 Summary of significant results for ST comprehension

Variable	TrtS	FixS	FPDurS
Token length	√	√	–
Token position	√	√	–
Token trigram probability	√	√	√

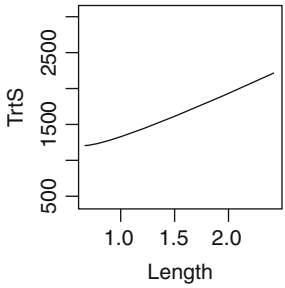


Fig. 11.4 Total reading time on ST Token (TrtS, in ms) vs. length of ST Token in the chain (character count naturally logarithmically transformed)

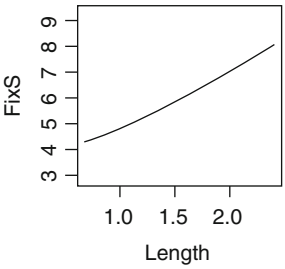


Fig. 11.5 Number of fixations on ST Token (FixS) vs. length of ST Token in the chain (character count naturally logarithmically transformed)

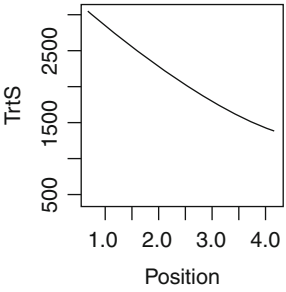


Fig. 11.6 Total reading time on ST Token (TrtS, in ms.) vs. position of ST Token (naturally logarithmically transformed)

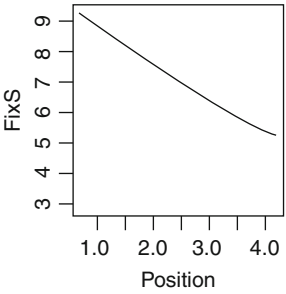


Fig. 11.7 Number of fixations on ST Token (FixS) vs. position of ST Token (naturally logarithmically transformed)

type of chain had no significant impact on the eye movements on the ST and the text was relatively short, the results seem to indicate a priming effect: the initial items in the chains receive more attention from participants because the beginning of the text is crucial for their orientation in order to understand the entire text (Halliday and Hasan 1976).

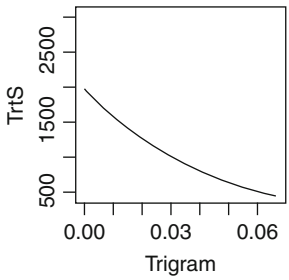


Fig. 11.8 Total reading time on ST Token (TrtS, in ms.) vs. trigram probability (naturally logarithmically transformed)

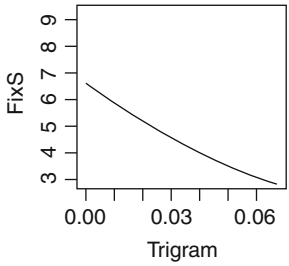


Fig. 11.9 Number of fixations on ST Token (FixS) vs. trigram probability (naturally logarithmically transformed)

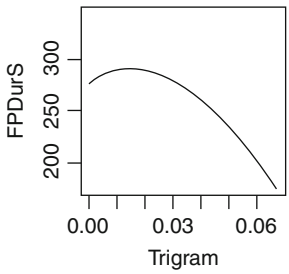


Fig. 11.10 First pass reading time on ST Token (FPDurS) vs. trigram probability (naturally logarithmically transformed)

The probability of word association, measured through trigram probability, had a significant effect on total reading time on ST Token (Fig. 11.8) and number of fixations on ST Token (Fig. 11.9). This confirms findings in the literature, such as Frisson et al.'s (2005) and McDonald and Shillcock's (2003), who found that the more probable or common a word or expression is, the shorter and the less often it is fixated.

The probability of a three-word combination also had a significant effect on first pass duration (Fig. 11.10), which may be indicative of processing of higher level information (see Staub and Rayner 2007, 329). According to our results, the less probable the occurrence of such combination, the more often and the longer it was fixated in subjects' first gaze on it (see the effect of trigram probability on FPDurS).

11.5.2 *Comprehension: Eye Movements along Chain A and Chain B in the TT*

Table 11.8 summarizes the effects (√) that had significant impact on the dependent variables related to TT comprehension. Each variable had an impact on only one of the dependent variables. The type of task had no impact on any of the dependent variables regarding eye movements on the TT. No variable impacted on first pass duration on TT Token.

Token length was significant for number of fixations on TT Token: the longer the token, the greater the number of fixations (Fig. 11.11). Unigram frequency was significant for total reading time on TT Token There seems to be a tendency in fixating more on both the most and the least frequent words; the reason for more fixations on frequent words may be related to their role in the chain, as we are observing instances of participant tracking (Fig. 11.12). Type of chain was significant for total reading time on TT Token: (Figure 11.12): tokens in chain A were fixated longer than those in chain B (Fig. 11.13).

Table 11.8 Summary of significant results for TT comprehension

Variable	TrtT	FixT	FPDurT
Token Length	–	√	–
Token unigram frequency	√	–	–
Type of chain (A)	√	–	–

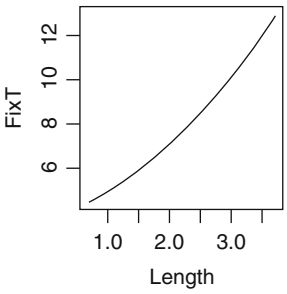


Fig. 11.11 Number of fixations on TT Token (FixT) vs. length of TT Token in the chain (naturally logarithmically transformed)

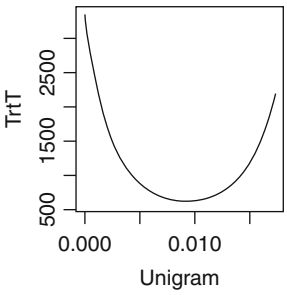


Fig. 11.12 Total reading time on TT Token (TrtT, in ms.) vs. unigram frequency of TT Token (naturally logarithmically transformed)

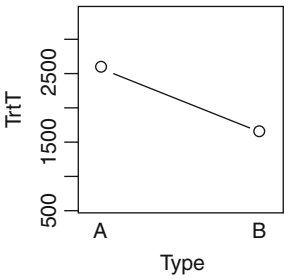


Fig. 11.13 Total reading time on TT Token (TrtT, in ms.) vs. type of chain (A and B)

Table 11.9 Summary of significant results for TT production

Variable	Duration of token production time
Character count	√
Correctness of token (wrong)	√
Type of chain (A)	√

11.5.3 Production: Keyboard Movements for Producing of Chain A and Chain B

Table 11.9 summarizes the effects (√) that had significant impact on the dependent variable "duration of token production time."

The more the participants inserted or deleted characters, the longer was the duration of their token production time (Fig. 11.14). The participants took longer to produce a wrong token in the chain than to produce a right item (Fig. 11.15), which may be related to the number of renditions that they provided while being uncertain to what would be an adequate solution. The participants also took longer to produce the items in chain A (Fig. 11.16).

The longer time for producing tokens in chain A may be indicative of hesitation, need for internal support to make decisions, as well as on-line revisions. These results seem to be consistent with our previous findings (Sect. 11.5.2). The participants' verbalizations also showed that they found it difficult to render some items in chain A. The type of task did not have any significant effect on the results. This suggests that processing of cohesive ties is similar in translation from scratch and post-editing.

11.6 Summary and Future Directions

We set out to answer two research questions: (1) Is it cognitively more demanding to understand and produce a cohesive chain that is built on participant tracking than a secondary cohesive chain? And (2), in dealing with cohesive chains, is it cognitively more demanding to translate than to post-edit?

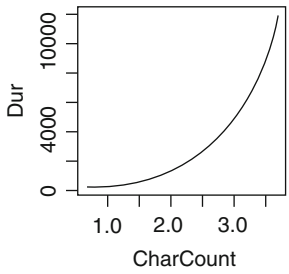


Fig. 11.14 Duration of token production time (Dur, in ms.) vs. character count (naturally logarithmically transformed)

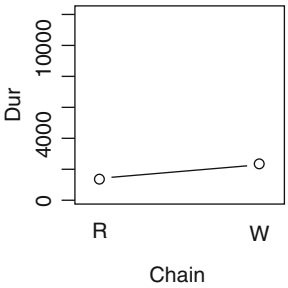


Fig. 11.15 Duration of token production time (Dur, in ms.) vs. correctness of token (*R* right, *W* wrong)

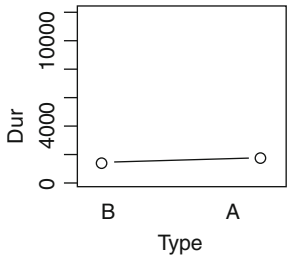


Fig. 11.16 Duration of token production time (Dur, in ms.) vs. type of chain

As for question 1, our results show a significant effect of the type of cohesive chain on eye movements on the TT, which let us infer that it is cognitively more demanding to produce a chain built on participant tracking when it comes to the TT, but no significant effect was observed for the metrics related to the ST.

We expected that the results were also significant for the ST, especially because we assumed that keeping track of participants in the main chain of a text would be challenging both to understand the ST and to produce the TT. However, we should be aware that both translation and post-editing involve transiting gaze from ST to TT, and therefore, the reading of the TT may be closely connected to the understanding of the ST. Future studies should include transiting from ST to TT and vice-versa as also a measure of effort and investigate if ST and TT comprehension should be addressed as one single event. Measures of global processing should also be developed and tried out to account for "relationships between pieces of text information that span relatively long distances in a text" (Hyönä et al. 2003, 314), especially across sentences.

Furthermore, in a larger scale study involving the four tasks for which we collected data, we intend to check if the order in which the task was carried out had an effect on the participants' processing. As we collected all data from each participant on the same day, there might have been an effect of fatigue on the results (e.g., the total time of the last sessions seems to be shorter than that of the first sessions).

As for question 2, the results pointed to no significant impact of type of task on the measures that we assessed in this chapter. We have two potential non-mutually exclusive explanations for this. One reason might be experience—none of the participants had PE experience. Another factor might have been that our design is between subjects, so differences in the results may have to do with differences in the groups. The other potential explanation is that the sample size is small. Considering previous studies that do show significant differences between post-editing and translation (e.g. Balling and Carl 2014; Mesa-Lao 2014), it is possible that either the impact of the type of task is on the transitions across both ST and TT areas or that type of task has an impact on the global processing of the entire text, rather than only on particular cohesive chains. A third possibility concerns to MT quality, as reported in the protocols, participants found the MT text ambiguous at some points and exophoric reference to what MT tokens refer also lacked.

To address some of the aforementioned limitations, in future work we intend to analyse more than one task and use a between subject design, which will allow us to have data for participants that both translated and post-edited. We also intend to compare the tasks considering the entire text and a larger volume of data (including four texts). Following Alves et al. (2014), we also intend to perform a more fine-grained analysis by qualitatively examining the renditions and their processing as shown, for instance, in scan paths.

Besides answering the research questions, the alternatives we had to come up with in order to cope with limitations to process Chinese language data are also a contribution of the present chapter. We hope that the procedures we reported herein contribute to facilitating further studies involving the Chinese language and that our results awake the interest of new scholars to approach language pairs other than those involving only alphabetic scripts. Although the tendency to use the same language pairs and scripts may have methodological advantages (e.g., one language, English, is kept as a standard for comparison's sake, and blank spaces undoubtedly delimit words), not only does it overlook the insights that other language pairs and scripts may add to understanding (non-)language and (non-)script specific cognitive aspects of post-editing and translation, but it also prevents future generalizations based on a comprehensive body of research encompassing multiple languages, language pairs, and scripts.

Acknowledgment The results reported in this chapter are part of a project sponsored by University of Macau Research Grant AuTema-PostEd MYRG058 (Y1-L1)-FSH12-ALL, and carried out with the kind cooperation of the Centre for Research and Innovation in Translation and Translation Technology (CRITT), at Copenhagen Business School, Denmark, the Laboratory for Experimentation in Translation (LETRA), at Federal University of Minas Gerais, Brazil (grants CNPq 307964/2011-6, and FAPEMIG SHA/PPM-00170-14), and the Translation Lab at Federal

University of Uberlândia, Brazil (grant CNPq 461054/2014-0). The authors are very grateful to the editors for their valuable comments.

Appendix 1: Summary of Mixed-Effects Analysis

Table 11.10 Fixed effects in the analysis of ST and TT tokens total reading time, and total production time with estimated effects size, standard error, t- and p-values

Variable	Estimate	Std. error	t	p
Total reading time on ST Token				
Intercept	7.750364	0.310935	24.926	<2e-16
Log token length	0.359864	0.092903	3.874	0.000164
Log token position	−0.223413	0.062151	−3.595	0.000450
Log trigram probability	−22.56563	3.176764	−7.103	5.78e-11
Task (translation)	0.269650	0.301090	0.896	0.387272
Type of chain (A)	−0.008212	0.149548	−0.055	0.956287
Number of fixations on ST Token				
Intercept	1.83671	0.21923	8.378	1.34e-11
Log token length	0.36931	0.07098	5.203	6.90e-07
Log token position	−0.16169	0.04754	−3.401	0.000879
Log trigram probability	−12.67565	2.43980	−5.195	7.18e-07
Task (translation)	0.30240	0.19646	1.539	0.148426
Type of chain (A)	0.01305	0.11462	0.114	0.909544
First pass reading time on ST Token				
Intercept	5.61694	0.09665	58.117	<2e-16
Log trigram probability	−1.39461	0.50621	−2.755	0.00664
Task (translation)	0.18261	0.13136	1.390	0.18839
Total reading time on TT Token				
Intercept	7.2591	0.2175	33.371	<2e-16
Log frequency (poly. 1)	−2.6958	1.1700	−2.304	0.0229
Log frequency (poly. 2)	4.7131	1.1569	4.074	8.18e-05
Task (translation)	0.2772	0.2543	1.090	0.2980
Type of chain (A)	0.4325	0.2063	2.097	0.0381
Number of fixations on TT Token				
Intercept	1.2405	0.2820	4.399	0.000176
Log length	0.3549	0.1034	3.431	0.000806
Type of chain (A)	0.3347	0.2449	1.367	0.219303
TT Token total production time				
Intercept	4.56944	0.19611	23.300	<2e-16
Log character count	1.29942	0.07501	17.323	<2e-16
Correctness of token (wrong)	0.51087	0.13583	3.761	0.000265
Type of chain (A)	0.20504	0.09570	2.142	0.034301

Table 11.11 Random effects in the analysis of ST and TT comprehension and production time

Variable	Random factor	Intercept/level	Standard deviation
Total reading time on ST Token	ST Token	Intercept	0
	Participant	Intercept	0.5097
	Residual		0.6155
Number of fixations on ST Token	ST Token	Intercept	0
	Participant	Intercept	0.3244
	Residual		0.4712
First pass reading time on ST Token	SToken	Intercept	1.565e-08
	Participant	Intercept	1.848e-01
	Residual		5.045e-01
Total reading time on TT Token	TT Token	Intercept	0
	Participant	Intercept	0.2783
	Residual		1.1424
Number of fixations on TT Token	TT Token	Intercept	0.2393
	Participant	Intercept	0
	Residual		0.8528
TT Token total production time	TT Token	Intercept	0
	Participant	Intercept	0.2401
	Residual		0.4782

Appendix 2: Source Text

O gigante asiático (5) continua com fome. O Instituto de Ouro da China divulgou que, em 2012, **o volume de produção (1)** de ouro no país (6) atingiu 403 toneladas. **Isso (2)** o (7) torna o maior produtor mundial desse metal. **O aumento (3)** da produção chinesa foi devido às políticas governamentais de apoio à indústria da mineração. É fato, contudo, que **essa quantidade (4)** não é suficiente para satisfazer a gigantesca procura pelo ouro na China (8). Ela tem crescido consideravelmente nos últimos anos.

References

Alves, F., Pagano, A. S., & da Silva, I. A. L. (2014). Effortful text production in translation. *Translation and Interpreting Studies, 9*(1), 25–51.

Angelone, E. (2010). Uncertainty, uncertainty management, and metacognitive problem solving in the translation task. In G. M. Shreve & E. Angelone (Eds.), *Translation and cognition* (pp. 17–40). Amsterdam: John Benjamins.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to Statistical Using R*. Cambridge: Cambridge University Press.

Balling, L. W. (2008). A brief introduction to regression designs and mixed-effects modelling by a recent convert. In S. Göpferich, A. L. Jakobsen, & I. Mees (Eds.), *Looking at eyes: Eye-*

tracking studies of reading and translation processing (Copenhagen studies in language, Vol. 36, pp. 175–192). Frederiksberg: Samfundslitteratur.

Balling, L. W. (2013). Reading authentic texts: What counts as cognate? *Bilingualism: Language and Cognition, 16*(3), 637–653.

Balling, L., & Carl, M. (2014). Production time across language and tasks: A large-scale analysis using the CRITT translation process database. In J. Schwieter & A. Ferreira (Eds.), *The development of translation competence: Theories and methodologies from psycholinguistics and cognitive science* (pp. 239–268). Cambridge: Cambridge Scholar Publishing.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4. R package version 3.1.2*. Available at http://CRAN.R-project.org/package=lme4

Bell, R. T. (1991). *Translation and translating: Theory and practice*. London: Longman.

Carl, M. (2012). Translog-II: A program for recording user activity data for empirical reading and writing research. In *Proceedings of the eighth international conference on language resource and evaluation* (pp. 4108–4112). Istanbul: European Language Resources Association.

Carl, M. (2013). Feature representation in the translation process research DB. In R. Bonk, V. Alabau, M. Carl, & P. Koehn (Eds.), *D5.3: Beta release of Casmacat workbench*. Available at http://www.casmacat.eu/uploads/Deliverables/d5.3.pdf

Carl, M., & Dragsted, B. (2012). Inside the monitor model: Process of default and challenged translation production. *Translation: Corpora, Computation, Cognition, 2*(1), 127–145. Special issue on the Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation.

Carl, M., & Jakobsen, A. L. (2009). Towards statistical modelling of translator's activity data. *International Journal of Speech Technology, 12*(4), 125–138.

Denver, L. (2009). Unique items in translation. In S. Göpferich, A. Jakobsen, & I. Mees (Eds.), *Behind the mind: Methods, models and results in translation process research* (pp. 125–148). Copenhagen: Samfundslitteratur.

Frisson, S., Rayner, K., & Pickering, M. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 862–877.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. New York: Longman.

Hasan, R. (1984). Coherence and cohesive harmony. In J. Flood (Ed.), *Understanding reading comprehension: Cognition, language and the structure of prose* (pp. 181–219). Newark, DE: International Reading Association.

Hatim, B., & Mason, I. (1990). *Discourse and the translator*. New York: Longman.

Hvelplund, K. T. (2011). *Allocation of cognitive resources in translation: An eye-tracking and key-logging study*. Published PhD thesis, Copenhagen Business School, Copenhagen.

Hyönä, J., Lorch, R. F., Jr., & Rinck, M. (2003). Eye movements measures to study global text processing. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 313–334). Amsterdam: North-Holland.

Jakobsen, A. L. (2011). Tracking translators' keystrokes and eye movements with Translog. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research: Integrative approaches in translation studies* (pp. 37–55). Amsterdam: John Benjamins.

Leong, K. S., Wong, F. D., Tang, C. W., & Dong, M. (2006). CSAT: A Chinese segmentation and tagging module based on the interpolated probabilistic model. In Z. H. Yuan & M. W. Yao (Eds.), *Computational methods in engineering and science* (pp. 1092–1098). Sanya: Tsinghua University Press/Springer.

Li, X. J. (Ed.). (2010). *现代汉语规范词典 [Modern Standard Chinese Dictionary]*. Beijing: Foreign Language Teaching and Research Press.

Li, X. S., Grimes, S., & Strassel, S. (2009). *Linguistic data consortium. Guidelines for Chinese-English word alignment, version 4.0.* Philadelphia, PA: Linguistic Data Consortium. Available via https://catalog.ldc.upenn.edu/docs/LDC2012T16/GALE_Chinese_alignment_guidelines_v4.0.pdf

McDonald, S. A., & Shillcock, R. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science, 14*(6), 648–652.

Mesa-Lao, B. (2014). Gaze behavior on source texts: An exploratory study comparing translation and post-editing. In S. O'Brien, L. W. Balling, M. Carl, M. Simard, & L. Specia (Eds.), *Post-editing of machine translation* (pp. 219–245). Newcastle upon Tyne: Cambridge Scholar Publishing.

Mossop, B. (2003). *An Alternative to "Deverbalization"*. http://www.yorku.ca/brmossop/Deverbalization.htm

O'Brien, E. J., Raney, G. E., Albrecht, J. E., & Rayner, K. (1997). Processes involved in the resolution of explicit anaphors. *Discourse Processes, 23*, 1–24.

O'Brien, E. J., Shank, D. M., Myers, J. L., & Rayner, K. (1988). Elaborative inferences during reading: Do they occur on-line? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 410–420.

Rayner, K. (1998). Eye movements in reading and information processing. *Psychological Bulletin, 124*(3), 372–422.

Rayner, K., Li, X., Juhasz, B. Z., & Yan, G. (2005). The effect of word predictability on the eye movements of Chinese readers. *Psychonomic Bulletin & Review, 12*(6), 1089–1093.

Sjørup, A. C. (2013). *Cognitive effort in metaphor translation: An eye-tracking and key-logging study.* Published thesis, Copenhagen Business School, Copenhagen.

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. In G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 327–342). Oxford: Oxford University Press.

Tian, L., Wong, F., & Chao, S. (2011). Word alignment using GIZA++ on Windows. *Machine Translation Summit, 13*, 369–372.

van Gompel, R. P. G., & Majid, A. (2004). Antecedent frequency effects during the processing of pronouns. *Cognition, 90*, 255–264.

Wong, D. F., & Chao, L. S. (2010). PCT: Portuguese-Chinese machine translation systems. *Journal of Translation Studies, 13*(1–2), 181–196.

Zang, C. L., Liversedge, S. P., Bai, X. J., & Yan, G. (2011). Eye movements during Chinese reading. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 961–978). Oxford: Oxford University Press.

Zeng, X. D., Wong, D. F., Chao, S., & Trancoso, I. (2013). Graph-based semi-supervised model for joint Chinese word segmentation and part-of-speech tagging. In *Proceedings of the 51st annual meeting of the association for computational linguistics (ACL 2013)* (pp. 770–779). Sofia, Bulgaria: Association for Computational Linguistics.

Chapter 12
The Task of Structuring Information in Translation

Bergljot Behrens

Abstract The present chapter compares and evaluates the merits of three recent studies dealing with the cognitive processes of structuring information in translations. The studies differ in taking a syntactic, a functional and a conceptual approach respectively. Correlation between structuring operations in translation and cognitive effort is found to be higher when a conceptual relevance-theoretic approach is taken, yet the results are somewhat inconclusive due to weaknesses in the operationalization of the relevance theoretic concept of procedural information. The syntactic parsing approach would also be improved by a more fine grained analysis. Functional categories as well as reallocation measures are found to be relevant for a more precise understanding of the effort related to structuring operations in translation.

Keywords Translation effort • Target text structuring • Re-distribution • Syntactic vs conceptual approach • Information structure

12.1 Introduction

Beyond choosing adequate lexical items for a target text, translators have to decide on a proper structure in their translation. Sometimes the structuring involves a pure mapping of the source text syntax into the target sentence string, with slight modifications on account of regular syntactic differences in the relevant language pair, but in most cases (Thunes 1998, 2011)[1] the translators have to or choose to restructure the information given in the source text. It may be assumed that

[1]This finding is based on the systematic analysis of a bidirectional English-Norwegian corpus of 68,000 words, including fiction and legal texts comprising about 4500 clause strings: 55.2 % of the data are classified as only pragmatically equivalent to their source strings (Thunes 2011: 257).

B. Behrens (✉)
Department of Literature, Area Studies and European Languages, University of Oslo, Oslo, Norway
e-mail: bergljot.behrens@ilos.uio.no

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_12

these restructuring operations are lexically motivated, in that the chosen target word or phrase comes with a different syntactic frame, they may be information structurally motivated in order for the target phrase to get the right focus, or it may be that the translator performs an unpacking of a source phrase only to re-pack the information in a more implicit or a more explicit form, possibly involving a complete redistribution of the information in the source. One assumption in cognitive translation studies is that the more alternatives the translator entertains before selecting her target expression, the more demanding the translation. Campbell (2000) hypothesized that multi-translation data, i.e. translations of the same source text by a number of translators, can be used to draw inferences about the cognitive processes during translation. His Choice Network Analysis (CNA) postulates that the more options and the more complex choices a translator has to consider, the more effortful is the translation of a particular item. Various measures of translation effort have been proposed to test this hypothesis, and different approaches have been suggested to isolate the relevant kinds of unit a translator considers. Among them, three papers (Dragsted (2012) and Carl and Schaeffer (forthcoming), see also Chap. 9) focus on the lexicon and the effect of target text variation on translator behavior. These studies demonstrate a significant correlation between reading times and the number of target lexical options available for a particular source word, indicating that translators entertain target alternatives already during reading the source text. Similarly, studies are beginning to appear that report on the cognitive effort of structuring translation segments (Chap. 10; Alves and Gonçalves 2013). The present paper takes up questions pertaining to the operationalization of structuring mechanisms and their relevance to the measure of cognitive load in translation. This involves two issues: What are the relevant (re-)structuring mechanisms in translation and how do we relate them to translation behavior?

The paper is structured into four parts. After presenting the types of measure used in translation process research on cognitive load (Sect. 12.2) and the assumptions forming the background for the studies to be discussed here (Sect. 12.2.1), the paper assesses three different analyses of structuring operations in translation (Sect. 12.3). Section 12.3.1 takes up in detail the merits and problems with a study in which shallow syntactic annotations form the basis of analysis, Sect. 12.3.2 discusses an alternative annotation system which makes use of a more complex syntactic annotation including functional categories. Section 12.3.3 assesses the approach by which the relevance theoretic notions of procedural and conceptual encodings are operationalized to investigate the cognitive load of structuring information in translation. In the final remarks in Sect. 12.4, an information structural approach to contrastive translation studies is suggested as a way ahead to get at the structuring mechanisms that involve cognitive translation load.

12.2 Measures of Cognitive Load in Translation

Previous studies that compare reading a text for comprehension and reading it for translation have shown that the two reading tasks are approached very differently; reading a source text for subsequent translation is slower, the saccades are shorter and the fixations are longer (Jakobsen and Jensen 2008). This indicates very clearly that the purpose of the reading task has an impact on the reader's processing behavior. Jakobsen and Jensen, among others, interpret this to mean that the translator co-activates both source and target language during reading, i.e., some (pre-)translation is going on in the reading process. This implies that reading time during a translation task is a potential measure of the cognitive load of translating. With eye tracking technology, temporal measures of fixations or gaze on particular words or the reading of larger strings can be used as behavioral indicators of translation difficulties, which we shall see below.

Another measure of cognitive effort in translation processes is the temporal logging of pauses taken by the translator during the production of target segments and the number of edits performed on target strings, which is done with keylogging technology. Jakobsen (2011) suggests that the interaction of the two measures should be taken into account for a better understanding of the cognitive operations at play in the process of translation.

The studies reported on below have each measured cognitive effort in different ways, one using reading time measures and key activity duration,[2] the other using edits as a measure of cognitive effort. Both are relevant for answering questions about cognitive translation processes. Source text reading time measures assume that some (pre-)translation is going on already before writing (see the introduction above), indicating co-activation of the source and target languages. Since editing measures relate to operations on the target text, (pre-)translation considerations are not taken into account on this approach; it measures cognitive effort in the production phase only.[3]

12.2.1 Lexical and Structural Translation Options

Translation options are of various kinds. Dragsted (2012), Carl and Schaeffer (forthcoming) and Chap. 9, studied the correlation between reading times and lexical options in translation. Dragsted's experimental study finds that when the same lexical item is chosen by all her (eight) participants, total reading time on the source text is significantly lower than in cases where each participant opts for a

[2] In the present paper I concentrate on the reading time measures only.

[3] This does not mean, of course, that the translator does not go back to reading the source text while editing. The TPR-DB shows that often ST reading and TT writing occur concurrently (see Chap. 9).

different word. This is interpreted as an indication that the subjects actually consider target lexical alternatives in the mind while selecting a final target word, and that the more alternatives are considered, the more effortful is the selection process. Carl and Schaeffer go one step further by weighting such alternatives and apply weighting measures on a larger set of data. The relative weighting of alternatives across a large set of translation data collected on the same set of texts across several languages is quantified and measured in what they call translation entropy: a measure of the effect of an item's relative likelihood to occur (Shannon 1951).[4] If the likelihood of a choice is small, i.e., there are many different translations to choose from, then the entropy is high. The cognitive effort of selecting a translation is deemed high when there are many equally likely alternatives to choose from. On the same account, translation should be facilitated when there are only one or two options, i.e., when the entropy is low. Their hypothesis is confirmed: The correlation between entropy values and reading times was high. High (weighted) variation in the target texts correlates with high source text and target text total reading times, measured in means across participants, text and language combinations per character. The conclusion drawn from the study is that translators activate and entertain several translation options (consciously or sub-consciously) while reading.

The interesting correlations found on the lexical level in the above mentioned studies have triggered questions relating to whether this correlation would carry over into structural choice in translation.

12.3 Structural Choice

Structural choice involves choosing an appropriate information structure in the target language, which is not identical with, but includes plain surface syntactic choice. Syntactic choices can be a choice between an active or a passive structure, a choice between an intransitive or a transitive structure, or a choice between a prepositional phrase or a clause, to name a few. To some extent syntactic choices are clearly lexically driven, given that lexical items come with a syntactic frame.[5] A correlation between lexical choice and cognitive effort should therefore find its parallel in syntactic choice, although a weaker correlation would be expected since many lexical alternatives come with the same syntactic frame. Information structural choices also involve focus structure, which may imply redistributions of semantic material into different syntactic slots without a change in the overt syntax of the clause. Target language style conventions also differ (see for example Behrens 2014). The interplay between syntax and focus structure in translation will be considered towards the end of the paper.

[4]For a more extensive account of translation entropy, see Chaps. 2, 9 or 10.

[5]This does not imply that syntactic priming cannot also affect lexical choice (see Chap. 10 for the study on syntactic priming).

12.3.1 Syntactic Translation Entropy Studies

A first attempt to measure the correlation between syntactic variability and translation effort across languages appears in Bangalore et al. (Chap. 10). This study involves data sets comprising translations of the same English source texts into three languages, collected in the TPR-DB (see Chap. 2).The data collection is based on a number of experimental translation process studies by various researchers,[6] and includes behavioral measures of the translators' process performances. In this study, the source and target text segments (sentences) have been manually annotated for the syntactic features valency, voice and clause type. The variants have been weighted according to their relative likelihood to appear (on the basis of the variants resulting from the syntactic annotation of each segment), and entropy values have been computed. High syntactic entropy values were expected to correlate with high total reading times. The correlations turned out positive across the languages, thus indicating that the syntactic variability measured in the studies is a relevant factor in the effort of structuring target text. The positive result was seen when correlated with the translators' source text reading time. The results thus support the hypothesis that translators entertain syntactic translation alternatives also during source text reading. The study furthermore support Hartsuiker et al's hypothesis (2004) that shared syntactic forms across language pairs have a priming effect.

One may ask whether the annotation system chosen is not optimal for teasing out all the relevant structuring alternatives actually entertained by the translators. The relatively small effects relative to the strong effects that were found in the studies on lexical choice mentioned in the introduction, may very well be due to the assumption suggested above that the lexical translation alternatives entertained very often come with the same syntactic frame.

Example (1) illustrates the system. Each data set has between 20 and 32 translations from English, albeit an unequal number of translations for each text in the various language experiments. The examples show but one of the choices for each language.

(1) a. ST: Only the attention of other hospital staff put a stop to him and the killings.
(transitive, active, independent: TAI)

b. DE: Nur die Aufmerksamkeit der anderen Krankenhausmiterbeider setze ihm und den Morden ein Ende.
(*Only the attention the-GEN other-GEN hospital staff set him and the murders an end*)
(transitive, active, independent: TAI)

c. DA: Det var udelukkende opmærksomhed fra andre hospitalsmedarbejdere, der fik stoppet ham og mordene.

[6]The studies from which the data was taken: SG12 for German, KTHJ08 for Danish, and BML12 for Spanish, for a description of these studies, see Chap. 2.

(*It was only attention from other hospital staff that got stopped him and the murders*)
(impersonal active independent: MAI, transitive active dependent: TAD = MAI-TAD)

d. ES: Solo el hecho de que el personal reparara en ello pudo parlarle los pies y detener los asesinatos. (TAD-DAI-TAI)
(*Only the fact that the personnel noticed him could stop his feet and end the murders*)

Valency (transitive(T), intransitive(I), ditransitive(D), impersonal(M)), Voice (active(A), passive(P)) and Clause Type (dependent(D), independent(I)) mark a triplet of syntactic features for each clause. The example shows that some translations retain the structure of the source segment, while others are more expansive, including a combination of clauses.

The annotation system allows us to see the variation in syntactic constellations for each language, as per translator. In Spanish, for example, the source segment in (1) yields several structures, alternating between the TAD-DAI-TAD (as in (1)), a simple active ditransitive (DAI) and an MAD-DAI combination. In Danish, the same segment shows over 10 different options, from a simple TAI structure or a TPI structure, to embedded structures of four clauses of various kinds (MAI-TAD-IAD-TPD or MAI-TAD-TPD-IAD). The syntactic entropy value is computed on the basis of each syntactic form's likelihood to occur, and then correlated with the translator's reading time on the source segment and the target segment.

The merit of the annotation system is that it captures clause-level syntactic features that are applicable across all the languages in the data set, which ensures comparability, and makes it possible to study syntactic variability on a much larger size corpus than we generally find in the translation process literature. This has not been done before. Its weakness is that it may be too general to capture the structural alternatives that correlate with the more demanding tasks, whether language specific or across target languages. For a better understanding of the choices available to a translator at a given point in a text, one would need a more fine-grained framework, although also one general enough to allow for comparison across the languages under study.

Structural choice involves a variety of operations that one would expect the translator to entertain and find difficult to decide on, such as category changes on the phrase level and the morpho-syntactic level as well as other syntactic restructurings and redistributions of information. Such choices may be driven by cross-linguistic differences at various levels, and may even be triggered by information structural and/or functional cues in the source texts that the translator makes use of to infer meanings that are only implicitly expressed in the source.

In the following some examples are looked into in more detail for an evaluation of factors the system can capture and factors that will be overlooked by it.

Phrase level encodings my cause effortful restructuring operations that are not captured by our annotation and thus not reflected in the analysis. Translators introduce a variety of changes, such as shifts in grammatical functions. One example

is the translation of a source text complex compound and its Danish translation:

(2) a. ST: To make matters worse, escalating prices are racing ahead of salary increases, especially those of nurses and ..., who have suffered from the government's insistence that those in the public sector have to receive below-inflation salary increases. (TAD)

b. DA: at de offentlig ansattes lønstigninger skal ligge under inflasjonsraten. (IAD)
(*that the public-sector employees' salary increases shall lie under the inflation rate*)

The source text has a syntactic structure of four clauses: the main clause follows a context connecting sub-clause, the apposition following the main clause is not registered in our system since it is not a clause, the subsequent relative clause picks up the referents of the apposition as subject, and the final clause of the sentence functions as a complement to a nominalization in the prepositional adjunct: MAD, IAI, IAD, TAD.

The object of the last clause, a complex compound, is unpacked and redistributed into other syntactic functions in the Danish translation: the head of the syntactic object 'salary increases' is made the head of the syntactic subject in the translation, while the modifier is partly recategorized into a verb, and partly encoded in a prepositional phrase. Such unpackings and re-allocations of information are thought to be cognitively demanding. The difference between the source and target structure in this clause is annotated as a change from a transitive to an intransitive structure in our system, which hardly reflects the many restructuring operations that have taken place, also syntactically. Although the changes are indirectly reflected in the annotation from a TAD structure to an IAD structure, and thus count as a variant in the entropy computation, the analysis obscures the many translation operations the translator has coped with.

Another type of change not reflected in the syntactic analysis is metaphorization as a re-categorization procedure.

Consider the Spanish translation in the following segment:

(3) a. ST: His withdrawal comes in the wake of fighting flaring up again in Darfur and is set to embarrass China, which has sought to halt the negative fallout from having close ties to the Sudanese government.

b. ES: Su retirada ha coincidido con una nueva intensificación armada en Darfur y sin duda significará para China una mella pública. China a su vez ha realizado un intento de no cortar los estrechos lazos que le unen al gobierno del Sudan.
(*His withdrawal has coincided with a new military intensification in Darfur and no doubt will signify for China a public dent. China, in turn, has effected an intent not to reduce the close ties that unites it to the Sudanese government*)

The example is a case of irregular, complex re-categorization found in the Spanish dataset P05_T3, (segment 3). The clausally postmodified nominalization 'fighting flaring up again in Darfur' is translated as a noun phrase 'una nueva

intensificación armada'—(a (new) military intensification)—with the relative clause information placed in the noun and the nominal information placed in the adjective. The metaphor 'una mella pública' (lit.: a public dent) is of interest here, considered creative relative to the source 'embarrass'. Note also the re-categorization of information in this clause; the semantic content of the main verb in the English source is re-categorized into the metaphoric noun phrase. The metaphor furthermore includes information inferred from the next (sub-)clause of the source: the negative fallout implies a negative response from public opinion. The re-categorization operation is irregular, unlike the general re-categorization operations such as nominalization or sententialization, and unlike general expansion or explicitation, viz.:

'is set to embarrass China' (TAI) → significará para China una mella publica (TAI)

In comparison, P02, P07, who spend less time on the segment, are closer to the source text formulation, retaining the relative clause structure and the verbal expression of the second conjunct[7]:

(3) c. P02: Su prostesta aparece en el momento en el que Darfur está más oprimida y sirve para avergonzar al gobierno chino . . . .
(His protest come at the momento in which Darfur is more squeezed and serves to embarrass the Chinese government . . .)

d. P07: Su rechazo se relaciona con los nuevos combates que han surgido en la región de Darfur y su objetivo es dejar en evidencia a China . . . .
(His withdrawal relates to the new fights that have risen in the region of Darfur and his objective is to unmask for China . . . .)

The restructuring operations chosen by P05 are not reflected in our annotation, yet the translator who produced this translation spent twice as long on this segment as the next highest, as can be seen in Table 12.1, showing the target text reading times per token on the five segments in the text:

In sum, then, the triplet annotation system captures all the solutions that affect the number of clauses used in the segments. This means that any restructuring involving

Table 12.1 Gaze time on the target text, measured per source text token in the Spanish data

	GazeT/TokS			
Segment	P05	P08	P07	P02
1	6766	8942	2553	2847
2	2243	3489	1258	2053
3	**4136**	1848	1044	1477
4	1780	1719	410	2206
5	976	1076	1132	1328

The measure for the segment discussed above is marked in bold. The measures are computed from the CRITT TPR1.7.1 tables

[7] P09 has misunderstood the segment, so her solution is irrelevant for my purpose here.

the re-categorization of information from a phrase to a clause or a clause to a phrase is captured. However, position changes may affect reading time differently among the languages, since for example an adjective (pre-posed) restructured into a relative clause (post-posed) may affect temporal measures for Spanish less than for German and Danish on account of the fact that adjectival modifiers appear in postnominal position in all unmarked cases in Spanish, while German and Danish translators have to consider the options of a preposed adjective or a post-posed relative clause. This language difference may affect cognitive load, yet is not captured in the entropy analysis (see Jensen et al. 2009; Ruiz et al. 2008).

Our annotation system also captures the syntactic changes of passivization and the transitivity choice. Some preliminary looks at the temporal measures, not presented here, indicate that passivization and the choice of a transitive verb are not the most relevant measures unless the restructuring also includes other syntactic operations. When going through some of the segments of each text in the data, I find that syntactic operations of the following kinds are not captured by the system:

a) information merging and information splitting within the clause, such as the unpacking of a compound into a noun phrase with a post-posed prepositional phrase, or a reallocation of the information given in an adjective into a verb or vice versa;
b) explicitations from pronominal form to a repetition or a re-formulation of nouns;
c) changes in the semantic role of the subject (captured only if the valency of the verb changes);
d) generalizations involving a simplification of the clause-internal structure (for example dropping modifiers)
e) sub-clause type: finite and non-finite clauses are not distinguished, nor are adjectival clauses and adverbial clauses kept apart.
f) sub-clause embedding and cross-over phenomena within the clause are not marked.

In sum, the merit of the system is that it is a relatively simple measure that can be used across languages and that can be carried out within a reasonable time even though it requires manual annotation. As was seen above, it also captures a number of syntactic operations indirectly. However, it seems that some of potentially effortful structuring operations that involve clause-internal reallocation operations may be obscured, which will affect the results of a statistical analysis of cognitive effort in translation. There are also indications that information structural aspects of translation are important for restructuring operations, and should be considered in future work.

12.3.2 An Alternative Annotation System: The CroCo Corpus of Translations

One very thorough annotation system is found in the CroCo corpus of English and German texts and their respective translations (Hansen-Schirra et al. 2012). CroCo is a product-based corpus of published translations and their sources. It includes annotations of aligned translations at the levels of word and phrase as well as syntactic functions. The alignment links cross-over phenomena at all levels (Alves et al. 2010). Consider one of their examples:

(4) a. ST: We mapped these three stages to our business strategy, [. . .]the third stage focusing on the four elements that we could influence or control as mentioned above.

b. DE: Wir haben unsere Geschäftsstrategie genau auf diese drei Phasen abgestimmt. [. . .] In der dritten Phase liegt der Schwerpunkt auf der Beeinflussung und Steuerung der bereits angesprochenen vier Faktoren. (Alves et al. 2010: 117)

The CroCo alignment system maps segments that do not find a partner to pair with, such as the modal auxiliary 'could' in the English source above. The system also captures low level links which belong to different syntactic functions, such as 'the third stage', which appears as the subject of the absolute construction in the English version, but as the complement to a preposition (in der dritten Phase) in the German target. The choice of retaining the noun phrase in the initial position, yet including it in a prepositional phrase, triggers a re-categorization operation that changes the information in the English verb to a subject noun phrase with an informationally weak verb in the German target: 'focussing' → '(liegt) der Schwerpunkt'. Furthermore, the information in the relative clause is re-categorized to a nominalization. Finally, the interpersonal comment clause, 'as mentioned before', is re-categorized to an adverbally modified adjective phrase in the target and placed before the noun: C → Adv + Adj: 'as mentioned before' → 'bereits angesprochener'.

These restructuring operations would be expected to affect translation effort, yet most of the restructuring operations would not be visible in the annotation system of the Bangalore et al. studies discussed in the previous section, according to which the English segment consists of four clauses: the main clause, the absolute clause, the relative clause, and the final comment clause: TAI-TAD-TAD-TPD. The German translation consists of two independent clauses: TAI.TAI. Admittedly, though, the simple triplet system captures a compression of the information, which means that it captures some of the restructurings, although only indirectly.

Cross-over phenomena are clear indications of re-structuring that would be of interest for correlations with measures of cognitive effort and annotations at all levels are needed to capture them.

12.3.3 A Cognitive Measure of Restructuring: Conceptual and Procedural Encodings

An alternative, and very different approach, is presented in Alves and Gonçalves (2013), who study the translators' consideration of alternatives in terms of the changes or edits translators perform on target text units. They investigate the relative cognitive load according to cognitively based encodings in language. Processing effort is measured relative to the relevance-theoretic distinction between conceptual and procedural encodings, thus disregarding syntactic units in the classical sense. In a relevance theoretic account of communication linguistic material is input to the inferential mechanism which constructs and manipulates conceptual representations. Utterances encode two types of information: conceptual information, which is representational, and procedural information, which is computational in the sense of encoding instructions on how to manipulate the conceptual representations encoded in the lexical entities (Wilson and Sperber 1993: 1). Relevance theory is less concerned with syntactic categories than with the kind of words that encode procedural information. However, closed classes of function words carry procedural information, such as pronouns and other anaphors as well as conjunctions and other connective function adverbials (Allott 2013; Blakemore 1987).

Translation units (TUs) in Alves and Gonçalves' framework are very different entities than the syntactic clause units used in the scheme discussed in the previous sections. TUs are units of fluent target text typing up to a pause in the production of 2.4 s or more. Within the TPR-DB, sequences of coherent typing are referred to as Production Units (PUs), which are defined by 1 s of inter-keystroke pause, (see Chap. 2).

The TUs can be whole clauses or shorter units such as single words or syntactic phrases. A distinction is made between a *micro*-unit, which equals the definition above, and a *macro*-unit. A macro-unit includes all the edits on the micro-unit up to the final version of the translation, i.e. correction and reformulations on the unit that take place right after it has been produced, or only in the revision phase of the translation process, are included in the macro unit. These units may well be more realistic measures of cognitive entities considered for alternative translation solutions than whole segments, although there seems to be more general consensus in the linguistic literature that the clause is a realistic measure.

Although based on a small set of data, comprising eight translators' production of two texts between English and Portuguese (in both directions), the methodological approach taken in this study is interesting as an alternative to the segmental syntactic approach.

Edits on the TUs, indicators of cognitive load, are counted according to types, and according to when they occur: Edits that occur during the production of a translation unit or take place during the production of the next unit both count as edits during the production flow. Edits may occur later, meaning the translator stops in a unit farther away from the unit to be edited, or it happens in the revision phase. Types of edits are more or less complex, ranging from typos (t) and breaks in the completion

Table 12.2 Edits on procedural and conceptual encodings in Alves and Gonçalves (2013)

Type of edits in A&G (2013)	Overall mean numbers
Typos (t)	46.38
Completions (c)	5.94
Lexical (l)	12.81
Morphosyntactic (m)	17.25
Complex phrasal (p)	6.63
SUM l + p (CE)	19.44
SUM m + p (PE)	23.88

of a word to be typed (c) to lexical edits (l), morphosyntactic edits (m) and complex phrasal structures (p). The edits are then related to whether the unit is a procedural or a lexical encoding or both.

Annotation of procedural and conceptual encodings is not clear cut, certainly. The function of procedural expressions is to activate procedures whose main function is to help the hearer understand an utterance by finding the intended combination of context, explicit content and cognitive effects. In the traditional account of Relevance Theory, procedural encodings do not contribute to the truth conditions of an utterance, but trigger the derivation of implicatures relating to the meaning meant to be conveyed by the speaker. Classical examples of linguistic categories encoding procedural information are discourse connectives and conjunctions, and we may add focus particles and other function words that are conceived of as presupposition triggers in classical semantics. Conceptual encodings, on the other hand, are lexical words such as nouns, verbs, adjectives and adverbs, used to convey concepts that are extendable to propositions, which denote truth conditions. The distinction is still a matter of debate. The parallelism between the truth-conditional vs the non-truth-conditional distinction and the conceptual/procedural distinction is given up on a number of accounts, and there is furthermore an indication that lexical categories also carry procedural information (Wilson 2011). Analyzing translation units according to the distinction is therefore still a challenge. Alves and Gonçalves are well aware of the problem. They solve it by annotating TUs with complex phrasal structure edits (p) as an overlap category, belonging to both conceptual encodings (CE) and procedural encodings (PE). On this measure they find that overall, editing procedures are significantly higher on PEs than on CEs. The overall means in their study is repeated in Table 12.2 for an overview:

12.3.4 Conclusion

Syntactic operations as well as procedural encoding operations are likely involved in the cognitive task of structuring information in translation. According to the results of the studies reported on in this paper, procedural encoding seems to be a stronger indicator of higher processing effort than shallow syntactic annotation can bring out.

The few examples that have been provided in the present paper, demonstrate that structuring operations go beyond syntax; they include a redistribution of content within phrases and clauses which is not captured by the syntactic measures alone, and which are not clearly defined as procedural encodings in the literature. As a final note, I would add information structural constraints to procedural information, since they clearly inform the hearer about how to update the message with context. If basic information structural markers can be annotated, they should be included among the procedural encodings.

12.4 A Way Ahead

Doherty (2002) has made a thorough study of how focus structural differences in English and German lie at the heart of translation revisions from a draft to an optimal output. She also shows how it interacts with syntax. Her main psycholinguistic assumption is that focus interpretations are first read off from the linguistic form of a sentence before they are integrated with the information of the preceding discourse. A distinction is made between structural focus (sentence focus marked by stress)—and contextual focus (focal marking of updating procedures), both of which affect translation choice. If an analogous translation[8] results in a mismatch between structural and contextual focus, a restructuring of the analogous version will have to take place which involves a paraphrase that secures optimal processing conditions, not least from an information structural perspective (Doherty 2002: 161). It would be reasonable to think that information structural options of this kind are entertained by the translator and alternative redistributions considered to secure an encoding which is optimal for contextual update.

Finally, on the assumption taken up at the beginning of the paper that some translation is already going on during first time reading of the source text, it would be interesting in future work to test potential correlations between procedural encodings and source text reading time. If such correlations are not found, we may conclude that any pre-translation in the source text reading phase on the whole involves lexical translation alternatives in shallow or primed syntactic representations (see also Chap. 9), and that a more fine grained parse is left for the formulation phase only.

[8] An analogous translation, in Doherty's view, is one which retains high similarity of form at every level. Grammatically acceptable analogous translations are seen as the starting point for the translator's search for an optimal translation (Doherty 2002: 166).

References

Allott, N. (2013). Relevance theory. In A. Capone, F. Lo Piparo, & M. Carapezza (Eds.), *Perspectives on pragmatics and philosophy*. Berlin: Springer. 12 pp.

Alves, F., & Gonçalves, J. L. (2013). Investigating the conceptual-procedural distinction in the translation process. *Target, 25*(1), 107–124.

Alves, F., Pagano, A., Neumann, S., Steiner, E., & Hansen-Schirra, S. (2010). Translation units and grammatical shifts. In G. Shreve & E. Angelone (Eds.), *Translation and cognition*. Amsterdam: Benjamins.

Behrens, B. (2014). Nominalization: A case study of linguistic text conventions in comparable and parallel texts: English and Norwegian. In G. Ebeling, K. Hauge, & D. Santos (Eds.), *Corpus-based studies in contrastive linguistics. Oslo Studies in Language, 6*(1), 143–160.

Blakemore, D. (1987). *Semantic constraints on relevance*. Oxford: Blackwell.

Campbell, S. (2000). Choice network analysis in translation research. In M. Olohan (Ed.), *Intercultural faultlines* (pp. 29–42). Manchester: St. Jerome.

Carl, M., & Schaeffer, M. (forthcoming). Literal translation and processes of post-editing. In *Translation in transition: Between cognition, computing and technology*. Amsterdam: Benjamins.

Doherty, M. (2002). *Language processing in discourse: A key to felicitous translation*. London: Routledge.

Dragsted, B. (2012). Indicators of difficulty in translation: Correlating product and process data. *Across Languages and Cultures, 13*(1), 81–98.

Hansen-Schirra, S., Neumann, S., & Steiner, E. (eds.) (2012). *Cross-linguistic corpora for the study of translations. Insights from the language pair English-German*. W de Gruyter.

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science, 15*(6), 409–414.

Jakobsen, A. L. (2011). Tracking translators' keystrokes and eye movements with Translog. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research. Integrative approaches in translation studies*. Amsterdam: Benjamins.

Jakobsen, A. L., & Jensen, K. T. H. (2008). Eye movement behavior across four different types of reading task. In S. Göpferich, I. M. Mees, & A. Lykke Jakobsen (Eds.), *Looking at eyes. Eye-tracking studies of reading and translation processing* (Vol. 36, pp. 103–124). Copenhagen: Samfundslitteratur. special issue of *Copenhagen Studies in Language*.

Jensen, K. T. H., Sjørup, A. C., & Balling, L. W. (2009). Effects of L1 syntax on L2 translation. In F. Alves, S. Göpferich, & I. M. Mees (Eds.), *Methodology, technology and innovation in translation process research: A tribute to Arnt Lykke Jakobsen* (pp. 319–336). Copenhagen: Samfundslitteratur.

Ruiz, C., Paredes, N., Macizo, P., & Bajo, M. T. (2008). Activation of lexical and syntactic target language properties in translation. *Acta Psychologica, 128*(3), 490–500.

Shannon, C. E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal, 30*(1), 50–64.

Thunes, M. (1998). Classifying translational correspondences. In S. Johansson & S. Oksefjell (Eds.), *Corpora and cross-linguistic research: Theory, method, and case studies* (pp. 25–51). Amsterdam: Rodopi.

Thunes, M. (2011). *Complexity in translation*. PhD thesis forthcoming to the University of Bergen, Norway.

Wilson, D. (2011). The conceptual-procedural distinction: Past, present and future. In V. Escandell-Vidal, M. Leonetti, & A. Ahern (Eds.), *Procedural meaning: Problems and perspectives* (pp. 3–29). Bingley: Emerald Group.

Wilson, D., & Sperber, D. (1993). Linguistic form and relevance. *Lingua, 90*(1), 1–25.

Chapter 13
Problems of Literality in French-Polish Translations of a Newspaper Article

Dagmara Płońska

Abstract The present paper is concerned with the question of literality of translations. The theoretical part presents the results of some think-aloud protocol (TAP) research on literal translation regarded as a translator's basic procedure. It also deals with the problem of operationalization of literality in translation, enumerating Carl and Schaeffer's (n.d.) criteria for an ideal literal translation and presenting Kielar's (2013) definition of literal translation. The empirical part describes the results of a study concerning French-Polish translations of a newspaper article, involving 60 participants and using Translog as a primary logging tool. The main aim of the study was to investigate the degree to which translators' construction of a full mental representation of the source text prior to translation and their translation experience affect the literality of produced translations. An analysis of the relationship between the literality operationalized according to Kielar's definition and one of the definitional criteria for literality proposed by Carl and Schaeffer, namely the translation entropy, is an additional element.

Keywords Literality • Literal translation • Translation procedure • Translation experience • Text representation • Translation entropy • Translog

13.1 Literal Translation as a Translator's Basic Procedure

Many findings suggest that replacing words of one language with those of another without more complex text analysis is the predominant strategy of individuals with little experience in translation, as I already argued in my previous paper (Płońska 2014). For instance, Lörscher notes that "most of the foreign language students … produce translations mainly by an exchange of language signs" (Lörscher 2005, p. 605). Königs and Kauffmann observe that translation procedures of foreign language students are vocabulary-centered and their mental activity is focused mostly on the vocabulary to the detriment of the grammar (Königs and Kauffmann 1996,

D. Płońska (✉)
University of Social Sciences and Humanities, Warsaw, Poland
e-mail: dplonska@swps.edu.pl

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_13

pp. 18–19). Tirkkonen-Condit remarks that "novices tend to approach a translation task as a series of lexical or phrasal problems that are to be solved in the order in which they appear in the text. In novices' performance, translation tends to proceed word by word, phrase by phrase, sentence by sentence" (Tirkkonen-Condit 2005, p. 408).

On the other hand, research carried out by Tirkkonen-Condit (2005) showed that a tendency to translate literally occurred in both beginner and experienced translators. This was visible not only in the translation process but also in the finished translations. The author claims that literal translation is the default procedure used by a person translating a text until that person notices a problem with the text of the translation. This finding is in line with the theoretical considerations of Newmark, who believes literal translation to be the basic translation procedure (Newmark 1988, p. 70). The tendency to apply literal translation as a default procedure was also noted by Mandelblit (1996). In her psycholinguistic experiment bilinguals were asked to translate idioms from French into English and vice versa, each participant translating into their mother tongue. According to the researcher's hypothesis, the idioms with a different cognitive mapping in the target language would be more difficult and thus take more time to translate. For instance, the French expression "trouver le temps long" [lit. "to find the time long"; this and further English translations and annotations in square brackets are my own], which can be translated to English as "time is passing slowly", would be more difficult to translate than the expression "perdre du temps", which is a literal equivalent of the English idiom "to waste time". In the first case, French uses the "time as space" metaphor, while English uses the "time is a moving object" metaphor. In the second case both French and English make use of the "time is a valuable object" metaphor. The results confirmed the author's hypothesis but also showed that "when translating DMC [different mapping condition] sentences, subjects tended to first suggest a word-to-word (and "same mapping") translation for the source sentence and only later propose the better translation" (Mandelblit 1996, p. 493).

At the same time, comprehension strategies of professional and non-professional translators seem to differ. Tirkkonen-Condit (2005) notes that beginner translators and amateurs focus on lexical units and seek information in external translation aids, while experts concentrate on the text itself, its semantic, pragmatic and intertextual aspects, trying to extract as much information as possible. In other words, the comprehension strategies of amateurs have a local orientation, while those of experts are global.

These findings are in consonance with those of Jääskeläinen (1996). The researcher discovered that the authors of mediocre and poor translations rely more on linguistic knowledge, while the authors of good translations tend to apply world knowledge. According to her, "in the good processes most of the attention is directed at text comprehension, at relating the text to the extra-textual world. The less successful processes seem to remain more exclusively at the linguistic surface level" (Jääskeläinen 1996, p. 69). Similarly, analyzing translation processes of foreign language students, Königs and Kauffmann note "l'énorme restriction de l'activité de contextualisation qui ne s'effectue qu'au niveau de la phrase, voire même du

syntagme" ["a huge restriction of contextualization activity which occurs only on the level of the phrase, or even of the syntagm"] (Königs and Kauffmann 1996, p. 19).

Accordingly, the mental representation of the text being translated seems to have significant importance for the translation process. In this chapter I investigate whether forming a full mental representation of the source text before taking up the task will influence participants' translation behavior. In particular, I want to find out whether non-professional translators and translation students would produce less literal translations if they had a mental representation of the source text prior to translating. My research is based on the text comprehension model proposed by van Dijk and Kintsch (1983). This model distinguishes three main levels of text representation: the superficial level of words and syntax, the text base level consisting of propositions, and the situational model level presenting the situation described in the text. The main objective of the comprehension process is to develop an accurate situational model.

In contrast with the studies mentioned above, based on TAPs, my research employs keystroke log data from Translog (Carl 2012).

13.2 Literal Translation: Problems of Operationalization

One of the main challenges in this field of research is the lack of a single commonly accepted definition of literal translation (see Carl and Schaeffer n.d.; Chesterman 2011). For the purposes of translation process research, Carl and Schaeffer (n.d.) propose three definitional criteria for an ideal literal translation. According to them, a translation is literal if the word order is identical in source and target texts, if source and target text items correspond one-to-one and if each source text word has only one possible translated form in a given context. This last criterion is operationalised in terms of translation entropy (see Chap. 2, Sect. 2.4.7). Using these criteria it is possible to measure how literal a translation is.

As the present study is a part of a larger research project concerning other more complex translation strategies too, I needed a definition which would allow me to identify precisely the passages translated literally. Carl's and Schaeffer's criteria for an ideal literal translation seemed too narrow to cover all the instances of what I intuitively identified as literal translation. That is the reason why I used a different definition of literal translation, formulated by Kielar (2013, p. 51). Kielar's definition is the one I refer to further in the text every time I talk about literal translation. According to this definition, in literal translation, the rules of the syntax of the target language are used to combine the words calqued from the source language as separate lexical units. This definition does not presuppose that the word order should be preserved in translation. In fact, French and Polish differ substantially in terms of word order. As noted by Gniadek (1979, p. 131–132), "en français l'ordre des éléments est fixé depuis la disparition de la flexion nominale, tandis qu'en polonais l'ordre des éléments est plus libre, parce que la forme du nom

indique sa fonction dans la phrase" ["in French the order of items is fixed since the disappearance of the nominal inflexion, while in Polish the order of items is freer because the form of the noun indicates its function in the sentence"]. Given these differences, in the present study I decided not to apply the identical word order as one of the criteria of literality of translation. However, I wanted to verify if the notion of literality operationalized according to Kielar's definition correlates with the notion of entropy proposed by Carl and Schaeffer. The value of entropy indicates how many different translations a given source text word has. If a word has only one possible translation, it has an entropy value of 0. I admit the possibility of a source text word having more than one literal equivalent in the target language. This intuition is based on my previous experience. At the same time I suppose that the number of literal equivalents of a given word is limited. Therefore, it seems to me that there should be a strong relationship between the literality of translation of a given word as defined by Kielar (2013), and translation entropy as defined by Carl and Schaeffer (n.d.). The entropy values should be significantly smaller for the words translated literally according to Kielar's definition.

The study was aimed at ascertaining whether professional translators, non-professional translators and translation students differ in terms of the literality of the translations they produce. Firstly, I expected that the tendency to translate literally would decrease with experience, i.e. professionals would produce less literal translations than students and students less literal ones than non-professionals. Secondly, I investigated the impact of constructing an initial mental representation of the source text on the literality of produced translations. I hypothesized that participants who did form a mental representation of the text prior to translating would translate less text literally than those who did not. Thirdly, I analyzed the relationship between literality as defined by Kielar (2013) and entropy as defined by Carl and Schaeffer (n.d.). As I already stated above, my assumption was that the entropy values would be significantly smaller for the words translated literally according to Kielar's definition.

13.3 Method

13.3.1 Participants

The study involved 19 professional French to Polish translators aged 28 to 61, 20 students of applied linguistics with French language aged 22 to 34 and 20 persons with advanced-level French language skills and without a background in translation aged 25 to 54. Further in the text I refer to these groups by the terms "professionals", "students" and "non-professionals". The professionals' work experience ranged from 5 to 38 years at the time of the study. Among the non-professionals, 9 persons had DALF certificate (Diplôme approfondi de langue française) confirming their advanced knowledge of French. The other 10 persons were teachers of French in

upper secondary schools (Polish: liceum) in Warsaw and one person was a teacher of French in a primary school (Polish: szkoła podstawowa) at the time of the participation in this research.

13.3.2 Materials and Procedure

The study was conducted on an individual basis. The task was to translate from French to Polish. The participants were assigned to translate an article for a magazine covering European issues, and were asked to prepare a text ready for publication without any need for further corrections. The task was preceded by brief technical instructions regarding the software. The process of translation was recorded using Translog. The participants had access to hard copies of a French-Polish dictionary and a monolingual French dictionary and their behavior during translation was filmed. No online dictionaries were put to use. After the task was finished I interviewed the participants about the completed task.

The article had been written for the purpose of the study by a French journalist having good command of Polish. In view of the study's objective, it was deemed important for the text to have a narrative structure and be easy to understand but nonetheless present some problems in translation: idioms, metaphorical expressions, "false friends" etc. The length of 365 words was specified so that the text was long enough to reveal some regularities in participants' behavior but not too long due to time constraints.

To investigate the role of forming an initial representation of the text, the participants were randomly assigned to two groups. In the experimental group, before taking up the translation task, the participants were given the following instruction (in Polish): "Please read the following text very carefully. In a moment you will be asked to answer some questions about its content and form". After having read the French text, without being able to refer back to it, the participants filled in a questionnaire with a sentence recognition test and instructions to write a summary in Polish. The sentence recognition test contained four types of samples: literal samples from the text, paraphrases, correct conclusions, i.e. sentences which were not in the text but which are consistent with the text meaning, and the incorrect conclusions, i.e. sentences which are not consistent with the text meaning. In the control group the participants did not read the text and filled in a shorter version of the questionnaire regarding their personal information only.

The study was conducted on the premises of the University of Social Sciences and Humanities, in several upper secondary schools and at the participants' homes. The time for each task was not limited.

13.3.3 Data Analysis

The dataset was added to the CRITT Translation Process Research Database (see Chap. 2). The translations were manually aligned using the YAWAT tool (Germann 2008). For the purposes of the analysis, the data were subsequently processed into a set of tables.

In line with Kielar's definition of literal translation, for every passage of the French text I tried to imagine all the possible literal translations by using dictionary equivalents of French words and by connecting them according to Polish syntax rules. I used "The Great French-Polish Dictionary" (Dobrzyński et al. 1996) and "The Great Polish-French Dictionary" (Frosztęga 1995–2008) as a reference material. I took into consideration all the possibilities of word order available in Polish syntax. It does not mean that I physically made an exhaustive list of the possible literal translations for every sentence of the French text. Such a list would be very long for two reasons. Firstly, because according to the dictionary most of the French words used in the text have more than one Polish equivalent. Secondly, because in Polish the word order is freer than in French. I don't think having a list of all the possible literal translations of all the sentences of the text would be necessary for the purpose of the subsequent analysis. Therefore, I looked for all the dictionary equivalents of the French words used in a given passage, and imagined how they could be connected according to Polish syntax rules to envisage how the word order could be changed in the sentences thus generated.

Afterwards, I compared the actual translations with the imaginary literal ones. I marked all the passages that matched literal translations in terms of word form as translated literally. It means that for every translation all of the source text words were labeled as translated literally or non-literally. An example of the labeling is provided in Table 13.1. The abbreviation "lit." stands for "literal" while the abbreviation "non-lit." stands for "non-literal". The first column presents the original passage as well as its English literal translation (based on the dictionary entries). The next three columns present three of the possible literal translations I imagined: one with the same word order as in the original text (Imaginary literal translation 1) and two with different word order (Imaginary literal translations 2 and 3). The next two columns present two translations provided by participants. These translations are only partially literal. In the translation by Participant 1, lexical changes in rows 4 and 6 result in changes in form of the words in rows 5, 7 and 8. Accordingly, all the words in rows 4–8 are labeled as translated non-literally. In the translation by Participant 2, lexical change in row 5 results in changes in form of the words in row 7.

Articles, subject pronouns and possessive adjectives omitted in translation in accordance with the rules of Polish grammar, as well as the French prepositions "de" and "à" in the phrases translated as nominal inflections or adjectives, were considered together with the following words and labeled accordingly. The annotation was blind, i.e. I did not know who had produced the translations. Initially, I also adopted the procedure of back-translation described by Ivir (1997) in order to

Table 13.1 Labeling of different versions of Polish translation of a French passage

	Original text	Imaginary literal translation 1	Imaginary literal translation 2	Imaginary literal translation 3	Participant 1 translation	Participant 2 translation
1	Vous [you]	Pani	Musi	Po	Musi lit.	Swoją lit.
2	devez [have to]	musi	pani	swoją	pani lit.	emeryturę lit.
3	personnellement [personally]	osobiście	osobiście	emeryturę	osobiście lit.	musi lit.
4	venir [come]	przyjechać	przyjechać	musi	**zjawić się** **non-lit.** **[appear]**	pani lit.
5	chercher [and fetch]	po	do Bułgarii	pani	**w Bułgarii** **non-lit.** **[in Bulgaria]**	**odbierać** **non-lit.** **[collect]**
6	votre [your]	swoją	po	przyjechać	**po odbiór** **non-lit.** **[to collect]**	osobiście lit.
7	retraite [pension]	emeryturę	swoją	osobiście	**swojej** **non-lit.** **[your]**	**w Bułgarii.** **non-lit.** **[in Bulgaria]**
8	en Bulgarie. [in Bulgaria]	do Bułgarii.	emeryturę.	do Bułgarii.	**emerytury.** **non-lit.** **[pension]**	

ensure that the target text words identified as such were indeed literal equivalents of the source text words. However, originally, this method was used to check the semantic content of translation segments of at least two words. When I tried to apply the method to separate words it turned out to be unavailing, because all the words appearing in the French-Polish dictionary as the equivalents of a given French word could be translated back by the means of the same French word using Polish-French dictionary.

13.3.4 Results

The total number of source text words translated literally was used as a measure of literality of translation. This variable was examined with a 3×2 (Experience [non-professionals, students, professionals] × Initial text representation [yes, no]) analysis of variance (ANOVA). The results showed a statistically significant effect of the main variable experience, $F(2, 53) = 6.32$, $p < 0.01$, $\eta^2 = 0.19$ (see Fig. 13.1). The students translated significantly less text literally ($M = 162.55$; $SD = 27.46$) than non-professionals ($M = 193.55$; $SD = 33.40$).

Furthermore, the effect of the interaction between the variables Experience and Initial text representation was statistically significant, $F(2, 53) = 5.78$, $p < 0.01$, $\eta^2 = 0.18$ (see Fig. 13.2). Simple effects analysis showed that students translated

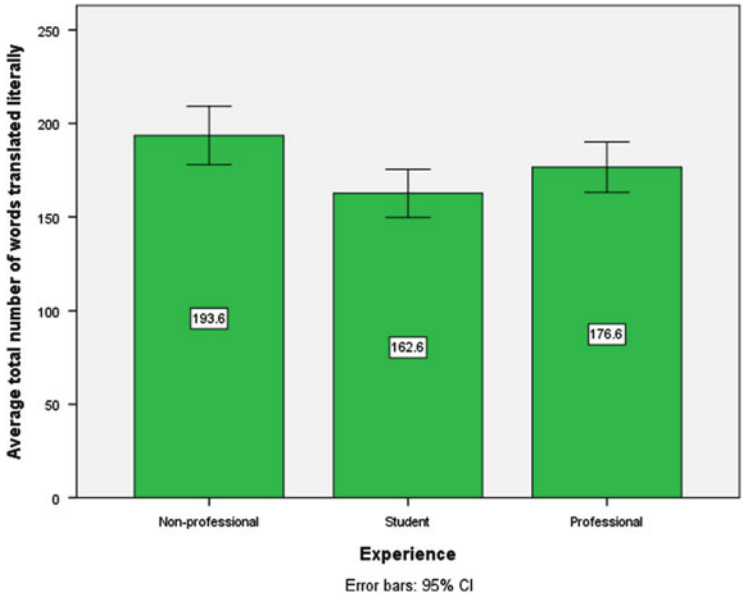


Fig. 13.1 Average total number of words translated literally depending on the experience

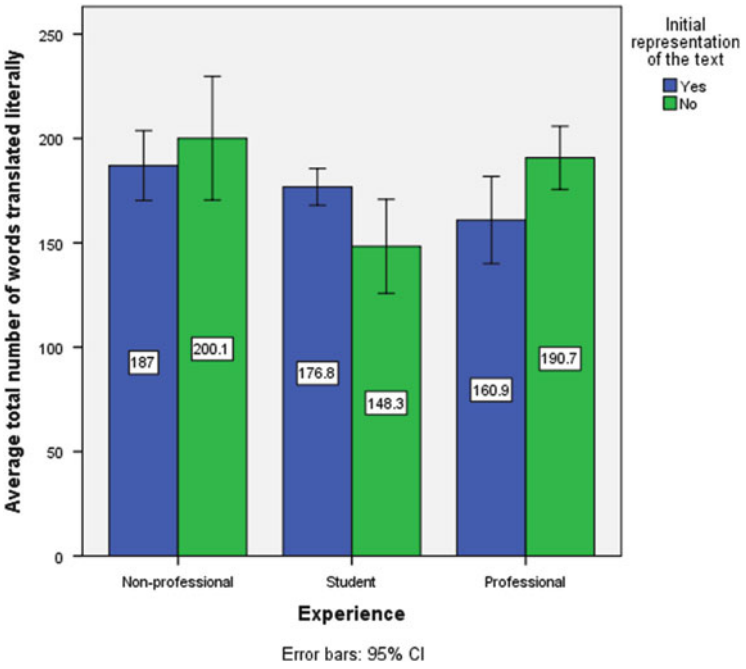


Fig. 13.2 Average total number of words translated literally depending on the experience and depending on whether an initial representation of the source text had been formed

significantly less text literally than both professionals and non-professionals but only when the participants did not form a mental representation of the text prior to commencing work. Creating an initial representation of the text significantly reduced the amount of text translated literally among professionals and significantly increased this amount among students.

I used the total number of source text words translated literally as a measure of literality of the whole translation. However, to investigate the relationship between the literality of translation of a given word and the entropy of translation alternatives I used a nominal variable literality concerning separate words. As I already stated above, for every translation, all of the source text words were labeled as translated literally or non-literally. The entropy values were also calculated for every source text word. In order to verify whether the nominal variable literality concerning separate words can be a good predictor variable of the entropy of translation alternatives, a one-way ANOVA was performed with literality as a factor. The effect of this variable was statistically significant, $F\ (1,\ 22772) = 7470.62$, $p < 0.001$, $\eta^2 = 0.25$. The entropy values were smaller in the case of source text words labeled as translated literally (see Fig. 13.3).

Another illustration of this relationship is provided by Fig. 13.4 presenting the number of occurrences of literal and non-literal translation depending on the entropy values.

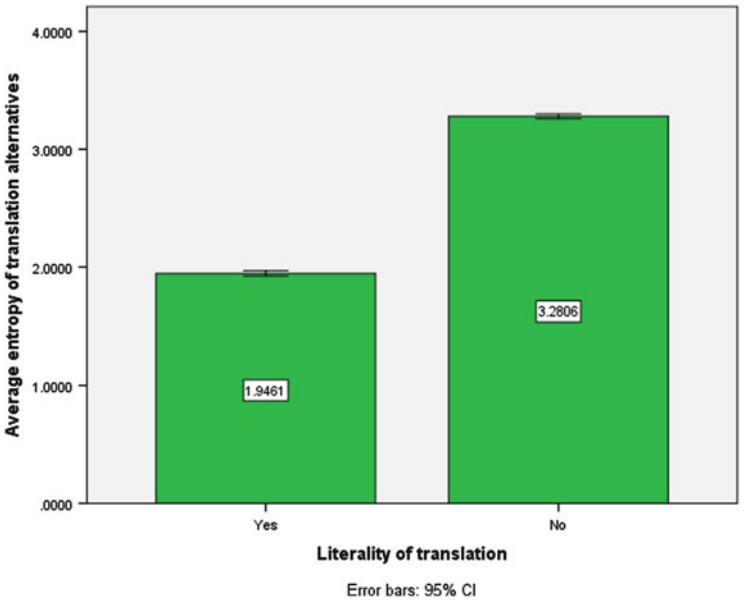


Fig. 13.3 Entropy of translation alternatives depending on whether the word was translated literally

13.3.5 Discussion

The results of the study are unanticipated. As it turns out, the students translated less text literally than professional translators. Moreover, there are no significant differences in the amount of text translated literally between professionals and non-professionals. The results also show that an initial mental representation of the source text has a substantial impact on the subsequent translation process in terms of the frequency of words translated literally. According to expectation, having an initial representation of the text made the differences between groups less substantial. However, the influence of this variable is different for the three groups. The students who did form an initial mental representation of the source text translated literally more text than those who did not. In contrast, the professionals who did construct a representation of the source text before taking up the task made less frequent use of literal translation than those who did not.

The results concerning the entropy of translation alternatives conform to my preliminary expectations. For the words labeled as translated literally according to Kielar's definition the entropy values are significantly lower. It means that the number of translations proposed by participants is significantly smaller in the case of words translated literally. This finding confirms the assumption that a word of one language has a limited number of literal equivalents in another language. It also shows that, to a certain extent, my operationalization of literal translation is

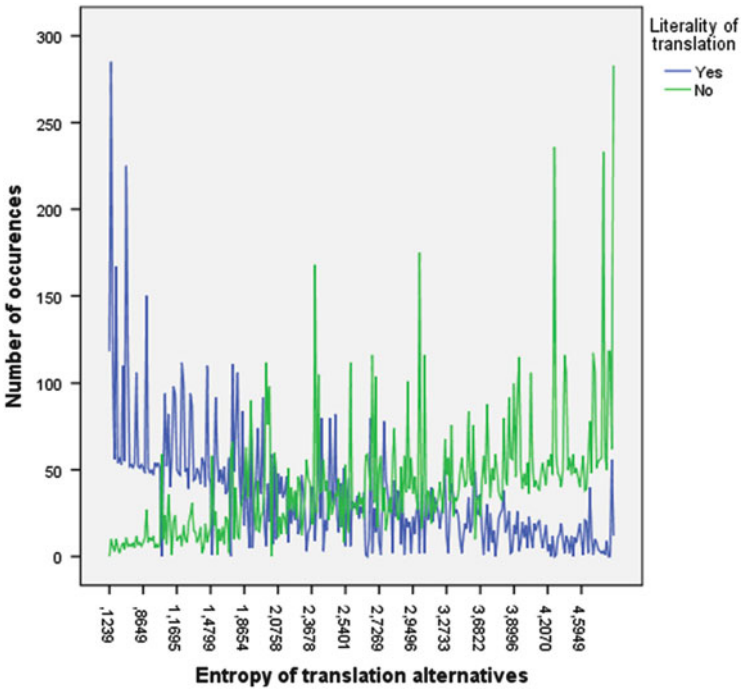


Fig. 13.4 Number of occurrences of literal and non-literal translation depending on the entropy of translation alternatives

compatible with Carl and Schaeffer's conception described in Chaps. 2 and 9. The entropy, one of the criteria used by Carl and Schaeffer to measure the literality of translations, is a continuous variable and the literality operationalized according to Kielar's definition is a nominal one. Nevertheless, considering the entropy criterion alone, the passages labeled as translated literally were significantly closer to an ideal literal translation as defined by Carl and Schaeffer than the passages marked as translated non-literally.

In my view students' reluctance to translate literally might have been the result of translators training. Students learn at a very early stage that the use of literal translation often results in translation errors and can be regarded as a sign of incompetence. This is the probable reason why they perceive this procedure as their last resort and try to avoid it by all means. As their experience grows, they learn to recognize the situations permitting the safe use of literal translation. In contrast, experienced translators can consciously use literal translation allowing them to provide translations that are both acceptable to target norms and adequate to the source. It is even easier to apprehend given that the use of this procedure requires less time and effort (cf. Schaeffer and Carl 2014). As the differences in the use of literal translation between professional and non-professional translators weren't statistically significant, it would be compelling to assess the quality of the

translations and to examine whether the use of this procedure was associated with specific translation errors in any of these groups.

The study confirms the importance of forming a mental representation of text before commencing translation. In light of the results, I believe that having a mental representation of the source text before taking up the task allows the translator to be freer in their choice of available translation procedures, including literal translation. This would help in explaining why, contrary to my preliminary expectation, the students who did form an initial representation of the text translated literally more text than those who did not.

The effect of experimental condition could also be interpreted in terms of a possible pre-translation during the initial reading and a probable priming effect (see Chaps. 9 and 10, this volume). Schaeffer et al. argue that reading for translation is substantially different from monolingual reading. However, when the participants were presented the original text for the first time, they were told the objective of the reading was to be able to answer the questions about the text form and content. The participants knew they were going to translate a text during the study. They might have supposed the text they were reading to be the one they would translate later. Nevertheless, they weren't explicitly told so. Certainly, some of the words of the original text were translated during writing the summary. On the other hand, the text was too long to be memorized, so at this stage the participants were writing a new text based on their recollection of the original text content rather than translating.

As far as the priming effect is concerned, the participants were presented not only the original text but also the sentence recognition test with different kind of samples, including paraphrases, correct conclusions and incorrect conclusions. They also wrote a Polish summary of the original text. As noted by Schaeffer et al. (see Chap. 9, this volume), in translation priming studies the priming from L1 to L2 was observed more often than priming from L2 to L1. Thus, it is legitimate to suppose that Polish words and syntactic constructions the participants used in their own summaries of the original text were more accessible to them during the subsequent translation. It might be an interesting concept for the future to examine the summaries written by the three groups of participants and to see how close they are both to the original text and to the translation text. An analysis of the time data from Translog could also shed a light on the role of a pre-translation and of a possible priming effect in the ulterior translation process.

A more complete picture of the translation process may be gained by researching the ways the three groups of participants apply more complex translation strategies. As a part of the current project, I also plan to take a closer look at the participants' errors by analyzing the entire process of making corrections.

Acknowledgment I would like to express my gratitude to the former director of the Center for Research and Innovation in Translation and Translation Technology at Copenhagen Business School, Arnt Lykke Jakobsen and to the current director of the center, Michael Carl, for their assistance and guidance in my work.

The research was supported by Polish National Science Centre (NCN); grant awarded by decision N° DEC-2013/09/N/HS6/02863.

References

Carl, M. (2012). Translog-II: A program for recording user activity data for empirical reading and writing research. In N. Calzolari (Ed.), *Proceedings of the eighth international conference on language resources and evaluation (LREC 2012)* (pp. 2–6), May 23rd–25th, 2012, Istanbul. European Language Resources Association.

Carl, M., & Schaeffer, M. (n.d.). Literal translation and processes of post-editing. In: Edited volume. *Translation in transition: Between cognition, computing and technology*, under review with Benjamins.

Chesterman, A. (2011). Reflections on the literal translation hypothesis. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research: integrative approaches in translation studies* (pp. 23–35). Amsterdam: John Benjamins Publishing.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Orlando: Academic.

Dobrzyński, J., Frosztęga, B., & Kaczuba, I. (Eds.). (1996). *Wielki słownik francusko-polski [The Great French-Polish Dictionary]*. Warsaw: Wiedza Powszechna.

Frosztęga, B. (Ed.). (1995–2008). *Wielki słownik polsko-francuski [The Great Polish-French Dictionary]*. Warsaw: Wiedza Powszechna.

Germann, U. (2008). Yawat: Yet another word alignment tool. In *Proceedings of the ACL-08: HLT demo session* (Companion Volume) (pp. 20–23). Columbus, OH: Association for Computational Linguistics.

Gniadek, S. (1979). *Grammaire contrastive franco-polonaise [Contrastive grammar of French and Polish]*. Warsaw: PWN.

Ivir, V. (1997). Formal/contrastive correspondence and translation equivalence. *Studia Romanica et Anglica Zagrabiensia, 42*, 167–180.

Jääskeläinen, R. (1996). Hard work will bear beautiful fruit. A comparison of two think-aloud protocol studies. *Meta, 41*(1), 60–74.

Kielar, B. Z. (2013). *Zarys translatoryki [The outline of translatology]*. Warsaw: Wydawnictwo Naukowe IKL@.

Königs, F. G., & Kauffmann, R. (1996). Processus mentaux étudiés chez des sujets allemands apprenant le français lorsqu'ils sont en train de traduire [Investigation of mental processes of French-learning German subjects involved in translating]. *Meta, 41*(1), 7–25.

Lörscher, W. (2005). The translation process: Methods and problems of its investigation. *Meta, 50*(2), 597–608.

Mandelblit, N. (1996). The cognitive view of metaphor and its implications for translation theory. In M. Thelen, & B. Lewandowska-Tomaszczyk (Eds.), *Translation and meaning, part 3. Proceedings of the Maastricht session of the 1995 Maastricht-Łódź Duo Colloquium on "Translation and Meaning", Held in Maastricht* (p. 483–495), The Netherlands, April 19–22, 1995 . Maastricht: Rijkshogeschool Maastricht, School of Translation and Interpreting.

Newmark, P. (1988). *A textbook of translation*. New York: Prentice Hall.

Płońska, D. (2014). Strategies of translation. *Psychology of Language and Communication, 18*(1), 67–74.

Schaeffer, M., & Carl, M. (2014). Measuring the cognitive effort of literal translation processes. In U. Germann, M. Carl, P. Koehn, G. Sanchis-Trilles, F. Casacuberta, R. Hill, & S. O'Brien (Eds.), *Proceedings of the workshop on humans and computer-assisted translation (HaCaT)* (pp. 29–37). Stroudsburg, PA: Association for Computational Linguistics.

Tirkkonen-Condit, S. (2005). The monitor model revisited: Evidence from process research. *Meta, 50*(2), 405–414.

Chapter 14
Comparing Translation and Post-editing: An Annotation Schema for Activity Units

Jean Nitzke and Katharina Oster

Abstract The current chapter introduces an annotation schema of TPR data that categorises post-editing behaviour into five different classes and compares general-language and domain-specific English-to-German translation and post-editing with respect to production times, key-logging (text production activity and text elimination activity) and eye-tracking data (total reading times on source text and on target text). The results support the hypothesis that post-editing is faster than translation from scratch for both domain-specific and non-domain-specific text types. When key-logging and eye-tracking data are taken into consideration, domain-specific texts require more effort when translating from scratch, but less effort, when the machine translation output is post-edited. It is hypothesized that the introduced annotation schema could provide more details about translation processes, and better insights into the differences between different domains.

Keywords Translation process research • LSP • Key-logging • Eye-tracking • Post-editing • Annotation schema

14.1 Introduction

The global demand for translated texts is constantly rising (De Palma 2009). Companies are increasingly using machine translation and editing the translations produced by the computer in order to improve the translator's efficiency (cf. O'Brien 2011). In regard to this development, translation process researchers started to investigate the advantages and disadvantages of post-editing and the differences between the translation process and the post-editing process (cf. Čulo et al. 2014; Carl et al. 2014; Winther Balling and Carl 2014). In this chapter we will present new findings and methods to analyse post-editing in order to help to find answers to the question how translation and post-editing differ.

J. Nitzke (✉) • K. Oster
Department for Language, Culture and Translation Studies Germersheim (FTSK), University of Mainz, Mainz, Germany
e-mail: nitzke@uni-mainz.de; osterk@uni-mainz.de

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_14

This chapter presents an annotation schema for the Translog data that has already been applied to some of the CFT13 study data and that categorises post-editing behaviour into five different classes. The main classes are the orientation and revision phases, which in turn are separated into two or three subclasses specified by indices. This annotation schema could also be adapted to translation behaviour and thus used to compare the two activities. Finally, we will introduce a possible extension of the database in regard to text type and domain-specific translation/post-editing. We believe that focusing on domain-specific texts is very interesting since they make up the biggest component of the translation market (Hommerich and Reiß 2011). We will thus use first findings to compare translation and post-editing behaviour in three domains: journalistic and sociology texts (classified as general-language—data already available in the database in study SG12), extracts from a refrigerator manual (domain-specific language) and extracts from patient information leaflets (domain-specific language). Then we will show the advantages and disadvantages of translation and post-editing in these fields.

14.2 A Novel Activity Unit Annotation Schema

In the investigation of the translation process, the main focus has been on different processing phases—namely orientation, drafting and revision (Dragsted and Carl 2013; Carl et al. 2014). We chose these phases as a starting point to identify different types of translators and post-editors, and to compare the two activities. We propose an annotation schema which will focus on the aforementioned phases and which is adaptable to both post-editing and translation.

This annotation schema is based on the sentence level and was developed for the CASMACAT study CFT13 which is part of the CRITT TPR database. CASMACAT presents aligned source-target segments in the user interface, where a translator can work independently. We therefore started in the annotation schema from the sentence level. Processing a sentence is considered a unit, which is then divided into several translation phases. A new sentence always starts with a new phase, although the translation phase (for example orientation) at the beginning of a sentence could be identical to the one at the end of the previous sentence.

We propose three groups of labels: orientation, drafting and revision (cf. Dragsted and Carl 2013). According to our definition, orientation and revision can be found in post-editing as well as in translation, while drafting is a typical of translation. We consider the machine translation output to be a first draft which has been produced by the computer; post-editing thus lacks manual drafting. Orientation and revision are therefore especially interesting when investigating differences between translation and post-editing. Different revision behaviour could, for example, be used to identify different types of translators and post-editors, and to investigate how their behaviour changes during the different tasks. In this chapter we will therefore

concentrate on orientation (O) and revision (R).[1] O and R have been specified by indices (orientation in source *or* target text *or* in source and target text: O_s, O_t, O_{st}; linear *or* scattered revision: R_l, R_s) which will be further described below.

14.2.1 The Orientation Phase

The orientation phase has been defined as several fixations on different parts of a sentence before any insertion or deletion takes place. When there are only fixations and no insertions or deletions in a segment, the orientation label is assigned to the whole segment. When we discovered a longer period of fixations between two editing activities within a segment, we decided to label this phase as orientation as well. The minimum duration of an orientation phase was 1230 ms, maximum duration 265,500 ms and the mean duration was 21,474 ms.

Indices were added in order to specify what the participant was looking at most of the time during the orientation phase (Table 14.1):

Figure 14.1 shows two examples of the orientation label assigned to the CFT13 data. The translation progression graphs represent activity data recorded with CASMACAT. The time line is represented on the x-axis, the numbers on the left y-axis represent successive words in the source text (ST). The grey lines define the segment borders. The dotted lines mark the beginning and end of a translation phase.

Table 14.1 Annotation labels for orientation

Label	Features
O_{st}	The participant spend time reading both source and target text (Fig. 14.1, left)
O_s	More than 80 % of the fixations were on the source text.
O_t	More than 80 % of the fixations were on the target text (Fig. 14.1, right)

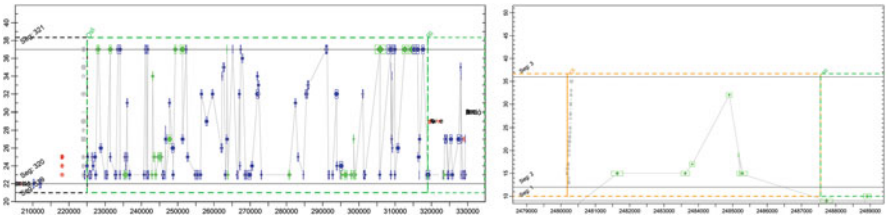


Fig. 14.1 (*left*) O_{st}—orientation in source and target text. The participant looked at both ST (*green dots*) and TT (*blue dots*); (*right*) O_t—orientation in target text. The participant looked more frequently at TT (*green dots*)

[1]The annotation can be retrieved from CFT13 study: https://svn.code.sf.net/p/tprdb/svn/CFT13/OR_cu/

Green dots represent fixations on TT words, blue dots are fixations on ST translations. Black letters are insertions, and red letters represent deletions. Since CASMACAT automatically inserts and deletes text depending on the input of the post-editor, we also find grey and light red letters in the progression graphs for the CFT13 data. These are automatic insertions and deletions made by the computer. When looking, for example, at Fig. 14.1 (left) from left to right, we encounter some blue dots on lines 23, 24, 25 at the beginning of the marked orientation phase (after the green dotted lines) which indicate the words that were read in the source text.

14.2.2 The Revision Phase

The revision phase has been defined as deletions and insertions in a sentence—in translations from scratch, it occurs after completing a first draft. Different fixation patterns have not been specified in the revision labels. When preceded by an orientation phase, the revision phase starts with the first deletion or insertion. When there are no, or almost no fixations on the text before the first deletion or insertion, the revision label is assigned to the whole segment. Minimum duration was 1300 ms, maximum duration was 681,861 ms and the mean duration was 80,933 ms. We divided the revision phase into linear revision and scattered revision (Table 14.2).

Figure 14.2 shows two examples of the revision labels. As in Fig. 14.1 (see above), green dots represent fixations in TT, blue dots are fixations in ST, black letters are insertions, red letters are deletions and grey/light red letters are automatically produced insertions/deletions. The dotted lines represent the manually annotated activity units. We also inserted red circles in order to highlight revisions

Table 14.2 Annotation labels for revision

Label	Features
R_l	Every word or phrase is processed only once. In Fig. 14.2 (left) for example, revisions 1–5 are on different parts of the text.
R_s	The participant works on a part of the text, moves on but jumps back later to readjust the parts (s)he already worked on. In Fig. 14.2 (right), for example, revision 1 is on the same part of text as revision 4.

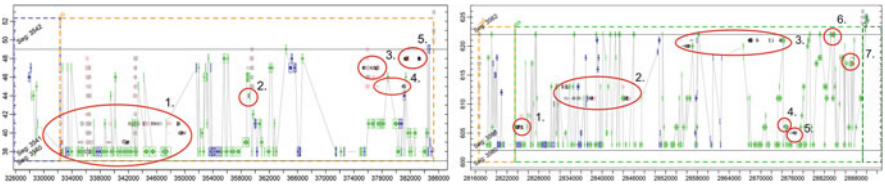


Fig. 14.2 (*left*) R_l—linear revision. The participant edited every word only once. Revisions 1–5 are on different words; (*right*) R_s—scattered revision. The participant edited one word twice. Revisions 1 and 4 are on the same word

made by the participant. The black numbers represent the order in which the revisions were performed.

14.2.3 Sequences of the Translation Phase

Figure 14.3 (left) depicts a case in which we decided to split up a revision phase and to label a longer period of fixations as O_{st}. Prior to adapting this annotation schema to other datasets or creating an algorithm to extract the phases automatically, there needs to be a discussion on whether fixations within a segment should be labelled as orientation or whether they are part of a revision phase.

Figure 14.3 right shows a sequence of several phases in two sentences. The sentences are processed in a linear manner which means that the participant first processes segment number 1415 and then continues to work on segment 1416.

In contrast to Fig. 14.3, Fig. 14.4 shows a sequence of several phases where the sentences are not processed in a linear manner. The participant first reads segment 1, then continues to read segment 2, and then jumps back again to segment 1 to edit the text.

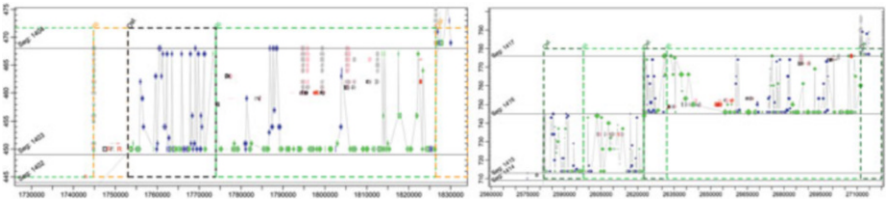


Fig. 14.3 (*left*) Orientation splitting up a long revision phase; (*right*) Sequence of several phases

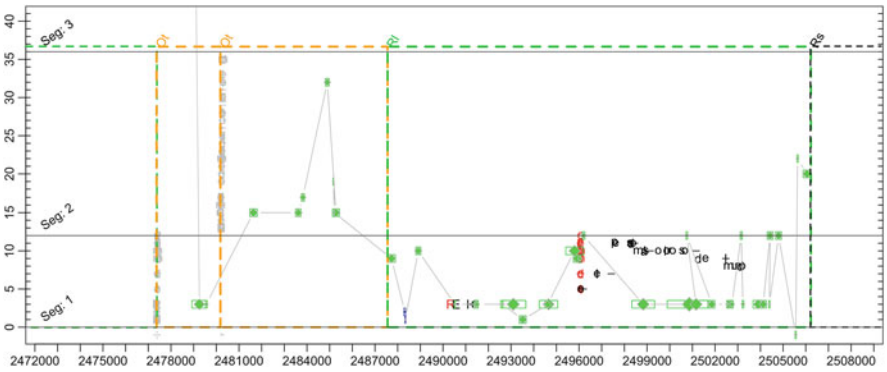


Fig. 14.4 Sequence of several phases

14.2.4 Translation Phases in CFT13

So far, 11 recordings have been annotated. They contain a total of 406 segments which were divided into 985 phases. Table 14.3 shows the distribution of the different labels for each participant. The table shows that there are only a few cases of O_s and that the numbers of O_t and R_s in particular vary between participants and tasks. We therefore believe that especially O_t and R_s could be interesting when comparing translation from scratch, post-editing, and different types of translators/post-editors on a large scale.

14.2.5 Activity Units in the TPR-DB

In this section we compare the segmentation of activity units (CUs), as available in the TPR-DB (see Chap. 2, Sect. 2.5.6) with our alternative segmentation. The TPR-DB CUs are automatically computed for all data sets in the TPR-DB. They provide information about user activity on a very fine granularity level, and distinguish between seven different categories:

- Type 1: Reading ST
- Type 2: Reading TT
- Type 3 Typing
- Type 4 Typing while reading ST
- Type 5 Typing while reading TT
- Type 6 Typing while reading ST and TT
- Type 7 No recorded activity

We used these annotations as a basis for our annotations. The granularity of the original TPR-DB activity units (CUs) is however quiet different from the one we suggest. Figure 14.5 reproduces the segments of Fig. 14.2 (right) with the original TPR-DB CU segmentation. The graph represents a duration of approximately 60 s, which is a single segment in the annotation (Fig. 14.2), but amounts to 44 segments in the TPR-DB annotation. Although we forfeit information about user activity when using the annotation schema presented in this paper, it is possible to distinguish immediately, for example, between different revision patterns. As Fig. 14.5 shows, it cannot easily be distinguish between scattered and linear revision behaviour, which however is possible with our annotation schema.

In contrast to the TPR-DB CU-units, which are computed automatically and which represent activity patterns on a very fine granularity level, the classification of the labels described above was created by visual analysis of the logging data. We believe that the advantage of the coarse granularity level of our annotation schema would allow users to distinguish between different types of post-editors and translators more easily. It might be easier to immediately discover a certain type of participant at first glance in our annotations, whereas the TPR-DB units need a

Table 14.3 Annotated data for study CFT13

Participant	O_{st}	O_s	O_t	R_l	R_s	Sum
P01_P11	28 (29.79 %)	0 (0 %)	19 (10.64 %)	43 (45.74 %)	4 (4.26 %)	94
P01_PI21	13 (26 %)	1 (2 %)	7 (14 %)	27 (54 %)	2 (4 %)	50
P01_PIA31	16 (22.86 %)	1 (1.43 %)	6 (8.57 %)	43 (61.43 %)	4 (5.71 %)	70
P03_P31	2[a] (3.57 %)	–	–	49 (87.5 %)	5 (8.93 %)	56 (101)[b]
P03_PI11	27 (42.19 %)	4 (6.25 %)	0	30 (46.88 %)	3 (4.69 %)	64
P04_P31	33 (33.67 %)	2 (2.04 %)	3 (3.06 %)	46 (46.94 %)	14 (14.29 %)	98
P04_PI11	45 (38.14 %)	2 (1.69 %)	6 (5.08 %)	51 (43.22 %)	14 (11.86 %)	118
P04_PIA21	32 (39.51 %)	1 (1.23 %)	5 (6.17 %)	34 (41.97 %)	9 (11.11 %)	81
P05_P11	33 (32.04 %)	3 (2.91 %)	8 (7.77 %)	49 (47.57 %)	10 (9.71 %)	103
P05_PI21	30 (38.96 %)	1 (1.30 %)	5 (6.49 %)	30 (38.96 %)	11 (14.29 %)	77
P05_PIA31	43 (33.33 %)	1 (0.76 %)	20 (15.50 %)	54 (41.86 %)	11 (8.53 %)	129
Sum	286	16	79	456	87	

[a]Some of the eye-tracking data were probably lost for this dataset after the first segments. We therefore decided not to use indices when annotating orientation phases. They were assigned the label O. A total of 44 phases in this dataset were labelled O.
[b]The number in brackets includes the label O

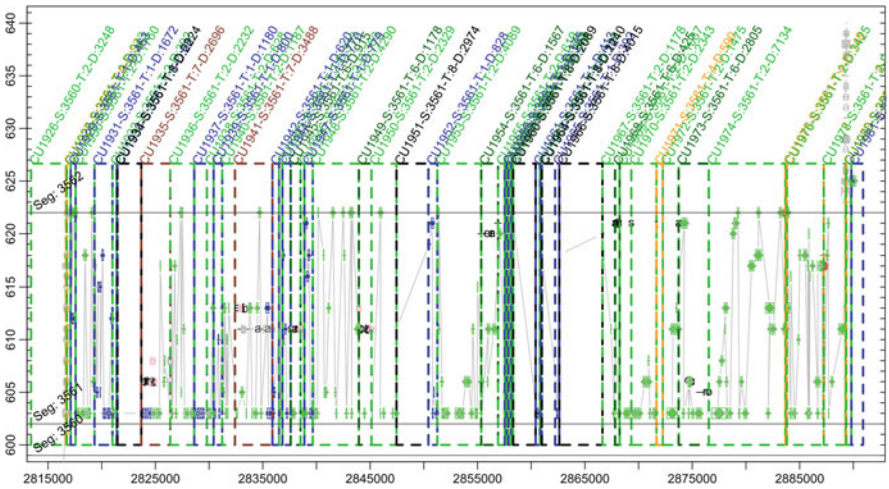


Fig. 14.5 R_s segment reproduction of Fig. 14.2 (*right*) with original TPR-DB CUs

more extensive analysis before patterns become visible to the researcher. However, our annotation schema is not yet based on quantifiable measures. A next step would thus involve setting thresholds to generate an algorithm to automatically extract such phases from the raw data. Läubli and Germann (see Chap. 8, this volume) are the first to attempt to use the above described annotation. In future work, the annotation schema should be applied to all post-editing and translation data in the database so that the behaviour can be compared between different participants, languages, tools and/or text types.

14.3 Translators' Behaviour in Different Text-Types/Domains

This section presents two studies in which translation process data were collected for general language texts and domain-specific texts. An analysis of the data will compare production times, key-logging, Text Production and Text Elimination, and eye-tracking data, Total reading time (*Trt*) on Source Text and on Target Text, from the two studies. A short summary of the results will be presented at the end of each subsection and Sect. 14.4 will discuss the results of all three analysed components. We address the following questions: Is post-editing time-saving? Do the translation and post-editing processes differ for the different text types? Is the machine translation more useful to the post-editor in the domain-specific texts?

14.3.1 The Translog Study SG12

In the first study, translators worked with Translog-II[2] and translated/post-edited journalistic and sociology-related texts of different complexity levels that were between 100 and 150 words long from English into German. Experiments were conducted on behalf of the Copenhagen Business School for the CRITT TPR database at the University of Mainz, Faculty of Translation Studies, Linguistics and Cultural Studies in Germersheim in 2012. 24 participants took part in the study, 12 of them professional translators (university degree and professional work experience) and 12 semi-professionals (students of the university with no or only little professional experience). Only one third of the participants had prior post-editing experience.

The participants were asked to translate two texts from scratch (T), bilingually post-edit (P) two machine translated texts—the English source text was available—and monolingually post-edit (E) two machine translated texts without the source text. Before and after the tasks, the participants had to complete questionnaires, which dealt with general information about the participant, his/her attitude towards machine translation (in general and in regard to the machine translation output for the tasks), and self-estimation of their task performance.

Gaze data was recorded with the Tobii TX300 eye-tracker, which also recorded the sessions, keystrokes, mouse activity and gaze data for evaluation in Tobii Studio. There were no time restrictions and the participants could use the Internet freely as a research tool. Printed aids were not provided. Finally, the machine translation for the post-editing task was produced by the free online MT system Google Translate.[3]

14.3.2 The Study OCT13

In the second study OCT13, 12 participants were asked to translate and post-edit 3 technical texts (extracts from a refrigerator manual) and 9 participants were to translate and post-edit 3 medical texts (excerpts from patient information leaflets). The volume of the source texts was about 150 words and they were post-edited in Translog-II (Carl 2012).

Two of the medical texts are snippets of texts that were used for CASMACAT-related studies in the database, e.g. study JN13 (English-German) or CEMPT13 (English-Portuguese). All participants were translation students. To make the studies more comparable in number and professionalism of participants, only the student group from SG12 will be taken into consideration in our analysis below.

[2] https://sites.google.com/site/centretranslationinnovation/translog-ii. Last accessed 28th February 2015.

[3] https://translate.google.com/ last accessed 28th February 2015.

In the SG12 study, each of the participants had to translate two texts. Therefore, every text was translated and post-edited eight times by student translators. In the set-up of the study OCT13, each participant had to translate one text from scratch, so that each technical text was translated four times and each medical text three times. Two texts had to be post-edited by each participant—one according to the guidelines of full post-editing, one according to the guidelines of light post-editing.

The guidelines of the light post-editing task are similar to the guidelines of the bilingual post-editing task in study SG12. Therefore, the processes and products of the light post-editing task in OCT13 are comparable to the post-edited texts from study SG12. In all, we have a corpus of 24 translated and post-edited texts from study SG12, and 21 translated and post-edited texts from OCT13.

14.3.3 Production Time

First, the production time of the studies will be analysed, because it has been claimed that one of the major benefits of post-editing is that the task is time-saving. Therefore, we will compare the two studies and the two tasks. Although OCT13 includes two different domains, the analysis will be general-language texts vs. domain-specific texts as the differences between the domains in study OCT13 are not significant.[4] The data are similar for medical texts ($M = 1275.4$, $SD = 538.4$) and for technical texts ($M = 1217.3$, $SD = 462.7$) in the translation task ($W = 56$, $p = 0.6556$). The data is not significant for the post-editing task either (technical texts: $M = 747.1$, $SD = 299.1$; medical texts: $M = 685.7$, $SD = 123.8$) ($t(15.5) = -0.64$, $p = 0.5304$). However, the descriptive measures already indicate that post-editing took much less time than translation.

As visualised in Fig. 14.6, most participants needed more time for the translation from scratch than for post-editing, especially in the second study. Studies SG12 and OCT13 are separated by the red line and participants are labelled according to the study, e.g. the first participant is labelled P1_01 for study SG12 and P2_01 for study OCT13. The figures reflect total text production time. In study SG12, three participants needed more time to post-edit the texts than to translate the texts. However, as the texts in study SG12 are quite different in length, it is useful to compare production time on a word level as done in Carl et al. (2014)—according to their analysis, only one participant was slower at post-editing than at translating. In the second study, only two participants needed more time for post-editing than for translation.[5] As mentioned above, the texts of study OCT13 are of about the same length (150 words).

[4]We conducted different undirected tests for significance. When the data was distributed normally, a *t*-test was conducted; when the data was not distributed normally, a Mann-Whitney-*U*-Test was conducted.

[5]Due to technical problems the data of P10 for translation are not considered.

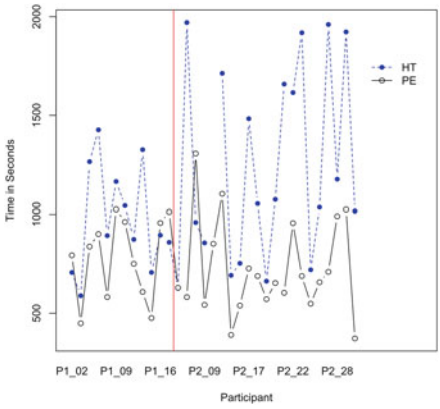


Fig. 14.6 Production times of both studies separated by task. Study SG12 and OCT13 are separated by a *red line*

The production times are much higher for the domain-specific texts ($M = 1243.4$, $SD = 485.4$) than for general-language texts ($M = 991.1$, $SD = 311.9$) in the translation task. The contrary applies to the post-editing tasks, though the difference is not as high (domain-specific $M = 720.8$, $SD = 237.3$; general-language $M = 779.6$, $SD = 227.5$). However, the differences between the domains are not statistically significant: $W = 295$, $p = 0.117$ for translations from scratch and $t(41.7) = -0.84$, $p = 0.4027$ for post-editing.

Without separating the production times by domain, participants need significantly longer for translation than for post-editing: $W = 1498.5$, $p < 0.0001$. The same applies, when the studies are separated. For study SG12, the difference is significant as well ($t(40.2) = 2.65$, $p = 0.012$), but not as strong as for study OCT13 ($W = 357$, $p < 0.0001$).

Conclusively, translation from scratch takes up more time than post-editing, irrespective of the domain the participants worked in. The post-editing task seems to be more time-saving for technical and medical texts than for general-language texts, however, the differences are not significant.

14.3.4 Key-Logging Data

This section focusses on key-logging data from the studies. The parameter considered for analysis (Text Production—TP—and Text Elimination—TE) are taken from the statistics in Translog-II. One problematic issue for these parameters is that Translog-II does not count characters, but keyboard/mouse events instead. This means, e.g. when eight characters are marked and erased, the software only counts it as one text elimination event and not as eight. For study SG12, the CRITT TPR database provides a parameter (e.g. *Mins* and *Mdel* report on manually inserted and deleted tokens) that would be more accurate for comparison in this case. However,

Table 14.4 Mean values and standard deviation of text production (TP) and text elimination (TE) according to study (SG12 and OCT13) and task, translation (T) and post-editing (P)

Study	Task	Mean TP	Standard deviation TP	Mean TE	Standard deviation TE
SG12	T	1095.2	194.1	100.4	56.1
SG12	P	306.3	146.2	128.5	82.5
OCT13	T	1297.4	158.0	181.9	89.1
OCT13	P	199.9	72.6	111.9	65.4

as these parameters are only available for study SG12 and not for study OCT13, we had to use the Translog-II output.

The descriptive statistics in Table 14.4 reveal that less text was produced and erased in study SG12 than in study OCT13 for the translation task. For the post-editing task, it is the other way around: Participants produced and erase more text in study SG12 than in study OCT13.

In translation, the difference between study SG12 and OCT13 for TP ($t(41.5) = 3.80$, $p = 0.0005$) and for TE ($W = 383.5$, $p = 0.0009$) is significant. For post-editing, the differences between the studies is only significant for TP ($W = 131$, $p = 0.0062$), but not for TE ($W = 230$, $p = 0.6246$).

To summarise the results, the differences in TP and TE in the studies for both tasks suggest that the MT output was more useful for domain-specific texts than for general-language texts, although the difference for TE is not significant. Another interpretation would be that the participants in the seconds study used the MT output more efficiently.

14.3.5 Eye-Tracking Data

In the following, total fixation duration on the source text and on the target text will be considered in order to analyse eye-tracking data and cognitive effort. For study SG12, the parameters were calculated from the database tables, which are in study SG12. The parameters are *TrtS* and *TrtT*: "[. . .], *TrtS* and *TrtT* represent the total reading time, i.e. the sum of all fixation durations on the source and target text respectively" (see Chap. 2, this volume). These labels will be used in the following for both studies.

In OCT13, Areas of Interest (AOIs) were defined in Tobii Studio—one in the source text, one in the target text—and the total reading time was automatically calculated for each AOI. The parts of the sessions in which the participants used the Internet for research were excluded to reduce noise in the data (see Chap. 7).

It is hypothesised that gaze in the translation task is (almost) equally divided between ST and TT while the focus is more on the TT during post-editing. Further, the effect is expected to be stronger for the domain-specific texts. First, we will evaluate some descriptive statistic data as shown in Table 14.5.

Table 14.5 Mean values and standard deviation of total reading time—or total reading time—on source (*TrtS*) and target text (*TrtT*) in seconds according to study (SG12 and OCT13) and task, translation (T) and post-editing (P)

Study	Task	Mean TrtS	Standard deviation *TrtS*	Mean *TrtT*	Standard deviation *TrtT*
SG12	T	360.29	116.62	507.67	229.77
SG12	P	206.97	81.31	474.15	136.64
OCT13	T	241.60	100.05	368.46	194.17
OCT13	P	149.69	44.34	377.88	144.70
Combined	T	304.90	123.47	442.70	222.84
Combined	P	180.24	72.00	499.22	147.09

In both studies and tasks, the *Trt* of the target text was higher than of the source text. Additionally, the *TrtT* is similar for both tasks in each study. Although the *TrtS* is always lower than of TT, the gap is bigger in post-editing than in translation. Finally, *Trt* is shorter for both ST and TT and for both tasks in study OCT13.

Initially, both studies were analysed separately. In study SG12 the difference between translation from scratch and post-editing is significant in terms of gaze on the ST ($t(41.1) = 5.28$, $p < 0.0001$), while there is no significant difference for gaze on the TT ($W = 301$, $p = 0.7984$). The same applies for study OCT13: Again, the difference between translation from scratch and post-editing is significant for gaze on the ST ($W = 351$, $p = 0.0008$), while there is no significant difference for gaze on TT ($W = 210$, $p = 0.8034$). Conclusively, there is no significant difference between *TrtT* for both tasks, but the difference is significant for *TrtS*, and according to the descriptive values gaze on ST is significantly higher for translation.

Combining the two studies and thereby increasing the number of texts ($n = 45$ per task) leads to clearer results: The difference between the tasks in *TrtS* is very significant ($W = 1636$, $p < 0.0001$), while there is no significant difference between the two tasks when considering *TrtT* ($W = 1011$, $p = 0.9936$).

Next, total fixations duration was compared between the tasks in different domains considering the two tasks. When looking at the translation from scratch data, the *TrtS* of the general-language texts was significantly longer than in the domain-specific texts ($t(43.0) = -3.67$, $p = 0.0007$). It has to be kept in mind that the texts were about the same length. The same applies to *TrtT*, but the effect is not as strong ($W = 151$, $p = 0.0212$). When looking at post-editing, the difference between gaze on the general-language ST and the domain-specific STs is again significantly higher for the general-language texts ($W = 142$, $p = 0.0117$). Similar results can be found for *TrtT* in post-editing ($W = 146$, $p = 0.0154$).

All in all, gaze on the ST decreases in post-editing, while it stays about the same for the TT in both tasks. Further, less gaze was spent on both texts in the domain-specific texts. It has to be kept in mind that the figures for total fixation duration on the source text and the target text were taken from different sources (tables in CRITT TPR database vs. Tobii Studio). This should not result in any differences, but it cannot be completely ruled out.

14.4 Conclusion

The present study aimed at providing a more in-depth investigation into the differences between translation and post-editing by analysing how texts from different domains are processed. The results showed that participants need more time for translation than for post-editing. However, no significant difference was found between the different domains but the mean values indicate a tendency toward translation taking longer for domain-specific texts, with translators working faster in post-editing for those domains. A similar result was observed in the key-logging data. While more key-logging activity was recorded in the domain-specific texts for the translation task, fewer keystrokes were necessary to post-edit these texts. The gaze behaviour on the TT is about the same for both tasks, while gaze on ST is significantly lower in the post-editing task than in translation. This applies for both studies. Conclusively, the gaze behaviour changes between the two tasks and the TT receives more visual attention in the post-editing task compared to the translation task. While the ST is the main information source in the translation task, and an entire translation has to be produced, in the post-editing task, the most important part is the MT output and the TT. In post-editing, the ST is only used for reference. Further, less gaze was spent on the domain-specific texts than on the general-language texts, in regard to both ST and TT.

The annotation schema presented in Sect. 14.2 of this chapter could reveal differences between post-editing and translation of domain-specific texts which we did not discover with the analysis presented in this paper. Our data failed to reach significance level but still showed a tendency towards a difference between different domains. It could therefore be worth analysing the recordings for the phases of our annotation schema. We believe that especially the phases Rl and Rs could give more insights into the differences between different domains. The revision behaviour was not covered by our measurements of key-strokes and eye-tracking data. As a future step, it might thus be worth annotating the data of study SG12 and study OCT13 and analysing the different phases.

We also believe that, although the TPR-DB already includes domain-specific (medical) texts for the CASMACAT project, it would be sensible to expand with additional domains. The analysis of the SG12 and OCT13 studies showed differences in translators' behaviour in non-domain-specific vs. domain-specific texts. These differences can be expected to be present for other domains as well. Further, domain-specific texts are far more relevant in professional translation than the translation of newspaper articles or similar general-language texts. An online survey (Hommerich and Reiß 2011) conducted for the BDÜ[6]—one of the leading German professional associations for interpreters and translators—reported that 49 % of the members that participated in the study (in total 1570) specialised in the field "Industry and Technology (general)", 45 % in "Law and Administration",

[6]Bundesverband der Dolmetscher und Übersetzer e.V.

41 % in "Economics, Trade, and Finances", 25 % in "Medicine and Pharmacy", and 23 % in "Information Technology". Only few specialised in fields that might require the use of general language like "Culture and Education" (13 %), "Sports, Recreation, and Tourism" (10 %), or "Media and Art" (9 %)—though most of these fields require domain-specific language and terminology as well.

14.5 Future Work

As shown in this paper, different techniques can be used to classify translation process data (e.g. the proposed annotation schema vs. automatically compiled CUs). It would be desirable to expand the visual annotation from Sect. 14.2 so that different types of translators and post-editors can be recognised in all datasets and comparisons can be made between languages, different tools, different text types etc. Further effort should be invested into automatising the annotation process.

Additionally, the TPR-DB should be further extended with texts from different domains, as domain-specific texts are more common in translators' practice than general language texts. Preferably, the texts would be the same or similar for the different languages to enable multilingual comparisons. Two of the medical texts from study OCT13, for example, are excerpts from medical texts that were used in CASMACAT studies as well. With similar experiment set-ups, data analyses could be conducted for various languages, tools, and text types which are all influential factors for translation studies.

Acknowledgement We would like to thank David Imgrund who helped conduct the experiments in study II and Anke Tardel who helped prepare the data for analysis.

References

Carl, M. (2012). Translog-II: A program for recording user activity data for empirical translation process research. In *Proceedings of the eighth international conference on language resources and evaluation*. Istanbul, Turkey.

Carl, M., Gutermuth, S., & Hansen-Schirra, S. (2014). Post-editing machine translation–a usability test for professional translation settings. In *Psycholinguistic and cognitive inquiries in translation and interpretation studies.* Newcastle Upon Tyne: Cambridge Scholars Publishing.

Čulo, O., Gutermuth, S., Hansen-Schirra, S., & Nitzke, J. (2014). The influence of post-editing on translation strategies. In *Post-editing of machine translation: Processes and applications*. Newcastle Upon Tyne: Cambridge Scholars Publishing.

De Palma, D. (2009). The business case for machine translation. *Common Sense Advisory*. Accessed March 30, 2015. http://www.mt-archive.info/MTS-2009-DePalma-ppt.pdf

Dragsted, B., & Carl, M. (2013). Towards a classification of translation styles based on eye-tracking and key-logging data. *Journal of Writing Research, 5*(1), 133–58.

Hommerich, C., & Reiß, N. (2011). *Ergebnisse Der BDÜ-Mitgliederbefragung*.

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine Translation, 25*, 197–215.

Winther Balling, L., & Carl, M. (2014). Production time across languages and tasks: A large-scale analysis using the CRITT translation process database. In *Psycholinguistic and cognitive inquiries in translation and interpretation studies.* Newcastle Upon Tyne: Cambridge Scholars Publishing.

Index

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4