DE GRUYTER
MOUTON

Gert De Sutter, Marie-Aude Lefer, Isabelle Delaere (Eds.)

EMPIRICAL TRANSLATION STUDIES

NEW METHODOLOGICAL AND THEORETICAL TRADITIONS

TRENDS IN LINGUISTICS

DE
G

Gert De Sutter, Marie-Aude Lefer, Isabelle Delaere (Eds.)

Empirical Translation Studies

Trends in Linguistics
Studies and Monographs

Volume 300

Empirical Translation Studies

New Methodological and Theoretical Traditions

Edited by
Gert De Sutter
Marie-Aude Lefer
Isabelle Delaere

DE GRUYTER
MOUTON

ISBN 978-3-11-045684-4
e-ISBN (PDF) 978-3-11-045958-6
e-ISBN (EPUB) 978-3-11-045729-2
ISSN 1861-4302

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Typesetting: RoyalStandard, Hong Kong
Printing and binding: CPI books GmbH, Leck
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

Table of contents

Gert De Sutter, Marie-Aude Lefer and Isabelle Delaere

Introduction

In corpus-based translation studies (CBTS), many scholars have conducted research based on the hypothesis that translated texts have certain linguistic characteristics in common which do not, or to a lesser extent, occur in original, non-translated texts. Baker's (1993) seminal paper described these characteristics as "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (Baker 1993: 243). Research of this kind has resulted in observations of, for example, how translations conform to the typical characteristics of the target language (normalization) (Bernardini and Ferraresi 2011; Scott 1998), how translated texts are linguistically more homogeneous than non-translated texts (levelling out) (Olohan 2004), how translated texts are more explicit than non-translated texts (explicitation) (Olohan and Baker 2000; Øverås 1998) or how translated texts exhibit fewer unique items (under-representation) (Tirkkonen-Condit 2004). In recent years, however, it has been shown that these detected characteristics are not only attributable to the difference between translated and non-translated texts, but co-vary with other (language-external) factors as well, such as text type, source language and the translator's educational background (see e.g. Bernardini and Ferraresi 2011; De Sutter, Delaere, and Plevoets 2012; Kruger and van Rooy 2012; Neumann 2011). As a consequence, linguistic behaviour in translations versus non-translations has to be considered a multifactorial phenomenon rather than a monofactorial one. Multifactorial investigations into the linguistic behaviour of translators compared to non-translators remain rather scarce though, and, as a result, standard multivariate statistical techniques which can be used to visualize, describe, explain and predict patterns of variation within translations and between translations and non-translations do not easily find their way into CBTS (e.g. multidimensional scaling, hierarchical cluster analysis, mixed-effect models). This type of multifactorial investigation, using highly advanced and adequate statistical techniques, is urgently needed in order to find out which factors simultaneously affect linguistic behaviour in translations compared to non-translations. Next to the (language-external) factors mentioned above, other possibly influencing factors include characteristics of the writing process (did the translator use translation software?, did the translator experience any time pressure?, what is the degree of editorial control?, what is the policy of the publishing house?), typological or usage differences between source and target languages, the sociological status of the source and

DOI 10.1515/9783110459586-001

target languages, the style of the translator or original author, the sociological status of different types of translators, etc.

Whereas the identification of the determining factors is a necessary first step to take, the ultimate goal of CBTS is to find out what these factors reveal – on a higher level – about underlying sociological, cognitive,... causes and motivations of linguistic choices in translations vs. non-translations. In recent years, several interesting high-level explanatory mechanisms have been developed, from different perspectives, but they have not been the object of extensive empirical testing yet. From a sociological point of view, Pym (2008) has introduced the idea of translators being risk averse: if they can choose between a safe option (e.g. a variant that is widely accepted as a standard variant), and a risky option (e.g. a variant that is considered restricted to informal conversations), translators will most often opt for the former option, depending on whether they get rewarded or not when taking a risk. From a cognitive point of view, Halverson (2003, 2010) has introduced the so-called gravitational pull hypothesis, which seeks to connect translation behaviour with underlying cognitive properties, such as salience and activation. The gravitational pull hypothesis states that translation characteristics such as under-representation can be explained by the structure of semantic networks and prototypes, i.e. the distance between the activated concepts in the semantic network of the bilingual or multilingual translator.

The present volume aims to push the frontiers of CBTS by presenting original and innovative research which is methodologically rigorous, descriptively adequate and theoretically relevant. Each of the chapters sheds new light on what constrains translational behavior – and to what extent – and how this all fits in an empirical theory of translation. More particularly, this book's aim is twofold: (i) to bring together advanced quantitative (multifactorial) studies of translated texts (compared to non-translated texts on the one hand and/or source texts on the other hand), building on large-scale, well-structured parallel or comparable corpora, which provide additional evidence for the effect of (language-external) factors on translation behavior, resulting in more fine-grained insights into translational tendencies, and which elaborate on explanatory devices uncovered in previous studies; (ii) to investigate to what extent other, complementary methods from related research fields or new data sources can improve the descriptive and explanatory accuracy of corpus-based results. By embracing other, complementary methods aiming at descriptive and theoretical progress, the field of Corpus-Based Translation Studies will eventually emerge as Empirical Translation Studies, in which different methods and models are confronted, ultimately leading to a more adequate and fully-fledged empirical theory of translation.

Overview of the chapters in this volume

Sandra Halverson's chapter is exemplary for the type of new-generation research envisaged in the previous paragraph, viz. theory-based, methodologically pluralistic and improving our understanding of the translational act. Starting out from a well-informed cognitive-linguistic model of bilingual language processing, Halverson investigates how translators deal with semasiological salience, using so-called converging empirical evidence (corpus data and elicited data). She distinguishes between three different types of salience, which might cause translations to be linguistically different from non-translations: a magnetism effect occurs when a translator is attracted to a prominent sense in the target language, a gravitational pull effect occurs when a translator is attracted to a prominent sense in the source language, and an effect of association strength occurs when two senses in the source and target language are often used as translational equivalents. In order to test which of these effects occur under which circumstances, Halverson develops a multi-stage and multi-methodological research design. First, an independent sentence generation test and a semasiological contrastive corpus analysis of the English polysemous verb *to get* and two of its Norwegian equivalents *få* and *bli* are conducted in order to establish a semantic network of these verbs, elucidating which senses are more salient and how strong the connection between the translation equivalents is. Then, a corpus analysis of Norwegian fiction and non-fiction translated into English is carried out in order to determine which of the above-mentioned salience effects occur. Her results show a.o. a clear magnetism effect for one of *get*'s most prominent senses, but other hypothesized effects remain unverified. Finally, an online keystroke experiment reveals that salience also affects revision behavior in that highly frequent verbs tend to be replaced more often than low frequent verbs during later stages of the translation process. Although much more research is needed along the lines sketched in this chapter, the research presented here clearly demonstrates how the effect of bilingual cognition can be studied within an empirical translation framework.

Stefan Evert and Stella Neumann present an advanced multivariate methodology for investigating differences and similarities between original and translated German and English. Starting out from no less than 27 lexicogrammatical features shared by both languages (frequency of finite verbs, passives, prepositions, etc.), they apply a series of multivariate techniques, such as principal component analysis, linear discriminant analysis and support vector machines, to discern visual patterns in the data. The results convincingly show that English and German originals have a clearly unique profile in terms of the lexicogrammatical bundles they display, and that translations shift to some extent towards the source language, which is interpreted as a shining-through effect. This effect, however, is more prominent in translated German (from English) than in translated English (from German). The authors connect this finding tentatively with Toury's hypothesis that less-prestigious languages are more tolerant towards interference (or shining through) than vice versa. In sum, this chapter does not only stand out because of its solid empirical foundations (27 features) and the use of a series of multivariate techniques, it is also remarkable because of the clear presentation of the methodology and the reasoning behind it (thereby enabling replication studies) while at the same time revealing clear patterns, thus contributing to a better understanding of translational behavior.

Isabelle Delaere and Gert De Sutter investigate three fundamental factors that can impact on the linguistic features of translated text, namely *source language*, *register* and *editorial intervention*. Relying on the Dutch Parallel Corpus, the authors apply two multivariate statistics (profile-based correspondence analysis and logistic regression analysis) to measure the exact effect of the three factors investigated on the variability of English loanword use in translated and non-translated Belgian Dutch. Their study, which draws on both comparable and parallel data, shows that *source language*, *register* and *editorial intervention* all influence the use of loanwords (vs. endogeneous alternatives) in translated Belgian Dutch. The findings are interpreted in relation to the normalization behavior of both translators and writers of original texts. Isabelle Delaere and Gert De Sutter's study compellingly illustrates the need to simultaneously consider a wide range of factors that can influence the linguistic make-up of translated language. As shown by their study, this can be done by relying on a combination of advanced multivariate statistics and careful qualitative analyses, which makes it possible to further our understanding of the cognitive and social mechanisms that shape translation.

Next, **Haidee Kruger** examines the under-researched effect of *editorial intervention* on the linguistic traits of texts. To do so, she relies on data extracted from a monolingual English parallel corpus of originally produced edited texts and their unedited counterparts, representing 4 registers (academic, instructional, popular writing and reportage). Looking at 8 features traditionally used as linguistic operationalizations of increased explicitness, simplification and conventionalization in CBTS (such as cohesive markers, sentence length and trigrams), she convincingly shows that revisers/editors make texts *more explicit*, *syntactically simpler* and *more conventional*, three features which, to date, have been attributed to the translation process itself. Haidee Kruger's study has far-ranging implications for CBTS and – more generally – for studies of language mediation and constrained communication (Lanstyák and Heltai 2012), as it demonstrates that features attributed to translation may very well, in fact, be features of editing/revision, or more general features typical of mediated and constrained language (some of these traits, for instance, have also been found to characterize New Englishes). This can only encourage translation scholars to take editorial intervention into account in their own work and to start collecting new types of corpora to tease apart features of translated language and edited language.

Adriano Ferraresi and Maja Miličević's chapter also addresses issues related to language mediation, as it adopts an intermodal approach, i.e. an approach where two *translation modes* (written translation and simultaneous interpreting) are compared, with the aim of identifying the typical features of translated language and interpreted language. Together with Silvia Bernardini, the authors have built the comparable and parallel *European Parliament Translation and Interpreting Corpus* (EPTIC), which contains four components: (1) speeches delivered at the European Parliament and (2) their interpretations, (3) verbatim reports of the proceedings (which are edited versions of the original speeches) and (4) their translations. In this study, the authors rely on four EPTIC subcorpora: interpreted Italian, translated Italian (both with English as source language), original spoken Italian and original written Italian. The study focuses on phraseology, which has been extensively studied in CBTS so far, mainly in relation to interference and normalization/conventionalization. More specifically, the study is devoted to infrequent, highly frequent and strongly associated collocations made up of a noun and a modifier. The results suggest that translations are more phraseologically conventional than interpretations, especially as regards strongly associated expressions, which require more time for processing. This trend, the authors argue, may be related to the cognitive and task-related constraints characterizing translation and interpreting. It clearly emerges from this chapter that CBTS can (and will) benefit from a broader research focus, where a.o. different translation modes are systematically compared (not only comparing written translation with simultaneous interpreting, but also considering sight translation, consecutive interpreting, voice-over, subtitling, dubbing, etc., provided comparable corpora can be compiled).

Oliver Čulo, Silvia Hansen-Schirra and Jean Nitzke focus on an underresearched, technology-related factor in CBTS, viz. the effect of computer-aided translation. More particularly, the authors investigate terminological variation across three types of translations: human translations, machine translations and post-edited translations. They contrast texts translated from English into German from two specific genres, which have been relatively overlooked in previous research, viz. manuals and patient information leaflets. To do so, they rely on the perplexity coefficient, a technique borrowed from the domain of Machine Translation, which, to date, has not been used in CBTS. Although the results suggest that post-edited translations are influenced by the initial machine translation output, further research is needed to determine the cause(s) of this trend. The authors put forward a number of hypotheses that require further investigation, such as the idea that post-editors might tend to focus on the micro-level rather than the overall text, thereby paying less attention to terminological consistency.

Along the same lines, **Ekaterina Lapshinova-Koltunski** is the first to shed empirical-quantitative light on the interplay between translation method and text register. In this study, she compares a number of linguistic features in 5 translation varieties, such as professional human translation and rule-based machine translation, and in seven written registers (including, for example, manuals, tourism leaflets and fiction). The lexico-grammatical patterns under investigation originate from the Hallidayan framework of field, tenor and mode and are linked to a number of well-known translation features such as explicitation, simplification and shining through. The author applies an unsupervised technique, i.e. hierarchical cluster analysis, to investigate (i) variation across translation methods, (ii) variation across registers, and (iii) the interplay between translation method and register. The results reveal that both dimensions are present in the clusters. Interestingly, an additional dimension emerges from the analysis, i.e. translation expertise, which certainly requires further research in the field.

The study presented by **Bert Cappelle and Rudy Loock** re-opens a discussion, which had been relegated to the periphery in Mona Baker's research programme (Baker 1993), viz. the effect of typological differences in source languages on translational products. The authors set out to determine whether there is a difference in usage of phrasal verbs in English translations from Romance languages and from Germanic languages. Their study relies on a monolingual comparable corpus made up of three components: the British National Corpus and two Translational English Corpus components, representing six Romance source languages and six Germanic source languages, respectively. The distribution of phrasal verbs with *up*, *down* and *out* reveals that source language family interference has a significant effect on translation. This leads the authors to dismiss normalization and levelling-out as translation universals. Additionally, a small-scale, more qualitative complementary study on *Le Petit Prince* and its English translation is carried out to determine what elements in the source text lead to phrasal verbs in the target text, revealing that morphologically complex verbs are much more likely to be translated with a phrasal verb than simplex source verbs.

Finally, **Kerstin Kunz, Stefania Degaetano-Ortlieb, Ekaterina Lapshinova-Koltunksi, Katrin Menzel and Erich Steiner** present the findings of a contrastive study of cohesive devices in German and English original texts. Their aim is to uncover contrastive trends that can help translators overcome language-pair specific pitfalls and make strategic choices with regard to the translation and use of cohesive devices. The distribution of cohesive features is analysed both in written and spoken registers in GECCo, a German-English corpus, which allows for deriving suggestions with regard to register-specific translation strategies as well. GECCo is analyzed by means of an exploratory data analysis technique, i.e. correspondence analysis, so as to uncover similarities and differences with regard to cohesive devices between the languages and the registers investigated. In addition, a supervised technique with support vector machines is applied to determine which cohesive features are distinctive and therefore contribute to the differences between the languages and registers under investigation. The results show, among others, that (i) register is an important variable when it comes to lexicogrammatical variation, and (ii) the differences between registers in the German subcorpus are more pronounced than those in the English subcorpus which, in turn, reflects the importance of the language variable.

Acknowledgments

Most of the chapters in this volume were first presented at the *New Ways of Analyzing Translational Behavior in Corpus-Based Translation Studies* workshop, held at the 46th Societas Linguistica Europeae (SLE) meeting in Split, Croatia in 2013. We would like to thank the organizers of the SLE meeting in Split for providing us with the most optimal circumstances to discuss the current state of the art in CBTS and identify some of the future directions that need to be explored. In editing the present volume we were supported by various reviewers. We would like to thank them all for sharing their insightful comments and advice. We are also deeply grateful to the authors for doing such a wonderful job and for not giving up on us after another round of critical remarks and suggestions. We are very well aware that it might have annoyed them (and perhaps even frustrated them at times), but we are convinced that it has greatly contributed to the quality of the present volume. Finally, we wish to thank most heartedly Julie Miess at Mouton for taking care of all practical details concerning this publication.

References

Baker, M. 1993. Corpus linguistics and translation studies. Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.), *Text and technology. In honour of John Sinclair*, 233–250. Amsterdam: John Benjamins.

Bernardini, S. & A. Ferraresi. 2011. Practice, description and theory come together: Normalization or interference in Italian technical translation? *Meta* 56(2). 226–246.

De Sutter, G., I. Delaere & K. Plevoets. 2012. Lexical lectometry in corpus-based translation studies. Combining profile-based correspondence analysis and logistic regression modeling. In M. Oakes & J. Meng (eds.), *Quantitative Methods in Corpus-based Translation Studies. A practical guide to descriptive translation research*, 325–345. Amsterdam/Philadelphia: John Benjamins.

Halverson, S. 2003. The cognitive basis of translation universals. *Target. International Journal of Translation Studies* 15(2). 197–241.

Halverson, S. 2010. Cognitive translation studies: developments in theory and method. In G. Shreve & E. Angelone (eds.), *Translation and Cognition*, 349–369. Amsterdam: John Benjamins.

Kruger, H. & B. van Rooy. 2012. Register and the features of translated language. *Across Languages and Cultures* 13(1). 33–65.

Lanstyák, I. & P. Heltai. 2012. Universals in language contact and translation. *Across Languages and Cultures* 13(1). 99–121.

Neumann, S. 2011. *Contrastive register variation. A quantitative approach to the comparison of English and German*. Berlin: Mouton de Gruyter.

Olohan, M. 2004. *Introducing corpora in translation studies*. Taylor & Francis.

Olohan, M. & M. Baker. 2000. Reporting that in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures* 1(2). 141–158.

Øverås, L. 1998. In search of the third code. An investigation of norms in literary translation. *Meta* 43(4). 557–570.

Pym, A. 2008. On Toury's laws of how translators translate. In A. Pym, M. Shlesinger & D. Simeoni (eds.), *Descriptive Translation Studies and beyond. Investigations in Honor of Gideon Toury*, 311–328. Amsterdam/Philadelphia: John Benjamins.

Scott, N. 1998. *Normalisation and readers' expectations: A study of literary translation with reference to Lispector's A Hora Da Estrela*. Liverpool: University of Liverpool doctoral dissertation.

Tirkkonen-Condit, S. 2004. Keywords and ideology in translated history texts: A corpus-based analysis. In A. K. Mauranen & P. Kujamäki (eds.), *Translation Universals. Do they exist*?, 177–184. Amsterdam/Philadelphia: John Benjamins.

Sandra L. Halverson

1 Gravitational pull in translation. Testing a revised model[1]

Abstract: The gravitational pull hypothesis was introduced as a possible explanation for some general features of translated language (Halverson 2003, 2010a), building on the cognitive semantic concept of semasiological salience in linguistic categories. The basic idea is that highly salient linguistic items (lexis or grammatical constructions) would be more likely to be chosen by translators and thus be overrepresented in translational corpus data. The hypothesis is being developed into a more comprehensive and detailed cognitive linguistic model to incorporate salience phenomena in both source and target language categories as well as the effects of entrenched links between translation pairs. This chapter presents preliminary investigations of central elements of the model using the polysemous verb *get* as a test case. Following a presentation of the revised model, the first stage of the analysis involves using independent empirical studies of *get* (Berez and Gries 2008; Johansson and Oksefjell 1996; Gronemeyer 1999) and of *get* and its Norwegian counterparts (Ebeling 2003) to establish a viable model of a bilingual (Norwegian-English) schematic network for this verb. In order to test this model in an online non-translation task, an elicitation test is run on Norwegian-English bilinguals. This provides further evidence of the salience structure within the target language category in these bilinguals. In the second stage, corpus data from the English-Norwegian parallel corpus and Translog performance data are analyzed to look for evidence of the hypothesized effects. The empirical results are discussed both in terms of the evolving cognitive model and in terms of the contribution of various data types to testing cognitive theoretic notions.

1 Introduction

Twenty-odd years have passed since Baker's (1993) call for corpus linguistic investigation of aggregate patterns in, or features of, translated language ('translation universals'). In that period, a body of research has provided evidence of

[1] Some of the work presented here was financed by a grant from the Meltzer Foundation (project number 11498) and by a joint grant from the Research Council of Norway and the Faculty of Humanities at the University of Bergen. I am indebted to these institutions for this support. I would like to thank Ingeborg Hitland for the coding of the corpus data, and Josep Marco, the editors and an anonymous referee for very helpful comments on an earlier version of this manuscript. All remaining errors are mine alone.

DOI 10.1515/9783110459586-002

some of the proposed features, e.g. simplification, generalization, normalization/conventionalization, interference (see Laviosa 2009, 2011; Chesterman 2011a for overviews). While the empirical results are not conclusive on all accounts, the viability of this research paradigm seems evident. As the field develops, the fundamental starting point of the research paradigm remains constant: the idea that translated language is in some way distinct, that it demonstrates characteristics that make it different from language that is not the result of a translation process. This idea underlies Toury's laws of translational behavior (1995) as well as Baker's universals (see Pym 2008). It is also related to the concept of so-called *translationese* (Gellerstam 1986; Santos 1995) and the notion of a third code (Frawley 1984) or hybrid text (Schäffner and Adab 2001).

As part of the emerging paradigm of Corpus-Based Translation Studies (CBTS), empirical investigations have been accompanied by work querying some aspects of the universals framework (see especially the papers in Anderman and Rogers 2008; Kruger et al. 2011, Mauranen and Kujamäki 2004; Oakes and Ji 2012; Xiao 2010). The question of whether or not the postulated features are unique to translated language is one that has also been the subject of some discussion, and this question too is receiving renewed attention (e.g. Halverson 2003, 2010b, 2015b; Lanstyák and Heltai 2012; Mauranen 2004/5), though the issue remains unresolved empirically.

Quite recently, corpus-based translation studies (CBTS) has emerged as the locus of a new phase of methodological innovation. This innovation is characterized by the use of advanced statistical methods (see e.g. current volume and Oakes and Ji 2012; Delaere et al. 2012; Cappelle and Loock 2013; Delaere and De Sutter 2013; Vandevoorde 2016) and mixed methods research (e.g. Alves et al 2010; Hansen 2003).

While much has happened since the late 1990s and early 2000s, it is fair to say that the most substantial gains have been empirical and methodological ones. The addition of individual studies has meant that more patterns have been studied across additional language pairs. More advanced statistical tools are facilitating more in-depth and robust investigations of the linguistic data. It is also fair to say, however, that these empirical gains have not been accompanied by equally striking developments in theory. Here progress has been more modest and incremental.

At present there are two main approaches taken to the problem of explaining translational patterns. These two are socially and cognitively oriented, respectively. In the former domain, Pym (2005) has suggested that translators are risk averse, and that this may account for some of the patterns demonstrated, e.g. explicitation. This is described as a socially motivated explanation because the propensity for risk aversion is motivated by employment conditions, status, features of the communicative situation, and other social contingencies (2005: 34). With a basis in systemic functional linguistics, Steiner (2001, 2012), Alves et al (2010), Neumann (2014), and Teich (1999, 2003) seek explanations in either register (in what might be considered a proxy for social forces) or in characteristics of a language system or a pair of systems. For instance, in Teich (2003), patterns of normalization and shining-through in translated text are linked to register characteristics and to particulars of the differences between German and English. Register is also a key variable in Kruger and van Rooy (2012) and Delaere and De Sutter (2013).

As regards cognitive explanations, there are two main alternatives: the relevance-theoretical account advanced by Alves and Gonçalves (2003, 2007) and the cognitive grammatical one proposed in Halverson (2003, 2007, 2010b). There are fundamental differences in the two approaches, despite their common cognitive orientation and a shared interest in developing a "psychologically plausible account of communication" (Evans and Green 2006: 463). The most important differences lie in a set of underlying assumptions concerning learning mechanisms (relevance theory adopts a nativist assumption, and cognitive grammar does not), and the relationship between language and general cognitive processes (relevance theory assumes a separate language module, while cognitive grammar does not). The two also differ in that relevance theory requires a distinction between linguistic and non-linguistic knowledge, while cognitive grammar assumes the converse (for comparison of the two approaches, see Evans and Green 2006: 463–465). For the present author, the case for a cognitive grammar approach is more compelling, and it is this framework that is adopted.

The aim of this chapter is to present preliminary tests of the expanded model that is emerging from the original gravitational pull hypothesis. The tests are to be considered preliminary in that the key relationships are not modelled in their full complexity, and in that new data types are being tried in this type of investigation (an elicitation test and keystroke logs). Finally, the statistical tests used are quite simplistic. The objective is to use these rather simple tools to inform more refined statistical modelling at a later stage.

In section 2, the current, expanded version of the gravitational pull hypothesis will be sketched out. In section 3, a test case is outlined, and a network structure is postulated for the selected bilingual verbal category on the basis of non-translational data. Section 4 presents predictions based on the posited structure and tests of these predictions using translational corpus and keystroke data. The results are discussed in section 4.4, and section 5 includes concluding remarks.

2 The gravitational pull hypothesis revised: three sources of translational effects

The gravitational pull hypothesis was originally derived from the theory of Cognitive Grammar and certain assumptions about how this theory could be extrapolated to make it compatible with relevant models of bilingual semantic and syntactic representation (Halverson 2003). Later revisions have also incorporated findings from studies of bilingualism (Brysbaert et al. 2014; Halverson 2010a; Hartsuiker et al. 2004; Hartsuiker 2013; Kroll and Stewart 1994; Pavlenko 2009). Particular emphasis is also placed on current knowledge of bilingual cognition and crosslinguistic influence (Bassetti and Cook 2011; Jarvis and Pavlenko 2008). As a consequence, the current version of the hypothesis is more firmly grounded in the multicompetence perspective (Cook 2003), which emphasizes that linguistic cognition in bilinguals is qualitatively different from that in monolinguals. On this view, "[...] people who know more than one language have different knowledge of both their first and second languages from monolingual speakers of either (Cook 2003) [...]" (Bassetti and Cook 2011: 144). Within this framework, it has been demonstrated that not only do linguistic categories in bilingual speakers differ from those of monolingual speakers, they also change structure throughout these speakers' linguistic life history. This dynamically developing competence is reflected in language performance at all linguistic levels (Bassetti and Cook 2011; Jarvis and Pavlenko 2008; Pavlenko 2009). The methodological consequence of this starting point is that in modelling linguistic categories in bilinguals, it is not sufficient to consider monolingual data alone.

As originally presented, the gravitational pull hypothesis assumed a cognitive grammatical model of semantic structure. In this account, all linguistic items constitute form-meaning pairings (Langacker 1987: 76), and both form and meaning are represented cognitively. Form is taken to be either graphemic or phonological, and meaning (conceptualization), in turn, is accounted for through reference to conceptual content and processes of construal (Langacker 1987: 99–146). Conceptualizations which have been used enough to become entrenched are ordered into networks of related meanings. For example, the network for a lexical item would link all of the senses of that item, and each individual sense would also be linked to synonyms (Langacker 1987: 385; Langacker 2008: 27–54). The features of the posited semantic networks that are of current interest are two: first, the relative prominence of specific elements within a network, and second, connectivity within the network, i.e. the existence and strengths of the links between network elements.

Prominence within the network, or salience, is a complex notion and can be understood to be related to a number of cognitive phenomena. For the purposes of this discussion, the term *salience* will be used to refer to the idea that some patterns of activation within schematic networks will be more prominent than others, due to their higher frequency of use over time. As a result of frequent use, these patterns are thus the "most entrenched and most readily activated" (Langacker 2008: 226), making the linguistic forms (words/constructions) associated with them more likely to be selected.[2] Asymmetries of this type within lexical categories are described in Geeraerts (2009), and for the purposes of the current discussion, one of the salience types he identified is most important: semasiological salience. According to Geeraerts, "Semasiological salience is a relationship among the various semantic possibilities of a given lexical item" (2009: 79). Geeraerts continues, "some of the values expressed by the lexical item may be more central than others, for instance because they occur more frequently within the range of application of the lexical element, [...] (2009: 80). I interpret this type of salience, which is one of many in Geeraerts typology, as much the same type suggested by Langacker (2008). This suggests that one of a word's many senses may be more prominent than the others, giving it greater cognitive weight and increasing its likelihood of being selected. It is important to note that cognitive salience may be impacted by a number of factors, including type of meaning, recency of activation, and various elements of the unfolding discourse representation. In the current context, however, we will be operationalizing salience solely as frequency of use. This is a more restricted use of the term than that presented by Geeraerts in 2009 and later work.

In terms of the bilingual networks activated in translation, semasiological salience (or frequency, for the current purpose) may be evident in both the lexical category activated in the source language and in the lexical category or categories that are being jointly activated for the target language[3]. It is important to remember that this form of salience is a gradable quality that is identified within the networks linked to individual lexical items or constructions and is thus manifested by one of a word's multiple senses in the case of a polysemous lexical item, for example.

[2] Note that the network is a visual metaphor that builds on the conceptual notion of spreading activation. Langacker has pointed out the problems inherent in taking the discreteness of the depicted elements too literally, and has proposed a 'mountain range' metaphor as an alternative (2008:227).

[3] There is a broad agreement in bilingualism research that a bilingual's two languages are both activated during language use, and that some kind of control mechanism is responsible for inhibiting the undesired language. A number of different models have been proposed to account for this process (de Groot 2011: 279–338).

Semasiological salience in the target language is the phenomenon that was originally discussed as *gravitational pull* in Halverson (2003). In the current version of the developing model, I propose that salience in the target language may be more clearly captured by the metaphorical term *magnetism*. This alternative term would be a more appropriate means of expressing the idea that in the cognitive search for a target language item, the translator is more likely to be drawn to a target language item with high salience/frequency. While this type of salience impacts all linguistic choice, the model for translation hypothesizes that this particular frequency effect will be greater in translation than in monolingual language production. Similar frequency effects have been identified in bilinguals, compared to monolinguals in various bilingual tasks (see Diependaele et al. 2013).

Prominence in the source language category, which may also impact translational choices, could then be metaphorically understood as a true form of cognitive gravity, i.e. a cognitive force that makes it difficult for the translator to escape from the cognitive pull of highly salient representational elements in the source language. This would cause what is referred to as interference/transfer or cross-linguistic influence in second language acquisition research.[4]

In addition to the two types of salience discussed above, an additional source of hypothesized translational effects is the nature and strength of links between elements in a bilingual's two languages. Let us call this *connectivity*. A helpful way of describing this is in terms of frequency: if salience patterns emerge due to the type frequency of source and target elements, this third source reflects the impact of high frequency co-occurrence of a translation pair, either in learning or in production tasks over time, or both.[5] Indeed, the links between translation pairs across languages are also strengthened through frequent activation of one member of the pair, given an assumption of joint activation at some representational level.[6] In earlier work (Halverson 2003), it was noted that this feature of the model might be relevant for the unique items hypothesis (Tirkkonen-Condit 2004, 2005), which claims that source language

[4] I have not factored in here the effect of discourse factors in increasing or decreasing salience. Langacker refers to the type of salience under consideration (linked to schemas and prototypes) as 'less transient cognitive salience' (1987: 430), while at the same time recognizing that contextual/discourse factors also impact salience. For the time being, we are attempting to isolate the less transient type to look for possible translational effects.

[5] This reasoning may be linked to the notion of the 'dominant translation', or the most frequently chosen translation for a given word (see Boada et al. 2013).

[6] The issue of the bilingual representational model is not dealt with here, but several of the alternative models share this fundamental assumption. See de Groot (2011: 129–144) for an overview.

(SL) lacunae in semantic networks could cause underrepresentation of target language (TL) items. Thus the more established (entrenched) a link is, the more likely it will be activated and used in translation, and vice versa.

The distinction between three different potential sources of translational effects, two based on prominence and one on the entrenchment of translation pairs ('equivalents') clarifies the account given in Halverson (2003). The original gravitational pull hypothesis is now split into three posited sources of translational effects: source language salience (gravitational pull), target language salience (magnetism), and link strength effects (connectivity). However, the basic thrust of the cognitive model remains the same: that specific characteristics of schematic bilingual networks are hypothesized to have translational effects, more specifically aggregate patterns of over- and underrepresentation in translated language.

3 Step one: developing a partial Norwegian-English bilingual schematic network

In order to test the three related hypotheses outlined above, a semantic network is needed to serve as a case. In selecting the case to be investigated here, two main criteria were adopted:

1. The network should involve a polysemous category in the target language.
2. The sense distinctions and their relationships should be described in existing (preferably corpus-based) studies, including the crosslinguistic relationships in one language pair.

Based on these criteria, the network selected for this investigation is the polysemous target language verb *get*. The bilingual network will include, for the purposes of this study, two Norwegian verbs, *få* ('get') and *bli* ('become'). Though these are not the only two Norwegian verbs that are of relevance to the entire *get* network, they overlap semantically with several of *get*'s senses, and are consequently the starting point for the development of a partial network model.

In the following, two types of non-translational data will be used to look for salience in the semantic network for *get* and its closest Norwegian counterparts, *få* and *bli*. In section 3.1, corpus data for both English and Norwegian will be considered, and in section 3.2, an elicitation task is used to investigate possible network effects in English produced by a group of bilingual Norwegian-English speakers. The problems associated with using corpus data in the investigation of cognitive theoretic notions are well known (see Gilquin 2008; Heylen et al. 2008; Tummers et al. 2005). In this chapter, we adopt the stance taken by Divjak and Arppe:

> Although corpus data do not reflect the characteristics of mental grammars directly, we do consider corpus data a legitimate source of data about mental grammars. Since the results of linguistic cognitive processes, e.g. corpus data, are not independent of, or unrelated to, the linguistic knowledge that is represented in the brain, we may assume with justification that characteristics observable in language usage reflect characteristics of the mental processes and structures yielding usage, even though we do not know the exact form of these mental representations. (2013: 229)

For the purposes of the current analysis, the corpus data is therefore complemented by performance data[7], presented in section 3.2. The two data types are used to postulate salience and connectivity patterns within the bilingual network, which will be tested in translational data in section 4.

3.1 Monolingual English and Norwegian: evidence from corpora

For the purposes of this study, semasiological salience will be operationalized in terms of rank order by frequency among the various senses of the verbs in question. In other words, the senses that occur most frequently are taken to be the most salient in the category. The sense distinctions for *get* have been described in Gronemeyer (1999) and in Johansson and Oksefjell (1996). Gronemeyer describes her analysis as “based on data from the Brown corpus” (1999: 2), and her analysis involves describing the syntactic environments associated with the various senses of the verb. In her classification, there is a very clear association between sense distinctions and syntactic frames, which makes the syntactic frame a useful indicator of the sense. Johansson and Oksefjell also used corpus data, including the Brown, LOB and London-Lund corpora of written and spoken British and American English, and their description was

[7] I have previously argued for a three-way classification of translational data, with the term ‘product’ reserved for observational data such as translated texts (either singular ones or corpora), ‘performance’ used about observational data that is not a translational product as such, and ‘process’ used about data types that can be immediately linked to theoretical notions regarding cognitive processing. Thus we would avoid the confusing use of ‘process’ to refer both to cognitive processing and to the observable behaviors involved in creating a translation, not all of which may be directly linked to cognitive theoretical accounts (yet).

primarily a syntactic one. A number of the same distinctions were later confirmed by Berez and Gries (2008) in a corpus analysis using the behavioral profile methodology[8].

In the analysis of the English data, the figures presented here are from the English original texts in the English Norwegian Parallel Corpus (ENPC) and a comparable subcorpus of the British National Corpus (BNC). The sense categories are those given in Gronemeyer (1999), with three simplifications: a) the collapsing of obligation/permission (which share the syntactic frame *get* + *to* infinitive), b) the grouping together of all of the causative senses and c) the addition of a category for idioms.

The ENPC (http://www.hf.uio.no/ilos/english/services/omc/enpc) consists of English originals, Norwegian translations, Norwegian originals, and English translations (c. 2.6 million words). The translational and non-traditional subcorpora are further subdivided into fiction and non-fiction subcorpora and the fiction subcorpus accounts for roughly 60 percent of the material. The texts in the non-fiction corpus represent primarily a type of popular science genre, though there are also a few legal texts. The non-fiction subcorpus is not further subdivided by genre. It is important to note that genre has not been incorporated as a variable in this investigation, due to both the relatively coarse descriptions of genre categories in the corpora and the relatively small size of the ENPC subcorpora. For the current purpose, an attempt has been made to alleviate the potential genre effect by matching the BNC subcorpus as closely as possible to the ENPC, in order to enable the comparison of translated to non-translated language. This was done by selecting written books and periodicals and written miscellaneous as the text categories for the BNC. Since the genre categories in both corpora are relatively imprecise, they do not really ensure complete comparability. This is only acceptable in this early stage of hypothesis testing and development. It is certainly possible that there are genre differences in the sense distributions, and this variable must be more carefully catered for in later investigations. For now, the attempt to ensure corpus comparability will have to

[8] The sense distinctions utilized by Berez and Gries (2008) were derived from WordNet 2.1, and the data was taken from ICE-GB. The authors classified the senses in their data and coded each for a set of semantic, morphological and syntactic information. They used this information in a cluster analysis to see whether the sense categories identified at the outset were confirmed by the cluster analysis. Their initial categories were the same as the ones identified in Table 1, with the exception of the causative senses, which were classified as either passive or inchoative, the ingressive sense, which was listed as a subsense of movement, and the combination of the obligation and passive senses. Two variants of cluster analysis confirmed the onset, possession, movement and must/passive and causative clusters.

serve as an interim solution, and caution should be exercised in interpreting the results.

Table 1 shows the distribution of the sense categories in the English original material. All instances of *get* in the ENPC were analyzed, and the selection of occurrences in the BNC is a random sample representing 1% of all of the occurrences. All corpus data was coded manually. The sense classifications were enabled by the high degree of isomorphism between semantic and syntactic classifications: in almost all cases the senses are distinguished also by syntactic means. For example, the onset sense involves *get* + NP, while stative possession is expressed by *have* + *got* + NP. Movement requires an adverbial, and permission/ obligation a *to*-infinitive complement. The only ambiguous case involved the distinction between an inchoative and passive reading given a participial complement (e.g. *attached, married*). This was resolved by classifying all cases involving participles classified as adjectives in the *Collins English Dictionary* as inchoatives. If the participle was only listed as a verb form, the instance was classified as a passive. Idioms were identified through a criterion of semantic opaqueness[9]. Movement also incorporated metaphorical movement.

The frequency rankings of the various senses correspond very closely across the two corpora. The three most frequent senses, in descending order, are movement, onset, and inchoative in both corpora. Rank orders 4–6 vary marginally across the two corpora, primarily due to the higher frequency of stative possession (*have got*) in the ENPC originals. The categories of passive and idioms are ranked slightly differently, but represent similar percentages of the respective corpus occurrences (5 v 6 percent passives and 9 v 8 percent idioms in the ENPC and BNC respectively). A Mann-Whitney test[10] demonstrates that the differences in the distributions of senses across the corpora are not significantly different (n = 2471, p = .931). Thus we see that the rank orders of the senses of *get* are largely the same across the two monolingual English corpora, which suggests that we may tentatively posit a higher degree of semasiological salience for the most frequent senses: get_3, get_1, and get_6, in that order[11]. The relative positions of the senses as illustrated here will serve as a basis for the construction of the verbal category in English.

[9] Occurrences of the verb were defined as idioms if they were semantically opaque, even if they exhibited the same syntactic frames as the other senses. Examples include: *get wind of, get on somebody's nerves, get something over with.*

[10] The Mann-Whitney is a non-parametric test that can be used to see whether two samples are independent with regard to one dependent variable (whether the differences between the two groups are statistically significant). In this test, the calculation makes use of ranks within groups, so that the rank order of the sense scores is not significantly different.

[11] Salience is posited for the three most frequent senses, which account for roughly 70 percent of the occurrences in both corpora.

Table 1: Sense categories for get with corpus frequencies in ENPC originals and BNC

Sense	Meaning	Example – from ENPC	N (%)	Rank	n (%)	Rank
			ENPC	ENPC	BNC	BNC
get$_1$	onset of possession	Would you like me to go out and *get* some croissants?	304 (23)	2	324 (28)	2
get$_2$	stative possession	*Have* you *got* any of those?	139 (11)	4	67 (6)	6
get$_3$	Movement	I don't want you to *get* there after dark.	418 (32)	1	340 (30)	1
get$_4$	permission/ obligation	You've *got to take into account* that I'm virtually single-handed here.	36 (3)	7	46 (4)	7
get$_5$	Causation	Despite his tuggings with the wrench he couldn't *get the screw to shift.*	28 (2)	8	26 (2)	8
get$_6$	Inchoative	Sit down and *get warm*	188 (14)	3	164 (14)	3
get$_7$	Passive	. . . and you really do have to be a winklebrain to *get ejected* from there . . .	66 (5)	6	70 (6)	5
get$_8$	Ingressive	I'm going to *get moving*.	22 (2)	9	23 (2)	9
get$_9$	Idioms	Like the Whistler, they *get* their kicks from watching people die.	122 (9)	5	88 (8)	4
Total			1323 (100)		1148 (100)	

The bilingual network in this study is a Norwegian-English one. The relevant Norwegian items are the two verbs *få* ('get') and *bli* ('become'). As we will see in the discussion below, several senses of the former verb overlap with senses of *get*. *Bli*, on the other hand, is a copular verb in Norwegian that expresses change of state. It corresponds to the inchoative sense of *get* (get$_6$). A cross-linguistic corpus-based analysis exists for this network in the dissertation by Ebeling (2003), who also based her analysis on the ENPC. Her analyses take the two Norwegian verbs as a starting point and investigate the translational relationships that pertain between the various senses and their English counterparts. In addition, the Norwegian *få* has recently been described by Askedal (2012), using a smaller corpus of Norwegian fiction and non-fiction material, comparable to the ENPC.[12] Both Ebeling and Askedal classified the Norwegian data on the basis

[12] Askedal's corpus consisted of eight novels and non-fiction writing, representing contemporary Norwegian fiction and non-fiction. It is comparable to the ENPC in text types. The total size of his corpus is not given, but the total number of occurrences of *få* equalled 779, or roughly half the number found in the ENPC. Askedal also has categorized his data on the basis of primarily syntactic properties, but has incorporated semantic information in certain of the categories. His discussion also allows for a recategorization into semantic categories. Any faults in this process are the responsibility of the present author alone.

of syntactic criteria. Even so, they both provide numerous corpus examples and they discuss the semantics of each syntactic category. With this verb also, there is a high degree of isomorphism between the sense distinctions and the set of verbal constructions. This information was used to reverse the classification and reclassify the data semantically using the semantic descriptions given in the two studies. The figures in the table are adapted from the studies by Ebeling (ENPC) and Askedal (JOA), and any errors are those of the present author alone.

Table 2: Corpus frequencies for sense categories for få in ENPC (based on Ebeling 2003: 207) and Askedal (2012). Examples from Ebeling (2003)

Sense	Meaning	example ENPC	N (%) ENPC	rank ENPC	N (%) JOA	Rank JOA
$få_1$	onset of possession	Gutten *fikk* et eple./ The boy got an apple	623 (40)	1	397 (51)	1
$få_2$	movement	Pengene de får inn/ The money they get in . . .	132 (9)	5	48 (6)	4
$få_3$	permission	*Får* jeg spørre deg om en sak?/May I ask you about something?	239 (15)	2	133 (17)	2
$få_4$	causation	Vi *fikk* henne i godt humør./We got her in a good mood.	212 (14)	3	86 (11)	3
$få_5$	reflexive	Flere og flere chokoner *hadde fått seg* ildvåpen/ More and more Chokonen had firearms . . .	18 (1)	8	39 (5)	6
$få_6$	Passive	. . . det ikke gjorde vondt *å få* en tann rotfylt/ it didn't hurt to get a tooth 'rootfilled'	138 (9)	4	29 (4)	7
$få_7$	passive resultative	. . . før vi får kommet oss avgårde/before we get ourselves off[13]	2 (0)	10	47 (6)	5
$få_8$	ingressive	Snart skulle hun *få vite*. . ./ Soon she would find out . . .	68 (4)	7		
$få_9$		Hun *fikk lyst* til. . ./ She wanted to	112 (7)	6		
pro form		Jeg *får* (gå) bort (I have to (go) . . .	3 (0)	9		
Total			1547 (99[14])		779 (100)	

[13] This example is from Askedal (2012: 1315).

[14] Does not equal 100 due to rounding.

As shown in Table 2, the three most frequent senses of *få* (onset, permission, and causation) correspond across the two corpora, while the less frequent senses show slightly different distributions. An important characteristic here is the overall frequency of the first ranked sense (onset), which is 40% of all occurrences in the ENPC Norwegian material and 51% in the JOA corpus. The second most frequent sense (permission) represents only 15 and 17% respectively. A Mann-Whitney test demonstrates that the differences in rank order of the sense distributions in the two corpora are not statistically significant (n = 2326, p = .165). Thus, the relative positions of the senses illustrated here will also inform the construction of the network model.

In order to fully describe the bilingual *get* network, we must also include frequency information for the Norwegian *bli*. Ebeling (2003) analyzed all occurrences of this verb in the ENPC, and once again, her syntactic categories are described in terms of their semantics. This description is used to reclassify the instances into semantic categories, as indicated in Table 3. In this case the only deviation from Ebeling's classification is the merging of her copular and intransitive categories, which constitute the inchoative here:

Table 3: Frequency of sense categories for bli, ENPC (adapted from Ebeling 2003)

Sense	Meaning	Example ENPC	N (%)
bli_1	Inchoative	Hun *ble* redd./She *grew* frightened.	1545 (57)
bli_2	Passive	Barnet *blir lagt* til brystet./The child is put to her breast.	884 (32)
bli_3	aspectual aux	Jeg *blir stående* her./I'll *stay standing* here.	177 (6)
bli_4	multiword verbs	Hvor var det *blitt av* lykken deres?/What had *become of* their happiness?	119 (4)
Total			2725 (99)

The frequencies of the senses of *bli* in the ENPC demonstrate that, according to the corpus data, there is one predominant sense, the inchoative one, which represents 57% of all uses. The verb is also frequently used as a passive auxiliary (32%). It is important to note that in the inchoative sense, the verb can also take an NP as its complement (e.g. [...] *å bli landets lys.../to become the light of the nation*; Ebeling 2003: 85). In this case this sense does not correspond to get_6, as inchoative *get* takes only an adjectival complement. So while the two verb senses correspond semantically, the verbs are not complete translation equivalents due to the non-congruity of the syntactic frames in which they can be used.

The corpus data presented here thus provides a basis for establishing semasiological salience among the senses of *get*, *få* and *bli*. Of the three most salient (frequent) senses of *get*, the first (movement) corresponds to a less frequent sense of *få*. The second (onset of possession) corresponds semantically to the most salient sense of *få*. This sense of *få* accounts for 40 and 51 percent of the occurrences, respectively, in the two corpora investigated. The third-ranked sense (inchoative) corresponds to the most salient sense for *bli*. This information will be incorporated in specific predictions for translation in section 4.1.

One final note on the status of this corpus data for the cognitive model under development: the corpus data analyzed here provides information on the relative frequencies of the various senses of the verbs in aggregate language use in a language community. We assume that there is a relationship between aggregate patterns and individual knowledge, though this is not a straightforward matter. For now, we will take the relative frequencies as a starting point in elaborating a cognitive model, and this model will be considered in light of other data types that provide better access to linguistic cognition.

The corpus does not provide information on the linguistic background of the language users who have produced the texts, and we are assuming, for the English data at least, that the authors are predominantly monolingual. As regards the Norwegian data, monolingualism is not as tenable an assumption, though most of the texts in this corpus were produced prior to the 1990s, when general levels of English proficiency in Norwegian adults were much lower in Norway than they are today. As argued elsewhere (Halverson 2015b), a cognitive model based on the language of monolinguals alone is problematic for the investigation of translational cognition. This is a further reason to investigate other forms of data.

3.2 Effects of semasiological salience – evidence from performance data

As mentioned in the introduction to this section, corpus data gives us only indirect evidence of cognitive linguistic structure. It is thus necessary to look for some other type of evidence of the posited salience of the senses that are most frequent in the corpus data. Moreover, the relationships posited for the English verb senses were derived based on what we assume is language produced by monolingual speakers. Given our assumption of multicompetence (see section 2) in bilinguals, it is also necessary to incorporate some information about potential salience effects in English produced by Norwegian-English bilinguals[15]. For this purpose, a sentence generation test similar to the one used

[15] The subjects in the test are classified as ‘bilinguals’, even though they may not be equally proficient in both languages. They were all native speakers of Norwegian who were highly proficient L2 speakers of English. In accordance with the multicompetence perspective, we assume

by Sandra and Rice (1995) was carried out. In this test, subjects are asked to spontaneously produce 10 sentences using the word *get*. In a paper and pencil version of this test, BA-level Norwegian university students of English, fluent L2 speakers, were given an envelope with 10 index cards in it. They removed the cards, and wrote one sentence on each of the cards, placing the cards back in the envelope as they finished each one. The cards were numbered, so that the order of production of the sentences could be retrieved. There was no time limit. There were 38 subjects, all highly proficient Norwegian-English bilinguals.

The elicitation data provides evidence of salience in the following way: salience implies ease of access, which is operationalized as frequency of selection and early selection. Thus the sentences generated by the bilingual subjects will be analyzed to see how often the various senses of *get* were selected, and in what order (rank among the ten sentences produced by each subject).[16]

Figure 1 illustrates the overall frequency of production of the various senses by this group of subjects on the test:

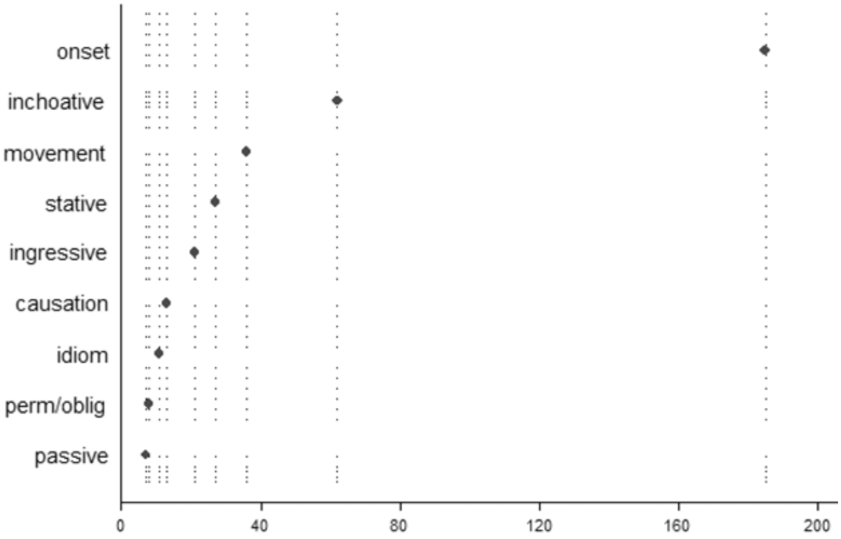


Figure 1: Sentence generation for ‘get’. Sense frequency.
Norwegian subjects. N = 370

that the cognitive impact of Norwegian-English bilingualism will be evidenced in this group. As Bassetti and Cook explain, ‘[. . .] it would be wrong to assume that any cognitive consequences of bilingualism only appear in maximal bilinguals who have acquired and used the language for many years; effects may manifest themselves at a comparatively low level of knowledge and use of the second language after a matter of hours.’ (2011: 144)

[16] See Gilquin (2008) for a similar use of the methodology.

As indicated in Figure 1, there is a very significant predominance of the onset of possession sense in the data. This sense accounts for 50 percent (185/370[17]) of the instances of *get*, while the second most frequent sense, the inchoative one, was produced in only 62 sentences (17 percent). The third most frequent sense, movement, was produced 36 times (10 percent). Examples of the three senses in the data are given in order of frequency (onset of possession, inchoative, and movement):

(1) *He got a lot of presents for his birthday.*
(2) *I usually get tired of running around.*
(3) *It took me thirty minutes to get to the airport.*

The sharp decline from the most frequent senses is clearly visible in the figure. The three most frequent senses are the same here as in the corpus data, though they are ranked quite differently: instead of movement, onset, inchoative, here the order is onset, inchoative, movement.[18]

The frequency with which each sense was selected at each rank in the order of production is demonstrated in Table 4. The order of production is indicated in the left column, and row percentages indicate the percentage per sense per rank. Total percentages do not equal 100 due to rounding.

The results presented in Table 4 further illustrate the salience pattern suggested by Figure 1, in that the most frequent sense (onset of possession) was selected first most often (55 percent of the first choices). If we consider the three most frequently produced senses (onset, inchoative and movement), then the three together account for c. 80 percent of both the first and second selections. In general, the data reveal that the two most frequent senses (onset of possession and inchoative) are selected in that order at every rank except the first. When it comes to first choices, the most frequent was onset of possession, the second most frequent first choice was the movement sense, and inchoative was the third most frequent.

Interestingly, the two senses that were most often produced are the two that most clearly correspond to the predominant senses of the Norwegian verbs *få* (onset of possession) and *bli* (inchoative), while the third most frequent sense

[17] The total N in this data set should be 380 (38 × 10), but several subjects did not write ten sentences. In other words, they stopped writing after they had produced fewer than ten sentences. The total is 370, and in Table 4 some of the rows add up to less than 38, because a subject did not produce a seventh, eighth, ninth or tenth sentence.

[18] As pointed out by Gert De Sutter (personal correspondence), one possible objection to this methodology is that the subjects are not producing isolated verbs; they might also be producing fixed expressions or collocations. Any fixed expressions would have been coded as idioms here, and would be captured as such. There is, however, a possibility that subjects produced frequent collocations, as this was not controlled for. Ideally this would be controlled for.

Table 4: Sentence generation task for get in Norwegian-English bilinguals. Senses produced by rank order. Row percentages

Order of Production	onset (%)	stative (%)	movem. (%)	perm/obl. (%)	caus. (%)	inchoative (%)	passive (%)	ingr. (%)	idiom (%)	Total (%)
1	21 (55)	1 (3)	7 (18)	1 (3)	1 (3)	3 (8)	1 (3)	2 (5)	1 (3)	38 (101)
2	18 (47)	5 (13)	3 (8)	0 (0)	1 (3)	9 (24)	1 (3)	1 (3)	0 (0)	38 (101)
3	19 (50)	3 (8)	2 (5)	1 (3)	3 (8)	4 (11)	0 (0)	3 (8)	3 (8)	38 (101)
4	18 (47)	3 (8)	5 (13)	1 (3)	1 (3)	8 (21)	1 (3)	1 (3)	0 (0)	38 (101)
5	16 (42)	4 (11)	6 (16)	0 (0)	1 (3)	6 (16)	2 (5)	1 (3)	2 (5)	38 (101)
6	17 (45)	3 (8)	3 (8)	1 (3)	2 (5)	10 (26)	0 (0)	2 (5)	0 (0)	38 (100)
7	19 (53)	2 (6)	2 (6)	1 (3)	0 (0)	5 (14)	1 (3)	6 (17)	0 (0)	36 (102)
8	19 (53)	0 (0)	3 (8)	1 (3)	2 (5)	4 (11)	1 (3)	3 (8)	3 (8)	36 (99)
9	18 (50)	3 (8)	4 (11)	1 (3)	1 (3)	6 (17)	0 (0)	1 (3)	1 (3)	36 (98)[19]
10	20 (59)	3 (9)	1 (3)	1 (3)	1 (3)	7 (21)	0 (0)	1 (3)	0 (0)	34 (101)
Total	**185 (50)**	27 (7)	**36 (10)**	8 (2)	13 (4)	**62 (17)**	7 (2)	21 (6)	10 (3)	370 (101)

[19] For one subject, the ninth sentence did not use any of the senses of *get* listed here.

(movement) is the one that is most frequent in the monolingual English data. This might suggest that the salience pattern among the English senses is not only being affected by magnetism, i.e. the frequencies of various senses in the target language, but also by the connectivity between the English verb senses and their frequent Norwegian equivalents. Unfortunately, a task such as this does not allow us to disentangle these two potential sources of impact. One source of further information would be to carry out a comparable elicitation test for the Norwegian verbs, to see what relationship the salience patterns in the subjects' two languages might have. This has not been done at present, though such a test will be considered as one of several alternative psycholinguistic tests to be incorporated at a later date.

If we take the results of this test as an indication of the relative salience of the senses of *get* for this group of bilingual Norwegian-English speakers, then we have corroborated the salience of the same three senses as dominated in the (monolingual) English corpus data, i.e. the onset of possession, movement and the inchoative sense. However, for the bilingual Norwegian-English speakers, it is likely that selection processes are also impacted by their knowledge of Norwegian.

For the purposes of this study, the quantitative information regarding the relative salience of the senses will not be utilized beyond identifying three salient category members and positing connectivity patterns. In future studies, additional psycholinguistic tests will be used to posit a refined network structure, including connectivity patterns. This information will feed into a more advanced statistical model for more robust testing. The present study represents a rather simplistic first attempt at utilizing different data sets to test the outlined cognitive linguistic hypotheses. Further elaboration of statistical testing is planned for a later stage.

4 Testing the hypotheses in two types of translational data

The previous section presented monolingual data for English and Norwegian as well as data from English language production by Norwegian L2 speakers. The data serves as the basis for positing a schematic network characterized by semasiological salience among the senses. In this section, the model will be used to hypothesize about translational effects of this posited semantic structure. As mentioned in section 3, corpus data provides only indirect evidence of the type of cognitive processes that are in focus here. For that reason, corpus data is brought together with a form of online performance data in this section, more specifically revision data from keystroke logs. This data provides evidence of the stages involved in translational decision-making. In the following, the predictions for this data are outlined in section 4.1, section 4.2 reports on corpus analyses and in section 4.3 keystroke log data is investigated.

4.1 Predictions for translation (and bilingual production)

Figure 2 given below is an idealized depiction of what a schematic network for the English verb *get* and the related Norwegian verbs *få* and *bli* could look like, based on the corpus data and the elicitation test outlined in section 3. The figure depicts the various senses as boxes, and the more thickly outlined boxes represent those senses that are taken to be more semasiologically salient in English and Norwegian. In the interest of clarity, some elements of the network are not depicted. For instance, not all senses of *get* are linked, though they should be understood to be so linked. Moreover, only the senses of *få* and *bli* that are relevant for the analysis of *get* are given, not all senses of the two Norwegian verbs. Finally, the thicker lines linking some of the English and Norwegian senses are taken to represent relatively greater connectivity, as suggested by the tendency of bilingual Norwegian-English speakers to produce the senses of *get* to which these Norwegian verbs are linked.

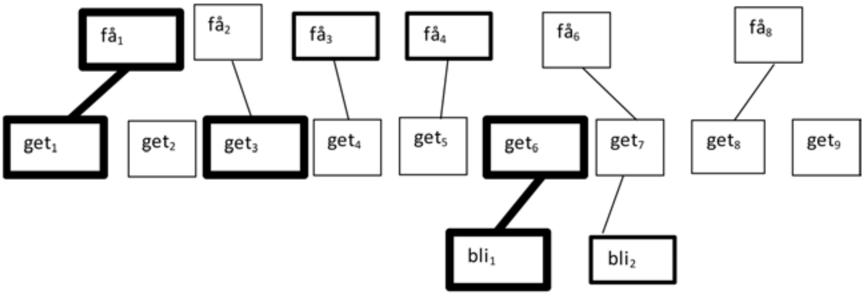


Figure 2: Semantic network for *get*/*bli*/*få*

The model outlined here is thus derived from independent corpus-based analyses of *get*, *bli* and *få*, as well as an elicitation test on Norwegian-English bilinguals. As indicated in the figure, get$_1$ (onset of possession), get$_3$ (movement) and get$_6$

(inchoative) are the most salient senses in English (in the corpus and the elicitation data) and are expected to serve as magnets in translation. The Norwegian *få*$_1$ (onset of possession) and bli$_1$ (inchoative) are by far the most salient senses of these two verbs respectively, and are expected to serve as centers of SL (source language) -derived gravitational pull. Similarly, we interpret the predominance of get$_1$ and get$_6$ in the sentence generation test as grounds for hypothesizing stronger connectivity between *få*$_1$ – get$_1$, and bli$_1$ and get$_6$ (this is also an expected consequence of high frequency for both the Norwegian and English verbs). The sentence generation test is only one means of testing for connectivity: in future work, targeted psycholinguistic tests of these relationships must be developed.

At this stage, it is not possible to predict how salience patterns and connectivity interact as the sources of translational effects. For this initial investigation, we shall assume that each of the potential sources of effect can work independently or jointly, though a future objective is to test for cumulative or interaction effects in a more advanced statistical model. For this study, we will look for translational effects and attempt to identify the sources of these effects, though these may in some cases represent more than one of the three types under consideration. Given the structure of this bilingual network, it is not possible to isolate the magnetism, gravitational pull and connectivity for the onset and inchoative senses. This must be done through the testing of different networks with different patterns of salience and connectivity.

The null hypothesis is that the relative distributions of the senses within a lexical category should be roughly the same in translated and non-translated language. For the configuration of this particular network, the predictions for translated text (corpus analysis) are as follows:

- get$_3$ (movement) will be overrepresented (magnetism alone)
- get$_4$ (permission/obligation) will be overrepresented (gravitational pull alone)
- get$_1$ (onset) and get$_6$ (inchoative) will be overrepresented (magnetism, gravitational pull and connectivity)

In the hypotheses outlined above, we are testing for effects of semasiological salience in translation. In other words, we are assuming that within a lexical category, some senses will be more salient than others, due to frequency of use. Thus, the predictions for magnetism and gravitational pull are based on high relative frequencies for some of the *get* and *få* senses.

In the context of translation, of course, it would be logical to look for the effect of onomasiological salience as well (Geeraerts, personal correspondence). This would involve looking at the range of translations of a given ST item and looking for salience effects there. Section 4.2.2 presents an analysis that builds on a related line of reasoning.

4.2 The translational corpus analysis

4.2.1 Relative frequency of get senses in English translated and non-translated text

The first analysis of the corpus data involves a test for overrepresentation of some of the senses of *get* in translated vs non-translated language. The corpora compared here are the translational and non-translational subcorpora of the English-Norwegian Parallel Corpus (ENPC). Table 5 illustrates the frequencies of the various senses across the ENPC subcorpora:

Table 5: Distribution of senses of get across ENPC subcorpora

	ENPC translated (%)	**ENPC originals (%)**
1. get$_1$ (onset of possession)	363 (23)	304 (23)
2. get$_2$ (stative possession)	60 (4)	139 (11)
3. get$_3$ (movement)	581 (37)	418 (32)
4. get$_4$ (permission/obligation)	34 (2)	36 (3)
5. get$_5$ (causation)	27 (2)	28 (2)
6. get$_6$ (inchoative)	290 (19)	188 (14)
7. get$_7$ (passive)	65 (4)	66 (5)
8. get$_8$ (ingressive)	31 (2)	22 (2)
9. get$_9$ (idiom)	114 (7)	122 (9)
Total	1565 (100)	1323 (101)

An overall frequency comparison indicates that *get* is overused in English translated from Norwegian (log likelihood critical value = 11.70, $p < .001$). More specific differences in the distribution have been investigated using two alternative tests, the Mann-Whitney rank sums and the Chi-Square. The results are not statistically significant using the former analysis ($p = .213$), which means that the overall distributions are not significantly different. A Chi-square test was carried out to determine whether the differences in frequency for individual senses were statistically significant. Three sense categories showed significant differences: stative possession ($p < .0001$), movement ($p = .002$), and inchoative ($p = .002$). All tests were run on 2 × 2 contingency tables, and the p-values are shown with the Yates correction. A chi-square test on the full table showed a small effect size ($\chi^2 = 71.347$, df = 9, $p = .000$, Cramer's V = .157).

As illustrated in Table 5, some of the predicted patterns of effects are found in the data, though not all. First, get$_1$ is, in fact, not overrepresented, contrary to the prediction. On the other hand, get$_3$ and get$_6$ are overrepresented in translated English, as opposed to non-translated in this dataset. Interestingly, get$_4$

(permission/obligation) is not overrepresented, as hypothesized. The remaining senses show quite similar frequencies across corpora.

4.2.2 Source and target concentrations

The ENPC data analyzed here overlaps to a certain extent with the data analyzed by Ebeling (2003) in her study of the Norwegian verbs *få* ('get') and *bli* ('become') and their correspondences in English. Ebeling carried out a detailed analysis of both the Norwegian originals and the English translational material in the corpus. By combining some of her analyses with my own, we may develop a corpus-analytical means of checking for what is referred to in section 2 as connectivity between lexical items in a bilingual's two languages. In order to do so, we will make use of an adaptation of one of the measures of association that have been proposed for corpus analysis. Such measures are most often used in corpus linguistics to test the strengths of the associations between words and their collocates or words and the constructions that they appear in. The logic of one such measure, however, may be applied for our purposes. The selected measure is Schmid's (2010) *attraction-reliance method*, which is a means of calculating the relative frequency of specific nouns occurring in so-called *shell-content* constructions (e.g., N + that clause, N + to-infinitive). The measure expresses the relative frequency of a particular noun in a construction relative to all occurrences of the construction ('attraction') and the relative frequency of a particular noun in the construction, relative to all occurrences of the noun in the corpus ('reliance'). The two figures thus express the strength of the relationship between specific nouns and the constructions they occur in.[20]

Our interest is in strength of the relationship between a particular verb and a specific translation of it. The two figures mentioned above will be retooled in order to capture what, for the time being, will be referred to as *source concentration* and *target concentration*. These two measures are taken as an indication of the strength of the translational relationship between items in a parallel corpus. Source concentration is thus operationalized as the percentage of all occurrences of a TL item that are translations of a specific SL item. For example, in this data, we will be considering the percentage of translated *get* in the inchoative sense which originated in a ST occurrence of Norwegian *bli*. Target concentration is operationalized as the percentage of a set of translations of an SL item that is comprised by a given TL item. In this data, for instance, this would be indicated

[20] As pointed out by one of the referees, this measure is a less familiar relative of collostructional analysis, as developed by Stefanowitsch and Gries 2003.

by the percentage of all instances of *få* (in the onset of possession sense) that are translated as *get*. Thus, these two figures express the relative frequency of a particular target language item as a translation of a specific source language item, from the perspective of the translations or the originals, as follows[21]:

$$\text{SC (source concentration)} = \frac{\text{frequency of item X as translation of Y} \times 100}{\text{all occurrences of X in T (target corpus)}}$$

$$\text{TC (target concentration)} = \frac{\text{frequency of item X as translation of Y} \times 100}{\text{all occurrences of Y in S (source corpus)}}$$

In the analysis of *get*, the data from Ebeling (2003) has been utilized to calculate the source and target concentrations for the senses of *get* that correspond to senses of the Norwegian verbs *få* and *bli* (*get* = X, *få* or *bli* = Y). The figures are given in Table 6, and the relevant Norwegian verbs are given in parentheses next to the English sense designation. Some of the senses of *get* are marked not applicable (n/a), as the Norwegian equivalents of these senses are not a form of either *få* or *bli*. Possession, for instance, must be expressed by *ha* (have), and movement through a variety of verbs.

Table 6: Source and target concentrations for senses of get corresponding to *få* and *bli*. Data adapted from Ebeling (2003)

	Source Concentration	**Target Concentration**
1 – onset (*få*)	**43.25**	**25.20**
2 – possession	n/a	n/a
3 – movement	n/a	n/a
4 – permission/obligation (*få*)	**35.29**	**5.02**
5 – causation (*få*)[22]	–	–
6 – inchoative (*bli*)	**54.48**	**10.15**
7 – passive (*bli*)/(*få*)	**29.23**/9.23	**2.15**/12.24

[21] These two measures together provide two different perspectives on the status of a specific TL item with regard to a specific SL item, and the combination is designed to speak to the status of a given *translational choice*. In other words, these measures are intrinsically directional. Another well-established measure, Altenberg's 'mutual correspondence' (MC) formula (1999: 254) also measures the strength of relationships across translational corpora. Calculating the MC for the various sense categories here would be a relevant addition to an extension of this study to incorporate the other direction, from English to Norwegian.

[22] There is a problem with the causative sense, in that the number of Norwegian *få* translated into *get i*n Ebeling's data exceeds the number of instances of causative *get* identified in the analysis presented in Table 5. This should not be possible if the causative meaning is retained in translation. One possible explanation for this might be that the causative meaning has been lost, while the verb *get* has been used in another sense (e.g. passive). This requires a more detailed analysis of the data than is possible here. But the discrepancy means that a calculation of the source and target concentrations would be based on misleading figures for this sense.

The figures in Table 6 illustrate that, with the exception of the passive use of *få*, the source concentrations are many times higher than the target concentrations. In other words, when *get*, in the senses that correspond to *få* and *bli*, is found in translated text, there is a good chance that the source text item was either *få* or *bli*. The source concentrations ranged from nearly 30 percent for passive *get* to over 50 percent for inchoative *get*. At the same time, from the starting point of these senses of *få* and *bli* in the Norwegian originals, the chances of their being translated as *get* were much lower, from c. 2 percent for passive *bli* to roughly 25 percent for onset *få*. The relationship between source and target concentrations for the individual senses ranged from a factor of 1.7 (43.25/25.20) for sense 1 (onset) to a factor of nearly 14 for passive (*bli*). These results are in some regards puzzling. They demonstrate that translators working out of Norwegian into English choose an array of translations for each of the senses of *få* and *bli*, and that *get* is not very prominent among them. On the other hand, if *get* is found in English translated from Norwegian, the likelihood is rather high that it is a translation of either *få* or *bli*.

Interestingly, the calculation of target concentrations actually involves the issue of onomasiological salience. The frequencies of various translations of an SL item in a sense reflect the onomasiological salience of the various alternatives. In this case, *get* is not a highly salient choice for any of the senses of *få* or *bli*, except, perhaps for get$_1$, which accounts for 25 percent of the translations of *få* in the onset sense. Of course, the selection of *get* in any given instance is affected by a range of contextual variables, and the choice between a set of alternatives is not always unconstrained. This is particularly relevant for the grammaticalized senses of these verbs, for instance the passive uses above. Thus, in-depth studies of the effects of onomasiological salience in translation will require much more detailed analyses of the semantic-syntactic constraints in operation in any given occurrence.

4.3 Performance data

The analyses presented in section 4.2 make use of corpus data, which, as has been stated repeatedly, only provide indirect evidence of both the knowledge of language users and the cognitive semantic patterns they access and use. In this section, therefore, a different data type will be investigated for translational effects resulting from the model of semantic structure outlined in section 2. In this section, we shall be reporting on keystroke data that may be classified as observational data, though not of the same type as corpus data. In previous work, I have referred to the type of data analyzed here as performance data, as opposed to the traditional categories of *product* and *process* commonly used in TS (see note 6).

In this small investigation, salience and connectivity effects, as discussed in the preceding sections, are assumed to impact revision behavior, as reflected in keystroke logs. The assumption is that items of high salience and/or connectivity are chosen more readily, and that they consequently would tend to be more subject to replacement later on. This type of first-choice revision is common in translation (Englund Dimitrova 2005; Malkiel 2009). This reasoning is similar to, but not the same as, Chesterman's (2011b) literal translation hypothesis. According to this hypothesis, "during the translation process, translators tend to proceed from more literal versions to less literal ones" (2011b: 26). *Literal* in terms of the cognitive semantic model adopted here is to be understood as a default choice, affected by salience and connectivity (see Halverson 2015a for further details). In terms of the hypotheses presented in section 2, this data will provide evidence of salience and connectivity, but will not be able to tease the three sources apart. This initial investigation is a coarse-grained first look to see whether there is, in fact, evidence of these cognitive features in online revision. The hypothesis is: the most frequent TT (target text) verbs will be more frequently replaced through revision than the less frequent ones.

As the type of data collected here involves a small number of instances of individual lexical items, such as a given verb, we will not be studying only one verbal category. Instead, the verbal category will be enlarged (from just *get*) to look for frequency effects in verb translation in general.

As a means of operationalizing salience for this study, we again make use of frequency by adopting the notion of *basic verbs*, as outlined in Viberg (2002). Viberg defines basic verbs as, "the most frequent verbs in an individual language" (2002: 53). As suggested by de Groot (1992), it is often the case that highly frequent verbs correspond across languages. Indeed, based on the word lists for the corpora used in this study (COCA/BNC/Norwegian newspaper corpus), the lists of the top 20 verbs in the two languages correspond completely semantically. This suggests that those ST verbs that are among the set of basic verbs will be likely to be translated into a high-frequency TT (target text) verb.[23]

As mentioned, for the study[24], frequency was operationalized through frequency lists for corpora of contemporary English and Norwegian (COCA/BNC/Norwegian newspaper corpus). In accordance with Viberg, the top 20 verbs for English and Norwegian were identified, and this group was labelled *basic*. All

[23] There is a risk that the cut-off between high and low frequency verbs in this study is an artifact of the (sub)corpora that were investigated.

[24] This analysis and additional related analyses were presented first in Halverson (2011).

verbs occurring in the ST and TT were identified as either *basic* or *non-basic*. The source text (see Appendix for an English translation of the ST) contained nine basic verbs (five types, *være, se, gi, få, ha*) and 27 non-basic verb occurrences (16 types).

The source text was an authentic text taken from a small pamphlet available to customers at Norwegian pharmacies. It was an informational text on the topic of sleep quality, and included no technical language. It was translated by 13 Norwegian undergraduate students who were participants in a one-semester, 15-credit introduction to translation course. The subjects translated the text using Translog (Carl 2012), and had no access to any materials. The data collection was done at a computer lab at the University in the spring of 2011. The students translated the brief informational brochure text into their L2, English. They were given two hours to complete the task, but were allowed to leave when they were done. The data consists of the logs and the final translated text for each subject.

The keystroke logs were examined, and all ST and TT verbs were classified as either basic (top 20) or non-basic. All revisions (replacements) of verbs were also analyzed and classified according to the category of the verb that was selected as a replacement (basic/non-basic). The results are given in Table 7. The rows include the revision categories, and in the first four rows, the first choice and the second choice constitute the type. Thus *basic to non-basic* means that the translator first selected a basic verb and then revised it to a non-basic one. *Non-basic to basic* is the reverse, and so on. The columns represent the two categories of ST verbs, low frequency (non-basic) and high frequency (basic):

Table 7: Revisions of verbs by basic/non-basic status

	ST non-basic N (%)	ST basic N (%)	Total N (%)
Basic to non-basic	4 (1.2)	**13 (10)**	17
Non-basic to basic	4 (1.2)	3 (2.3)	7
Basic to basic	0	1 (.8)	1
Non-basic to non-basic	13 (3.8)	1 (.8)	14
Paraphrase	2 (.6)	2 (1.5)	4
Omit (delete)	**12 (3.6)**	**0**	12
No revisions	303 (89.6)	110 (84.6)	413
Total	338 (100)	130 (100)	468

It is striking that these translators made so few revisions of the verbs (leaving nearly 90 percent of the non-basic verbs un-revised). This is, however, in line with other research (Jakobsen 2002) suggesting that verbs may not be overly vulnerable to editing in the translation process. It is interesting, however, that the revisions that were done primarily involved basic verbs, and that the revision types were as expected, from basic to non-basic (see bold above). The hypothesis is thus supported. Another interesting finding is that the two frequency categories showed different results with regard to omission in the TT. As shown in bold in Table 7, no basic verbs were omitted, while nearly 4 percent of the non-basic verbs were. The effect of frequency on revision type is statistically significant and the effect size is small (N = 468, df = 6, Fisher's exact $p \leq .001$, Cramer's V = .233).

4.4 Discussion

In the first step of the analysis in section 3, a model for a Norwegian-English semantic network was developed on the basis of corpus data from two corpora for each language (except for *bli*) and the results of an elicitation test carried out on Norwegian-English bilinguals. This model was used to predict three particular patterns in translational corpus data on the basis of suggested patterns of semasiological salience (magnetism and gravitational pull) and connectivity between Norwegian and English senses:

- get$_3$ (movement) will be overrepresented (magnetism alone)
- get$_4$ (permission/obligation) will be overrepresented (gravitational pull)
- get$_1$ (onset) and get$_6$ (inchoative) will be overrepresented (magnetism, gravitational pull and connectivity)

The second part of the analysis (section 4) involved investigations of corpus and keystroke data. The corpus analyses were presented in section 4.2.1, and demonstrated that not all of these predictions were supported. The most frequent sense (get$_3$ movement) was overrepresented, which was in line with the prediction based on magnetism of the TL item. In this case, there was no influence within the bilingual network from a Norwegian SL item, as there is not specific Norwegian verb which corresponds to this sense of *get*.

The result for get$_4$ (permission/obligation) did not demonstrate the hypothesized overrepresentation based on a relatively strong position for this sense in the Norwegian *få* category (i.e. gravitational pull, or a type of interference effect). This result may have to do with the relative infrequency of this sense of *get*. As regards the less frequent senses in general, while get$_{4,}$ get$_5$ and get$_8$ were all relatively equally frequent across corpora, get$_2$ showed underrepresentation in the translational corpus data. The significance tests run on the data in Table 5 illustrate the overall pattern here, which is that frequent senses (at least 2 of 3) are overrepresented in translational data, while relatively infrequent ones are not. While this pattern does not hold for get_1, the broader pattern does seem to be visible in this data.

A different set of frequency effects in verb translation were demonstrated in the keystroke data. There was a difference between highly frequent and less frequent verbs in the frequency with which the verbs were replaced or omitted, and in the type of replacing verbs that were selected. In other words, the keystroke data suggested that frequency has effects on translational revisions of verbs, and the corpus data demonstrated that relatively more frequent verbal senses seem to be subject to different cognitive constraints in translation than is the case for the less frequent senses.

The third prediction was that get_1 (onset) and get_6 (inchoative) would be overrepresented in the translational corpus data, and that this could be the result of all three cognitive characteristics (magnetism, gravitational pull or connectivity). Surprisingly, however, the second most frequent sense overall (get_1 onset) was not overrepresented, though get_6 was.

The second corpus analysis, involving the suggested source and target concentration measures, showed some interesting results. In this analysis, it appears *get* is not predominantly chosen as a translation for any of the senses of *få* or *bli*. This relatively low translatability is also suggested for the Swedish *få* and *get* in Viberg (2002). On the other hand, when *get* is found in translated English, in the senses that correspond to the Norwegian senses of *få* and *bli* retrievable in the data used here (get_1, get_4, get_6 and get_7), it is quite likely to be a translation of *få* or *bli* (from 25 percent for get_1 to 35 percent for get_4 and nearly 55 percent for get_6). In other words, from the vantage point of *get* in English translated from Norwegian, the *få/bli* – *get* relationship is much stronger than it appears to be if we look at it from the vantage point of the relevant senses of Norwegian *få/bli* and their English translations, of which *get* is one. In terms of the cognitive semantic model being developed here, this suggests that it will be important to get a better grasp of the role played by the links within the network and their relationship to frequency, among other things. In other words, in future work, it will be necessary to measure the strength of the connectivity patterns within a network in order to investigate the translational effects of more or less entrenched connections. The source and target concentration shown in Table 6 are highest for the two most frequent senses (onset and inchoative), suggesting that frequency is an important part of this overall connectivity issue. Furthermore, the issues of language dominance and potential directionality effects must be factored into later studies of connectivity patterns[25]. In addition, the strength of activation of connectivity within the network will

[25] Thanks to the editors for raising this point.

most likely be task specific. It is likely that the effect of connectivity in an overtly bilingual task such as translation will be different from the effect in a more monolingual production mode. All of these issues require closer attention.

The anomalous result here is that for get$_{1.}$ This sense is the second most frequent one in English, it is by far the one that is most frequently produced by Norwegian-English bilinguals, and its semantics match the most frequent sense of its close Norwegian counterpart *få*. For all of these reasons, we would expect overrepresentation of this sense in translated English. It may be, however, that the results here are an effect of the decision to operationalize the status of the various senses of *get* purely in terms of frequency. In the first presentation of the gravitational pull hypothesis in Halverson (2003), reference was made to studies of learner language, particularly a study by Ijaz (1986), based on a cognitive linguistic framework. Ijaz studied the acquisition of certain English prepositions and related learner patterns to the prototype structure of the lexical categories. Ijaz found that where L1 and L2 prototypes were the same, learners were able to achieve native-like usage more quickly. This was only one of a number of category structure effects. Something like this prototype similarity may be at work in the case of get$_1$, where translated text is quite similar to non-translated text in this data. Given the salience of the onset of possession sense for *få*, given its overall frequency in Norwegian (it is a basic verb), and given the high association of *få* with get$_1$ in the elicitation test, one line of reasoning might suggest that this is a case where close cross-linguistic similarity leads to translational patterns that closely match the original English figures (as has been demonstrated for learner language). One might also wonder whether the relationship between frequency and overrepresentation in translation, in some cases, might be U-shaped, rather than linear. It may also be that frequency effects interact in some way with the cross-linguistic semantic relationships that pertain. This requires more study.

With regard to the original hypotheses concerning magnetism, the counter-example of get$_1$ suggests that frequency effects interact with other characteristics of the bilingual network in ways that are not yet captured in the model. This may also be the case with respect to *get*$_2$, which occurred nearly twice as frequently in the non-translational data as in the translational data. The difference was highly significant, and is the only evidence of significant underrepresentation in this translational data. This sense is relatively infrequent in the monolingual data, and was not frequently produced in the elicitation task: this suggests relatively low salience, which would be a relevant factor. An additional source of possible influence here might be the rather particular status of the construction in question here, i.e. *have got*. This construction is pleonastic and alternates onomasiologically with the formally simpler alternative, *have*. This suggests that further developments of the model should incorporate the factor of onomasiological salience, as it is probably interacting with other factors in interesting ways.

Finally, we recall from the discussion in the introduction that the pattern of effects demonstrated here has not been posited as being unique to translated language, and it has also been suggested that the same patterns may be found in L2 language production (Halverson 2003: 225). Indeed, Ringbom (1998) presents evidence of overuse of high frequency items by language learners in his study of patterns in the International Corpus of Learner English (ICLE). Of particular relevance to the present study is his discussion of high frequency verbs. In addition to identifying a correlation between verb frequency and overuse by learners, Ringbom (1998: 44–45) also looks at *get* in particular, and compares the main uses of this verb across L1 groups (native speakers of English, French, Spanish, Finnish, Finnish Swedish, Swedish, Dutch and German). Ringbom classifies the uses by type of complement, meaning that not all of the categories are directly comparable to the semantic categories in use here. The ones that can be compared are the most frequent ones, however, i.e. the ones that we have identified as get_1 (onset), get_3 (movement), get_6 (inchoative). Ringbom's final category corresponds in part to our category of permission (get_4). Ringbom's data illustrates that there are both similarities and differences in the usage patterns across L1s. Interestingly, all learner groups overuse get_1 relative to native speakers. The speakers of Nordic languages (and Spanish) do so to a much larger degree, however. The speakers of Nordic languages are also noted for their overuse of get_6 (inchoative). As regards get_3 (movement), usage across all L1 groups is close to that of native speakers, with the exception of the German group, who overuse it.

Of course, Norwegian is not one of the L1 groups analyzed by Ringbom, but Swedish and Norwegian share all of the most frequent sense categories, and the results he presents above are thus a relevant comparison to Norwegian. They serve as a further demonstration of the need to consider particular L1/L2 relationships in attempting to understand bilingual language production of all types.

5 Concluding remarks

This is only the first relatively simple account to bring together online and offline data to look for effects of cognitive semantic patterns in translation. The tests were used to look for preliminary support for and to further develop the emerging cognitive model, but also to illustrate the ways in which the various types of data can address complementary questions and to lead to more detailed and testable cognitive semantic hypotheses.

A number of reservations must be mentioned regarding the corpus data used here. The ENPC was one of the first corpora constructed to allow for cross-linguistic and translational studies. It has done valuable service in numerous contrastive linguistic and translation studies investigations. However, it has shortcomings that limit its utility for further investigations of the type conducted here. One is the age of its texts and another is its size. In the Norwegian context, it is fair to say that the level of English among Norwegian adults has changed in the 20–30 years since these texts were published. The English produced by Norwegian adult translators today is probably different in some ways from what is found here. By contemporary standards, the corpus is also very small. More serious, however, is the lack of detailed metadata on the translators and their bilingual histories. As argued earlier (Halverson 2010a), it is vital that such information be provided so that such aspects as language dominance and directionality may be controlled for. These factors, which are important for the workings of bilingual cognition, have not been considered in the present study due to lack of information on the individual translators.

An additional issue that must be mentioned is the role of genre or register. As mentioned in the introductory remarks, a number of recent studies have suggested that normalization and interference play out differently in different text registers in translation. In the current study, both fiction and non-fiction were included, but the genres were not separated in the material due to the size of the data set at the outset. In later studies, this variable should be included.

The case study involved the use of three different types of data: corpus data, a sentence generation test, and Translog data. All three data types were investigated to look for evidence of frequency effects in language production tasks. Corpus data was used to construct the bilingual network model, using corpus frequency to posit cognitive salience. This salience was checked against the sentence generation task in bilinguals, and the results here suggest both Norwegian and English language impact in the order and frequency of the senses produced. In looking for the hypothesized patterns in translated text, a frequency effect was found, though it would seem that reducing the notion of prototype to frequency alone might be premature. Frequency effects were also found in a broader investigation of verb translation in keystroke data, and both revision and omission were related to frequency. Other data (e.g. pause data) could be added to this analysis.

Much work remains to be done to achieve a robust and viable empirical approach to the testing of the cognitive linguistic model being developed here. While some initial support was found for the hypotheses presented here, this is not to be taken as conclusive. The results suggest that further work is required to develop the model, and more advanced statistical methods are required to tease apart the likely interactions of several of the factors discussed here. In addition, a natural extension of the type of analysis done here would be to test translation in the inverse direction (English-Norwegian) in different bilingual groups. The same data types as used here could be adapted to this task.

Perhaps one of the most glaringly obvious questions that should be mentioned in closing has to do with the utility of looking for effects from semantic networks in isolation when it is eminently obvious that translation does not involve the isolated translation of words. Indeed, psycholinguistic models of translation have long been criticized in Translation Studies on the grounds of insufficient ecological validity. The methods used here are ecologically valid, and the model as such builds on a theory that assumes contextual variation and discourse flexibility. At this stage, the modelling of semantic networks aims to reveal what Langacker has referred to as "less transient" features (1987: 430), features of semantic organization at a level where translational effects would be subtle, yet significant in aggregate. While translators translate words in context, they still translate words. It is not impossible that some of the characteristics of semantic organization at this level should percolate up to the surface of translated texts.

References

Altenberg, Bengt. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In Hilde Hasselgård & Signe Oksefjell (eds.), *Out of corpora. Studies in honour of Stig Johansson*, 249–268. Amsterdam: Rodopi.

Alves, Fabio & José Luiz Gonçalves. 2003. A relevance theory approach to the investigation of inferential processes in translation. *Triangulating translation: Perspectives in process-oriented research*, 11–34. Amsterdam: John Benjamins.

Alves, Fabio & José Luiz Gonçalves. 2007. Modelling translator's competence. Relevance and expertise under scrutiny. In Yves Gambier, Miriam Shlesinger & Radegundis Stolze (eds.), *Doubts and directions in Translation Studies*, 41–55. Amsterdam: John Benjamins.

Alves, Fabio, Adriana Pagano, Stella Neumann, Erich Steiner & Silvia Hansen-Schirra. 2010. Translation units and grammatical shifts: Towards and integration of product- and process-based translation research. In Gregory M. Shreve & Erik Angelone (eds.), *Translation and cognition*, 109–142. Amsterdam: John Benjamins.

Anderman, Gunilla & Margaret Rogers. 2008. *Incorporating Corpora. The linguist and the Translator*. Clevedon: Multilingual Matters.

Askedal, John Ole. 2012. Norwegian *få* 'get': A survey of its uses in present-day Riksmål/ Bokmål. *Linguistics* 50(6). 1289–1331.

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds.), *Text and Technology: in honour of John Sinclair*, 233–250. Amsterdam/Philadelphia: John Benjamins.

Bassetti, Benedetta and Vivian Cook. 2011. The second language user. In Vivian Cook & Benedetta Bassetti (eds.), *Language and bilingual cognition*, 143–190. New York: Psychology Press.

Berez, Andrea & Stefan Th. Gries. 2008. In defense of corpus-based methods: A behavioral profile analysis of polysemous *get* in English. Paper presented at *The 24th NWLC*, Seattle, WA, 3–4 May.

Boada, Roger, Rosa Sánchez-Cases, José M. Gavilán, José E. García-Albea & Natasha Tokowicz. 2015. Effect of multiple translations and cognate status on translation recognition performance of balanced bilinguals. *Bilingualism: language and cognition* 16(1). 183–197.

Brysbaert, Marc, Eef Ameel & Gert Storms. 2014. Semantic memory and bilingualism: A review of the literature and a new hypothesis. *Foundations of bilingual memory.* New York, NY: Springer Science and Business Media. Available at: crr.ugent.be/papers/Brysbaert_et_al_2014_semantic_memory_and_bilingualism.pdf. Accessed on 3 April 2014.

Cappelle, Bert & Rudy Loock. 2013. Is there interference of usage constraints? A frequency study of existential *there is* and its French equivalent *il y a* in translated vs. non-translated texts. *Target. International Journal of Translation Studies* 25(2). 252–275.

Carl, Michael. 2012. Translog – II: A program for recording user activity data for empirical translation process research. *IJCLA* 3(1). 153–162.

Chesterman, Andrew. 2011a. Translation universals. In Yves Gambier & Luc van Doorslaer (eds.), *Handbook of Translation Studies* 2, 175–179. Philadelphia, PA: John Benjamins.

Chesterman, Andrew. 2011b. Reflections on the literal translation hypothesis. In Cecilia Alvstad, Adelina Hild & Elisabet Tiselius (eds.), *Methods and strategies of process research*, 23–35. Amsterdam: John Benjamins.

Cook, Vivian J. (ed.). 2003. *The effects of the second language on the first.* Clevedon: Multilingual Matters.

Collins English Dictionary. 2005. Glasgow: HarperCollins.

De Groot, Annette M.B. 1992. Determinants of word translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18(5). 1001–1018.

De Groot, Annette M.B. 2011. *Language and cognition in bilinguals and multilinguals. An introduction*. New York: Psychology Press.

Delaere, Isabelle, Gert De Sutter & Koen Plevoets. 2012. Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring distance between language varieties. *Target. International Journal of Translation Studies* 24(2). 203–224.

Delaere, Isabelle & Gert De Sutter. 2013. Applying a multidimensional, register-sensitive approach to visualize normalization in translated and non-translated Dutch. *Belgian Journal of Linguistics* 27. 43–60.

Diependaele, Kevin, Kristin Lemhöfer & Marc Brysbaert. 2013. The word frequency effect in first and second language word recognition: a lexical entrenchment account. *Quarterly Journal of Experimental Psychology* 66. 843–863.

Divjak, Dagmar & Antti Arppe. 2013. Extracting prototypes from exemplars. What can corpus data tell us about concept representation? *Cognitive Linguistics* 24(2). 221–274.

Ebeling, Signe Oksefjell. 2003. *The Norwegian verbs* bli *and* få *and their correspondences in English. A corpus-based contrastive study.* Oslo, Norway: University of Oslo, Faculty of Arts.

Englund Dimitrova, Birgitta. 2005. *Expertise and explicitation in the translation process*. Amsterdam: John Benjamins.

Evans, Vivyan & Melanie Green. 2006. *Cognitive linguistics. An introduction*. Edinburgh: Edinburgh University Press.

Frawley, William. 1984. Prolegomenon to a theory of translation. In William Frawley (ed.), *Translation, Literary, Linguistic and Philosophical Perspectives*, 159–175. London: Associated University presses.

Geeraerts, Dirk. 2009. *Words and other wonders. Papers on lexical and semantic topics*. Berlin: Mouton de Gruyter.

Gellerstam, Martin. 1986. Translationese in Swedish novels translated from English. In Lars Wollin & Hans Lindquist (eds.), *Translation Studies in Scandinavia*, 88–95. Lund: CWK Gleerup.

Gilquin, Gaëtanelle. 2008. Taking a new look at lexical networks. *Lexis* 1. 25–39.

Gronemeyer, Claire. 1999. On deriving complex polysemy: the grammaticalization of *get*. *English Language and Linguistics* 3(1). 1–39.

Halverson, Sandra. 2003. The cognitive basis of translation universals. *Target. International Journal of Translation Studies* 15(2). 197–241.

Halverson, Sandra. 2007. A cognitive linguistic account of translation shifts. *Belgian Journal of Linguistics* 21. 105–119.

Halverson, Sandra. 2010a. Cognitive translation studies. Developments in theory and method. In Gregory Shreve & Erik Angelone (eds.), *Translation and Cognition*, 349–369. Amsterdam: John Benjamins.

Halverson, Sandra. 2010b. Translation universals or cross-linguistic influence: conceptual and methodological issues. Paper presented at 6th EST Conference, Leuven, 23–25 September.

Halverson, Sandra. 2011. Schematic networks in translation: bringing together process and corpus data. Paper presented at Text-process-text, Stockholm, 17–19 November.

Halverson, Sandra. 2015a. Cognitive Translation Studies and the merging of empirical paradigms: The case of literal translation. *Translation Spaces* 4(2). 310–340.

Halverson, Sandra. 2015b. The status of contrastive data in Translation Studies. *Across languages and cultures* 16(2). 163–185.

Hansen, Silvia. 2003. *The nature of translated text*. (Saarbrücken Dissertations in Computational Linguistics and Language Technology. Volume 13). Saarbrücken: DFKI and Saarland University.

Hartsuiker, Robert J. 2013. Bilingual strategies from the perspective of a processing model. *Bilingualism: Language and cognition* 16(4). 737–739.

Hartsuiker, Robert J., Martin J. Pickering & Eline Veltkamp. 2004. Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological science* 15(6). 409–414.

Heylen, Kris, José Tummers & Dirk Geeraerts. 2008. Methodological issues in corpus-based cognitive linguistics. In Gitte Kristiansen & René Dirven (eds.), *Cognitive sociolinguistics. Language variation, cultural models, social systems*, 91–127. Berlin: Mouton de Gruyter.

Ijaz, I. Helene. 1986. Linguistic and cognitive determinants of lexical acquisition in a second language. *Language Learning* 36(4). 401–451.

Jakobsen, Arnt Lykke. 2002. Translation drafting by translation professionals and translation students. *Copenhagen Studies in Language* 27. 191–204.

Jarvis, Scott & Aneta Pavlenko. 2008. *Crosslinguistic influence in language and cognition*. New York: Routledge.

Johansson, Stig & Signe Oksefjell. 1996. Towards a unified account of the syntax and semantics of *get*. In Jenny Thomas & Mick Short (eds.), *Using Corpora for Language Research. Studies in the Honour of Geoffrey Leech*, 57–75. London: Longman.

Kroll, Judith F. & Erika Stewart. 1994. Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language* 33. 149–174.

Kruger, Alet, Kim Wallmach & Jeremy Munday. 2011. *Corpus-based Translation Studies. Research and applications*. London: Bloomsbury.

Kruger, Haidee & Bertus van Rooy. 2012. Register and the features of translated language. *Across Languages and Cultures* 13(1). 33–65.

Langacker, Ronald. 1987. *Foundations of Cognitive Grammar. Volume I*. Stanford: Stanford University Press.

Langacker, Ronald. 2008. *Cognitive Grammar. A basic introduction*. Oxford: Oxford University Press.

Lanstyák, István & Pál Heltai. 2012. Universals in language contact and translation. *Across Languages and Cultures* 13(1). 99–121.

Laviosa, Sara. 2009. Universals. In Mona Baker & Gabriela Saldanha (eds.), *Routledge Encyclopedia of Translation Studies*, 2nd edn. 306–310. London: Routledge.

Laviosa, Sara. 2011. Corpus-based Translation Studies: where does it come from? Where is it going? In Alet Kruger, Kim Wallmach & Jeremy Munday (eds.), *Corpus-based Translation Studies. Research and applications*, 13–32. London: Bloomsbury.

Malkiel, Brenda. 2004/2005. Contrasting languages and varieties with translational corpora. *Languages in contrast* 5(1). 73–92.

Malkiel, Brenda. 2009. From Ántona to My Ántona: Tracking self-corrections with Translog. *Copenhagen Studies in Language* 37. 149–167. Copenhagen: Samfundslitteratur.

Mauranen, Anna & Pekka Kujamäki. 2004. *Translation universals: Do they exist?* Amsterdam: John Benjamins.

Neumann, Stella. 2014. Cross-linguistic register studies. *Languages in Contrast* 14(1). 35–57.

Oakes, Michael & Meng Ji. 2012. *Quantitative methods in corpus-based Translation Studies*. Amsterdam: John Benjamins.

Pavlenko, Aneta. 2009. Conceptual representation in the bilingual lexicon. In Aneta Pavlenko (ed.), *The Bilingual Mental Lexicon. Interdisciplinary Approaches*, 125–160. Bristol: Multilingual Matters.

Pym, Anthony. 2005. Explaining explicitation. In Krisztina Károly & Ágota Fóris (eds.), *New trends in Translation Studies: In honour of Kinga Klaudy*, 29–43. Budapest: Akadémiai Kiadó.

Pym, Anthony. 2008. On Toury's laws of how translators translate. In Anthony Pym, Miriam Shlesinger & Daniel Simeoni (eds.), *Beyond Descriptive Translation Studies*, 311–328. Amsterdam: John Benjamins.

Ringbom, Håkan. 1998. Vocabulary frequencies in advanced learner English: A cross linguistic approach. In Sylviane Granger (ed.), *Learner English on Computer*, 41–52. London: Longman.

Sandra, Dominiek & Sally Rice. 1995. Network analyses of prepositional meaning: Mirroring whose mind – the linguist's or the language user's? *Cognitive linguistics* 6(1). 89–130.

Santos, Diana. 1995. On grammatical translationese. Short paper presented at the Tenth Scandinavian Conference on computational Linguistics. Helsinki. http://www.solo.sintef.no/portg/Diana/download.no (accessed 10 August 2014).

Schäffner, Christina & Beverly Adab. 2001. The idea of the hybrid text in translation: Contact as conflict. *Across languages and cultures* 2(2). 167–180.

Schmid, Hans-Jörg. 2010. Does frequency in text instantiate entrenchment in the cognitive system? In Dylan Glynn (ed.), *Quantitative methods in cognitive semantics: Corpus driven approaches*, 101–133. Berlin: Mouton de Gruyter.

Steiner, Erich. 2001. Intralingual and interlingual versions of a test – how specific is the notion of 'translation'? In Erich Steiner & Colin Yallop (eds.), *Exploring translation and multilingual text production. Beyond content*, 161–190. Berlin: Mouton de Gruyter.

Steiner, Erich. 2012. Methodological cross-fertilization: Empirical methodologies in (computational) linguistics and translation studies. *Translation: Corpora, Computation, Cognition* 2(1). 3–21.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Teich, Elke. 1999. System-oriented and text-oriented comparative linguistic research. Cross linguistic variation in translation. *Languages in Contrast* 2(2). 187–210.

Teich, Elke. 2003. *Cross-linguistic variation in system and text.* Berlin: Mouton de Gruyter.

Tirkonnen-Condit, Sonja. 2004. Unique items – over- or underrepresented in translated language? In Anna Mauranen & Pekka Kujamäki (eds.), *Translation univerals. Do they exist?*, 177–184. Amsterdam: John Benjamins.

Tirkonnen-Condit, Sonja. 2005. The monitor model revisited: Evidence from process research. *META* 50(2). 405–414.

Toury, Gideon. 1995. *Descriptive translation studies and beyond.* Amsterdam/Philadelphia: John Benjamins.

Tummers, José, Kris Heylen & Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus linguistics and linguistic theory* 1(2). 225–261.

Vandevoorde, Lore. 2016. On semantic differences: a multivariate corpus-based study of the semantic field of inchoativity in translated and non-translated Dutch. Ghent: Ghent University doctoral dissertation. http://hdl.handle.net/1854/LU-7899148.

Viberg, Åke. 2002. Polysemy and disambiguation cues across languages. In Bengt Altenberg (ed.), *Lexis in contrast. Corpus-based approaches*, 119–150. Philadelphia: John Benjamins.

Xiao, Richard. 2010. *Using Corpora in Contrastive and Translation Studies*. Newcastle: Cambridge Scholars Publishing.

The ENPC: http://www.hf.uio.no/ilos/english/services/omc/enpc/.

The BNC: http://www.natcorp.ox.ac.uk/.

Translog: https://sites.google.com/site/centretranslationinnovation/translog-ii.

Appendix

Sleep good!

A good night's sleep is absolutely vital if you want to have the energy to cope with the challenges of your day. If you often wake up tired and listless, you should make an effort to follow the sleep tips below and also consider whether you might need to make some lifestyle changes. Following this advice will enable you to either counteract sleeping problems or to avoid them altogether.

Better sleep tips

How long and how well you sleep is the result of a complex interaction between your need for sleep, your daily rhythm and your habits. All of these are important, so the following list of practical tips should help contribute to better habits and thus better sleep.

* Get up at the same time every day, also on the weekend.
* Get at least a half-hour of daylight every day, as early in the day as possible.
* Cut down on activity in the evening and don't be active at night.
* Drink less coffee, tea and cola products during the day, and never drink anything containing caffeine after 6:00pm.
* Do not nap during the day.
* Avoid strenuous exercise or intense emotional experiences in the evening.
* Do not light bright lights if you must get up at night.
* Try not to be very hungry or overly full after 8:00 pm.
* Avoid alcohol. Alcohol is detrimental to sleep quality.
* Do not use your bedroom for other purposes, e.g. as a study or TV room.
* Keep your bedroom dimly lit and quiet, with a moderate temperature and smelling nice.
* Don't try too hard to get to sleep. Concentrate on relaxing.
* Don't look at the clock if you wake up at night.
* Get up if you can't sleep, and then go to bed again a bit later.

Revised from: www.apotek1.no/sovny

Stefan Evert and Stella Neumann

2 The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German

Abstract: This chapter investigates the influence of the source and target language on translations in a selection of 150 pairs of source and target texts from a bidirectional parallel corpus of English and German texts, applying a combination of multivariate analysis, visualization and minimally supervised machine learning. Based on a procedure developed by Diwersy, Evert and Neumann (2014), it investigates the way in which translations differ from comparable original texts depending on the translation direction and other factors. The multivariate approach enables us to detect patterns of feature combinations that cannot be observed in conventional frequency-based analyses, providing new evidence for the validity of interference or shining through in translation. We report a clear shining through effect that is more pronounced for translations from English into German than for the opposite translation direction, pointing towards a prestige effect in this language pair.

1 Introduction

The specific properties that are claimed to distinguish translations from non-translated texts have been the object of research in corpus-based translation studies for almost 30 years. We now have evidence for specific properties of translated versus non-translated text for various language pairs and for various properties (cf. e.g. contributions in Mauranen and Kujamäki 2004; Hansen-Schirra, Neumann, and Steiner 2012 and various individual studies). Many studies however are limited to a restricted set of features: Olohan and Baker (2000), for example, investigate the complementizer *that*, Mauranen (2004) investigates frequencies of lexis, Hansen-Schirra, Neumann and Steiner (2007) analyze cohesive devices. The use of statistical techniques to draw inferences from the observed patterns in a corpus to the underlying population is still not very well established in translation studies. If a statistical analysis is carried out at all, it is often limited to univariate techniques, e.g. comparing the frequencies of

DOI 10.1515/9783110459586-003

individual linguistic features between translations and originals with Student's t-test or a similar method. A typical example is Neumann (2013), who carries out t-tests comparing translated texts to a reference corpus, focusing on a single linguistic feature at a time. Such univariate methods are suitable for studies examining the effect of a single feature, but they become insufficient when a whole feature catalogue is analyzed.

Systematic properties of text – and that is how translation properties can be characterized – are hardly ever observable on the basis of just a single feature. Most likely, such properties are expressed through a combination of features. Register properties, for instance, may sometimes appear obvious by one individual feature (e.g. imperatives in instruction manuals), but the property of a text serving instructional goals only really emerges if the imperative mood is combined with other features such as short sentences, the use of appropriate terminology, a specific iconic order of clauses in temporal or causal relations, etc. By the same token, individual features hardly ever function in terms of a single property. It is much more likely to assume that one feature contributes to several properties. A high frequency of second person pronouns, for example, can be indicative of reduced social distance and at the same time of the spoken (as opposed to written) medium. Studies that analyze individual features cannot assess correlations between features. Furthermore interactions between different factors that influence the concrete realization of the features are missed. Therefore the use of multivariate techniques appears to be essential for a systematic investigation of translation properties.

Recently, scholars have adopted this approach to profiling translations as compared to non-translated texts. Delaere, De Sutter, and Plevoets (2012) analyze register-related lexical variation as an operationalization of norm-conforming behavior of translators with the help of profile-based correspondence analysis. Contributions in Oakes and Ji (2012) introduce various approaches to the quantitative investigation of translations. Related work by Kruger and van Rooy (2012) draws on analysis of variance (which is not a multivariate technique) to analyze operationalizations of the different translation properties discussed by Baker (1996) across different translated registers in comparison to non-translated texts.

The role of source language interference, one of the features identified as a potential property of translated texts and the main focus of this chapter, was ruled out as a relevant factor on translation by Baker (1993) arguing that it is not related to translation but rather pertains to all kinds of language use where more than one language is involved, such as second language text production. She also argued that a corpus design that collects translations from a wide range of different source languages would level out the influence of the individual source language. However, strictly speaking we would claim that it is methodologically impossible to determine differences between translated and non-translated texts without comparing the realization of a feature in the matching source text: the observed differences might be introduced by other factors than translation effects, e.g. a register divergence between the translations and originally written texts in the same language. Only differences between text pairs aligned at a level appropriate for the respective feature can reliably be claimed to represent properties of *translations* (see Steiner 2012a: 73–75, and on explicitation see Steiner 2012b: 59; for an extensive discussion of aligned pairs of source and target texts see Serbina 2013). Originally described in second language learning as an influence of the L1 on the L2, interference could simply be a general feature of using language in a context where both language systems are activated and trigger choices from both systems in text production (cf. Mauranen 2004). This would mean that, regardless of the specific type of language use (L2 writing and translating into the L1), features from the second activated language system would be likely to interfere with the language in which the text is produced. Interference in this case would not represent a translation-specific phenomenon. While the effect of both types of interference has not yet been sufficiently investigated, we would claim that it is most likely not the same (and also caused by different factors). Interference in second language production involves transfer from the mother tongue into the L2, whereas, at least in the default case of translation into the mother tongue (L1), interference in translation refers to transfer from the L2 into the L1 (see also Steiner 2008 on the directionality of language contact). The comparison between interference in L2 writing and in translation into the L1 is outside of the scope of this chapter. Suffice it to say that the specificity of the translation task justifies analyzing interference – or more specifically: shining through – in translation in its own right.

Teich (2003) describes a special case of L2 interference she calls shining through: this property refers to cases where the diverging frequencies of options existing in both languages are adapted in translated texts to those of the source language, thus resulting in a frequency difference between translations and comparable non-translated texts in the target language. It is this special case of source language-induced divergence of translations that is the focus of this chapter. One of the potential factors affecting the extent of L2 interference or shining through could be the diverging prestige of the languages involved (Toury 2012: 314). Toury draws on the sociolinguistic concept to argue that an unequal status of languages and cultures could affect the tolerance of interference. If his claim is right, a difference in prestige between two languages should lead to an asymmetric tendency, in which translations from the more prestigious language into the less prestigious one show more tolerance towards interference than in the opposite translation direction. Mauranen (2004) compares Finnish lexis translated from Russian, a presumably less prestigious culture in the Finnish context, with translations from English, a culture she assesses as more prestigious, and does not find a prestige effect. The claim has been made that the impact of Global English exerts an asymmetric influence on German also by way of translation, to the effect that target culture norms may no longer be maintained which in turn results in convergence with English norms (House 2002: 199–200). Testing this claim, Becher, House and Kranich (2009) report inconclusive evidence that modality is not affected by the contact with English whereas the use of sentence-initial concessive conjunctions seems to converge with English in a diachronic corpus comparison.

On a more general level, Hansen-Schirra and Steiner (2012: 272) describe the relationship between different types of translation-related behavior towards source and target language norms (which in frequency terms can be read as usage preferences) as a continuum ranging from shining through, i.e. orientation towards source language norms, to normalization, orientation towards target language norms.

It is still a matter of debate in translation studies whether such properties are caused by translation-inherent or general factors (cf. Becher 2011 on explicitation). We would claim that the debate could be decided with the help of more comprehensive corpus-based research designs that account for more factors simultaneously: Rather than controlling for register, register variation needs to be assessed as a factor on translation properties. Rather than focusing on individual features at a time, studies should include as many linguistic features as possible and use appropriate statistical techniques to assess these diverse factors and their interaction. Based on the evidence we now have, for instance, on the effect of register on variation in translation (Neumann 2013; Delaere 2015), it is obvious that studies concentrating on individual features and controlling for register must inevitably yield contradictory evidence. Finally, rather than excluding the source language, aligned text pairs should be investigated that take into account whether features in a translated text deviate from those in the aligned source text element. The question of the translation inherence of properties can only really be decided on the basis of such improved research designs.

We would claim that the fact that machine learning classifiers are able to distinguish translations from non-translated text with high accuracy provides strong evidence that there are specific traits of translations which need to be explained within the framework of translation studies. In the context of computational approaches, such traits are usually referred to as *translationese*, i.e. some form of distinctive language use in translations. Baroni and Bernardini

(2006), for example, report a classification accuracy of 86%, outperforming human annotators. Volansky, Ordan and Wintner (2015) combine the computational approach to translationese with a corpus-linguistic interest in translation properties.[1] They define a set of linguistic features operationalizing translation properties and show that classifiers do not perform equally well across all properties. A finding relevant for our study is that features related to shining through yield the highest accuracy. Despite their success at identifying translated language, these approaches are not geared towards pinpointing factors that might explain the specific make-up of translated texts, or towards detecting hidden structures, e.g. related to differences between translation directions.

In this chapter, we use exploratory multivariate techniques to analyze the influence of the source and target language on translations, based on the frequency patterns of different linguistic features in a bidirectional parallel corpus of German and English texts from a range of different registers.[2] To this end, we make the following distinctions. (i) We identify "genuine" shining through of properties of the source language into translations as a general tendency of translators to introduce feature patterns that are typical of the source language into the target texts, quantified in terms of the relative frequencies of comparable lexico-grammatical features. This is distinguished from (ii) text-specific, i.e. individual shining through of idiosyncratic properties of the source texts, reflecting author style, tone, topic domain, etc. In this case, certain linguistic properties of the specific source text are carried over in the translation process. In other words, translators do not adjust their linguistic patterns based on the source language, but simply translate texts in a relatively literal way. Shining

[1] The paper is also useful in providing a comprehensive overview of the state of the art of machine learning approaches to translationese.

[2] Implicitly, machine learning approaches – as well as our approach – adopt adherence to target language norms as the basis of comparison. However, from the point of view of translation studies it is not obvious to which norms translators should adhere, i.e. which translation strategy they adopt. Mimicking, as it were, original texts written in the target language is but one option, others being foreignization (e.g. induced by the perceived prestige of the source language), register norms (which are not the same as general source or target language norms), cultural expectations towards translations, specific translation briefs etc. For obvious reasons an individual, text-specific influence of the target language is impossible, but a more general influence of the target language could, for instance, mean that a translation in an aligned text pair displays a tendency to replace features of the source text untypical of the source language by features more typical of the target language. While this chapter concentrates on the part of the variation in translation linked to shining through, our results also suggest a normalization effect in the translation direction German-English, which might be linked to target language influence (see Figure 3).

through could also be a side-effect of register divergences between the German and English parts of the corpus. Since this is a special case of individual shining through – the relevant linguistic property being the sub-register of a text rather than e.g. author style – we do not consider this case separately. (iii) These two types of shining through are distinguished from other forms of translationese that cannot be traced back to the respective source language or to individual source texts.

Given the claims about the influence of English on German noted above, we believe that this language pair is a good example for exploring assumptions about source language shining through and more specifically about the impact of translation direction under a hypothesized prestige effect. Our approach is geared towards the type of norm-related translation properties Hansen-Schirra and Steiner (2012) discuss. We will argue that visualization plays a crucial role for understanding the multidimensional structure of the data set.

After a brief introduction of the data and procedure in the next section, we will examine the steps of the multivariate analysis in section 3. Section 4 is devoted to the detailed interpretation of the results of the analysis before these are discussed in section 5 in light of their meaning for translation studies. The chapter is rounded off by some concluding remarks and an outlook on future work.

2 Method

2.1 The data

The data used for this study comprise a subset of the CroCo Corpus (Hansen-Schirra, Neumann and Steiner 2012). We discarded the three most extreme registers (novels, instruction manuals and, to a lesser extent, tourism brochures), which accounted for most of the variation in Diwersy, Evert and Neumann (2014) and dominated the unsupervised multivariate analysis, obscuring more subtle, but important patterns such as variation between the remaining registers. We further excluded one text pair as an outlier because the PCA and LDA techniques used by our approach are sensitive to such outliers and give them undue weight in the analysis. In total, we used 298 texts from the five registers political essays (‘essay’), popular-scientific texts (‘popsci’), corporate letters to shareholders (‘share’), prepared political speeches (‘speech’) and websites (‘web’). These registers are similar in their focus on factual rather than fictional matters.

The study draws on lexico-grammatical indicators of underlying functions derived in the context of register theory (Neumann 2013). Of the indicators used by Neumann, we included only those which not only exist in both languages but are considered to be comparable, so that original texts and the corresponding translations can be meaningfully compared. We also discarded collinear features, resulting in a final set of 27 indicators which were obtained with a mixture of automatic and manual extraction procedures.[3] A full list of features and their extraction methods is contained in the Appendix. All frequency counts are given in relation to an appropriate unit of measurement, e.g. proportion of nouns among all tokens, finites among all sentences, passives among all verbs, imperatives among all sentences, adverbial themes among all themes, contracted forms among all tokens, etc. Additional features are lexical density, the lexical type-token ratio (TTR) and average sentence length (tokens/sentences). To account for the large frequency differences between the various indicators, all values were standardized (z-transformed). The z-transformation also ensures that each feature makes the same overall contribution to the distances between texts described in section 2.2. Every text is thus represented as a feature vector in multidimensional space consisting of the z-scores of 27 lexico-grammatical indicators.

2.2 The approach to multivariate analysis

We adopt the geometric approach of Diwersy, Evert, and Neumann (2014), which makes the assumption that Euclidean distances between feature vectors provide a meaningful measure of the dissimilarity between the corresponding texts and which emphasizes the use of orthogonal projections in order to visualize the geometric configuration of data points in a high-dimensional feature space from different perspectives. This approach has many advantages: First, the position of a text along an orthogonal second dimension does not affect its interpretation with respect to the first dimension. Second, the total variance of the data set – i.e. the average (squared) Euclidean distance between two texts – is the sum of its variances along a set of orthogonal dimensions. We can use the respective proportion of variance (R^2) as a quantitative measure of how much of the geometric configuration is captured by a particular orthogonal projection. Third, the angle between two non-orthogonal axes indicates the amount of overlap between the information provided by these axes about the data set. If the angle is small, the second axis offers little additional information over the first; if the axes are orthogonal at an angle of 90 degrees, they provide

[3] In comparison to Diwersy, Evert and Neumann (2014) one additional feature was discarded because of collinearity. Another feature (the frequency of infinitives) had to be discarded because the automatically obtained frequency counts turned out to be unreliable.

complementary information (cf. the first point made above). Diwersy, Evert, and Neumann (2014) propose the following steps for the multivariate analysis:

1. Apply unsupervised Principal Component Analysis (PCA) to obtain a perspective that captures the overall shape of the data set. PCA yields a ranked list of orthogonal latent dimensions, chosen to maximize the proportion of variance (R^2) preserved by orthogonal projection into the first PCA dimensions.
2. Visualize this perspective with two- and three-dimensional scatterplots, using meta-information such as language, translation status and register to highlight interesting patterns and facilitate the interpretation. The visualization can also reveal methodological problems such as outliers.
3. Introduce a minimal amount of theory-neutral knowledge in order to find a perspective that throws into relief aspects of the geometric configuration which are relevant to the research question. In our case, this leads to a perspective that shows a clear separation of English and German originals even though its R^2 is smaller than for the PCA dimensions.
4. A suitable perspective can be determined automatically using Linear Discriminant Analysis (LDA), a machine learning procedure that maximizes the distance between two (or more) groups while minimizing within-group variability. The LDA discriminant can be used as a dimension for the orthogonal projection and is usually combined with a PCA analysis of the orthogonal complement space for visualization.
5. Validate the LDA model on separate test data to ensure that it has not been overfitted to individual data points. This is usually carried out by cross-validation using Support Vector Machines (SVM) or a similar machine learning classifier. Diwersy, Evert and Neumann (2014: 185) emphasize the importance of this step to avoid circularity and deductive bias. Latent dimensions are identified based on their proven ability to distinguish categories introduced in step 3, rather than on the analyst's subjective interpretation.
6. If necessary, repeat from step 2 in order to improve the analysis. In this chapter, we only report the final analysis obtained after several iterations of visualization and interpretation.
7. Develop a linguistic interpretation based on visualizations, quantitative validation, and the (constellations of) feature weights of the LDA discriminant or other latent dimensions. In section 4, the interpretation of feature weights is scrutinized more thoroughly and further developed compared to the discussion in Diwersy, Evert and Neumann (2014).

2.3 Characterization of the approach

In comparison to conventional linguistic approaches, our method does not only support the choice and interpretation of features based on register theory but also gives a global perspective on feature combinations and correlations where, for instance, Neumann (2013) only analyzes the behavior of individual features. Comparing our approach to related work using unsupervised multivariate analysis – in particular Biber's multidimensional analysis (e.g. Biber 1988) – both approaches identify latent dimensions based on feature correlations and thus facilitate the visualization of the high-dimensional distribution of a data set. However, our approach assumes a geometric perspective by focusing on orthogonal projections, in contrast to the Factor Analysis (FA) used by Biber (1988). A key difference is the introduction of weakly supervised information in order to discover more delicate patterns of interest beyond the main dimensions of variation found by an unsupervised analysis (see our discussion in section 3). Our work can also be compared to studies that apply machine learning approaches to translationese (cf. section 1). Our approach goes beyond these by combining machine learning (LDA) with unsupervised multivariate analysis (PCA). We do not operationalize indicators for translationese or translation properties and test their usefulness in machine learning experiments (Volansky, Ordan, and Wintner 2015), but investigate the behavior of indicators derived independently of our translation-related research question (namely in the context of register studies). Finally, unlike studies based purely on machine learning, our analysis emphasizes the importance of visualization, especially since a direct interpretation of feature weights can be misleading (see section 4). Furthermore, visualization allows us to appreciate each data point individually rather than interpreting a summarized and thus inevitably idealized version of the data represented by means.

We use scatterplot matrices, as exemplified by Figure 1, to visualize high-dimensional vector spaces. Each panel in such a matrix shows a different two-dimensional perspective on the full space. In Figure 1, for example, the top-left, top-center and center panels display three side views of a three-dimensional cube. However, even trained analysts sometimes find it difficult to discern more complex structures that are not aligned with one or two of the dimensions, and overlapping data points in 2D plots further obscure important patterns. Therefore, we provide 3D animation videos as well as colored versions of some plots in an online supplement to this chapter at http://www.stefan-evert.de/PUB/EvertNeumann2016/. The animation for Figure 1 shows a 3D view of the first three PCA dimensions and rotates through the three side views seen in the scatterplot matrix.

3 Multivariate analyses

Following the procedure described in section 2, we begin with a Principal Component Analysis (PCA) in order to understand the overall geometric shape of the data set. Since it is unsupervised, PCA does not make use of any information on language, translation status or register of the texts, but these attributes can help to highlight structure in a visualization of the data set. Figure 1 shows the first four PCA dimensions in the form of a scatterplot matrix. Together, they account for R^2 = 41.9% of the variance of the data set, capturing major aspects of its overall structure. In the plot, German texts are represented by circles, English texts by crosses; originals are shown in black and translations in grey (a color version and animation can be found in the online supplement). The top-left panel, for example, shows the first PCA dimension on the vertical axis and the second dimension on the horizontal axis. The top center panel also shows the first dimension on the vertical axis, but the third PCA dimension on the horizontal axis.

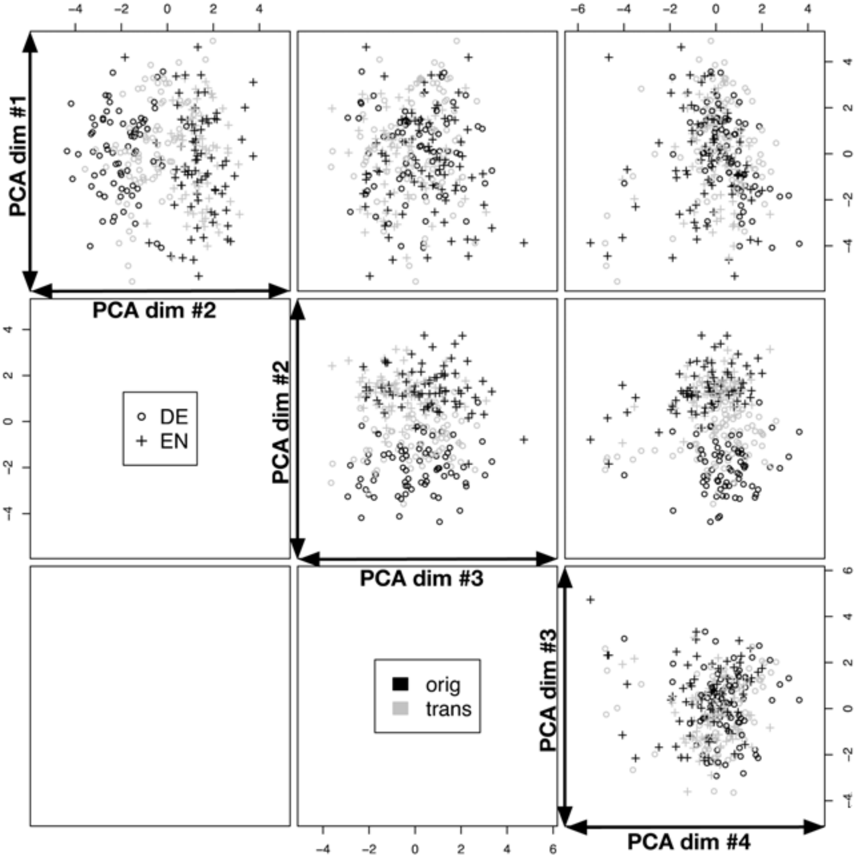


Figure 1: Scatterplot matrix showing the first four PCA dimensions

The main differences between German and English are captured by the second PCA dimension (horizontal axis of the top-left panel, vertical axis of the two panels in the middle row), which separates the two languages quite well (almost perfectly if translations are excluded). Dimensions 1 (vertical axis of top row) and 3 (horizontal axis of panels in center column) mainly account for register variation, as can be seen from the top-center panel of the register-coded scatterplot matrix in the online supplement. Dimension 4 separates some of the web texts, which appear to be markedly different from the rest of the corpus.

Figure 1 also shows that German translations are shifted towards the English side of the second PCA dimension, while English translations occupy the same range as English originals. This trend can be seen more clearly by plotting the distribution of texts from the four categories (Germans vs. English, original vs. translation) along this dimension. Figure 2 shows density curves, which can be thought of as smoothed histograms, with individual data points indicated by the marks at the bottom.

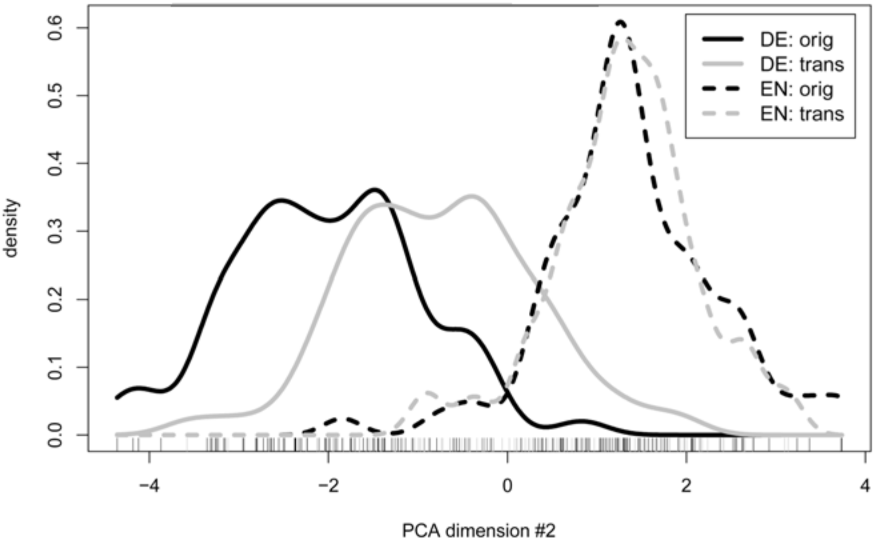


Figure 2: Distribution of texts along the second PCA dimension

The plot shows an identical distribution for English originals and translations (dashed lines on the right-hand side of the plot), while the German translations are shifted to the right compared to German originals (solid lines on the left-hand side of the plot). While there is more variability among the German texts – shown by a flatter and wider shape of their density curves – German translations and originals follow a very similar distribution, which is merely shifted for the translations. Focusing on the black curves, it is obvious that German and English originals are separated almost perfectly: original texts with a positive coordinate score are mostly English, those with a negative score are mostly German. The central range (roughly from −1 to +0.5) contains very few original texts, but a substantial number of translations into German: apparently, they tend to fall between the originals in both languages.

These observations strongly suggest a shining through effect for translation into the lower-prestige language (German), but not for the opposite translation direction. However, there are a number of issues that need to be taken into consideration before we can draw such a far-ranging conclusion. First, the four-dimensional projection on which our interpretation is based so far accounts for less than half of the total variance of the data (R^2 = 41.9%). While this is sufficient to give a general idea of the geometric shape of the data set, the remaining 58% – which are entirely invisible in Figure 1 – may contain further differences between German and English that put the observed shining through pattern in a different light. The characteristic differences between translations and originals that allow machine learning approaches to achieve high classification accuracy must also be hiding in these invisible orthogonal dimensions (especially for English, which shows no evidence for any form of translationese so far).

Second, there is still considerable variability along PCA dimension 2 within each language. In Figure 2, many of the German translations fall into a plausible envelope of variation for original German texts, so the observed shift cannot unambiguously be attributed to translation effects. One possible explanation are register divergences between the English and German originals. The German translations might simply represent sub-registers that are not covered by the German originals.

Third, the unsupervised PCA is based on the full data set containing both original and translated texts. It thus captures not only genuine differences between the two languages, but also translation effects, register divergences, etc. If dimension 2 is not based purely on the language contrast between English and German originals, the observed shift cannot directly be interpreted as a shining through effect. Let us clarify this point with a thought experiment: imagine that there is a dimension that captures the language contrast for original texts and a second, completely different dimension that captures a form of translationese introduced by the German translators which is independent of the source language. The PCA might have collapsed these two dimensions into a single axis, so that the shift of German translations from German originals reflects their position on the language-independent translationese dimension rather than actual, i.e. language-dependent, shining through.

In order to focus on the genuine language contrast, we apply supervised Linear Discriminant Analysis (LDA) between the German and English originals, temporarily excluding the translated texts. This procedure is in line with step 3 of Diwersy, Evert and Neumann (2014), adding a minimal amount of external information; since the learning algorithm is entirely unaware of the translations, there is no risk of biasing the results of the analysis with respect to the shining through hypothesis. The LDA discriminant aims to maximize the separation between German and English originals, while minimizing variability within each language at the same time. Speaking in geometric terms, the discriminant finds a perspective that reveals the most clearly articulated structure, resulting in a clear gap between the German and English originals. It does not account for all differences between the two languages, though, excluding weak tendencies towards higher or lower frequency in favor of characteristic properties separating the languages. As a consequence, the discriminant only captures 6.5% of the total variance of the data, compared to 11.1% for the second PCA dimension. We believe that this approach allows for a better interpretation with respect to the shining through hypothesis: any texts located in the gap between the two groups of originals have properties that are atypical of either language. Forms of translationese which are independent of the source language (type (iii) in section 1) are very implausible as an explanation for these observations. Note that our focus is not on disproving the existence of (universal) properties of translations, but rather on providing evidence for the existence of genuine shining through in translations. Type (iii) translationese may well exist in addition to shining through, but it does not explain the effect we found.

We can now carry out an orthogonal projection of all texts (both originals and translations) into the one-dimensional focus space defined by the discriminant. For visualization (step 4), the discriminant is extended with PCA dimensions from the orthogonal complement space in order to put the characteristic difference between German and English into perspective. The scatterplot matrix in Figure 3 shows that the characteristic difference between German and English – i.e. the spread of originals along the vertical axis in the top row – is noticeably smaller than register variation and other effects captured by the PCA dimensions – exemplified most clearly by the wider spread of data points in the panels of the middle row. A scatterplot matrix colored by register and a corresponding 3D animation can be found in the online supplement. Quantitatively, the LDA discriminant accounts for $R^2 = 6.5\%$ of the variance, compared to 15.6%, 8.1% and 7.9% for the first three PCA dimensions.

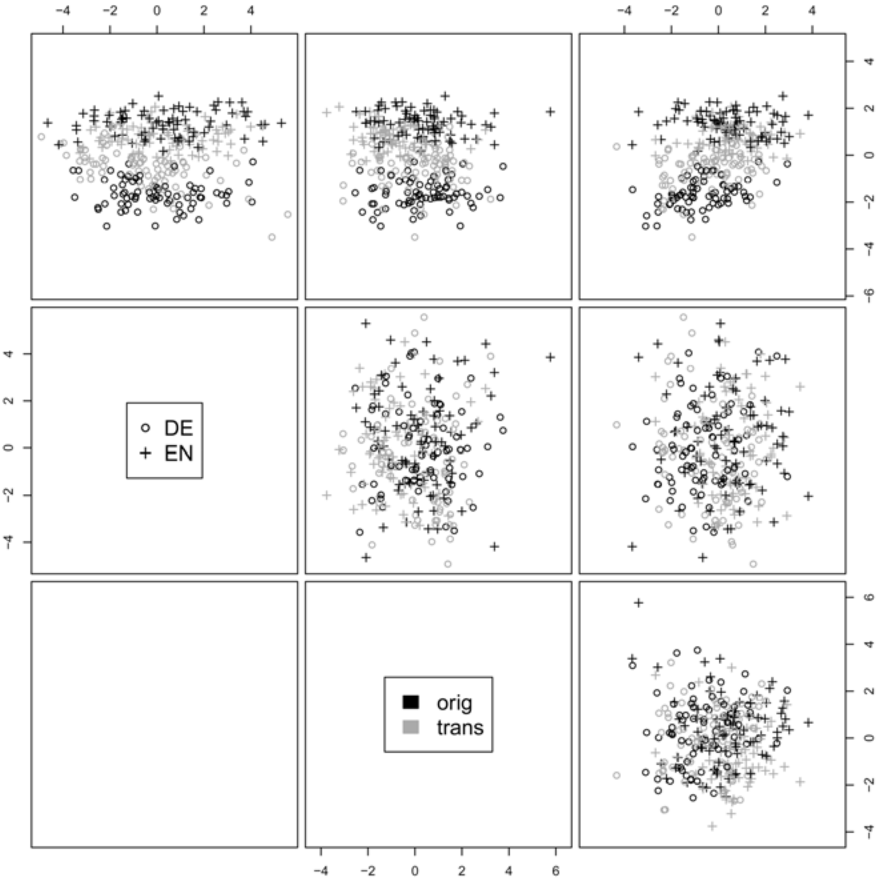


Figure 3: LDA discriminant for German vs. English originals (vertical axis of top row) with additional PCA dimensions from the orthogonal complement space

Focusing on the original texts (black points in Figure 3), we see a clear separation of German and English along the LDA discriminant, with only a few "outlier" texts in the gap region. This becomes even clearer in the online supplement where translations are shown in red. Translations in both languages (grey points) extend well into the gap, on the other hand, providing further evidence for a shining through effect, which seems stronger for translation from English into the less prestigious language German. As pointed out above, the LDA discriminant does not capture all differences between the German and English originals, since it focuses on bringing out the most distinctive structure. Dimension 4 (horizontal axis of top-right panel in Figure 3) shows a slight shift between German and English originals: most German originals (black circles) fall in a range from −4 to +2 on this dimension, whereas most English originals (black crosses) range from −3 to +3. However, variability within each language is much larger than along the LDA axis and the shift is a matter of degree rather than categorization. Like the second PCA dimension in the original analysis, it cannot be used to argue conclusively for or against the shining through effect.

Before taking a closer look at the distribution of translations along the LDA discriminant, we need to validate the supervised LDA (step 5 of Diwersy, Evert and Neumann 2014). We use ten-fold cross-validation to test whether the LDA axis is overfitted to the relatively small sample of 149 original texts. In each fold, 90% of the texts are used as training data to compute an LDA discriminant, and the remaining 10% are projected onto this dimension and classified as German or English. With a cross-validated classification accuracy of 97.3% (cf. the confusion matrix in Table 1), the distinction between German and English originals is excellent. Discriminant scores of the originals obtained by cross-validation correlate almost perfectly with the scores obtained by the single LDA on all 149 texts carried out above (Pearson correlation *r* = .989). This shows that it is valid to draw conclusions about the language contrast and shining through from the LDA dimension in Figure 3.

Table 1: Confusion matrix for cross-classification of originals in LDA

LDA prediction	**true category**	
	German	**English**
German	68	1
English	3	77

For a linguistic interpretation of the LDA discriminant, the feature weights will play a central role. Our findings are only meaningful if these weights are not affected by individual texts in the data set. We can quantify the robustness of feature weights by computing the angle between the full-data LDA and each of the ten LDA discriminants obtained from the cross-validation procedure (see Table 2). With an average angle of 9.9 degrees, there is some "wobble" in the LDA dimension, but the general direction of the vector of feature weights remains stable.

Having confirmed the validity and stability of the LDA discriminant, we can now interpret it as a characteristic difference between English and German originals. Because of the low variability within each language any texts that fall outside these relatively narrow bands have to be considered markedly non-German or non-English. If this holds for translations, these texts exhibit feature

Table 2: Angle between LDA discriminant from each cross-validation fold and the full-data discriminant

	fold 1	**fold 2**	**fold 3**	**fold 4**	**fold 5**	**fold 6**	**fold 7**	**fold 8**	**fold 9**	**fold 10**
angle	17.6°	14.6°	7.0°	9.7°	4.9°	9.0°	5.5°	11.3°	10.9°	8.8°

patterns that are atypical of the target language, deviating towards typical patterns of the source language: a clear case of shining through. The top-left panel in Figure 3 already gives a strong indication that this may in fact be the case, Figure 4 displays the distribution of texts along the LDA discriminant in order to confirm this impression.

There is a clear shining through effect for both translation directions, which is more pronounced for translation into German. Note that the two small peaks at the left-hand side of the density curves for German translations are caused by two texts from the web register. Disregarding such individual outliers, the distribution of translations is similar to the distribution of originals in the same language, but shifted by a certain amount towards the source language. The black curves show that German and English originals are separated perfectly by the LDA discriminant (without cross-validation). There is a clearly visible gap at the center that contains hardly any original texts. By contrast, a substantial proportion of the translations (grey curves) are located in this gap and thus are clearly different from originals in either language.

These visual impressions now have to be confirmed with a quantitative evaluation (step 6 of Diwersy, Evert, and Neumann 2014). The shift between originals and translations is validated by Student's t-test for independent samples, which shows highly significant shining through in both languages (German: $t = 9.2378$, df = 141.54, $p = 3.4 \times 10^{-16}$; English: $t = -6.6111$, df = 145.83, $p = 6.7 \times 10^{-10}$). The effect size (Cohen's d) is 1.5 standard deviations for German, but only 1.1 standard deviations for English, confirming the asymmetry of the effect. Note that the discrepancy between German and English may appear much larger visually, but the higher variability of the German data reduces the relative effect size.

The real test of the shining through hypothesis is whether it is able to account, at least in part, for the marked difference between originals and translations found by supervised machine-learning experiments; i.e., whether we can discriminate between originals and translations based on their LDA scores. Note that the LDA dimension is not overtrained for this purpose because it was determined exclusively based on the originals, without any knowledge about the translated texts. Close inspection of Figure 4 suggests that LDA scores below −1.1

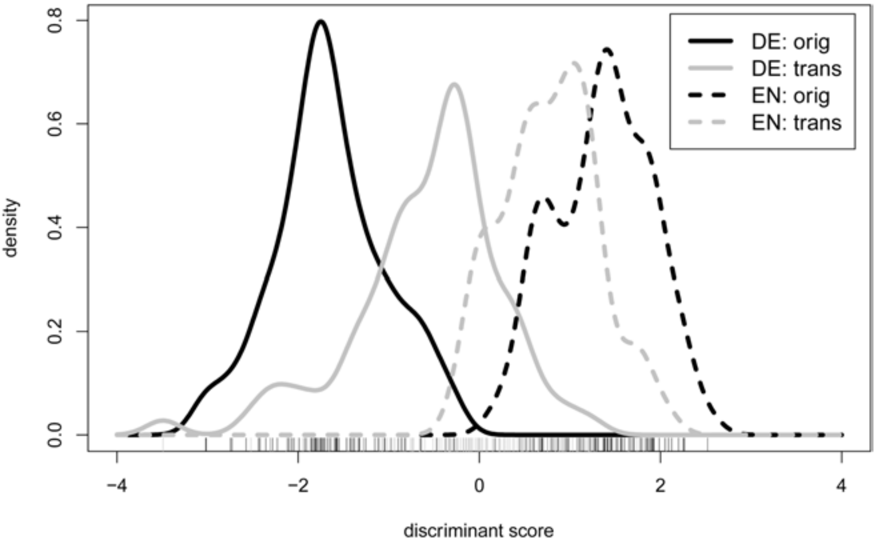


Figure 4: Distribution of texts along the LDA discriminant for German vs. English originals

indicate German originals, scores between –1.1 and +1.3 indicate translations (both into German and into English), and scores above +1.3 indicate English originals. A classifier using these manually determined thresholds is able to distinguish between originals and translations with 76.8% accuracy, which compares favorably against results reported in the literature (e.g. Baroni and Bernardini 2006), especially considering that those classifiers include translations as supervised training data. In order to exclude the possibility that our thresholds may be overfitted to the data set, we carry out ten-fold cross-validation, using a support vector machine (SVM) with quadratic kernel to select thresholds in each fold. This results in a classification accuracy of 75%–77%, depending on the random split into folds.

We have thus established a clear case of shining through and consequently ruled out other forms of translationese (see section 1), but there are still two possible explanations for this effect: Rather than showing genuine, i.e. language-specific shining through, the effect could be caused by individual, i.e. text-specific shining through. Note that individual shining through does not necessarily imply that translations are inherently different from originals. The LDA discriminant may have picked up incidental differences between the source texts in the two languages (e.g. because they were sampled from authors with different styles or because of register divergence) that are preserved in the translation and reflected by the shifts in Figure 4.

In order to test whether individual shining through is plausible, we compare the LDA scores of source and target texts in aligned text pairs. If there is individual shining through, we should find a strong correlation between the source and target text. For example, a German text with a very low LDA score should be translated into an English text with a relatively low LDA score and fall into the gap between the originals. A less typically German text with a relatively high LDA score should be translated into an English text with a very high LDA score, overlapping with the English originals. For genuine shining through, this is not the case: a translation tends to exhibit properties of the source language, but its particular LDA score does not depend on the corresponding original text and its LDA score.

Figure 5 and Figure 6 visualize the correlation between source and target texts. Each point represents a text pair: its horizontal position corresponds to the LDA score of the source text, and its vertical position to the LDA score of

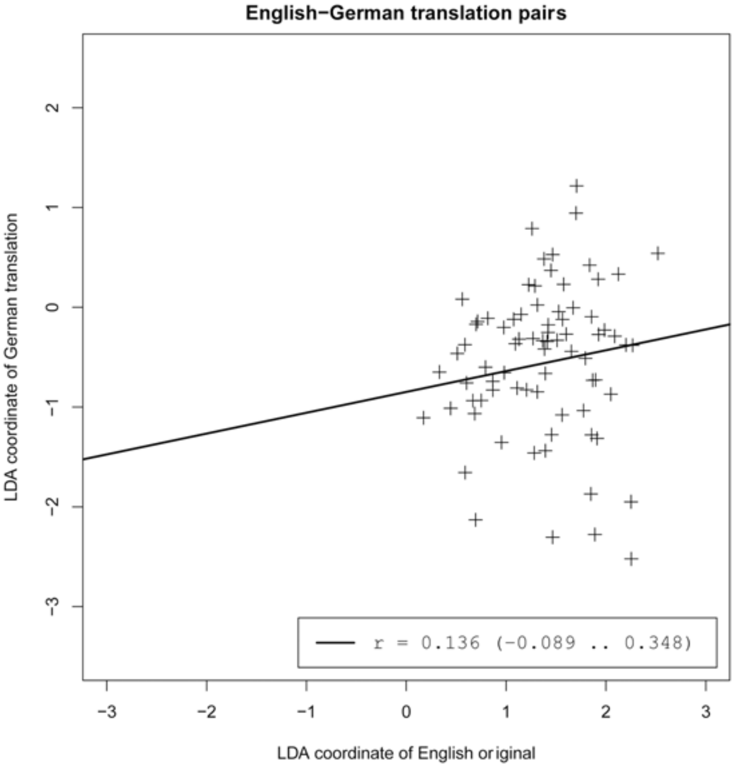


Figure 5: Correlation between LDA score of English originals (x-axis) and their German translations (y-axis), with regression line

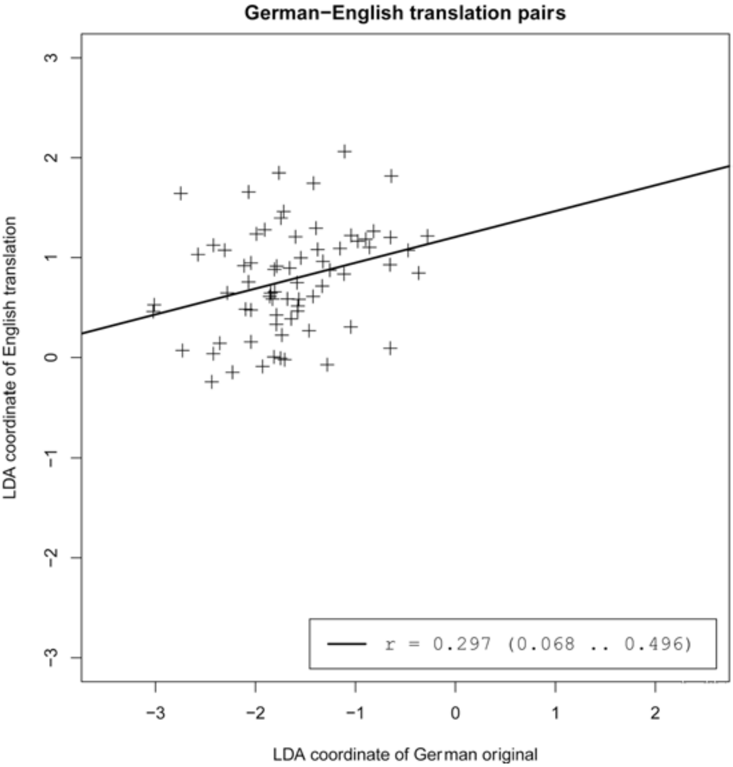


Figure 6: Correlation between the LDA scores of German originals (x-axis) and their English translations (y-axis), with regression line

the target text. If there is a strong correlation, the points should cluster along a diagonal line. The plots show a difference between the two translation directions, which simply reflects the different ranges on the LDA discriminant occupied by English and German originals (x-axis) as well as English and German translations (y-axis). However, there is no significant correlation between English originals and their German translations (Figure 5; note that the confidence interval includes the possibility of no correlation, *r* = 0), and only a weak, marginally significant correlation for the opposite translation direction (Figure 6). Therefore, individual shining through of any kind can be ruled out with high confidence.

Similar plots for the complement PCA dimensions (not shown here for space reasons) show strong evidence for individual shining through. This does not come as a surprise because the complement PCA dimensions mainly capture register variation, which we expect to be preserved in the translation (e.g., a popular science text should be translated into a text from the corresponding target register rather than an entirely different register). However, the correlation is much stronger than can be explained merely by register effects, in particular along the first complement PCA dimension (dimension 2 in Figure 3). We interpret this as evidence for individual shining through of linguistic properties of the source texts that are related to register and style, but are orthogonal to the contrast between the two languages and thus independent from the genuine shining through effect.

Having established a clear type (i) shining through effect in the LDA dimension and verified it with a quantitative evaluation, we can now proceed to the linguistic interpretation and general discussion of our findings.

4 Interpretation of the discriminant

The first step of the linguistic discussion is to determine which lexico-grammatical indicators contribute to the LDA discriminant and hence the observed shining through effect (step 7 of the procedure described in section 2.2). The traditional interpretation of latent dimensions in multivariate studies (e.g. Biber 1988 and related work) focuses on feature weights – as shown in Figure 7 for our LDA discriminant – and typically applies a cutoff threshold, disregarding features with absolute weights below the threshold.

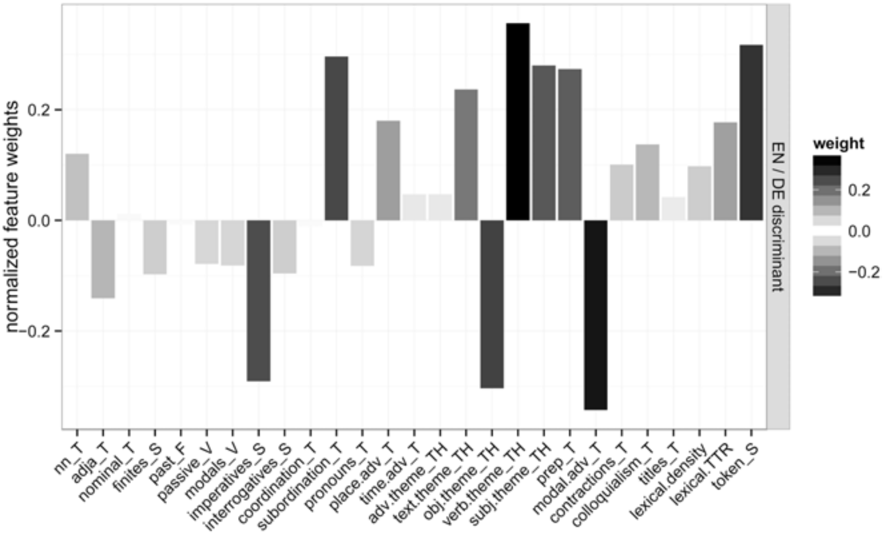


Figure 7: Feature weights contributing to the LDA discriminant (normalized for orthogonal projection)

At face value, positive weights indicate features that are characteristic of English originals (since the English originals are positioned on the positive side of the discriminant axis) and negative weights indicate features that are characteristic of German originals. The traditional interpretation would thus conclude that English originals are characterized by high proportions of textual themes[4], verbal themes, subject themes, subordinations and place adverbs, as well as long sentences (tokens / S) and a high lexical type-token ratio (TTR). German originals are characterized by high proportions of object themes, modal adverbs and imperatives. While such an interpretation may be acceptable for the first few PCA or FA dimensions with their strong correlational patterns, it does not do full justice to the multivariate nature of the analysis because each feature is assessed independently as an indicator of English or German. In our case, this amounts to little more than a traditional univariate language comparison. Consider the boxplots in Figure 8, which show the contribution each feature makes to the positions of texts on the LDA axis (i.e. standardized feature values multiplied by the corresponding feature weights), separately for German and English originals. A feature with a positive contribution pushes texts to the English side of the axis, a feature with a negative contribution pushes them to the German side of the axis. Note that positive contributions correspond to above-average feature values if the feature weight is positive, but to below-average feature values if the weight is negative (indicated by "(–)" in front of the feature name).

Diamond symbols indicate the average contribution of each feature to the positions of German and English originals, respectively. The further its two diamonds are apart, the more a feature pushes the English and German texts away from each other. However, this does not necessarily mean that the feature improves the discrimination between the two groups: it also adds within-group variability, indicated by boxes and whiskers around the diamonds in the plot. Several features have a very strong effect, including object themes, modal adverbs, subordinations and sentence length (token / S). Other features have a much smaller effect (e.g. textual, verbal and subject themes) or hardly any effect at all (imperatives) despite their large weights. Only one feature (object themes) is highly discriminative by itself, i.e. the boxes for German and English do not overlap: with very few exceptions, only German texts allow themes to be realized as objects. Modal adverbs, prepositions, subordinations and sentence length also contribute well to the language discrimination, while features such as textual, verbal and subject themes seem to add primarily to the within-group

[4] Note that only the first element in the sentence is analysed as the theme.

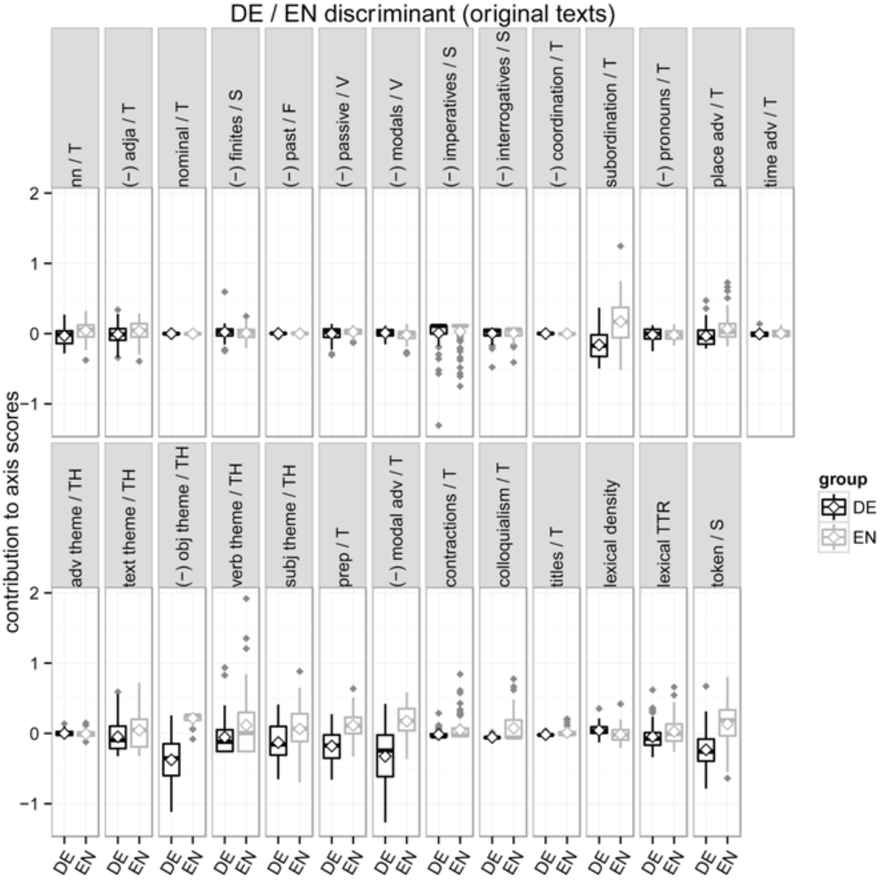


Figure 8: Boxplots showing the contribution of each feature to the position of German and English originals on the LDA discriminant

variability. Two features (lexical density and modals) even have a counter-intuitive effect: they nudge English originals towards the German side of the discriminant and vice versa. These observations show that an interpretation in terms of feature weights is too simplistic and can be outright misleading in some respects.

As we have already pointed out in section 1, multivariate analysis assumes that features are multi-functional, i.e. they reflect a mixture of several systematic text properties. Ideally, the linguistic interpretation should focus on such underlying properties rather than individual lexico-grammatical indicators, determining which properties account for the language contrast and have thus been

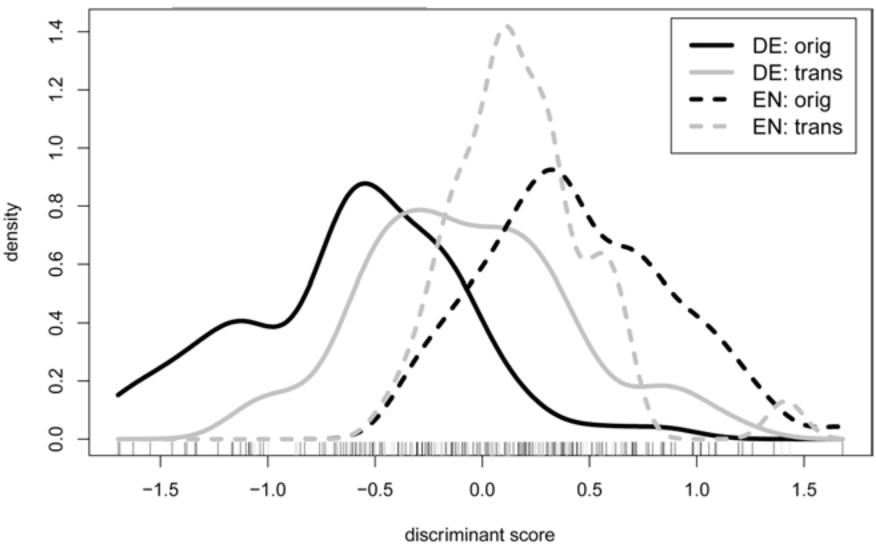


Figure 9: Distribution of texts along a simplified discriminant that represents characteristic patterns in the realization of themes

found to “shine through” into the target language. The LDA discriminant provides an excellent starting point for this purpose. Since it aims to minimize within-group variability (i.e. among the originals in each language), feature weights are adapted so that the effects of other, irrelevant text properties are cancelled out. This also explains why some features that have a small effect on the separation of the two groups but large within-group variability (e.g. textual and verbal themes) have nonetheless been included in the discriminant: their main purpose is to help cancel out irrelevant properties.

Due to this complex interplay between features in the underlying structure of the data set a detailed discussion of individual lexico-grammatical indicators will not be attempted here, with one exception. Prompted by the high discriminativity of the single feature object themes, we defined a simplified discriminant based on the four theme-related features with high LDA weights: object, subject, textual and verbal themes. If our assumptions hold true, this discriminant should represent patterns of theme realization that are characteristic of English and German texts, respectively.

Figure 9 displays the distribution of originals and translations along the simplified discriminant. There is still a significant shining through effect, which is stronger for translations from English into German (Cohen’s d = −1.08 for German vs. d = +0.65 for English). However, the originals are no longer clearly separated (88% classification accuracy) by the new discriminant. As a result, the translations are not located in a gap that would mark them as clearly distinct from originals in either language. Classification accuracy for translation status is reduced from 76% (for the full discriminant) to 61% (for the theme-related discriminant).

Our conclusion from these observations is that the realization of themes plays an important role in the language contrast between English and German. It is a major factor behind the observed shining through effect. Low classification accuracy shows that this picture is only partial. A full understanding of the LDA discriminant and shining through can only be achieved by exploring the genuinely multivariate patterns of correlations and interactions between the individual features. This is beyond the scope of the present chapter, however, and will be addressed in future work.

5 Discussion

Linking our findings to the general discussion in corpus-based translation studies about the character of translation properties, we might ask whether the observations might not support a more generalized claim that shining through is a *universal* feature of translation.[5] The quantitative validation confirmed by the t-test suggests that, at least in the language pair English-German, translated texts can be systematically separated from non-translated texts. This lends additional support to the results obtained by computational studies of translationese (see section 1), now based on more informative lexico-grammatical indicators. Moreover, the interpretation of the visualizations showed that translations in general tend to orient towards the target language, but are still distinctively different in their tendency to accommodate features of the source language.

This finding would support the universals hypothesis. However, the analysis also revealed differences in effect size for the two translation directions thus contradicting this hypothesis because it would require comparable results for both directions. This does not only let the universals hypothesis appear implausible but also makes parallel activation of both language systems and consequently a similar context as in L2 writing (see section 1) less likely because this scenario, too, would require the effect to be similar in both translation directions. Rather, we have to find additional factors that explain the differential

[5] As laid out in section 1 we include shining through as one of the properties of translation thus opposing to Baker's (1993) exclusion of source language interference.

situation for both translation directions. The fact that the effect is stronger for German translations than for English translations can be tentatively interpreted in terms of the differences in prestige discussed by Toury (2012). Note, however, that our study did not test for prestige so that this is just one possible factor that could explain why the translations into German seem to accommodate more characteristics of English as the source language than translations in the opposite direction. The influence of additional factor(s) also provides an argument why the universals hypothesis cannot be upheld. The translator works in too complex a context in which a whole range of factors influences the specific outcome of the translation process. These will interact in various ways depending on their respective strength. At the same time, this finding also further corroborates our initial claim that L2 writing and translation into the L1 are likely to yield different effects in terms of interference. Incomplete learning of the L2 can be assumed to be an important factor in writing in the foreign language, however, this is a less likely factor for translation – at least into the L1. By the same token, diverging prestige of the languages involved is a plausible explanation for the directionality effect in translation, but cannot be assumed to be a cause of L1 interference in L2 writing.

The analysis in section 3 focused on shining through. Nevertheless, we might also be interested in other properties. Hansen-Schirra and Steiner (2012) describe normalization as being linked to shining through on an assumed norm continuum. Consequently, our study should also reveal this property. Over-normalization, the exaggeration of target language norms, could have become observable in the visualizations if, for instance, the translations had been located on the remote side of the target language originals. While the exact definition of normalization is still a matter of debate (e.g. would not perfect alignment with target language norms be exactly what one would expect?), our study did not yield clear indications of the generalized type of normalization. This could tentatively be interpreted as a reduced importance of generalized normalization, but clearly requires more in-depth analyses in future work. Note that normalization would also be observable, if only part of the translations, say from a register which is particularly prone to covert translation, were located in the expected area. This would be in line with Delaere's (2015) evidence for register-specific target language orientation.

Levelling-out refers to the tendency of translations to converge towards unmarked features at the expense of more marked features that are observable in non-translated texts (Baker 1996). The methodology of this study would also allow us to observe levelling-out, but our experiments did not reveal any notable indications for this assumed property. Cursory examination of individual registers in the PCA dimensions suggests that some registers might display levelling-out, but, again, this has to be relegated to future work. The contentious properties of explicitation and simplification are difficult to investigate with our research design. They could be indirectly included in patterns of shining through, but would probably have to be investigated on the basis of dedicated operationalizations which in turn lead to a risk of circularity in the investigation.

What is the contribution of our study to (corpus-based) translation studies beyond what has already been shown by univariate studies of individual features and registers? Previous studies used differential linguistic features in order to operationalize properties such as shining through. By contrast, this study focused on features that are actually comparable across the two languages involved. Consequently, the study could have very well produced a quite different outcome showing, for instance, systematic normalization rather than shining through. It provides evidence for the intricate interplay between linguistic features: the overall pattern in the data emerges from a complex combination of features suggesting that findings based on the (cumulative) interpretation of individual features may lead to spurious results that could be counteracted by other features not included in the study. Moreover, our study shows that similar distributional patterns apply across registers, even though we also obtained indications of register-specific behavior in higher PCA dimensions. This will have to be examined in more detail on the basis of a broader coverage of texts and registers in future work.

The results are also of interest from a contrastive linguistics perspective, providing multivariate evidence that the difference between two languages is not only observable in features that only exist in one language but also emerges from the distributional patterns of comparable features.

Against this background, the study also complements claims about the assumed obligatory character of shifts due to contrastive differences. The shining through effect established in our study shows that translators do not necessarily adjust for differences between languages that only consist in usage preferences of comparable features, i.e. differences in their frequencies. In such cases, they do not always adapt the text to match the usage preferences of the target language (see section 1 and Teich 2003).

While the results of our study look very promising, there are also some clear limitations. As is usual in multivariate analyses, the choice of features and texts heavily impacts the results. This requires eliminating correlated features, computing relative frequencies with respect to appropriate units of measurement as well as avoiding features which cannot be quantified in the same way as the ones discussed here. Especially lexical features, which nevertheless shed light on language variation, can only be included in a quantified, i.e. abstracted form (e.g. in the form of lexical density). More specifically, an analysis of the type presented here requires a large number of lexico-grammatical features, which should be as informative as possible and which need to be extracted in a rather costly procedure. Drawing on automatic analyses makes the extraction of data more efficient but comes at the price of inheriting the inaccuracies of the annotation tools. While each step of the analysis involved great care to ensure reliable data – from selecting appropriate tools in Hansen-Schirra, Neumann and Steiner (2012) and establishing comparability of the features in Neumann (2013) to representing the data in our multivariate analysis – the final selection of features is still prone to undue influences. The results reported here need to be read against this background.

Text selection as well as number of texts included in the corpus play an important role. The inherent circularity of sampling texts to be representative of a given set of registers is an important issue that limits the outcome of our analysis. Furthermore, a well-known problem of comparable corpora is the assumption that comparable texts are indeed from comparable registers where, in fact, registers may be slightly diverging. One way of improving this situation is to carry out an annotation of the registers based on external parameters as shown by Delaere, De Sutter and Plevoets (2012). While this does not remedy the potential incomparability between registers in a bilingual corpus, it does facilitate the analysis of the incomparable registers because it helps narrowing down the exact area(s) in which the registers diverge. Furthermore, texts from extreme registers may obscure the behavior of the bulk of the corpus. This was shown to be the case in Diwersy, Evert and Neumann (2014). It is possible to mitigate the effect by eliminating outlier registers, as we did in the work reported here. However, this is not a perfect solution either.

Standardization was claimed to be essential to our approach, so that each feature is given the same weight regardless of the scale of numerical values (see section 2). However, to some extent this may also be a problem because it may increase the influence of individual features. We may be overemphasizing the importance of features that have relatively little variability in language. Passive may, for instance, be relatively frequent across the board; small differences between individual texts are then exaggerated by our approach.

6 Conclusion and outlook

In this chapter, we hope to have shown the intricate interplay between languages as well as originals and translations that emerges from the interpretation of latent dimensions of multivariate analysis. More specifically, we reported evidence for a generalized shining through effect of the source language in a corpus of originals and translations from the language pair English-German. To this end, we used a sequence of steps consisting of PCA, LDA, visualization and cross-validation. The interpretation of the analyses relied heavily on inspecting visualizations that proved to be very informative, throwing light especially on the assumed directionality effect that gives the chapter its title. One of our main results is that shining through manifests to differing degrees in the two translation directions, suggesting a tentative interpretation in terms of diverging prestige of the two languages involved.

This hypothesized role of prestige is one of many things that should be examined in more detail in future work. In addition to the aspects already mentioned in the previous sections, this also includes further investigating patterns that might emerge for other translation properties and an in-depth look at the interpretation of feature weights. Given the limitations of the study in terms of text and feature selection, repetition of this analysis on a different corpus such as the Dutch Parallel Corpus (Macken et al. 2011) would further support our exploratory findings.

We believe that the multivariate approach adopted here is not only very useful for understanding the nature of translations – because it supports the simultaneous investigation of a whole range of features that might affect the make-up of translations – but is also very promising for various other areas of the study of language variation.

Acknowledgments

The authors would like to thank Sascha Diwersy for extensive discussions of this research, which builds on joint work reported in Diwersy, Evert, and Neumann (2014). We would also like to thank the audience at the workshop “New ways of analyzing translational behavior in corpus-based translation studies” as well as the reviewers for valuable comments. Part of the research reported here was funded by the German Research Council under grants no. STE 840 / 5-2 and HAN 5457 / 1-2.

References

Baker, M. 1993. Corpus Linguistics and Translation Studies. Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.), *Text and technology. In honour of John Sinclair*, 233–250. Amsterdam: John Benjamins.

Baker, M. 1996. Corpus-Based Translation Studies: The challenges that lie ahead. In H. Somers (ed.), *Terminology, LSP and translation. Studies in language engineering in honour of Juan C. Sager*, 175–186. Amsterdam: John Benjamins.

Baroni, M. & S. Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21(3). 259–274.

Becher, V. 2011. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts*. Hamburg: Universität Hamburg doctoral dissertation.

Becher, V., J. House & S. Kranich. 2009. Convergence and divergence of communicative norms through language contact in translation. In K. Braunmüller & J. House (eds.), *Convergence and divergence in language contact situations*, 125–152. Amsterdam: John Benjamins.

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

Delaere, I. 2015. *Do translators walk the line? Visually exploring translated and non-translated texts in search of norm conformity.* Ghent: Ghent University doctoral dissertation.

Delaere, I., G. De Sutter & K. Plevoets. 2012. Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target. International Journal of Translation Studies* 24(2). 203–224.

Diwersy, S., S. Evert & S. Neumann. 2014. A weakly supervised multivariate approach to the study of language variation." In B. Szmrecsanyi & B. Wälchli (eds.), *Aggregating dialectology, typology, and register analysis. Linguistic variation in text and speech*, 174–204. Berlin/New York: Mouton de Gruyter.

Hansen-Schirra, S., S. Neumann & E. Steiner. 2007. Cohesive explicitness and explicitation in an English-German translation corpus. *Languages in contrast* 7(2). 241–265.

Hansen-Schirra, S., S. Neumann & E. Steiner. 2012. *Cross-linguistic corpora for the study of translations – Insights from the language pair English-German*. Berlin: Mouton de Gruyter.

Hansen-Schirra, S. & E. Steiner. 2012. Towards a typology of translation properties. In S. Hansen-Schirra, S. Neumann & E. Steiner (eds.), *Cross-linguistic corpora for the study of translations – Insights from the language pair English-German*, 255–279. Berlin: Mouton de Gruyter.

House, J. 2002. Maintenance and convergence in translation – Some methods for corpus-based investigations. In H. Hasselgård, S. Johansson, B. Behrens & C. Fabricius-Hansen (eds.), *Information structure in a cross-linguistic perspective*, 199–212. Amsterdam: Rodopi.

Kruger, H. & B. van Rooy. 2012. Register and the features of translated language. *Across Languages and Cultures* 13(1). 33–65.

Macken, L., O. De Clercq & H. Paulussen. 2011. Dutch Parallel Corpus: A balanced copyright-cleared parallel corpus. *Meta: Journal des traducteurs* 56(2). 374–390.

Mauranen, A. & P. Kujamäki, eds. 2004. *Translation Universals. Do They Exist?* Amsterdam: John Benjamins.

Mauranen, A. 2004. Corpora, universals and interference. In A. Mauranen & P. Kujamäki (eds.), *Translation universals. Do they exist?*, 65–82. Amsterdam: John Benjamins.

Neumann, S. 2013. *Contrastive register variation. A quantitative approach to the comparison of English and German*. Berlin: Mouton de Gruyter.

Oakes, M.P. & M. Ji (eds.). 2012. *Quantitative methods in corpus-based translation studies. A practical guide to descriptive translation research*. Amsterdam: John Benjamins.

Olohan, M. & M. Baker. 2000. Reporting *that* in translated English. Evidence for subconscious processes of explicitation?. *Across Languages and Cultures* 1(2). 141–158.

Serbina, T. 2013. Construction shifts in translations: A corpus-based study. *Constructions and Frames* 5(2). 168–91.

Steiner, E. 2008. Empirical studies of translations as a mode of language contact – 'Explicitness' of lexicogrammatical encoding as a relevant dimension. In P. Siemund & N. Kintana (eds.), *Language contact and contact languages*, 317–346. Amsterdam: John Benjamins.

Steiner, E. 2012a. A characterization of the resource based on shallow statistics. In S. Hansen-Schirra, S. Neumann & E. Steiner (eds.), *Cross-linguistic corpora for the study of translations – Insights from the language pair English-German*, 71–89. Berlin: Mouton de Gruyter.

Steiner, E. 2012b. Generating hypotheses and operationalizations. The example of *explicitness/explicitation*. In S. Hansen-Schirra, S. Neumann & E. Steiner (eds.), *Cross-linguistic corpora for the study of translations – Insights from the language pair English-German*, 55–70. Berlin: Mouton de Gruyter.

Teich, E. 2003. *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. Berlin/New York: Mouton de Gruyter.

Toury, G. 2012. *Descriptive Translation Studies – and beyond: Revised edition*. 2nd ed. Amsterdam: John Benjamins.

Volansky, V., N. Ordan & S. Wintner. 2015. On the Features of Translationese. *Digital Scholarship in the Humanities* 30(1). 98–118.

Appendix

The linguistic features in alphabetical order

(cf. Neumann (2013), see Hansen-Schirra, Neumann and Steiner (2012, chapter 3) for a full description of the annotation referred to here)

adja_T: attributive adjectives, per no. of tokens

English: all tokens receiving the part-of-speech tag "JJ.*" (general adjective), computed as the proportion of all tokens per text.

German: all tokens receiving the part-of-speech tag "ADJA" (attributive adjective), computed as the proportion of all tokens per text.

colloquialism_T: colloquialisms per no. of tokens

English: all strings like *yeah, bloody, damned, bitch, sissy, crap, buddy* etc., computed as the proportion of the total number of tokens per text

German: all strings like *toll, spitze, geil, bekloppt, bescheuert, Weichei, Blödmann, Klamotten* etc., computed as the proportion of the total number of tokens per text.

contractions_T: contractions per no. of tokens

English: all strings like *'m, 's, 't* etc. and, where applicable, a part-of-speech tag like "P.*" (pronoun) followed by a string like *'s, 'll* etc., computed as the proportion of the total number of tokens per text.

German: all strings like *gibts, willste, biste, guck, kuck, mal, drauf, runter, sagste, rüber, aufs, ums, nebens, ʻn* etc., computed as the proportion of the total number of tokens.

coordination_T: coordinating conjunctions, per no. of tokens

English: all tokens receiving the part-of-speech tag "CC" (coordinating conjunction), computed as the proportion of the total number of tokens per text.

German: all tokens receiving the part-of-speech tag "KON" (coordinating conjunction), computed as the proportion of the total number of tokens per text.

finites_S: finite verbs, per no. of sentences

English & German: all items receiving the tag for finite verb (chunk_gf = "fin") in the manual grammatical annotation, computed as the proportion of the total number of sentences per text.

imperatives_S: imperative mood, per no. of sentences

English: all sentences starting with the part-of-speech tag "VV0" (manually verified), computed as the proportion of all sentences per text

German: all sentences (manually verified) starting with the part-of-speech tag "VVIMP" (for the German imperative verb mood), and "VVFIN" ending on *-en* (for the plural form) followed by the personal pronoun *Sie* (for polite imperatives), and the part-of-speech tag "VV.*" ending on *-n* at the end of a sentence (represented by a punctuation mark) as the only verb in the sentence, computed as the proportion of all sentences per text.

interrogatives_S: interrogative mood, per no. of sentences

English & German: all sentences (manually verified) ending with a question mark, computed as the proportion of all sentences per text.

lexical.density: lexical density

English & German: all lemmatized items assigned a part-of-speech tag for nouns, full verbs, adjectives and adverbs computed as the proportion of the total number of tokens per text.

lexical.TTR: lexical type token ratio

English & German: all lemmatized items assigned a part-of-speech tag for nouns, full verbs, adjectives and adverbs computed as the proportion of the total number of items assigned a part-of-speech tag for noun, full verb, adjective and adverb tokens per text.

modal.adv_T: modal lexis per no. of tokens

English: all strings *very, highly, fully, completely, extremely, entirely, strongly, totally, perfectly, absolutely, greatly, altogether, thoroughly, enormously, intensely,* *utterly, only, almost, nearly, merely, hardly, slightly, partly, practically, somewhat, partially, scarcely, barely, mildly, just, really, most, more, quite, well, anyway, anyhow* in combination with the part-of-speech tag “R.*” (all adverbs), computed as the proportion of the total number of tokens per text.

German: all strings *sehr, ziemlich, recht, ungewöhnlich, höchst, außerordentlich, ziemlich, fast, nahezu, ganz, aber, vielleicht, denn, etwa, bloß, nur, mal, nun, nunmal, eben, ruhig, wohl, schon, ja, doch, eigentlich, auch, lediglich, allein, ausschließlich, einzig, ebenfalls, ebenso, gleichfalls, sogar, selbst, gerade, genau, ausgerechnet, insbesondere, erst, schon, noch* in combination with the part-of-speech tag “ADV” (adverb) and where applicable following the “V.*FIN” (finite verb), computed as the proportion of the total number of tokens per text.

modals_V: modal verbs per no. of verbs

English & German: all items receiving the part-of-speech tag “VM.*” (modal verb), computed as the proportion of the total number of verbs per text.

nn_T: nouns per no. of tokens

English & German: all items receiving the part-of-speech tag “N.*” (all nouns), computed as the proportion of all tokens per text.

nominal_T: nominalizations per no. of tokens

English: all tokens receiving the part-of-speech tag “N.*” and ending on *-ion, -ism, -ment, -ness* and the respective plural endings, computed as the proportion of all tokens per text.

German: all tokens receiving the part-of-speech tag “N.*” and ending on *-ung, -heit, -keit, ismus* and their respective plural endings, computed as the proportion of all tokens per text.

passive_V: passive voice, per no. of verbs

English: the results for the query for the part-of-speech tag “VB.*” followed by “VVN” (manually verified) with up to 3 intervening tokens, computed as the proportion of the total number of verbs per text.

German: the results for the query for strings of the auxiliary *werden* followed (or preceded) by “VVPP” (manually verified) with up to 8 intervening tokens, computed as the proportion of the total number of verbs per text.

past_F: past tense, per no. of finite verbs

English: all items receiving the tag for finite verb (chunk_gf=“fin”) in the manual grammatical annotation in combination with the part-of-speech tag “V.D.*” anywhere within the chunk, computed as the proportion of the total number of finites per text.

German: all items receiving the tag for finite verb (tns="past") in the morphology annotation, computed as the proportion of the total number of finites per text.

place.adv_T and time.adv_T: place and time adverbs, per no. of tokens

English: all tokens receiving the part-of-speech tag "RL" (adverb of place or direction) and "RT" (adverb of time), computed as the proportion of all tokens per text

German: all strings (and their variants in upper case) *hier, da dort, oben, unten, rechts, links, vorn.*, hinten, vor, dorthin, herab, herbei, dahin, jenseits, hinein, hierhin, hinunter, hierher; heute, jetzt, zuletzt, bald, sofort, morgen, derzeit, einst, früher, gestern, später, heutzutage, soeben, kürzlich, nachher, demnächst, jüngst, vorgestern, unlängst*, etc., computed as the proportion of all tokens per text.

prep_T: prepositions per no. of tokens

English: all items receiving the part-of-speech tag "I.*" (all prepositions), computed as the proportion of all tokens per text.

German: all items receiving the part-of-speech tag "AP.*" (all prepositions), computed as the proportion of all tokens per text.

pronouns_T: personal pronouns, per no. of tokens

English: all strings *I, me, mine, you, yours, he, him, his, she, her, hers, it, we, us, ours, they, them, theirs* in combination with the part-of-speech tag "PP.*" (all personal pronouns), computed as the proportion of all tokens per text.

German: all strings *ich, mir, mich, du, dir, dich, er, ihm, ihn, sie, ihr, ihn, es, wir, uns, euch, ihnen, Sie, Ihnen* in combination with the part-of-speech tag "PPER" (personal pronoun), computed as the proportion of all tokens per text.

subordination_T: subordinating conjunctions, per no. of tokens

English: all tokens receiving the part-of-speech tag "CS.*" (subordinating conjunction), computed as the proportion of the total number of tokens per text

German: all tokens receiving the part-of-speech tag "KOU.*" or "KOKOM" (subordinating conjunction), computed as the proportion of the total number of tokens per text.

adv.theme_TH, obj.theme_TH, subj.theme_TH, text.theme_TH and verb.theme_TH: specific themes per no. of themes

English & German: the first grammatical function in each sentence (adv.theme: all adverbials, obj.theme: all objects and predicatives, subj.theme: subjects, text.theme: conjunctions and other types of connectives, verb.theme: verbs) in the manual grammatical annotation, computed as the proportion of the total number of themes per text.

titles_T: titles per no. of tokens

English: all strings like *Doctor, Professor, Sir, President, Senator, Chairman* etc., computed as the proportion of the total number of tokens per text.

German: all strings like *Doktor, Professor, Präsident, Minister, Botschafter* etc., computed as the proportion of the total number of tokens per text.

token_S: tokens per sentence

English & German: all token segments per text in proportion to all sentence segments per text.

Isabelle Delaere and Gert De Sutter

3 Variability of English loanword use in Belgian Dutch translations. Measuring the effect of source language and register

Abstract: In this chapter we want to highlight the importance of taking the factors *source language* and *register* into account when trying to make sense of linguistic differences between translated and non-translated texts. More specifically, we investigate how the aforementioned factors affect so-called normalization behavior of both translators and writers. This will be achieved by verifying how translated and non-translated texts in the Belgian Dutch context deal with (accepted) English loanwords when there is a synonymous, more endogenous alternative available. Furthermore, we draw attention to the added value of applying multivariate statistics in corpus-based translation studies together with more qualitative analyses. Therefore, three complementary analyses were carried out, viz. a correspondence analysis, a qualitative analysis of the source text lexemes, and a logistic regression analysis, which not only allows us to determine how the various factors under investigation *behave*, but also if and how they *affect* one another. Our results show that all factors under investigation do indeed have an influence on whether a loanword or an endogenous alternative is used and should therefore not be ignored in future inquiries.

1 Introduction

In this chapter, we aim to illustrate how corpus-based translation studies can benefit greatly from a broader research focus on the one hand and advanced statistical analyses on the other. More particularly, we will show how examining additional factors besides translation status, i.e. whether a text is translated or not, can lead to a better understanding of the mechanisms which cause certain translation properties, such as standardization or shining through, to emerge in some translations but not in others. As such, we are continuing the line of our previous research, in which we have repeatedly demonstrated that language use in translations is influenced by the factors register and source language, and that the alleged normalization or standardization universal[1] cannot be retained,

[1] Scott (1998: 112) describes normalization as “a term generally used to refer to the translator’s sometimes conscious, sometimes unconscious rendering of idiosyncratic text features in such a way as to make them conform to the typical textual characteristics of the target language”.

DOI 10.1515/9783110459586-004

at least not in the very strict sense that it occurs in every translation irrespective of register, source and target language, time period etc. In De Sutter et al. (2012), we investigated standardization through the use of formal lexemes vs. corresponding neutral lexemes (e.g. *indien* versus *als;* concept: if), and found no evidence whatsoever for a generalized standardization behavior in translations as it was not the case that, overall, translations make more use of formal lexemes than non-translations. In particular, we found that (i) the hypothesized phenomenon was observed in only one of the registers (external communication), but not in the others; and (ii) source language does play a role as translations from English and French behave differently. The same was observed in Delaere et al. (2012) on the use of non-accepted Belgian Dutch linguistic features (vs. accepted General Standard Dutch features) and Delaere and De Sutter (2013) on the use of accepted Belgian Dutch features (vs. accepted General Standard Dutch features). Two important conclusions can be drawn from these investigations. First, translations do *not* differ from original, non-translated texts in the sense that a translator, like any other author, uses different linguistic forms in different registers. Second, individual linguistic choices made by translators sometimes *do* differ from the linguistic choices made by other authors, and some translated registers are therefore linguistically different from corresponding non-translated registers. However, it is – at least for the time being – unclear what causes translators to choose differently on those occasions.

These main findings of our own research are very robust, as they occurred in every single case study we have performed. Furthermore, they are also supported by independent research, focusing on different languages with different linguistic features and using different corpus-based methodologies. Kruger and van Rooy (2012), for instance, compare 7 different types of linguistic features in a comparable corpus of English translations and originals produced in South Africa, including frequency of linking adverbials, frequency of lexical bundles and lexical diversity. The very detailed analysis of each of these features led the authors to the conclusion that there is no evidence for the hypothesis “that translated language, overall and in all dimensions, is simpler, more explicit and more conservative than non-translated language” and that “the distribution and prevalence of linguistic realisations that may be linked to these features of translated language are variable and subject to the influence of a variety of factors, amongst others register” (Kruger and van Rooy 2012: 62; see also Diwersy et al. 2014; Kruger 2012; Lefer 2012; Neumann 2013 for similar conclusions).

In sum, it has become unquestionable that translated texts, like non-translated texts, show register differences, as well as traces of source language influence, and that the attested differences between translated and non-translated registers are not attributable to any kind of absolute translation universal whatsoever.

This obviously does not imply that the translation scholar's quest for theoretically plausible patterns should be ended. As Chesterman puts it: "What ultimately matters is perhaps not the universals, which we can never finally confirm anyway, but new knowledge of the patterns, and patterns of patterns, which helps us to make sense of what we are looking at" (Chesterman 2004: 11). And that is exactly what this chapter aims to do: continuing the line of research set out in our previous papers, this chapter investigates how translators (vs. other authors) working in the Belgian Dutch context deal with (accepted) English loanwords such as *unit*, *team* or *service* when they have a more endogenous (Dutch) variant at their disposal such as *afdeling, ploeg* or *dienstverlening*. It should be stressed that the features under investigation are English loanwords which are accepted in the Dutch language and which are not labeled as non-standard language, an aspect which was verified by means of normative reference works on language usage[2]. This topic was selected as it allows us to look at normalization[3] or conservatism from yet another angle in addition to the perspectives adopted in our previous case studies. Normalization is defined here as the significant inclination to conform to either (i) patterns which are typical for the target language as a whole, the target register or target culture (i.e., norms laid bare via descriptive comparative research), or (ii) to prescriptive norms (as explicitly mentioned in style books, prescriptive dictionaries, language proficiency syllabi etc.).

Previous research on the use of English loanwords in translated texts by Bernardini and Ferraresi (2011) showed that translators are more conservative than authors of original texts. The authors used a parallel and comparable corpus of technical texts to conduct a thorough quantitative and qualitative analysis of the dispersion of overt lexical borrowings, adapted borrowings, semantic loans and morphosyntactic calques, revealing that translators were more averse to loanwords compared to authors of original texts. Similarly, Laviosa found that applying a corpus-driven teaching approach in translation training made her students aware of the fact that "in translational language there seems to be a preference for native equivalents" (2006: 272).

These studies investigated the use of English loanwords on the basis of single-register corpora with only one pair of languages (English-Italian). As research has shown that English loanwords are very frequent in registers which

[2] Van Dale: Groot Woordenboek der Nederlandse Taal (den Boon and Geeraerts [2010–2013]) and www.taaladvies.net

[3] We thus treat normalization as synonymous with conservatism or norm conformity (although we are aware that there is at least some variation in the definition of the terminology; cf. Delaere 2015: 26–28).

deal with domains like internet-related technology, economy and advertising (Booij 2001: 3; Laviosa 2006: 267), but less so in other registers, it seems reasonable to suppose that the assumed conservative behavior of translators is register specific. Additionally, given the evidence of source language interference or shining-through (e.g. Teich 2003; Cappelle and Loock, this volume), it seems equally reasonable to suppose that the assumed conservative behavior of translators is language pair specific. Given this assumed importance of both register and language pair, it appears we need a corpus containing translations from more than one language into at least one other language which is also stratified for different (balanced) registers. The Dutch Parallel Corpus (further: DPC; Macken et al. 2011) meets these criteria, as it consists of Dutch non-translations as well as Dutch translations from both French and English source texts, subdivided in five different registers. As in our previous research, we will adopt a profile-based approach to the study of loanwords in Dutch translations. By means of a profile-based correspondence analysis, we will visually inspect the distribution of the selected loanwords together with their Dutch endogenous variants as a function of the different registers and the different source languages in the corpus. Although it may seem counterintuitive to investigate the use of English loanwords in texts translated from French, preliminary analyses showed that these texts do indeed contain English loanwords. Therefore, translations from French were not discarded from our analyses, which allowed us to investigate the influence of the factor source language on the use of English loanwords versus endogenous alternatives (see section 2.2).

Contrary to our previous papers, we will not use the registers the way they were originally defined in the DPC, but we propose a feature-based bottom-up register reclassification of the DPC in order to overcome some fundamental problems regarding internal consistency. This will allow us to interpret the visual patterns more reliably than before. The methodology we adopted for this reclassification might also be useful for other existing corpora or future corpus compilation projects. Moreover, we will perform an additional multivariate statistical analysis (logistic regression analysis), in addition to profile-based correspondence analysis, which allows us to measure the exact effect of register and source language on the preference for loanwords vs. endogenous variants.

The next section is devoted to the methodological foundations of this corpus-based study. It describes and motivates the variables selected for this study, it presents the Dutch Parallel Corpus as well as the register reclassification which was deemed necessary. An overview of the statistical analyses used in this chapter closes the methodological section. Section 3 presents and discusses the attested patterns of this loanword study, whereas section 4 concludes this chapter with a summary and an outlook.

2 Data and methodology

The methodology underlying the research presented in this chapter can be characterized as multi-feature, multivariate and profile-based. The first characterization, multi-feature, refers to the simultaneous investigation of multiple linguistic features to describe the use of English vs. endogenous variants in translated and non-translated registers. By aggregating multiple linguistic features, we reduce the risk that the results are not representative due to the potential divergent behavior of only one ill-selected, potentially non-representative feature. The second characterization, multivariate, refers to the statistical quantification of the simultaneous effect of different explanatory variables (i.e., source language, register) on the choice between English loanwords and endogenous variants. Finally, profile-based refers to the corpus methodology developed by Speelman, Grondelaers and Geeraerts (2003) which builds on the idea that, while conveying a message, language users constantly choose between different formal alternatives to express a given concept or idea, e.g. *car* vs. *automobile*. Hence, a key aspect of this methodology is that words are not studied in isolation but always in relation to their (synonymous) naming alternatives. A profile then is defined as a set of synonymous lexical variants that can be used to express a single given concept. Using this method in a corpus-based lexical variation study thus allows us to uncover different lexical preferences per variety in a reliable manner (e.g. variety A uses *car* more often than variety B, which uses *automobile* more often). One of the main advantages of the profile-based method is that it produces more reliable descriptions of lexical preferences compared to corpus-based methods which quantify linguistic features in isolation. More particularly, it has been shown repeatedly that results based on the latter methodology are more likely to be distorted by topical differences between varieties instead of differences in lexical preference (e.g. Speelman et al. 2003; Ruette et al. 2014). For instance, when studying a lexeme as *car* in isolation, a higher frequency of *car* in variety A compared to variety B can be due either to the fact that variety A contains more texts about cars (which is a topical difference) or to the fact that variety A prefers the lexeme *car* over alternative lexemes. As the profile-based methodology quantifies the use of a word in relation to one or more naming alternatives, differences between varieties have the advantage that they can only be interpreted in terms of lexical preference[4].

[4] For a more elaborate discussion of the advantages of the profile-based method, see De Sutter et al. 2012: 330–332. Also, Heylen and Ruette 2013 present an empirical investigation of the different results produced by these different methodologies.

Against this background, the remainder of this section is devoted to the selection of linguistic profiles for this study, the extraction of the profiles from the Dutch Parallel Corpus (containing a proposal to reclassify the texts in the DPC) and an overview of the statistical analyses that will be conducted in section 3.

2.1 Profile selection

In order to find out which translated registers differ from their corresponding non-translated registers in their use of English loanwords vs. endogenous alternatives, we first have to define the concepts *English loanword* and *endogenous word*. In this study, we adopt the approach described in Zenner (2013), who uses the information with regard to etymology and pronunciation provided in Van Dale's Great Dictionary of the Dutch Language (den Boon and Geeraerts 2010–2013) to determine whether a given word can be considered an English loanword or not. Zenner suggests the following procedure: "young loanwords such as *software* are by default considered to be English, and older loanwords (borrowed before 1945) are only considered to be loanwords if the pronunciation of the word is not what naive speakers of Dutch would anticipate based on the spelling of the word" (2013: 103). To illustrate this, Zenner provides the following examples: a naive pronunciation based on the Dutch spelling and pronunciation rules of the word *manager* would sound like /mɑ’nɑːɣər/ instead of /’mɛnədʒər/. As the Dutch pronunciation of *manager* follows the English pronunciation, *manager* is considered an English loanword. A naive pronunciation of the word *film* in Dutch on the other hand is very close to how it is actually pronounced in English, and *film* is therefore not considered an English loanword, but an endogenous lexeme.

Next, we started a bottom-up query procedure in the Dutch Parallel Corpus (DPC) in order to build a representative profile set, each of which should consist of an English loanword and at least one (near-)synonymous endogenous word. First, we extracted a case insensitive list of words and lemmas from the DPC, which we then analyzed according to the criteria mentioned above, leading to the conclusion that the only words of English origin were nouns and verbs (n = 132). As our technique is based on the use of profiles, we subsequently selected the English words which have an endogenous counterpart, narrowing down the list to 98 profile candidates. Some examples of English words which do not have an endogenous alternative are *server* and *computer*. The next phase consisted of extracting all instances of these remaining candidate profiles from the corpus and manually verifying whether the loanwords could be replaced by their more endogenous counterparts (and vice versa) given the context they were used in. For example, the loanword *director* can be replaced by the endogenous *directeur*. However, our data showed that *director* is very often used in longer job titles such as *Managing Director, Technical Director*, or *Purchasing Director*, rendering it impossible to replace *director* by *directeur* without replacing the other parts of the job titles as well, which would go against the principle of interchangeability of our profile-based approach. This manual annotation task is based on a guideline which states that only those attestations which contain a variant which can be replaced by the other variant(s) in the profile are included in the data set. In other words, if one variant cannot be replaced by the other variant(s) in the profile, the attestation is discarded. This can be due to a difference in meaning, figurative use of a given word, fixed terminology, etc. This sifting process resulted in a data set of 17 profiles, an overview of which is provided in Appendix 1. Finally, in order to produce statistically reliable results, we used a minimal frequency threshold per profile and per explanatory variable (register and source language), resulting in a data set of 7 profiles (or 14 alternatives) with an overall frequency of n = 2331 (see section 2.3 for a quantitative overview and Appendix 2 for a full description and illustration of each of the profiles). It should be noted that for all naming alternatives within the profiles, both the singular and the plural forms were taken into account. Furthermore, the corpus queries were carried out in a case-insensitive manner and for the profile RESEARCH AND DEVELOPMENT both the abbreviations (*r&d*; *o&o*) and the full forms (*research and/& development*; *onderzoek en ontwikkeling*) were taken into account.

2.2 Dutch Parallel Corpus

Given this study's research objectives, the corpus requirements were highly specific: the corpus should be parallel, contain various registers, have translations from at least two source languages, and should have Dutch as the central language. There is one existing corpus which meets these requirements: the Dutch Parallel Corpus (Macken et al. 2011). It is a 10-million-word, bidirectional parallel corpus for three languages, viz. Dutch, French and English, and it consists of six registers (journalistic texts, instructive texts, administrative texts, non-fiction, literature and external communication). All texts in the corpus have been pre-processed (lemmatized and pos-tagged) and have been cleared of copyright by means of written agreements between publishers or authors and the DPC team (De Clercq and Montero Perez 2010), which makes it readily available for research.

Texts from the corpus for which the source language was unknown were excluded from our analyses as well as texts which were produced in the Netherlands, as the amount of data in the corpus turned out to be too low for obtaining reliable and generalizable results for Netherlandic Dutch (n = 1,409,721). Consequently, only data for the Belgian Dutch variety were included in the analysis, an overview of which is provided in Table 1 below.

Table 1: Overview of the Belgian Dutch subcorpus

	Original Belgian Dutch		**Belgian Dutch translated from French**		**Belgian Dutch translated from English**	
	N° tokens	**N° texts**	**N° tokens**	**N° texts**	**N° tokens**	**N° texts**
Fictional literature	0	0	116178	4	0	0
Non-fictional literature	412712	15	96688	6	0	0
Journalistic Texts	485876	356	272429	240	295039	253
Instructive texts	106640	27	45371	20	0	0
Administrative texts	428391	229	339826	177	237579	25
External communication	372256	255	261640	116	337978	377
Total	1805875	882	1132132	564	870596	655

2.2.1 Main problem area: DPC register classification

Although the Dutch Parallel Corpus is undoubtedly the best corpus available for our research objectives, there is at least one major problematic aspect which needs to be addressed, i.e. the limited background information on the registers in the corpus and how they were defined. Of course, there is some register information available in the DPC manual[5] and in Macken et al. (2011: 3), where it is mentioned that the main registers were subdivided according to the prototype approach suggested by Lee (2001) and that "the labels for the subtypes were chosen from cognitively tangible categories, most of them are encountered in everyday use" (Macken et al. 2011: 3). Nevertheless, it goes without saying that these descriptions are far too vague in order to capture the exact circumstantial differences between for instance administrative texts and external communication, which in turn has a negative effect on the interpretability of the results based on the DPC data.

[5] https://www.kuleuven-kulak.be/dpc/manual/DPC.pdf

Given the importance of the factor register in this chapter, we decided to thoroughly inspect the texts in the six DPC registers in order to find out (indirectly) how the different registers were defined and, consequently, how the texts in the DPC were classified in one of these registers. By doing so, several problems emerged, both with regard to the registers' contents and our research goals. These issues will be discussed below per register. Given the arguments provided above, we will only focus on the Belgian Dutch texts in the corpus.

– **Literature**: this register only contains 4 literary works translated from French: *Gelukkige dagen*, *De schreeuw*, *De man die op reis ging* and *Een vrouwelijke Odysseus, N'zid*. Although these novels obviously fit in this register, the fact that there is only a very limited set of text material available, severely restricts the generalizability of the results for this register. Moreover, the linguistic features that typify this register cannot be compared to the linguistic features of non-translated literature or literature translated from English, as the corpus does not contain material for these varieties, making it unusable for our research objectives.
– **Non-Fiction**: this register contains 21 different texts from three different publishers (Ons Erfdeel, Lannoo and The Flemish Government). The texts in this register are either non-translated or translated from French, so here too, there are no translations from English available, making it once again impossible to study the effect of the source language on the linguistic features in this register. Besides the fact that texts from only three publishers restricts the generalizability of the results, these publishers are also quite different in nature. Whereas the Flemish Government is the institution that governs the Flemish Community in Belgium, both Ons Erfdeel and Lannoo are institutions of a completely different nature: Ons Erfdeel is a cultural institution whose main aim is to promote cultural cooperation between the Dutch-speaking communities, and Lannoo is a Belgian publishing house which "wants to be a leading, creative and flexible knowledge enterprise pursuing both cultural and economic value"[6]. This makes this register rather heterogeneous, which also becomes apparent in the type of texts produced by each of these publishers: whereas the texts from Lannoo and Ons Erfdeel are essayistic in nature, the texts from the Flemish Government are essentially expository works of a general nature on cultural and historical topics.
– **Journalistic texts**: this register contains texts from six rather heterogeneous publishers: a university (KU Leuven), a publishing house (Roularta), two banks (ING and Fortis), and two newspapers (De Morgen and De Standaard). The text material in this register is highly diverse as well: various texts from

[6] http://www.lannoo.be/international

Roularta, for example, are clearly instructive texts such as step-by-step instructions on how to make curtains, felt accessories, cushions; various recipes, and step-by-step instructions on how to prepare a picnic.

- **Instructive texts:** this register contains texts from two text publishers: the FOD Sociale Zekerheid (the Federal Public Service of Social Security) and RIZIV (the Belgian National Service for Medical and Disablement Insurance), which are both governmental institutions. As in the previous register, text material here is heterogeneous: texts from the FOD Sociale Zekerheid are of a clearly legislative or contractual nature, such as Royal Decrees, employment contracts, rules and regulations, and agreements, whereas other texts are of a rather administrative nature such as the written report of a workshop, an informative document with regard to job regulations for students, and assignment specifications. The texts provided by RIZIV are heterogeneous too: procedural manuals, agreements, highly specific documents regarding the nomenclature of medical services, and technical documents regarding budgetary goals. In other words, many of these texts might not fit the intuitive idea one has of the register instructive texts as a directive register, enabling the addressee to do something (see e.g. Werlich 1982).
- **Administrative texts**: texts in this register were provided by 13 different publishers which belong either to the private sector (e.g. Barco, Bekaert and Melexis) or the public sector (e.g. the Federal Public Service of Social Security, Federal Public Service of Justice, the European Parliament and the Belgian Chamber of Representatives). Again, this results in a register containing rather heterogeneous texts. Remarkably, this register is the only register which contains material whose outcome was written to be spoken or written reproduction of spoken language, which stresses the heterogeneity. Similar to the previous register, the administrative register also contains technical documents regarding budgetary goals, documents regarding the nomenclature of medical services, rules and regulations, and agreements. The fact that these highly specific texts occur in both registers clearly exemplifies that the register classification procedure in the DPC is not watertight and obviously raises the question as to whether these two registers can be considered different enough.
- **External communication**: finally, external communication consists of press releases and newsletters from both the private and the public sector. Other examples of this register's contents are information brochures (e.g. from the Federal Public Service of Justice or from the RIZIV); medicine-related articles (e.g. from Transmed); and tourist information (e.g. from the Belgian National Railway Company or the Tourism Federation of Namur). No immediate problems were found where this particular register is concerned.

2.2.2 Bottom-up register reclassification of the Dutch Parallel Corpus

As became clear in the previous section, the original register classification in the DPC suffers from three types of problems: register-internal heterogeneity, partial overlap between the administrative and instructive registers and the small range of publishers within some of the registers. Given the strong focus on the variable register in this chapter, these problems called for a reorganization of the registers in the DPC.

Similar to the typology suggested by Lee (2001), we wanted to divide the corpus into various registers by looking at a number of objective non-linguistic characteristics. Building on Biber and Conrad's methodology (2009: 40), we first annotated each text in the (Belgian Dutch part of the) DPC for four situational characteristics based on the available metadata in the original DPC, viz. addressor, addressee, channel and communicative purpose. We opted for these situational characteristics because they allow us to redefine the register classification without interfering with our linguistic research goals. Based on the metadata provided by the DPC, the possible values which were assigned to the different characteristics were the following:

- addressor = { (commercial) company, research and education, public service, media, public enterprise }.
- addressee = { internal target audience, broad external audience, specialized external audience }.
- mode = { written to be read, written to be spoken, written reproduction of spoken language }.
- communicative purpose = { inform, persuade, instruct, activate, inform/persuade }.

The exact guidelines for the annotation of these four situational characteristics are provided in Appendix 3. The annotation was primarily done by the first author but in order to verify the consistency and generalizability of the annotation, a set of 43 texts from 17 unique providers was independently annotated by the first and second author. Afterwards, the inter-annotator agreement was calculated using Cohen's kappa statistic (Carletta 1996), resulting in an average kappa score of 0.86, which led us to conclude that the guidelines were specific enough to reliably annotate the texts for all four situational characteristics.

After annotating the texts, there were fifty unique combinations of situational features, each of which could be considered as a separate register. Obviously, fifty registers would render any interpretation and statistical evaluation impossible. For that reason, we decided to reclassify the DPC into seven registers: specialized communication, instructive texts, political speeches, journalistic texts, broad commercial texts, legal texts and tourist information. The first five registers are characterized by one decisive situational feature that is unique for this register, viz. communicative purpose = instruct (instructive texts), addressee = specialized target audience or internal target audience (specialized communication), addressor = media or research and education (journalistic texts), mode = written to be spoken or written reproduction of spoken language (political speeches) and addressee = broad external target audience (broad commercial texts). The two remaining registers were determined by a combination of features: the register legal texts is determined by communicative purpose = activate or persuade and addressor = public service and the register tourist information is defined by communicative purpose = activate or inform/persuade and addressor = public enterprise or media and addressee = broad external. Although we are aware that this final step of narrowing down the list of 50 potential registers to 7 registers is to some extent idiosyncratic, and hence could be reconsidered in future research, it did allow us to interpret the patterns emerging from the corpus in a much more reliable way (see section 3).

Table 2: Overview of the registers in the Dutch Parallel Corpus based on the bottom-up reclassification task

	N° Texts	**Content examples**
Broad commercial texts	508	(Self-)presentations of organizations, projects, events; press releases and newsletters; promotion and advertising material
Specialized communication	331	Internal and external correspondence; scientific texts; yearly reports
Political speeches	62	Official speeches; proceedings of parliamentary debates
Instructive texts	61	Manuals; DIY guides
Journalistic texts	901	Comment articles (columns); essayistic texts; news articles
Tourist information	47	Informative documents of a general nature
Legal texts	150	Internal legal documents; legislation, descriptions of legal procedures

2.3 Statistical analyses

In a first step, we will visually inspect to what extent non-translated and translated registers from English and French differ in their use of English loanwords vs. endogenous variants. As in our previous papers (e.g. De Sutter, Delaere and Plevoets 2012), we use profile-based correspondence analysis. More particularly,

Table 3: Overview of the distribution of the selected profiles across registers and source language varieties

Profile set	Label	Concept	Legal	Special	Political	Broad	Journal	DU_orig	DU < EN	DU < FR	Total
ploeg	endogenous	TEAM	2	5	0	3	130	70	10	60	140
team	loanword		63	115	2	148	270	268	103	227	598
dienstverlening	endogenous	SERVICE	18	75	9	117	28	111	43	93	247
service	loanword		0	14	1	57	18	33	46	11	90
baan	endogenous	JOB	0	4	35	21	74	57	44	33	134
job	loanword		6	13	45	24	144	142	25	65	232
afdeling	endogenous	DIVISION	23	76	3	62	172	208	65	63	336
unit	loanword		0	2	0	9	11	12	10	0	22
o&o	endogenous	R&D	1	84	11	23	15	44	90	0	134
r&d	loanword		0	15	0	28	15	25	32	1	58
partnerschap	endogenous	PARTNERSHIP	1	15	7	105	3	75	44	12	131
partnership	loanword		0	10	1	59	19	24	50	15	89
hulpmiddel	endogenous	TOOL	12	30	1	15	10	33	9	26	68
tool	loanword		0	20	0	11	21	18	12	22	52
Total			126	478	115	682	930	1120	583	628	2331

we used the R-library corregp[7] (Plevoets 2015) to conduct the profile-based correspondence analysis. It is a modified version of the more widely used correspondence analysis (Greenacre 2007), which treats all row variables in the data frame (i.c. the different lexemes) as autonomous entities, whereas our profile-based approach requires that some of these are grouped in a profile. Table 3 below provides an overview of the profiles under investigation and their absolute frequencies across registers and source language varieties. As can be seen in Table 3, our dataset only contains data for 5 of the 7 registers mentioned above. This is due to a threshold which was set with regard to the total sum of the occurrences of all variants which should be at least 100 per factor so as to ensure a certain level of statistical relevance. For example, the total sum of all occurrences of the features under investigation only amounted to 63 for the register Tourist Information and this register was therefore not added to the analysis. As a result of this threshold, the number of registers investigated in this particular case study differs from the number of registers which is available in the DPC.

In a second step, we will measure the exact explanatory and predictive effect of the explanatory variables *register* and *source language* on the use of English loanwords vs. endogenous words as well as the simultaneous effect of both variables. This will be done by means of a binary logistic regression analysis. All statistical analyses are carried out with the statistics software R 3.1.3 (R Development Core Team 2015).

3 Results and discussion

3.1 Profile-based correspondence analysis

The main goal of this chapter is to explore how translations, in comparison to non-translations, deal with loanwords when there is a more endogenous alternative available. Therefore, we carried out a first, exploratory analysis which resulted in a number of biplots such as the one displayed in Figure 1 below, which shows the dispersion of the endogenous lexemes (in light grey) and loanwords (in black italics), the position of which was determined by their frequencies in the various (source) language varieties and registers in the corpus.

[7] Contrary to our previous papers in which we used profile-based correspondence analysis, the package is now freely available to the research community via https://cran.r-project.org/web/packages/corregp/.

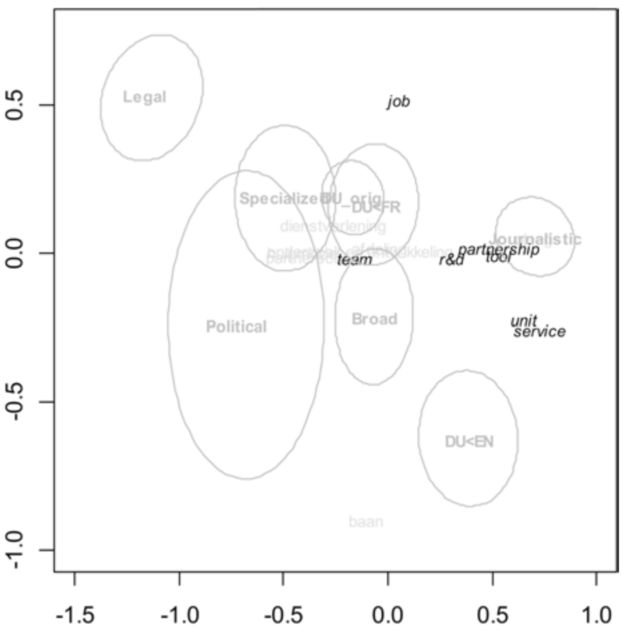


Figure 1: Biplot of the global results with the endogenous lexemes (light grey), the loanwords (black italics), the genres, and the (source) language varieties

The following information can be retrieved from the biplot: first of all, the ellipses around the variables, viz. the registers and the (source) language varieties, are two-dimensional representations of the usual confidence intervals and represent a level of statistical certainty. For each analysis, the confidence level was set at 95%, which means that under resampling 95% of all confidence ellipses will contain the true population position of the variable in question. As a consequence, the distance between two variables (e.g. Legal Texts or Dutch translated from English) can be considered statistically significant (p < .05) if their ellipses do not overlap. Secondly, the positioning of the variables vis-à-vis the features reveals whether a given variable (e.g. Journalistic Texts), in comparison to the other variables, makes more use of a given feature (e.g. *partnership*), or less.

The dispersion of these linguistic features clearly reveals that the most important dimension of our plot, i.e. the X-axis, is defined by the opposition between more endogenous lexemes on the left-hand side of the plot and English loanwords on the right-hand side of the plot. There are a few features which immediately attract attention, viz. *job, team, baan* and *ploeg* (which coincides with Journalistic), as they go against this left-right division of endogenous lexemes versus loanwords. The profile *job-baan* shows a more vertical dispersion, which reflects the high frequency of *job* in Political Speeches (in comparison to the other features' frequencies in that register) in combination with *baan*'s absence in Legal and *job*'s low frequency in Broad. With regard to the second profile which is stretched out in the opposite direction along the X-axis in comparison to the remaining profiles, we found that *ploeg* occurs substantially more often in Journalistic in comparison to the other registers, whereas *team* is one of the only two English loanwords which occurs in the register Legal Texts, hence its position on the left-hand side, away from the other loanwords. The striking behavior of these alternatives adds weight to our call for applying a multifeature approach which aggregates the variables, leading to results which are not based on the behavior of one single feature, for which we cannot know whether it is representative for the behavior of the other features as well. The remainder of this section will provide more details with regard to the linguistic preferences of the source language varieties and the registers.

a) Source language varieties

Figure 1 shows that there are significant differences between Dutch translated from English (DU < EN) and non-translated Dutch (DU_orig) and between DU < EN and Dutch translated from French (DU < FR), as their ellipses do not overlap. There is, however, no significant difference between DU_orig and DU < FR, as these ellipses clearly overlap. More specifically, the plot shows that non-translated Dutch as well as Dutch translated from French appear to make more use of endogenous lexemes while Dutch translated from English makes more use of English loanwords. These results are particularly interesting as they differ from the results of previous research on the use of English loanwords which suggested that there are indeed "contrasting tendencies in terms of the use of anglicisms on the part of translators and authors originally writing in Italian" and that translators make less use of anglicisms than authors of comparable, non-translated texts (Bernardini and Ferraresi 2011: 230; see also Laviosa 2006). Our investigation clearly shows that translators do not necessarily make more use of so-called native equivalents in comparison to authors of non-translated texts. Moreover, while the two studies mentioned above only investigated translations from one source language, i.e. English, the present study shows that the specific source language does seem to have an influence on the linguistic preferences, and should therefore not be disregarded as an explanatory variable. Section 3.2 will elaborate on this source-language effect.

b) Registers

In addition to the dispersion of the endogenous lexemes and the loanwords in the horizontal direction, i.e. the X-axis, this plot shows a similar dispersion between Political Speeches, Legal Texts, and Specialized Texts on the one hand and Journalistic Texts on the other, while Broad Commercial Communication is situated in the middle. This left-right division of the registers can be freely interpreted and one possible interpretation could be based on the level of specificity per register: while Political, Legal and Specialized address more technical topics using technical vocabulary and while these registers are mostly targeted at a specialized audience, this is not the case for the general register Journalistic, which makes use of general, non-specific terminology more often. The position of this broader register with respect to the more specific registers coincides with the dominance of loanwords for Journalistic, whereas Political, Legal and Specialized make less use of loanwords and are therefore situated further away from this group of features. Although additional research and alternative methods are needed so as to fully explain the observed differences, we would like to put forward the following, however speculative, explanatory hypotheses based on these observations. It seems not entirely implausible that for Legal Texts, for example, there is a more cautious attitude towards the use of English loanwords as the language used in this register is often rather conservative and in some cases even archaic. In contrast, there might be a tendency for the language used in a register like Journalistic Texts to be more susceptible to trends, which might explain why this register makes more use of English loanwords (Zenner et al. 2012 arrive at similar conclusions). At first sight, this tentative explanation may seem to imply that Journalistic Texts in general are less sensitive to normalization. However, previous case studies have shown that this is not entirely the case (see, e.g. Delaere and De Sutter 2013). Additionally, Journalistic Texts have shown to be a register which strongly conforms to its own, descriptive norms which are (mostly) implicit register-specific guidelines, picked up inductively by language users of this register (see Delaere 2015).

3.2 Analysis of the source text lexemes

Our profile-based analysis clearly revealed that translations from English into Dutch use significantly more English loanwords in comparison to the other varieties. The obvious explanation for this observation would be that Dutch translators who are translating from English are more inclined to use English loanwords because of the direct exposure in the source texts. Nevertheless, such an explanation can only hold if one can show for each English lexeme in the Dutch translations that it directly corresponds to an identical English source word. Hence, in this section we investigate whether an English source lexeme is transferred to the Dutch translation or translated as an endogenous Dutch lexeme. For example, an English source text may contain a lexeme (e.g. *unit*) which can either be transferred to the Dutch translation (*unit*) or translated as an endogenous lexeme (*afdeling*). Similarly, we investigate the dispersion of English loanwords in translations when the source text contains no such trigger word. Moreover, this section also investigates whether some registers are more inclined to transfer English source lexemes than others. In order to study this, we extracted the DU < EN data from the dataset, as the English source texts are obviously more likely to contain trigger words (i.e. English words that might trigger translators to transfer them to the Dutch translations).[8] We obviously only maintained those attestations for which the source text contained a possible trigger word (n = 491), viz. *team*, *service*, *job*, *unit*, *research & development*, *partnership*, or *tool*. We then verified to what extent these possible trigger words were either translated by a more endogenous lexeme or borrowed, resulting in a loanword. More specifically, we calculated the relative frequencies of each translation strategy per register, an overview of which can be found in Table 4. As the corpus contains no Legal Texts translated from English, Table 4 provides no information with regard to this register.

Table 4: Overview of the relative and absolute frequencies of each translation strategy per register with an English trigger word in the source text

	Borrow (loanword)	**Translation (endogenous word)**
BROAD COMMERCIAL TEXTS	65% (n = 146)	35% (n = 78)
JOURNALISTIC TEXTS	66% (n = 48)	34% (n = 35)
POLITICAL SPEECHES	0% (n = 0)	100% (n = 21)
SPECIALIZED COMMUNICATION	34% (n = 58)	66% (n = 115)

As can be seen in Table 4, whether a possible trigger source word is either borrowed or translated more often, does indeed depend on the register. More specifically, within Broad Commercial Texts and Journalistic Texts, the English source word is often borrowed whereas in Political Speeches and Specialized

[8] Admittedly, French translations can contain English trigger words too (or French words that are identical to English trigger words, like *service*). It turns out that the French source texts contain only 7% of such trigger words (n = 46), and the probabilities of transferring a trigger word to the Dutch translations is consequently much lower. Nevertheless, some interesting register differences can be noted (see Table 4 for the results on English source texts): in Dutch legal texts translated from French, 31% of the trigger words were transferred to the Dutch translation (only one lexeme: *team*). The percentages of transfer into the Dutch translations are much lower for the other registers, ranging from 3% to 15%. In total, only 4 trigger words were found in the French source texts (out of 7; see section 2.3): *service* (n = 10), *team* (n = 30), *job* (n = 5) and R&D (n = 1).

Communication, the English source word is more often translated into a more endogenous word. Although further research is needed to determine the exact causes for the linguistic preferences per register, we would like to put forward a number of speculative hypotheses. For example, this observation could be due to specific guidelines with regard to the use of loanwords, but could also be due to (a lack of) translator expertise as less experienced translators might be more inclined to transfer English source words (Lapshinova [this volume]; Göpferich and Jääskeläinen 2009). Or it could be due to editorial revision as editing might lead to conservative behavior and thus a preference for endogenous variants (Kruger [this volume]; Kruger 2012).

In a first attempt to assess these potential explanations, we contacted all text publishers who have contributed to the Dutch Parallel Corpus, asking them for more information on specific guidelines regarding the use of loanwords, translator expertise, and degree of editorial revision. Most of the publishers who replied to our message were not aware of specific guidelines regarding the use of loanwords. When cross-classifying the information on translator expertise and editorial revision on the one hand and register on the other, we were able to conclude that neither of these potential explanations can be used to explain the attested pattern in Table 4, as almost all translations in all registers are revised. The same applies for translator expertise, which we operationalized by means of education: almost all translators had a degree in translation, except for the translators of the journalistic texts. In sum, it appears that neither editorial revision nor translator expertise can be considered relevant as explanatory factors for the observation in Table 4, as this would require journalistic texts to behave differently from all other registers – quod non. Clearly, future, more in-depth research will have to verify this conclusion.

Obviously, the fact that the text publishers stated that there were no specific guidelines as to the use of loanwords does not rule out the possibility that implicit rules guide translators' behavior. One way of finding out the validity of implicit register-specific guidelines for explaining the observed behavior, is by investigating those contexts in our data set where translators chose an English loanword in their translations without there being a direct trigger in the source text. In other words, if the source text contains a lexeme which is not *team, service, job, unit, research & development, partnership*, or *tool*, how often does the target text contain one of these loanwords?[9] Based on the observation in

[9] Besides null-references in the source text where the English loanword was an addition in the target texts, these are some examples of the corresponding source text lexemes which resulted in a loanword in the target text: *bench, group, side, squad* (for *team*); *solution* (for *service*); *occupation, work* (for *job*); *alliance, commitment* (for *partnership*).

Table 4, we would expect journalistic texts, broad communication and – albeit to a lesser extent – specialized communication, to use loanwords more often than political speeches. Starting from the DU<EN subset, we removed all attestations for which the source text lexeme was one of the seven possible trigger words, and we then calculated the relative frequencies of the endogenous lexemes and the loanwords per register, which are represented in Table 5. Once again, this table provides no results with regard to Legal Texts as there are no data available for this register in this translation direction.

Table 5: Survey of the relative and the absolute frequencies of the more endogenous lexemes and the loanwords per register without a trigger word in the source text

	loanword	**endogenous lexeme**
BROAD COMMERCIAL TEXTS	21% (n = 7)	79% (n = 27)
JOURNALISTIC TEXTS	47% (n = 14)	53% (n = 16)
POLITICAL SPEECHES	0% (n = 0)	100% (n = 2)
SPECIALIZED COMMUNICATION	29% (n = 10)	71% (n = 25)

Although Table 5 shows that there seems to be a (strong) preference in all registers to use more endogenous lexemes if there is no trigger term in the source text, it also emerges that political speeches indeed are the only register that entirely refrains from using English loanwords, thereby potentially confirming the explanation in terms of implicit register rules.

In sum, we could see that the translation strategy depends on the register if the source text contains a trigger term. More specifically, Broad Commercial Texts and Journalistic Texts tend to borrow the source lexeme and use a loanword in the translations, whereas Political Speeches and Specialized Communication tend to use a more conventional option, i.e. a more endogenous lexeme, in the translation. If the source text does not contain a trigger word, we could see that for all registers, the endogenous options are used more often than the loanwords. Political Texts, however, show a tendency to use only endogenous words.

3.3 Logistic Regression Analysis

Whereas profile-based correspondence analysis gave us a first visual insight into the use of endogenous words and their alternative English loanwords across registers and source language varieties (which is especially useful when little is known about the underlying data and when the hypothesis to be tested is relatively vague, i.e. how does normalization work in translations?), logistic regression modelling will allow us to directly assess the exact explanatory impact of register and source language on the use of endogenous lexemes vs. English loanwords. Correspondence analysis as a visualization technique and logistic regression analysis are thus perfectly compatible.

In a first step, we fitted a logistic regression analysis on the complete dataset with endogeneity as a binary response variable (taking two values: endogenous lexeme and English loanword) and source language variety (translations from English into Dutch, translations from French into Dutch and Dutch non-translations) and register (journalistic texts, broad communication, specialized communication, legal texts and political speeches) as predictor variables. While fitting the logistic regression model, we used reference coding, signifying that one value of each of the variables was used as the reference value, viz. specialized communication for the register predictor and translations from English for the source language predictor. The effect of these reference values is set at 0, so that the effect of all other values within that predictor variable can be measured against that reference value (cf. Baayen 2008: 195–208). It is important to note that the choice for a given reference value is to some extent arbitrary and does not affect the results at all. We were not able to include the interaction between both predictor variables because of multicollinearity issues (yielding an unstable statistical model).

Table 6: Odds ratio values for register and source language in the logistic regression model (all data)

Predictor	Odds ratio (exp(β))
Broad commercial texts	1.49 ***
Journalistic texts	1.79 ***
Legal texts	1.82 ***
Political texts	n.s.
Translations from French	n.s.
Non-translations	n.s.

The logistic regression model with source language and register as predictors is significantly better than the null model, signifying that the predictors in the model significantly contribute to the explanation of the variation in the dataset (Model L.R. = 36.36, df = 6, p < .0001). Moreover, model diagnostics reveal that the regression model does not suffer from multicollinearity (measured by means of variation inflation factors, which are all below 1.96). When we look at the effects of the individual predictors in Table 6, we see that three out of six predictors significantly contribute to the choice between English loanwords and endogenous alternatives (p-values are represented by asterisks: *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$ and n.s. = not significant). Parallel to the results of the profile-based correspondence analysis, broad commercial texts, journalistic texts and legal texts significantly influence the choice between endogenous words vs. loanwords, whereas political speeches do not (compared to the reference value specialized communication). We observe once again that political speeches and specialized communications exhibit similar linguistic behavior (leading to non-significance in the regression analysis), whereas the previously attested differences between specialized communication and political speeches on the one hand and the other registers on the other correspond to significant effects in the logistic regression analysis. When looking at the significant effects in more detail, we can use the odds ratios in Table 6 to learn more about the directionality and the size of the effect. More particularly, the odds ratios show that the odds of having an English loanword (vs. an endogenous word) increases with 49%, 79% and 82% for broad communication, journalistic texts and legal texts respectively.

Contrary to the correspondence analysis, however, we observe that the source language predictors do not significantly influence the choice between endogenous words and English loanwords[10]. Recall that we observed in section 3.2 that Dutch translations from English differed significantly from non-translations on the one hand and Dutch translations from French on the other. The reason for this remarkable difference in the output of two multivariate analyses is that logistic regression analysis, unlike correspondence analysis, is able to weigh the exact effect of a certain predictor variable while simultaneously taking into account the effect of other predictor variables. So, what our logistic regression analysis tells us here is that the register predictor affects the use of English loanwords and endogenous variants to such an extent that it renders the source language predictor redundant (which is in fact confirmed by a stepwise logistic regression analysis we ran afterwards). In sum, whereas a correspondence analysis is the ideal technique to get visual insight into complex data patterns and the behavior of individual features, logistic regression analysis outperforms correspondence analysis in simultaneously weighing different predictor variables' effects on a response variable. The combination of both inferential multivariate

[10] This observation is true for source language as a main effect. We cannot rule out the possibility that source language does play a role to some extent, viz. via the register variable. This could be studied by including the interaction effect between source language and register. However, multicollinearity prevented us from doing so: we inspected a logistic regression model with interaction effects, but some of the main effects had completely different effects compared to both the regression model without interaction effects and the correspondence analysis plots. Therefore, we decided to only present the analysis with main effects.

statistics (logistic regression) and visual data mining (correspondence analysis) seems especially appropriate in emerging scientific fields like multivariate corpus-based translation studies, where relatively little is known about the underlying distribution of data.

In a final step we fit two logistic regression analyses, one on the basis of the non-translated texts in our dataset and one on the basis of the translated texts in our dataset[11]. By doing so, we are able to find out whether the effect of the register predictor is identical in those two varieties. The results in Table 7 clearly demonstrate that this is not the case.

Table 7: Odds ratio values for register in the logistic regression models of non-translated and translated texts

Predictor	Odds ratio (exp(β)) Non-translations	Odds ratio (exp(β)) translations
Broad commercial texts	n.s.	2.27 ***
Journalistic texts	n.s.	3.45 ***
Legal texts	n.s.	4.35 ***
Political texts	n.s.	n.s.

It emerges from Table 7 that register plays no role at all in the non-translated texts, whereas it does in translated texts. This implies that the observed significant effects in the previous regression model (as shown in Table 6) are completely due to the translated texts in the dataset, and, as a consequence, the effects of the register predictors are much stronger in the current model, without the non-translated texts. More particularly, the odds ratios show that the odds of having an English loanword (vs. an endogenous word) increase more than 2 (viz. 2.27), 3 (viz. 3.45) and 4 (viz. 4.35) times for broad communication, journalistic texts and legal texts respectively.

Although it is rather difficult to provide a comprehensive explanation for this significant pattern, one could wonder whether the exposure to another language, such as French or English, makes translators more aware of loanword use in the target language in comparison to authors of non-translated language. Although further research is needed, psycholinguistic investigations already provided evidence of parallel language activation in bilinguals during different

[11] We decided to fit a logistic regression model both for translations from English and from French, as preliminary analyses have pointed out that the differences are very small. When fitting two separate models for each of the two source languages, only one noteworthy difference can be observed, viz. the effect of Broad remains in the Dutch translations from English but not in the Dutch translations from French. All other effects are very similar.

types of bilingual experimental tasks (e.g. word association task, lexical decision task etc.; Blumenfeld and Marian 2005). In the case of highly trained and experienced bilinguals, such as translators, this parallel language activation might lead to a conscious and strategic decision regarding the use of an endogenous (non-cognate) word or a (cognate) loanword.

If one takes a closer look at the relative proportions of loanword use in each of the registers, one can clearly see that non-translators do not vary much in loanword use, as the percentages across the different registers range from 40% to 53% loanword use; in translated texts, on the other hand, percentages for loanword use range from 0% to 66% (Table 8).

Table 8: Survey of the relative and the absolute frequencies of loanword use per register in translated Dutch and non-translated Dutch

	Non-translations	Translations
BROAD COMMERCIAL TEXTS	44% (n = 123)	53% (n = 213)
JOURNALISTIC TEXTS	47% (n = 240)	62% (n = 258)
POLITICAL SPEECHES	53% (n = 49)	0% (n = 0)
SPECIALIZED COMMUNICATION	49% (n = 88)	34% (n = 25)
LEGAL TEXTS	40% (n = 22)	66% (n = 47)

That the variation range in translated texts is much larger than in non-translated texts should not come as a surprise, as in previous research we have demonstrated that some translated registers are much more linguistically conservative than other registers, and this might be the underlying explanation here too.

4 Conclusions

Similar to Bernardini and Ferraresi (2011) and Laviosa (2006), we wanted to investigate the assumption (based on the normalization hypothesis) that translators tend to make less use of loanwords in comparison to non-translators, an assumption which was confirmed by the results of the aforementioned authors. Moreover, as our own previous research showed promising results with regard to the influence of the factors *register* and *source language*, we wanted to investigate their influence on the use of loanwords in translations and non-translations. Furthermore, we investigated the matter in rather specific settings: we wanted to verify how translations (versus non-translations) in the Belgian Dutch context deal with (accepted) English loanwords when there is a more endogenous alternative available. We thus interpreted the normalization hypothesis as being similar to concepts such as conservatism or norm conformity, and hypothesized that, overall, translations make more use of endogenous lexemes, i.e. the conservative option, in comparison to non-translations.

Our first, exploratory analysis by means of profile-based correspondence analysis revealed that not all translations in the dataset behave differently from non-translations, as there was no significant difference between translations from French and non-translations and both varieties appear to make more use of endogenous lexemes. In other words, and contrary to Bernardini and Ferraresi's findings (2011), we did not detect "contrasting tendencies" between non-translations and translations as a whole, as Dutch translated from French behaves similarly to non-translated Dutch. Although these results are not corroborated by the logistic regression model, which showed no significant differences at all between the three source language varieties (non-translated Dutch, translated Dutch from French and translated Dutch from English), it still proves the point that there is not always a clear distinction between translations and non-translations, and that factors as register might interact with or even overrule the text's translation status (i.e. translation or not).

Our second, qualitative analysis of the source text lexemes showed that the presence of a trigger term in the source text such as *unit, job, service* or *team*, results in a register-determined preference between a loanword and an endogenous synonym. Moreover, the factor *register* includes various aspects which were not investigated in this study, but which might help explain the observed differences. For example, English loanwords might enjoy high status in one register, while this is not the case for another. Similarly, while conventions with regard to the use of loanwords might exist within one register, this is not necessarily the case for another. However, further research is needed to explore such additional, potentially explanatory aspects.

Our third analysis, the logistic regression analysis, allowed us to determine the exact effect of register and source language on the use of a loanword versus an endogenous alternative. Among other things, this additional analysis revealed that the effect of the factor register is so strong that it cancels out the effect of source language. In other words, source language (even if English is the source language) plays no significant role in translators selecting English loanwords vs. synonymous endogenous variants in Dutch translations. What does play a role is the register of the translation. We need to stress, however, that one major drawback of the regression analysis presented above, is the fact that we could not include the interaction effect between source language and register due to technical reasons. It is very plausible that a larger dataset would have revealed interesting interactions between register and source language, thus reinstating source language to some extent (via the register variable). When we combine the various source language varieties and compare translated texts with non-translated texts the logistic regression model showed that register only has an effect on translated texts. Similar to the results of our previous research, these current results show once again that the factor register should not be ignored when searching for patterns or tendencies in translated language (see also Kruger and Van Rooy 2012; Neumann 2013; Diwersy et al. 2014).

In this chapter, we hope to have shown that both our results and the results of previous research by other scholars have proved that quite a few of the assumed universal tendencies such as normalization, conservatism and explicitation are, in fact, not universal. Although there certainly are patterns, tendencies, etc. which can be linked to translated texts, these patterns are not independent of additional contextual factors as it was shown that certain patterns can be detected when investigating these linguistic phenomena over the factors *register* and *source language*. We would therefore like to urge our colleagues to certainly include these, and potentially other, factors in their analyses when investigating translated and non-translated language in search of linguistic phenomena, ultimately arriving at a better understanding of the social and cognitive mechanisms that affect the linguistic shape of translated texts.

Acknowledgements

We would like to thank our colleagues Pauline De Baets and Lynn Prieels for interviewing the text providers of the Dutch Parallel Corpus, and Koen Plevoets for co-annotating part of the dataset and introducing us to the wonderful world of correspondence analysis. We would also like to thank Kris Heylen for his helpful comments and his suggestion to carry out a logistic regression analysis in addition to the correspondence analysis.

References

Baayen, R. H. 2008. *Analyzing linguistic data. A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Bernardini, S. & A. Ferraresi. 2011. Practice, description and theory come together: Normalization or interference in Italian technical translation? *Meta* 56(2). 226–246.

Blumenfeld, H. K. & V. Marian. 2005. Covert bilingual language activation through cognate word processing: An eye-tracking study. *Proceedings of the Twenty-Seventh Annual Meeting of the Cognitive Science Society*, 286–291. Mahwah, NJ: Lawrence Erlbaum.

Biber, D. & S. Conrad. 2009. *Register, genre, and style.* Cambridge: Cambridge University Press.

Booij, G. 2001. English as the lingua franca of Europe, a Dutch perspective. *Lingua e Stile* 36. 351–361.

Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22(2). 249–254.

Chesterman, A. 2004. Hypotheses about translation universals. In G. Hansen, K. Malmkjaer & D. Gile (eds.), *Claims, challenges and changes in translation studies*, 1–13. Amsterdam/ Philadelphia: John Benjamins.

De Clercq, O. & M. Montero Perez. 2010. Data collection and IPR in multilingual parallel corpora: Dutch parallel corpus. Paper presented at the *International Conference on Language Resources and Evaluation*, Malta.

De Sutter, G., I. Delaere & K. Plevoets. 2012. Lexical lectometry in corpus-based translation studies. Combining profile-based correspondence analysis and logistic regression modeling. In M. Oakes & M. Ji (eds.), *Quantitative methods in corpus-based translation studies. A practical guide to descriptive translation research*, 325–345. Amsterdam/Philadelphia: John Benjamins.

Delaere, I. 2015. *Do translations walk the line? Visually exploring translated and non-translated texts in search of norm conformity.* Ghent: Ghent University doctoral dissertation.

Delaere, I. & G. De Sutter. 2013. Applying a multidimensional, register-sensitive approach to visualize normalization in translated and non-translated Dutch. *Belgian Journal of Linguistics* 27. 43–60.

Delaere, I., G. De Sutter & K. Plevoets. 2012. Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target. International Journal of Translation Studies* 24(2). 203–224.

den Boon, C. A. & D. Geeraerts (eds.) 2010–2013. *Van Dale: Groot woordenboek der Nederlandse taal*, 14 edn. Utrecht: Van Dale Lexicografie.

Diwersy, S., S. Evert & S. Neumann. 2014. A weakly supervised multivariate approach to the study of language variation. In B. Szmrecsanyi & B. Waelchli (eds.), *Aggregating dialectology, typology, and register analysis. Linguistic variation in text and speech*, 174–204. Berlin: Mouton De Gruyter.

Frankenberg-Garcia, A. 2005. A corpus-based study of loan words in original and translated texts. *The Corpus Linguistics Conference Series*. Birmingham.

Göpferich, S. & R. Jääskeläinen. 2009. Process research into the development of translation competence: Where are we, and where do we need to go. *Across Languages and Cultures* 10(2). 169–191.

Greenacre, M. 2007. *Correspondence analysis in practice*, 2nd edn. Boca Raton: Chapman & Hall/CRC.

Heylen, K. & T. Ruette. 2013. Degrees of semantic control in measuring aggregated lexical distances. In L. Borin, A. Saxena & T. Rama (eds.), *Approaches to measuring linguistic differences*, 361–382. Berlin: Mouton De Gruyter.

Kruger, H. & B. van Rooy. 2012. Register and the features of translated language. *Across Languages and Cultures* 13(1). 33–65.

Kruger, H. 2012. A corpus-based study of the mediation effect in translated and edited language. *Target. International Journal of Translation Studies* 24(2). 355–388.

Lefer, M.-A. 2012. Word-formation in translated language: The impact of language-pair specific features and genre variation. *Across Languages and Cultures* 13(2). 145–172.

Laviosa, S. 2006. Data-driven learning for translating Anglicisms in business communication. *IEEE transactions on professional communication* 49(3). 267–274.

Lee, D. Y. W. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language learning & technology* 5(3). 37–72.

Macken, L., O. De Clercq & H. Paulussen. 2011. Dutch Parallel Corpus: a balanced copyright-cleared parallel corpus. *Meta* 56(2). 374–390.

Neumann, S. 2013. *Contrastive register variation. A quantitative approach to the comparison of English and German*. Berlin: Mouton de Gruyter.

Plevoets, K. 2015. *corregp: Functions and Methods for Correspondence Regression*. Ghent: Ghent University. [https://cran.r-project.org/web/packages/corregp/]

Ruette, T., D. Speelman & D. Geeraerts. 2014. Lexical variation in aggregate perspective. In A. Soares da Silva (ed.), *Pluricentricity: linguistic variation and sociocognitive dimensions*. 95–116. Berlin: Mouton De Gruyter.

Scott, M. N. 1998. *Normalisation and readers' expectations: A study of literary translation with reference to lispector's A Hora da Estrela*. Liverpool: University of Liverpool doctoral dissertation.

Speelman, D., S. Grondelaers & D. Geeraerts. 2003. Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37(3). 317–337.

Teich, E. 2003. *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. Berlin/New York: Mouton De Gruyter.

Werlich, E. 1982. A text grammar of English. Heidelberg: Quelle & Meyer.

Zenner, E. 2013. *Cognitive contact linguistics. The macro, meso and micro influence of English on Dutch*. Leuven: KU Leuven doctoral dissertation.

Zenner, E., D. Speelman, & D. Geeraerts. 2012. Cognitive sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. *Cognitive Linguistics* 23(4). 749–792.

Appendix 1

This appendix provides an overview of the 17 profile candidates for this case study. This list was reduced to a final profile set of 7 profiles after applying a minimal frequency threshold per profile in order to produce statistically reliable results.

1	database
1	databank
2	team
2	ploeg
3	article
3	artikel
4	service
4	dienstverlening
5	job
5	baan
6	display
6	beeldscherm
6	scherm
7	unit
7	eenheid
8	unit
8	afdeling
9	r&d
9	o&o
10	director
10	directeur
10	bestuurder
11	partnership
11	partnerschap
12	buy-outs
12	overname
13	tendens
13	trend
14	begroting
14	budget
15	spanning
15	stress
16	tool
16	hulpmiddel
17	club
17	ploeg

Appendix 2

In this appendix, a description of all selected profiles is provided, together with an example. The Van Dale's Great Dictionary of the Dutch Language (den Boon and Geeraerts, 2010–2013) was used for the concept descriptions. The example sentences were extracted from the Dutch Parallel Corpus and their origin is presented by means of the file number. Each Dutch sentence contains one of the profile's variants which is matched by its translation or source text sentence in English.

Ploeg – team

Concept: TEAM – "a group of people who are or belong together for a certain goal"
Example: DU: De ploeg met het schrijnendste verhaal moet die uit Liberia zijn.
EN: The team with the most poignant story must be the one from Liberia [. . .].
dpc-ind-001653-nl

Dienstverlening – service
Concept: SERVICE – "the offering of services"
Example: DU: Ten eerste wilden we controleren hoe tevreden onze Europese industriële klanten zijn over onze producten en dienstverlening.
EN: First, we wanted to evaluate satisfaction levels among our European industry customers with regard to our products and services.
dpc-arc-002042-nl

Baan – job
Concept: JOB – "a post, a position"
Example: DU: Dit cijfer van 113 voltijdse banen houdt ook rekening met [. . .].
EN: This figure of 113 full-time jobs also takes into account [. . .].
dpc-bco-002447-nl

Afdeling – unit
Concept: DIVISION – "separate unit of people who pertain to a larger association, society or organization"
Example: DU: De afdeling leverde LED-visualisatieoplossingen aan klanten [. . .].

EN: As such the division has provided LED visual solutions to clients [...].
dpc-bco-002433-nl

Onderzoek en ontwikkeling – research and development
Concept: RESEARCH AND DEVELOPMENT – “research and (product) development”
Example: DU: [...] zullen beide bedrijven alle kosten voor onderzoek en ontwikkeling in gelijke mate dragen.
EN: [...] both companies will equally share all research and development costs.
dpc-aby-002285-nl

Partnerschap – partnership
Concept: PARTNERSHIP – “the aspect of being partners”
Example: DU: Een partnerschap gebaseerd op totaal vertrouwen, is [...].
EN: A partnership like this, based on total mutual confidence, is [...].
dpc-arc-002047-nl

Hulpmiddel – tool
Concept: TOOL – “means which facilitate reaching a goal”
Example: DU: Aanvankelijk was twinning een hulpmiddel om de integratie te ondersteunen [...].
EN: Twinning began as a tool to support integration [...].
dpc-arc-002046-nl

Appendix 3

In this appendix, the exact guidelines are provided for the annotation of the four situational characteristics which were used for the register reclassification of the Dutch Parallel Corpus.

Addressor

The first characteristic, addressor, refers to the person or the institution who produced the text.

– (Commercial) Company: a private enterprise whose main purpose is to produce or to trade certain goods or services.
  For example: Inbev, a brewing company.

- Media: publications in newspapers or magazines that were published by a publishing house with commercial intent. The articles cover topics which are not related to the publishing organization.
  For example: De Standaard, a newspaper.
- Research and education: publications from educational or research institutions which discuss the institution's own activities, products or organization.
  For example: University College Ghent.
- Public service: a governmental organization which does not engage in educational or commercial activities. The texts discuss the organization's own activities, products or organization.
  For example: RIZIV, the Belgian National Service for Medical and Disablement Insurance
- Public enterprise: a governmental organization which does engage in commercial activities. The texts discuss the organization's own activities, products or organization.
  For example: De Post, a postal company.

Addressee

The second characteristic, *addressee*, is the message's intended reader or listener.
- Internal target audience: in order to be able to understand the text's message, the addressee must have prior and/or expert knowledge. The information is confidential and intended for a limited audience only and is not supposed to be shared outside the organization. The available metadata are consulted to determine whether this label applies.
  For example: staff of the RIZIV, the Belgian National Service for Medical and Disablement Insurance.
- Broad external target audience: in order to be able to understand the text's message, a rather broad and/or general knowledge suffices.
  For example: newspaper readers.
- Specialized external target audience: in order to be able to understand the text's message, the addressee must have prior and/or expert knowledge, but the information is not confidential.
  For example: shareholders in a company.

Channel

The third characteristic is *channel* or *mode* and focuses on the differences between written and spoken material.
- Written to be read: the text was produced to be read.
  For example: a newspaper article.

- Written to be spoken: the text was produced to be delivered orally.
  For example: a political speech.
- Written reproduction of spoken language: the original message was delivered orally and transcribed? afterwards. For example: interviews.

Communicative purpose

The final characteristic is the text's main *communicative purpose*.

- Inform: for texts whose main purpose is to inform. The addressor has no personal interest in the text and is unrelated to the text and/or its message. The label *inform* is only chosen when the text has no other clear purpose. After reading the text, the addressee knows more about its subject and the addressor has nothing to gain from this subject.
- Persuade: the text's main purpose is to convince the target audience and to change its opinion. After reading or hearing the text, the addressee's opinion might have changed, which does not necessarily imply that the addressee has the intention to actually take action because of this new state of mind.
- Instruct: the text's main purpose is to instruct and is meant to help the addressee with regard to a certain act by means of a step-by-step guide, a manual, a recipe, etc. Vague descriptions of procedures or rules and regulations are not considered to be Instructive Texts, as these are informative rather than instructive.
- Activate: in addition to being informative, the text contains an explicit or implicit call for action. After reading or hearing the message, the addressee might actually have the intention to take action, which he or she did not intend before receiving the message. Furthermore, it is in the addressor's own interest that the addressee takes action.
- Inform/Persuade: in addition to the text's main purpose to inform, there is a second dimension, which is to persuade. Indirectly, the addressor has something to gain from the message. The addressor can either gain indirect commercial success or obtain a more positive reputation. The text provides the addressee with additional background information on a given topic and, ideally, results in the addressee having a positive image of the addressor.

Haidee Kruger

4 The effects of editorial intervention. Implications for studies of the features of translated language

Abstract: Various researchers, working both in corpus-based and sociological studies of translation, have suggested that editorial intervention may play a greater role in the ultimate appearance of published translations than is generally considered. This chapter presents an investigation of whether editorial intervention may, in principle, play a role in what is typically perceived as the features of translated language, specifically increased explicitness, conventionalisation and simplification. It is first demonstrated how the tasks associated with editing and revision may potentially lead to an increase in explicitness and conventionalism, and a decrease in complexity. Using a register-controlled parallel corpus of originally produced (untranslated) edited texts and their unedited counterparts, the study then compares the frequency of a number of linguistic operationalisations of explicitation, conventionalisation and simplification in edited texts and their unedited equivalents. The findings demonstrate that editing has significant effects on features that index all three dimensions. While the study has several limitations, there is nevertheless sufficient evidence to support the claim that the (highly variable and often invisible) role of editing, in the production of both original and translated texts, is a factor that needs to be taken into consideration in corpus-based studies of the features of translated language. Specifically, it is hypothesised that at least some of the conventionalising, explicitating and simplifying tendencies observed in studies of translated language may be the consequence of editorial processes subsequent to translation, a hypothesis that requires testing using a parallel corpus of non-revised and revised translations.

1 Introduction

Prototypically, a published text is associated with a particular individual, its author. This author is seen as being primarily responsible for the creation of the text, which is conceptualised as taking place through a complex process of planning, drafting and revision (see Hayes 2012) that eventually culminates in a completed text. While the question of authorship is considerably more fraught in the case of translated texts, it is also the case that translations are seen as

DOI 10.1515/9783110459586-005

primarily the product of a translator's work, established through a similar complex, phased translation-as-writing process (see Jakobsen 2003). This obvious association of text and author (or translator) masks the fact that in the vast majority of instances a number of people play a role in the production of a published text. Some degree of editorial intervention, at the level of content, structure, style and language, is a given for all published texts, in both print and online media, and whether original or translated. However, this editorial intervention is both highly variable in nature and scope (see Mackenzie 2004; Mossop 2014: 29–31 for some discussion of this variability), and generally remains a covert part of the publishing process, since a basic tenet of the editing profession is that "editorial skills, properly applied, do not draw attention to themselves" (Mackenzie 2004: xi). The invisibility of editing further entrenches the notion that a text is the result of a single text producer's work. However, the covert influence of editorial intervention raises the possibility that textual features that are ascribed to an author or a translator's conscious or unconscious decisions during the text production process may, in fact, have been the result of another text producer's subsequent intervention.

In recent work on the features of translated language, a number of researchers have speculated that features associated with explicitation, conventionalisation and simplification may (at least in part) be the consequence of editorial intervention, rather than of translation as such, as is usually assumed. Kruger (2012a) observes that editing may have its own effects that require consideration in studies of the features of translated language. At issue is the fact that both the original texts and translated texts typically used in corpus-based studies of the features of translated language undergo some process of editing – but this process is invisible, highly variable, and likely very different for original and translated texts. In essence, this means that some of the features viewed as the consequence of translation may, at least in part, be the consequence of editorial intervention. This point is also made by Delaere, De Sutter, and Plevoets (2012), who find that standardisation in their corpus of translated Belgian Dutch appears to be more pronounced in text types with a high degree of editorial control, raising the question of whether this feature may be ascribed to the translational process, or to the external effects of editorial intervention (Delaere, De Sutter and Plevoets 2012: 221).

To investigate this question, ideally, a corpus of the different versions of translated and original texts, as they progress through the different phases of the publishing process, is required. However, in the absence of such a corpus, this chapter aims to answer the question of whether editing may affect linguistic features indexing explicitation, conventionalisation and simplification by presenting an analysis of a parallel corpus of originally produced (i.e. non-translated) edited texts and their unedited counterparts, along the same parameters as in Kruger (2012a). The aim is to determine whether, in principle and typically, editing effects changes to texts (regardless of whether they are original texts or translations) that have previously been ascribed to translation – given that both original and translated texts are typically edited, but to unknown, highly variable, and probably not comparable degrees. This study is therefore a first step in the direction of determining whether what is conventionally interpreted as the features of translated language, in particular increased explicitness, simplification and conventionalisation, might not be the consequence of subsequent editorial processes rather than translation itself by investigating what the effects of editorial intervention are.

The following two sections first provide some necessary background to the research question. Section 2 focuses on the basic clarification of terms such as *editing* and *revision*, and their relation to translation. It also provides an overview of research on editing and revision in the context of translation. Section 3 attempts to answer the question of whether editorial work can, theoretically, influence what is regarded as the features of translated language, here conceived of primarily in terms of explicitation, conventionalisation and simplification. In order to do so, it briefly turns its attention to three areas of existing research: research on revision in the context of writing, translation revision, and research on the work of professional editors. Following these two sections, the chapter then presents the corpus methodology used for the study (section 4), followed by the findings (section 5), discussion (section 6) and conclusions (section 7).

2 Revision, editing, translation: definitions and relationships

Different disciplines have divergent definitions of what is understood by terms like *editing* and *revision*. Writing research generally narrows the focus to the writing process itself. In this context, the term *revision* is most commonly used to designate the final, evaluative part of the writing process as it is performed by the writer, which includes editing but is not limited to it (see section 3). It is mostly applied to the writer's revision of her own work, although Rijlaarsdam, Couzijn and Van den Bergh (2004: 191) point out that revision can also be done on other people's texts, in which case "the task of the reviser is to adapt the text to the demands of the communicative situation". This shows a large degree of overlap with what is generally understood as the task of the professional editor, whose work is

> an activity that consists in comprehending and evaluating a text written by a given author and in making modifications to this text in accordance with the assignment or mandate given by a client. Such modifications may target aspects of information, organization, or form with a view to improving the quality of the text and enhancing its communicational effectiveness (Bisaillon 2007: 296).

Mossop (2014: 1) points out the similarity between editing and translation revision (including both self- and other-revision): "They all involve checking linguistic correctness as well as the suitability of a text's style to its future readers and to the use they will make of it. Much of what you do when revising is identical to what you do when editing." However, in translation studies, the term *revision* has a broader meaning than its typical usage in writing research, which is the consequence of the fact that a translated text, by definition, stands in a relationship to a previously produced text in a different language – which is not the case for originally produced texts. Translation revision therefore involves a comparative, interlingual editing phase (to check for accuracy of transfer) as well as a non-comparative, unilingual editing phase, which includes the usual parameters of editing for style, structure, content and consistency (Mossop 2014). Self-revision is, of course, the final step of the translation process, but as Künzli (2005) points out, it is common practice for another person to also revise the translation before publication. However, Robert and Van Waes (2014: 307) emphasise that there is a great deal of variability in revision practice. The amount of self- and other-revision translated texts undergo is therefore variable, and frequently only one of the two steps (comparative or unilingual checking) is carried out. Furthermore, preferences for the order in which the two steps are performed also differ widely (see Robert 2008), which introduces further variability into the process of translation revision.

While some contexts impose more constraints on editorial intervention than others, in principle, revision and editing are extremely broad in scope, ranging on a continuum from minimal copyediting; to substantive editing of content, style and structure; blurring into rewriting (see Butcher, Drake, and Leach 2006: 1–2; Mackenzie 2004: 138–155; Mossop 2014). Most reference works on editing draw a distinction between copyediting (regarded as the core of the editorial process) and more substantive forms of editing, though the distinction is not categorical, but rather graded (Mackenzie 2004: 138). Mossop (2014) distinguishes among copyediting, stylistic editing, structural editing and content editing. Butcher, Drake, and Leach (2006: 1–2) propose that the editor's task basically has four components: substantive editing, detailed editing for sense, checking for consistency, and clear presentation of the material for the typesetter. In their estimation, copyediting includes only the latter three tasks. Mackenzie (2004:

138) refers to copyediting as the heart of the editorial process – but sees it as blurring into language editing and more generally into substantive editing.

In essence, copyediting (which includes checking for consistency) can be defined as the correction of a manuscript to ensure that it conforms to certain rules, which include grammar and spelling rules, usage rules, and the publisher's house style. In addition, copyediting involves editing for consistency in terms of terminology, layout, numbering and heading formatting (Mossop 2014: 30–31). Substantive editing, in Mackenzie's (2004) and Butcher, Drake, and Leach's (2006) terms, encompasses what Mossop (2014) defines as stylistic editing, structural editing and content editing. This is a more comprehensive, interventionist type of editing, which aims to "improve the overall coverage and presentation of a piece of writing, its content, scope, length, level and organization" (Butcher, Drake, and Leach 2006: 1). Stylistic editing has two dimensions: tailoring and smoothing (Mossop 2014: 31). Tailoring means to ensure that vocabulary and sentence constructions are suited to the readership and the purpose of the text, while smoothing is the process of improving the readability of a text. This can be accomplished by, for example, removing ambiguities, simplifying sentence structure, and inserting or correcting conjunctions marking the relationships between ideas (Mossop 2014: 31). Structural editing, which can be done on the micro-scale as well as the macro-scale, involves reorganising material to ensure logical presentation, as well as using discourse and layout features to signal structural relationships among parts of the text (Mossop 2014: 31). Content editing, too, can be done on the micro-scale or the macro-scale. On the micro-scale it involves checking facts, numbers and logic. However, it may also involve macro-scale work such as suggesting additions or deletion of material, and even writing additional material (Mossop 2014: 31). All these dimensions of editing applicable to originally produced texts are also applicable to the revision of translations, as Mossop (2014: 1) points out – with the added dimension that in translation, the question of accuracy and fidelity to the source text also arises.

Within translation studies, revision and editing have been investigated in a number of frameworks, though, as Robert and Van Waes (2014) point out, research is limited. In the field of translation process research, self-revision has been researched descriptively as part of the quest to understand the cognitive process of translation, particularly as related to expertise (see Antunović and Pavlović 2011; Englund Dimitrova 2005; Jakobsen 2002, 2003; Malkiel 2009). However, much revision research, and particularly research on the revision of others' translations, has been pedagogically inflected, either in proposing principles for revision and editing (see Mossop 2014), or in investigating best practices for revision empirically, making use of qualitative (see Künzli 2006, 2007) or quantitative approaches (see Robert 2013; Robert and Van Waes 2014).

Interest in the role of (other-) revision and editing has also emerged from sociological approaches to translation that emphasise that translation takes place within complex networks of agents, and emerges from the interaction between these agents, which is profoundly shaped by power relations (see Fawcett 1995; Milton and Bandia 2009). In this context, the often invisible role of the editor, among other gatekeeping agents, has been raised as a factor that has not received sufficient attention in studies of translation. Fawcett (1995: 189) points out that any translation "is usually submitted to a copy editor or other translation reviser, who normally exercises considerable influence in shaping the final product".

A last angle of investigation has been the idea that translated and edited language may share certain similarities as a consequence of the fact that both forms of language are the product of a cognitive and social mediation process, constrained in particular ways (Chesterman 2004, 2014; Kruger 2012a; Ulrych and Murphy 2008). Chesterman (2004, 2014) proposes that explanatory hypotheses for the features of translated language, usually described by terms such as explicitation, normalisation or conventionalisation, simplification, and homogenisation (see Zanettin 2012 for an overview, and section 3 for the definitions used in this chapter), may have been undergeneralised in ascribing explanations only to translation; rather, he proposes that similar features may be observed in other forms of discourse mediation (Chesterman 2014: 87), which includes, amongst others, editing. In this view, the features of translated language may, in fact, be the consequence of cognitive and/or social constraints that operate in a variety of forms of mediated discourse production.

Lanstyák and Heltai (2012) propose the most comprehensive framework thus far from within which to view these ideas. Their basic proposition is that translation universals and the features of bilingual communication may be regarded as language contact universals, which, in turn, are a subset of universals of constrained communication. In this view, all communication is constrained in some way – but some communication situations are more constrained than usual because of conditions "where one or several of the potential limiting factors play a greater than average role" (Lanstyák and Heltai 2012: 100). These constraints may be of various kinds, but they focus on two dimensions: the bilingual/ monolingual dimension, and the degree of task-dependent constrainedness (in other words, whether there is a pre-existing text on which the communication depends). These distinctions yield the classification of different types of language use set out in Table 1.

Table 1: Dimensions of constrainedness in language production (adapted from Lanstyák and Heltai 2012: 101).

	Single language	**Two languages**
Descriptive (independent text production)	Monolingual communication	Bilingual communication
Interpretative (dependent text production)	Quotation Paraphrase Intralingual translation	Interlingual translation

Intralingual translation, of course, also includes editing. Kruger (2012a) reports a first attempt to investigate the potentially shared features of translated and edited language, by utilising a comparable corpus design of edited, unedited, and translated texts in English. The discussion of dimensions of editing above clearly suggests that edited texts may be more explicit, more conventional and more simplified than their unedited counterparts – just as translated texts are frequently found to be more explicit, more conventional, and more simplified than either their source texts or comparable texts in the same language. Editors' concern with clarity of communication may lead them to increase the explicitness of lexicogrammatical encoding of texts, while their concern with ease of communication may cause them to simplify texts to improve accessibility. Copy-editing, with its strong emphasis on normative usage, self-evidently leads to greater conventionalisation. Based on this, Kruger (2012a) investigates eight linguistic features indexing explicitness, conventionality and complexity across the three subcorpora. Overall, the findings do not support the original hypothesis of similarity between the translated and edited subcorpora – but they do raise alternative hypotheses. Of the eight features investigated, four demonstrate a statistically significant difference in frequency among the three subcorpora. Two of these are explicitation features that also index conventionalisation tendencies (the omission of the complementiser *that*, and the frequency of full versus contracted forms), one a conventionalisation feature (the frequency of trigrams) and one a simplification feature (lexical diversity as measured by standardised type-token ratio). However, post-hoc tests demonstrate that the differences do not occur systematically between the two corpora of constrained or mediated text production, and the corpus of unmediated or unconstrained production, and as such her findings provide almost no evidence for a similar mediation effect in translated and edited language across the parameters investigated. She ascribes this to the fundamental difference between monolingual and bilingual language processing involved in the two tasks, as well as the differences in the degree of constrainedness of text production (while both translation and editing are interpretative text production, editing is even more constrained than translation). She finds a translation-specific effect, but this effect does not replicate exactly for the edited and unedited subcorpora in her corpus.

The findings of Kruger (2012a) raise further questions about the potential effects that editing may have on what is conventionally viewed as the features of translated language. The following section discusses these potential effects in more detail, in relation to specific features of translated language.

3 The scope of revision and editing, and the features of translated language

In the domain of writing research, revision (most commonly viewed as self-revision but also as other-revision) is regarded as the last phase of the writing process. While there are a number of different models of revision (see Hayes 2004 for an overview), there is general agreement that revision involves a process of reviewing (Allal and Chanquoy 2004: 2), during which the reviser reads the text, examining it for both errors and opportunities for improvement. The reviewing process leads to reflections on possible changes – though not all these changes are necessarily effected. Allal and Chanquoy (2004: 2) explain that once a decision to amend has been made, two broad categories of strategies may be followed:

> editing, which entails error correction and modifications designed to improve the adequacy of text without changing its general meaning; and rewriting, which entails transformations of text content (addition or deletion of segments), changes in text organization (sequencing of segments), and modifications of the meaning conveyed by a segment.

Revision, from this perspective, is more than error correction, but may involve changes at all levels of the text, from language, style and structure, to content. The motivation for these changes is the reviser's comparison of the instantiated text with her cognitive representation and assessment of, amongst others, the author's intent, the audience's needs, linguistic norms for correctness and clarity, discourse conventions, pragmatic knowledge, and conceptual background.

Professional editing involves a similar coordination of linguistic, textual, communicative and pragmatic representations in deciding which alterations to make to a text. Bisaillon (2007: 298) inductively arrives at two conceptions of editing, which she terms a normative versus a communicational conception. In the normative conception, the editor conceives of her work as primarily concerned with enforcing conformance with linguistic rules, whereas in the communicational conception, the editor is also, additionally, concerned with optimising the communicative value of the text for the reader. The normative dimension is what is typically conceived of as copyediting, whereas the concerns of the communicative dimension are subsumed under what is alternately termed substantive editing, developmental editing, content editing, structural editing and stylistic editing, as discussed in section 2 (see Butcher, Drake, and Leach 2006; Mackenzie 2004; Mossop 2014). The normative dimension of editing, associated particularly with copyediting, is strongly tied to the prescriptive linguistic tradition. Peters (2006: 775) makes the point that the "publishing industry itself, and the editorial profession, are not neutral parties in maintaining public awareness of usage sanctions... They have a gatekeeper role in enforcing selected usage practices..." She also reflects on the role of house styles, which are more than just a means to ensure consistency and a "corporate identity", but also encode and authorise normative positions on so-called correct usage (Peters 2006: 775).

As already discussed, the revision of translations shares much of the scope of writing revision and editing; however, it is set apart by the additional task of comparison reading to check for the accuracy of transfer. It is therefore, in some respects, more constrained than the editing of original text, since there is an expectation of fidelity to the source text.

Against this background, explicitation, conventionalisation and simplification may all, in principle, potentially be the consequence of editorial intervention, for both original and translated texts. The definition of these three features is not always straightforward, and they are related in complex ways. For the purposes of this chapter, they are defined as follows. Explicitation is the tendency of translated texts to "spell things out rather than leave them implicit in translation" (Baker 1996: 180). It may take the form of a preference for non-redundancy in the inclusion of optional elements, as well as for the inclusion of overt marking of propositional relationships. Conventionalisation is also often referred to as normalisation, standardisation or conservatism (Zanettin 2012: 19), and refers to the tendency for typical features and patterns of the target language to be exaggerated in translation, so that translations conform more closely to convention and reflect what is more routine in the target language (Kenny 1998: 515). As such, conventionalisation may take various forms – which may even, at times, appear to contradict each other. It may take the form of an overly conservative adherence to normative injunctions, or to the discourse conventions for a particular genre. It may also take the form of an imposition of formal discourse norms on other registers. In this form, there is frequently some overlap with explicitation, as it is often the case that more explicit forms are associated with more formal registers. Conventionalisation may also take the form of a preference for what is most common in the target language, such as, for example, a preference for more typical rather than more unusual collocations. Lastly, simplification reflects the tendency for translations to be less complex than non-translated texts, both lexically and syntactically (Baker 1996: 182).

Editing may, potentially, play a role in all these features. Editors may choose to explicitate the text at the formal or propositional level, in order to meet perceived discourse conventions for formality, or to make the text more accessible for the reader by clarifying the relationships between propositions. Editors may replace less conventional with more conventional lexical items or collocations in order to meet various normative conceptions (of standard usage, or of register-based conventions), or to improve the accessibility of the text for the audience, depending on their assessment of the audience's abilities and needs. Lastly, simplification at the lexical or syntactic level may also be the consequence of editors' attempts to remove impediments to successful communication. Some of these tendencies are illustrated in examples (1a) and (1b), from a schoolbook text on agriculture (not currently included in the corpus). Example 1(a) is the unedited version, and 1(b) the edited version.

(1a) Unedited

The procedure of using the triangle is as follows:

- *The percentages of silt is shown on the right side of the triangle. The percentage of clay is located on the left side of the triangle. The percentage of sand is shown on the base of the triangle.*
- *The different soil types are shown by bold lines.*
- *To find out the kind of soil you draw lines from the percentage point showing how much of each particle there is in the soil inwards. The area within which these lines meet gives the texture or soil type.*

(1b) Edited

This is how you use the textural triangle:

- *Clay percentages are shown by the dashed lines that go from left to right across the triangle. Silt percentages are shown by the light dotted lines that go from the upper right to the lower left of the triangle. Sand percentages are shown by the solid lines that go from the lower right towards the upper left of the triangle.*
- *The different soil types are separated by bold lines.*
- *To find out what type of soil a sample is, find the three lines that show the percentage of sand, silt and clay in the sample and follow them into the triangle. The point where these three lines cross in the triangle will tell you what the soil type is. For example, if you have a soil with 20% clay, 60% silt and 20% sand the soil type is called silt loam.*

Explicitation is evident in the addition of premodification to nouns to create a higher degree of specificity, so *triangle* becomes *textural triangle*. In the first bullet point, there is extensive explicitation through elaboration – for example, instead of just specifying ... *shown on the right side of the triangle*, the description is amended to ... *shown by the dashed lines that go from left to right across the triangle*. In the last bullet point, similar elaboration occurs, and there is also the addition of an example to illustrate the procedure in the edited version.

Simplification is evident in the replacement of complex, abstract, noun-based formulations with simpler, more concrete and verb-based formulations that also make use of direct address, as in the introductory sentence, where *The procedure of using the triangle is as follows* is edited to *This is how you use the textural triangle*. Conventionalisation is evident in syntactic changes to create a more "natural" formulation, which also remove ambiguity and ameliorate processing difficulty, as evident in the first sentence of the third bullet point. The use of more formal language (such as *separated* rather than *shown* in the second bullet point) is an example of another form of conventionalisation, namely normalisation to the written standard, though here there is additionally probably also an attempt by the editor to avoid the overuse of the word *shown*, and to introduce lexical diversity. Of course, *separated* is also, simultaneously, semantically more specific than *shown*, so there is also an explicitating effect associated with this choice.

There is scant existing research that investigates how editorial intervention (or other-revision) systematically affects texts (translated or otherwise). Most studies on editing, as well as writing and translation revision proceed from a process-oriented paradigm (typically utilising small numbers of participants), as part of the attempt to understand how revisers and editors go about their task under particular conditions (Bisaillon 2007; Lutz 1987); what the cognitive costs of revision and editing are, also in relation to the detection and correction of particular types of errors (Hacker et al. 1994; Largy et al. 2004; Piolat et al. 2004); the role of working memory (Hayes and Chenoweth 2006); and how particular revision or editing processes are tied to the optimisation of efficiency and/or quality (Künzli 2006, 2007; Robert and Van Waes 2014). Some of this research does make use of taxonomies of editing or revision changes, and report on the frequencies of such changes. However, these taxonomies do not usually consider the function of a particular change, rather focusing on aspects such as "the meaning-preserving or meaning-transforming nature of a modification, the level of language affected by the change, the operations used to carry out a revision (addition, deletion, substitution, reordering), the effect of the revision (positive, neutral, negative)" (Allal and Chanquoy 2004: 3). As a consequence, measured against the description of what editors and revisers can (potentially) do in amending texts, there is little comprehensive evidence of what they (typically) do.

Research on the types of changes that editors typically make is therefore required to determine whether editorial intervention is a factor that may influence what is understood as the features of translated language. From within translation studies research, there has been some evidence to this effect. Lanstyák and Heltai (2012: 109) refer to research by Horváth (2009) that demonstrates that translation revisers "are even more conservative and hyperpuristic than translators", while Munday (2008) points out that editors working on translations are prone to normalisation to conservative target-language standards, particularly when they are not familiar with the source language. While Kruger (2012a: 382) did not explicitly intend to investigate the effects of editing, her findings suggest that "in amending texts editors introduce collocational variety, rather than reducing variation in favour of more consistently explicit and standardised language" – but she cautions that these findings are difficult to interpret in the comparable design she utilised. The study reported on in this chapter aims to determine more conclusively whether editors' amendments may influence linguistic features of explicitness, conventionalisation and simplification, and whether editorial intervention is therefore a factor that ought to be taken into consideration in studies of the features of translated language.

4 Methodology

4.1 Corpus composition

Compiling a parallel corpus of unedited texts and their edited counterparts is complicated by access to such texts, which authors and publishers are not always willing to make available. The corpus used in this study is therefore a work in progress. It consists of 208 English texts, written in South Africa, in the period 1997 to 2012, with an unedited and edited version of each (for a total of 416 texts in the corpus).[1] Most of the texts are published texts, including books, academic articles, magazine articles, research reports, annual reports and corporate news reports. The majority of these texts were prepared for print publication. The unpublished texts in the corpus are primarily dissertations and theses.

The texts are all full texts, varying in length from 300 tokens to 60,000 tokens. The texts were collected from two language service-agencies in South

[1] These include the texts used in Kruger (2012a), with additional texts added.

Africa, as well as from individual editors affiliated with the Professional Editors' Group of South Africa. Because of the wide variety of clients for whom the editorial work was done, it was not possible to collect the style guides used. All texts were received with tracking provided, from which both the original and the edited versions were recovered. The corpus is therefore a parallel corpus, although in its current form still unaligned.

Various text types are represented, organised into four registers: academic, instructional, popular writing and reportage, utilising the standard register labels devised for the International Corpus of English (ICE) (see http://ice-corpora.net/ice/index.htm). However, the representation of these four registers is currently not balanced – largely because of the fact that the texts typically edited by the agencies and freelance editors willing to supply texts for the corpus mostly fall in the academic and instructional registers. The composition of the corpus is shown in Table 2.

Table 2: Corpus composition

	Edited word count	**Unedited word count**	**Total**
Academic (115 texts)	1,046,589	1,043,802	**2,090,391**
Instructional (67 texts)	413,072	389,209	**802,281**
Popular (10 texts)	42,035	41,725	**83,760**
Reportage (16 texts)	25,334	25,446	**50,780**
Total	1,527,030	1,500,182	3,027,212

It is evident from Table 2 that the academic register (consisting of academic articles, dissertations and theses) is overrepresented in the corpus. The innate formality of this register is a factor taken into consideration in the interpretation of findings, and an inter-register comparison is carried out for each of the analyses.

The instructional register is made up of a combination of administrative writing, schoolbooks, teacher's manuals, and sets of instructions, and together with the much smaller popular and reportage registers do provide some counterbalance for the formality of the academic register. The reportage register currently consists primarily of news reports from organisational newsletters, rather than newspapers, while the popular register is made up of popular religious and historical articles and book extracts, as well as some tourist brochures. Text collection for the expansion of both these registers (as well as a creative register) is an ongoing project.

Similarly, the collection of metadata for the texts in the corpus is ongoing. All the editors whose work is represented in the corpus are professional editors whose primary employment is as language practitioners, either within publishing houses, or as freelancers. The freelance editors whose work is represented are affiliated with the Professional Editors' Group, and/or they meet the professional and training requirements of the language agencies that provided the texts. The editors are either native speakers of English, or are Afrikaans-English balanced bilinguals.

Language-biographical information about the original text producers is in the process of collection. The texts in the corpus were all originally produced by proficient adult users of English, in preparation for publication or as part of academic work. However, South Africa is noted for its linguistic complexity (Schneider 2007), which is evident on the social as well as individual level. Like most postcolonial African countries, it is characterised by a high degree of individual bi- and multilingualism, with most speakers utilising more than one language on a regular basis for different functions. In this context, English has a particularly privileged position. It is functionally the major language, acting as lingua franca and language of formal public contexts (Webb 2002), dominating the educational and publishing landscape (see Kruger 2012b). Most South Africans therefore use English as part of their language repertoire, in both formal and informal contexts. However, the distinctions between speakers' first and second (and other) languages are highly variable and often unclear – making it difficult to apply designations such as *native speaker*, *mother-tongue speaker*, *second-language speaker* or *non-native speaker* to speakers in the South African context (see Mesthrie 2010; Schneider 2007: 13). Furthermore, different varieties of English are used in South Africa, including what would be identified as the native variety of white South African English, language-shift varieties (also spoken by native speakers of English), such as Indian English (see Mesthrie 2010: 599) as well as varieties such as Black South African English and Afrikaans English, where English is used as part of a bilingual community repertoire (Mesthrie 2010: 599). Mesthrie (2010: 600) argues that since 1990 Black South African English has developed into a native variety "in terms of fluency of usage and confidence of speakers' linguistic intuitions". In other words, while these varieties of English do have distinctive features, these are not necessarily features associated with the use of English as a second language.

The texts included in the corpus are produced by users of these varieties of South African English – who cannot straightforwardly be identified as either L1 or L2 users. This variable is therefore not considered in further detail in this chapter. However, some suggestions for further research that considers the implications of these varieties of English are raised in the concluding section of the chapter.

4.2 Data collection and processing

4.2.1 Linguistic features used as operationalisations

The linguistic features used as operationalisations of explicitation, simplification and conventionalisation largely replicate those used in Kruger and Van Rooy (2012), Kruger (2012a) and Redelinghuys and Kruger (2015), with some adaptations for the parallel corpus design. While there are numerous linguistic features that could potentially be used, these were selected for comparability with existing studies. More comprehensive multidimensional analyses with larger sets of features are foreseen as a future research possibility. All data were collected using WordSmith Tools 6.0 (Scott 2013).

(a) Frequency of the optional complementiser *that*

The frequency of the optional complementiser *that* in introducing verb complements was investigated to determine whether editors prefer the more complete surface realisation of constructions that offer the possibility of a reduced form – a tendency frequently ascribed to translators, and replicating fairly consistently across studies (see Kruger 2012a; Kruger and Van Rooy 2012; Redelinghuys and Kruger 2015; Olohan and Baker 2000). Preference for the full form may be seen as signalling a tendency towards conventionalisation, in the sense that full, rather than reduced forms, are associated with the formal written standard (see Biber et al. 1999: 680–681). However, Torres Cacoullos and Walker (2009: 6) point out that where the question of the use of the complementiser is raised in the work of prescriptive grammarians, it is mostly to make the point that the complementiser should be retained in order to ensure clarity, particularly in complex sentences. In this respect, addition of the complementiser may be seen as a structural marker that serves to explicitly mark a clausal relationship of subordination, otherwise left unmarked.

Verbs taking a *that* complement clause were used as search nodes. All the verbs classified by Biber et al. (1999: 663–666) as notably common and relatively common verbs controlling finite declarative complement clauses were analysed. A list of these verbs is presented in Table 3.

(b) Frequency of full forms rather than contracted forms

The same verb contractions and *not*-negation contractions investigated in Kruger (2012a) were also investigated in this study (see Table 4). The assumption is that the preference for the full form again indicates a preference for both greater explicitness in the avoidance of the reduced structure, as well as an imposition

Table 3: Verb lemmas investigated for that omission

	Mental verbs	Speech act verbs	Other communication verbs
Notably common (more than 100 instances per million words)	BELIEVE FEEL FIND GUESS KNOW SEE THINK	SAY	SHOW SUGGEST
Relatively common (more than 20 instances per million words)	ASSUME CONCLUDE DECIDE DOUBT EXPECT HEAR HOPE IMAGINE MEAN NOTICE READ REALIS(Z)E RECOGNIS(Z)E REMEMBER SUPPOSE UNDERSTAND WISH	ADMIT AGREE ANNOUNCE ARGUE BET INSIST TELL	ENSURE INDICATE PROVE

of the norms of formal written language (see Olohan 2003). While contractions are not usually viewed in terms of the omission of redundant or optional features, they do involve a case of "less" rather than "more" overt linguistic form to process, and in this sense may be said to be the formally less explicit or more synthetic alternative to their more explicit and more analytic uncontracted forms (see Hawkins 2003; Mondorf 2014). Generally, contracted forms are associated with informal, spoken language, and style guides often instruct the writers (and editors) of formal and informational texts to avoid them (Peters 2004: 127). For this feature, as in the previous, there is therefore a correlation between increased explicitness, normalisation to the written standard, and increased formality.

Table 4: Verb contractions and *not*-negation contractions investigated

Contracted form	Full form	Contracted form	Full form
aren't	are (*) not	shouldn't	should (*) not
can't	can (*) not	that's	that is
couldn't	could (*) not	there's	there is, there has
didn't	did (*) not	they'll	they will
doesn't	does (*) not	wasn't	was (*) not
don't	do (*) not	we'd	we would
haven't	have (*) not	we'll	we will, we shall
he's	he is, he has	we're	we are
I'd	I had, I would	weren't	were (*) not
I'll	I will, I shall	we've	we have
I'm	I am	who're	who are
isn't	is (*) not	who's	who is
it's	it is	won't	will (*) not
I've	I have	wouldn't	would (*) not
let's	let us	you'd	you had, you would
mustn't	must (*) not	you'll	you will
needn't	need (*) not	you're	you are
		you've	you have

(c) Frequency of linking adverbials

A selection of linking adverbials from the six categories defined by Biber et al. (1999: 875–879) was investigated (see Table 5). The addition of linking adverbials creates more explicit relations, of various kinds, between conceptual propositions in the text, and may be regarded as indicative of a tendency to mark conceptual relationships overtly to increase the ease of text processing for the reader. This explicitation of conceptual relationships is one of the primary tasks of structural editing (see section 2).

(d) Frequency of lexical bundles

Lexical bundles may be regarded as conventionalised stretches of language that commonly occur together in natural language (Biber et al. 1999: 990). As such, higher frequencies of lexical bundles may reflect greater conventionalisation, normalisation or conservatism in language. A number of existing studies have used lexical bundles to investigate the tendency of translated language to rely more heavily on these prefabricated stretches of language, rather than freer, less conventionalised combinations (see Dayrell 2008; Xiao 2010).

To investigate the frequency of lexical bundles, the list of trigrams generated by WordSmith Tools was used as search set. To investigate lexical bundles, word

Table 5: Linking adverbials investigated

Enumeration and addition	Summation	Apposition	Result and inference	Contrast and concession	Transition
firstly	in sum	in other words	therefore	on the other hand	by the by
secondly	to conclude	that is	consequently	in contrast	incidentally
thirdly	in conclusion	i.e.	thus	alternatively	by the way
lastly	to summarise	that is to say	as a result	anyway	
first of all	to summarize	which is to say	hence	however	
to begin with	overall	namely	as a consequence	conversely	
in addition	all in all	to be exact	in consequence	instead	
further		to be precise		on the contrary	
furthermore		to be more exact		by comparison	
likewise		to be more precise		anyhow	
moreover				besides	
similarly				nevertheless	
				still	
				in any case	
				at any rate	
				in spite of that	
				after all	

clusters of different lengths may be used. However, as Xiao (2010: 7) points out, the frequency of clusters is inversely related to their length. In this study, trigrams were selected for investigation, to ensure sufficiently representative but manageable amounts of data. For the initial computation of trigrams, the default settings of WordSmith Tools were used, which means that trigrams had to occur with a minimum frequency of 5 per million words to be identified as such.

After an initial quantitative comparison of all the trigrams identified, the most frequent trigrams, with a coverage rate of greater than 0.01% of each subcorpus (as calculated by WordSmith Tools; see also Xiao 2010) were identified, and used for further statistical analysis.

(e) Lexical diversity

Lexical diversity, or vocabulary range, was measured using standardised type-token ratio (TTR) per 300 words, as computed by WordSmith Tools, in order to investigate whether edited texts tend to be lexically less diverse, and hence simpler, than unedited texts.

(f) Mean word length

Word length is an indicator of morphological complexity, as well as lexical specificity. Shorter words are more frequent and more general, while longer words are less frequent and tend to have more specific and specialised meanings (Biber 1988: 104; Westin 2002: 75). Longer words are also typical of written, planned, informational and formal registers (see Biber 1988, 1995), and a preference for longer, more complex and more specific words may therefore also be viewed as indicative of a tendency to adjust to the formal written standard, indexing a particular form of conventionalisation. Mean word length, as calculated by WordSmith Tools, was therefore used to investigate the degree to which editing simplifies texts at the lexical level.

(g) Mean sentence length

Similarly, mean sentence length is generally an indicator of syntactic complexity (Szmrecsânyi 2004), and was analysed to determine whether editors simplify texts at the syntactic level.

(h) Hapax legomena created by editorial changes

In previous studies (see Bernardini 2011; Bernardini and Ferraresi 2011; Kruger and Van Rooy 2012; Kruger 2012a; Redelinghuys and Kruger 2015), coinages and unlexicalised loanwords were investigated as an indicator of idiosyncratic and unconventional language use. In some of these studies, hapax legomena (words that occur only once in a corpus) were used as a first search set, to circumscribe the search for unusual, infrequent lexical items.

As a consequence of this particular corpus design, this method could not be used to directly investigate the frequency of innovative, idiosyncractic single-use forms in this study. Since the texts in the unedited and edited subcorpora that comprise the corpus are identical, except for the editorial changes that have been made, a hapax exists, by definition, as the consequence of some editorial change – it is a word that exists in one of the two subcorpora, but not the other. Examples (2) and (3) illustrate editorial changes creating a hapax in the edited and unedited corpus, respectively (hapaxes are marked in bold).

(2a) *This creates a problem of the dominant classes or social groups attempting to reproduce their values and ideologies through the manipulation of the curriculum.* (I-010-O)

(2b) *This creates the problem of the dominant classes or social groups attempting to inculcate their values and ideologies in others by* ***perverting*** *the curriculum.* (I-010-E)[2]

[2] The extension of each filename indicates its edited or unedited status (O = unedited, E = edited). The first letter of each filename indicates the register of the text. As far as possible, permission was sought from authors to cite texts in their anonymised form.

(3a) [Redacted] ***alludes*** *to the fact that the school policy must be developed to ensure a consistent and equitable approach* [. . .] (A-098-O)

(3b) [Redacted] *states that the school policy must be developed to ensure a consistent and equitable approach* [. . .] (A-098-E)

In this study, therefore, hapaxes were used to investigate editorial changes leading to single-occurrence lexical items, with the aim of investigating whether the editorial amendments provide evidence of a move towards more or less conventional lexical and collocational options. All word forms occurring as hapaxes were included separately in the analysis.

The list of hapaxes was first sorted to remove words in languages other than (South African) English, proper nouns, acronyms, spelling errors, abbreviations, and parts of e-mails. A manual comparative analysis of the remaining data was done to identify the editorial change that led to the creation of the hapax. In other words, the parallel texts were investigated at the point where the hapax occurs in either the edited or unedited subcorpus, in order to identify the type of change that created the hapax. Subsequent to this, all hapax forms that were the consequence of obligatory editorial changes (specifically, error correction (e.g. of concord errors or incorrect word choices) and spelling variation, most commonly of *-ise/-ize*) were discarded, leaving only hapaxes created by a non-obligatory editorial intervention (a total of 1,559 items). Obligatory changes are typically copyediting changes to correct errors or ensure consistency, as in example (4), whereas non-obligatory changes are related to stylistic editing (see examples 2 and 3), which depends on the editor's assessment of the appropriateness and effectiveness of the lexis for the genre and target audience.

(4a) *However, an evaluator should be* ***weary*** *of smoke and mirrors* [. . .] (A-039-O)

(4b) *However, an evaluator should be wary of smoke and mirrors* [. . .] (A-039-E)

These data were used for statistical analysis. Following this, categories of editorial changes were investigated qualitatively to determine whether the processes leading to the creation of the hapax provide evidence of a conventionalising tendency, or instead of an editorial preference for introducing less conventional lexis or collocations.

4.2.2 Statistical analysis

All values for individual variables, except standardised TTR and mean word and sentence length, were standardised to a frequency per 1,000 words. TTR was standardised to 300 words – to accommodate the shorter texts in the corpus. A standardised frequency of each variable for each individual text was obtained. In other words, calculations are based on standardised values per text, which removes the risk of imbalance as a consequence of text length.

Statistical processing was done in Statistica 12 (Statsoft Inc. 2013). The normality of the data was first determined, using the Shapiro-Wilk test. Since none of the variables were normally distributed, descriptive statistics are reported as medians and interquartile ranges, and the Wilcoxon matched pairs test was used for significance testing.[3] Because many features investigated in the study are register dependent (see Biber 1995), factorial ANOVA was used to test for potential interactions between the independent variables register and corpus.

5 Findings

5.1 Frequency of the optional complementiser *that*

The use of the optional complementiser *that* needs to be viewed against the background of the total number of declarative complement clauses (where a choice between the full or reduced form could potentially have been made) in the corpus. The frequency of clauses with *that* present was therefore calculated as a ratio of the total *that* verb complement clauses (i.e. clauses with *that* realised / (clauses with *that* realised + clauses with *that* omitted)). This ratio forms the basis of the descriptive statistics in Figure 1.

Figure 1 demonstrates that in the edited subcorpus, half of the texts prefer the full form without exception, because the median value is a ratio of 1.00, which means that half (or even more) of all the edited texts contain only full forms with no omission at all. In contrast, in the unedited subcorpus, with a median of 0.96, at least half of all unedited texts had 96% or fewer full forms, and therefore by implication contain complement clauses that omit the complementiser in 4% or more cases. In some instances, therefore, editors add the complementiser *that* in the process of editing, as is illustrated by example (5), or they introduce *that* when they rewrite sections of text as part of substantive editing.

[3] The Wilcoxon matched pairs test is a rank-order test, used as a nonparametric alternative to the *t*-test for dependent, or paired, samples. While medians, as a measure of central tendency, are reported to facilitate the interpretation of the data, the Wilcoxon test is a rank sum test, not a median test.

(5a) *Remember a region is an area that has one thing that makes it different to the surrounding area.* (I-035-O)

(5b) *Remember* ***that*** *a region is an area that is different from the surrounding areas.* (I-035-E)

There is also considerably more variability in the unedited subcorpus, with both the interquartile range and the non-outlier range much larger, indicating greater dispersion of this feature in this subcorpus than in the edited subcorpus, which is more homogeneous in its preference for the full form.

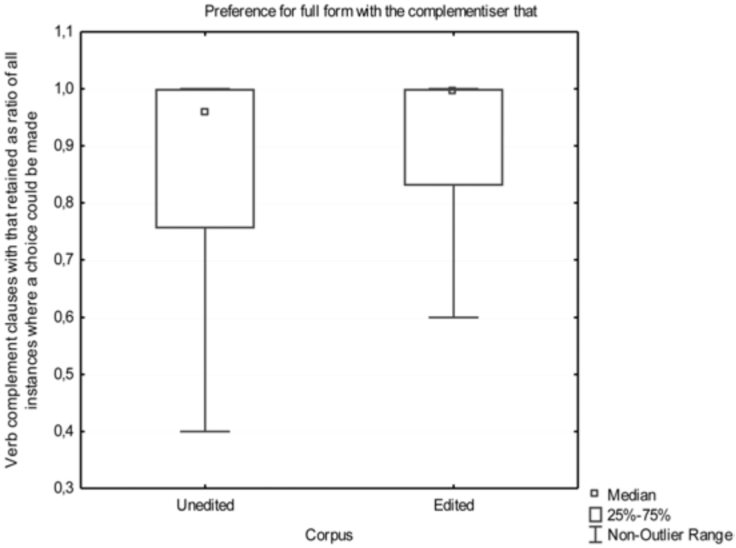


Figure 1: Preference for full form with the complementiser *that* as a ratio of all instances where a choice exists, by subcorpus

A Wilcoxon matched pairs test shows that the difference between the two subcorpora is statistically significant (T = 612.5, Z = 4.40, p < 0.001). There is no statistically significant effect for the interaction between corpus and register (F(3, 408) = 0.1, p = 0.96), with the distribution of preference for the full rather than reduced form close to identical across the four registers.

5.2 Frequency of full forms rather than contracted forms

As is the case for the use of the complementiser *that*, the frequency of full forms is reported as a ratio of all instances where a choice between the reduced form or the corresponding full form could have been made. In the edited subcorpus, more than three quarters of the texts prefer the full form up to 100% of the time, with no tolerance for variability, as is evident in Figure 2 from the fact that the median and both upper and the lower interquartile margin are all at 1.00. In the unedited subcorpus, at least half of the texts also prefer the full form without exception, shown by the fact that the median value is still 1.00, but at least a quarter of the texts do use contracted forms 2% or more of the time, shown by the lower interquartile value at 0.98.

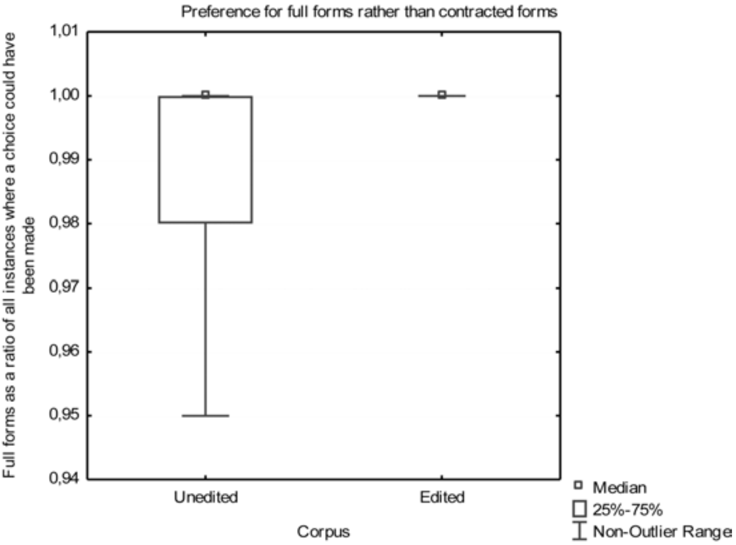


Figure 2: Preference for full rather than contracted forms, expressed as a ratio of all instances where a choice exists, by subcorpus

The difference between the two subcorpora is statistically significant (T = 321.5, Z = 4.25, p < 0.001), which indicates that editors prefer the full rather than the contracted form. While the specific style guides the editors used are currently unknown, many style guides caution that contracted forms are suited to informal rather than formal genres. Einsohn (2006: 92) points out that while there has generally been an increase in the acceptability of contracted forms, there are still strong sentiments among publishers and authors that “contractions have no place in formal writing”. Nevertheless, Leech, Smith and Rayson (2012: 72–74) demonstrate that the use of contractions has been increasing steadily in all genres from 1931 to 1991, as part of a general tendency of colloquialisation and de-formalisation in language change, with written registers moving closer to spoken registers (see also Biber and Gray 2012, 2013). It appears, then, that editors have a conservative impulse to curb this tendency towards colloquialisation. This tendency towards conventionalisation and an increase in formality is evident particularly in the popular register (see Figure 3), although there are no significant interaction effects for corpus and register (F(3, 408) = 1.84, p = 0.14).

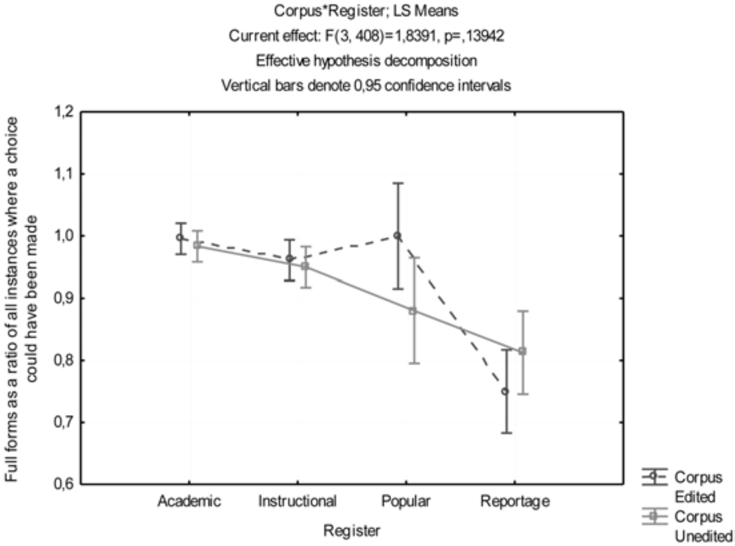


Figure 3: Frequency of full forms as a ratio of all instances where a choice exists, in the two subcorpora and four registers

Example (6) demonstrates a typical correction in this register (see also example (16) for a correction in the academic register):

(6a) *You **don't** need to worry about vintages for most South African wines, though 2001 and 1998 are good for top reds.* (P-005-O)

(6b) *You **do not** need to worry about vintages for most South African wines, although 2001 and 1998 are good for the top reds.* (P-005-E)

This suggests that the normalising and explicitating effects of editorial intervention may become visible only in less formal, less informational registers – being masked by the intrinsic formality of other registers, such as the academic register, where contracted forms are hardly used even in the unedited subcorpus.

5.3 Frequency of linking adverbials

Differences in frequency across the two subcorpora were found for only one of the subtypes of linking adverbials, namely those marking contrast and concession (T = 1932.5, Z = 5.67, p < 0.001). These linking adverbials are more frequent in the edited subcorpus (0.62 per 1,000 words) than in the unedited subcorpus (0.54 per 1,000 words). Example (7) shows a typical case of addition.

(7a) *Gradually a small number of academics started defining, questioning and informing research foci* [. . .] (A-008-O)

(7b) *Gradually, **however**, a small number of academics started defining, questioning and informing research foci* [. . .] (A-008-E)

The difference in this subtype of linking adverbials accounts for the overall difference in frequency of linking adverbials: when all six groups are considered together, linking adverbials occur more frequently in the edited subcorpus (a median of 3.02 occurrences per 1,000 words) than in the unedited subcorpus (a median of 2.81 occurrences per 1,000 words). The overall difference between the two subcorpora is statistically significant ($T = 6039.5$, $Z = 3.11$, $p < 0.05$). There is no statistically significant effect for the interaction between corpus and register ($F(3, 408) = 0.06$, $p = 0.98$), with both subcorpora following identical register-related preferences for the use of linking adverbials.

5.4 Frequency of trigrams

Overall, WordSmith Tools identifies more trigram types (occurring at least five times) in the edited subcorpus (27,206) than in the unedited subcorpus (26,256). There are also more trigram tokens in the edited subcorpus (344,922) than in the unedited (328,650). Trigrams can be grouped into four frequency bands (see Table 6). In both subcorpora, there are only five identical trigrams that occur more than 700 times per million words: *in order to*, *as well as*, *in South Africa*, *HIV and AIDS* and *in terms of*. The frequency band 300–699 contains just 10 and 11 distinct trigrams in the two subcorpora (shown in Table 7), with dramatic increases in the number of trigram types in the lower frequency bands – particularly the least frequent band (5–99 occurrences per million words).

Table 6 demonstrates that it is particularly in the less frequent ranges of trigrams that the edited corpus contains more trigrams, with the higher frequency ranges very similar – and even suggesting somewhat less conventionalised language in the edited subcorpus in these ranges. In the two upper ranges, the number of trigram types is almost identical in the two subcorpora. However, in the category of most frequent trigrams, the edited subcorpus has a lower number of instances of trigrams than the unedited subcorpus. In other words, the five most frequent trigrams occur somewhat less frequently in the edited subcorpus (4,147 times) than in the unedited subcorpus (4,260 times), suggesting that at the upper end, the most hackneyed forms are reduced slightly in order to introduce some stylistic variation. This kind of correction is shown in example (8), where two of the most frequent trigrams are removed in editing.

(8a) ***In order to*** *provide more clarity* ***in terms of*** *what riba is, [redacted] provided examples* [. . .] (A-030-O)

(8b) [Redacted] *provided examples of what could constitute riba* [. . .] (A-030-E)

In the two lower ranges, however, the edited subcorpus contains both more trigram types and more tokens of these types than the unedited subcorpus, indicating editors' tendency to, overall, opt for conventionalised lexical combinations.

Table 6: Frequency of trigram types and tokens in the two subcorpora

	Trigram types		Trigram tokens	
Frequency band	**Unedited subcorpus**	**Edited subcorpus**	**Unedited subcorpus**	**Edited subcorpus**
More than 700	5	5	4260	4147
300–699	10	11	4057	4379
100–299	201	226	29,981	33,546
5–99	26,040	26,964	290,352	302,850
Total	26,256	27,206	328,650	344,922

Table 7: Trigrams in the frequency band 300–699, by subcorpus

Frequency ranking of trigram	Unedited subcorpus	Edited subcorpus
6	be able to	restraint of trade
7	restraint of trade	be able to
8	one of the	one of the
9	part of the	based on the
10	based on the	part of the
11	the use of	the South African
12	the South African	the use of
13	of trade covenants	of trade covenants
14	the purpose of	the fact that
15	of the study	of the study
16		of South Africa

To reduce the data to a manageable amount for significance testing, the trigram types with a coverage of at least 0.01% of each subcorpus were selected as search set to calculate the number of tokens of these trigrams in each text. Figure 4 represents the results of this analysis. In the edited subcorpus, these most frequent trigrams occur with a median frequency of 13.41 per 1,000 words, whereas in the unedited subcorpus these trigrams occur with a median frequency of 11.82 per 1,000 words. The difference between the two subcorpora is statistically significant (T = 4043.5, Z = 7.79, p < 0.001). There is no interaction effect for corpus and register (F(3, 408) = 0.14, p = 0.93), with the register distribution very similar in the two subcorpora, and just slightly higher frequencies of trigrams across the board in all registers for the edited subcorpus (with the exception of the popular register, where the frequency in the edited and unedited corpus is almost identical).

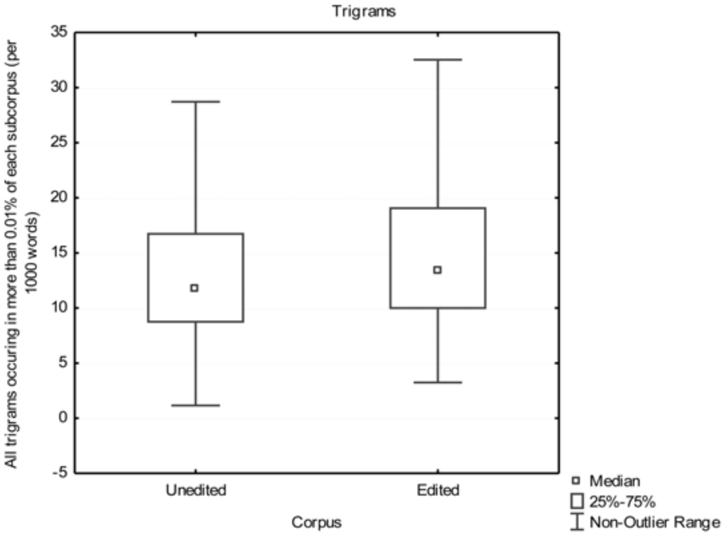


Figure 4: Occurrence of most frequent trigrams per 1,000 words, by subcorpus

5.5 Lexical diversity

Standardised TTR is almost identical in the edited and unedited subcorpora, with a median value of 50% unique lexical items per 300 words in both subcorpora. The difference between the two subcorpora is, however, statistically significant (T = 6603.5, Z = 4.56, p < 0.001). While there is no difference between the median values of the two subcorpora, there is a slight downward adjustment in the interquartile range for the edited subcorpus, which accounts for the significant difference and indicates that a slightly larger number of texts in the edited subcorpus have lower TTR values. There is no interaction effect for corpus and register as far as lexical diversity is concerned (F(3, 406) = 0.23, p = 0.88).

While it is possible that editors simplify the range of vocabulary, keeping in mind readers' needs or the injunctions of the plain language movement, which has been influential in the South African editing context, it is also possible that the perceived simplification effect is the consequence of the removal of spelling and other errors, which may lead to inflated TTR values for the unedited texts.

5.6 Mean word length

Mean word length is near identical in the two subcorpora, at around 5 characters per word, and there is no significant effect for corpus (T = 10307, Z = 0.53, p = 0.60), or for the interaction between corpus and register (F(3, 408) = 0.03, p = 0.99). Editing therefore effects no changes to lexical specificity or complexity.

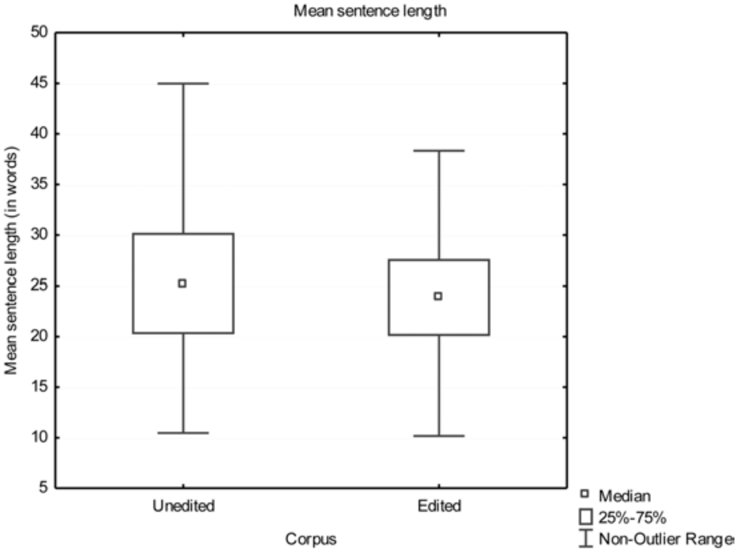


Figure 5: Mean sentence length, by subcorpus

5.7 Mean sentence length

Figure 5 illustrates the comparison of mean sentence length in the two subcorpora. In the edited subcorpus, the mean sentence length is 23.91 words, while the unedited subcorpus has a longer mean sentence length at 25.15 words per sentence. The difference between the two subcorpora is statistically significant (T = 5254, Z = 6.39, p < 0.001), but there are no significant interaction effects for corpus and register (F(3, 408) = 0.26, p = 0.86). This finding therefore supports the hypothesis that editors simplify texts at the syntactic level.

5.8 Hapax legomena created by editorial changes

As is evident from Figure 6, hapax legomena occur more frequently in the edited corpus (at a median value of 0.49 per 1,000 words) than in the unedited subcorpus (at a median value of 0.29 per 1,000 words). The difference between the two subcorpora is statistically significant (T = 4422.5, Z = 3.85, p < 0.001).

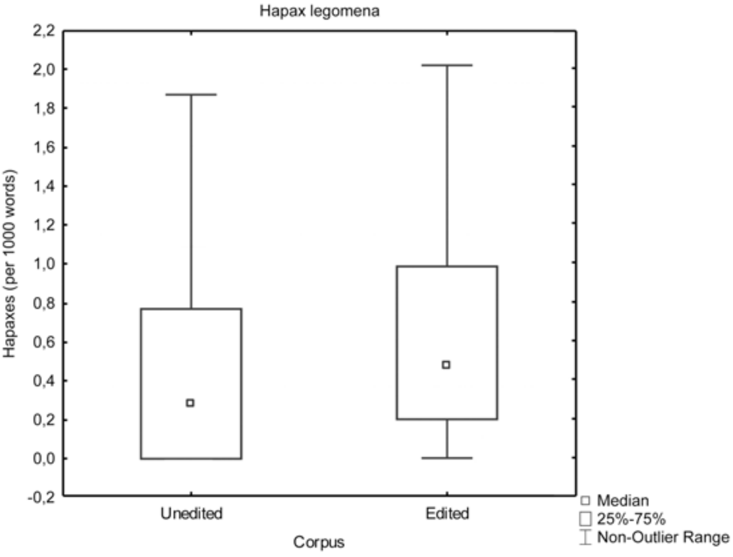


Figure 6: Frequency of hapax legomena (with hapaxes occurring as result of obligatory changes removed) per 1,000 words, by subcorpus

As a consequence of the corpus design, interpreting this finding is not straightforward – since in this corpus, a hapax exists as a word that occurs in one subcorpus, but not the other, as the result of an editorial change. Hapaxes are therefore not necessarily indicative of unusual or rare lexical choices.

To interpret the data, a classification of the kind of corrections causing hapaxes was done. Based on a manual, qualitative classification of 1,462 hapax-producing amendments,[4] two broad categories of editorial correction giving rise to hapaxes were distinguished. The first category is hyphenation or spacing changes, related to copyediting, accounting for 44% of changes. The second category of correction involves reformulation corrections at the stylistic level. Within this category, a further subclassification was made of (a) general reformulation changes that had no overt effect on the degree of conventionalism of the text (29%), (b) changes at the lexical or collocational level in favour of distinctly more conventional choices (26%) and (c) changes at the lexical or collocational level in favour of distinctly less conventional choices (1%). Where necessary, the latter two categories were also validated by cross-referencing usage on the Internet.

Copyediting produces hapaxes as a consequence of hyphenation or spacing changes (44% of instances). Hyphenation and spacing may mark an attitude towards conventionalisation in primarily two ways. Firstly, there is the normative rule, set out in style guides, that compound modifiers preceding a noun ought to be hyphenated (Ritter 2003: 134). Where such hyphenation occurs, it therefore indicates a conventionalising, normative tendency, as when *nutrient diluting effect* (A-057-O) is changed to *nutrient-diluting effect* (A-057-E). Secondly, hyphenation may mark the degree of conventionalisation of a compound word and a derivative:

> A compound term may be open (spaced as separate words), hyphenated, or closed (set as one word). In general the tendency is for new or temporary pairings of words to be spaced, and for new or temporary linkages of a prefix, suffix, or combining form with a word to be hyphenated. As the combination becomes fixed over time, it may pass through the hyphenation stage and finally come to be set as one word (Ritter 2003: 133–134).

While the data on hyphenation and spacing changes were not categorised at this fine-grained level, there is evidence of a tendency towards conservatism in the acceptance of unusual compounds and derivatives by the introduction of hyphens in edited versions, as in *co-vary* (A-055-E) instead of *covary* (A-055-O), and *mark-downs* (A-107-E) instead of *markdowns* (A-107-O). However, there is also some indication of change in the opposite direction, for example where *non-replacement* (A-098-O) is changed to *nonreplacement* (A-098-E). Generally, however, it appears that hyphenation and spacing changes resulting in hapaxes tend

[4] The number of items in this analysis is reduced by the fact that some hapaxes occur as part of the same correction (e.g. *whole-cereal* in I-019-O and *whole-grain* in I-019-E are both hapaxes, but since they occur as part of the same correction the two occurrences are subsumed under a single correction in the analysis – in this instance a classification as lexical conventionalisation).

to index a conventionalising tendency – a hypothesis which requires further testing.

As far as stylistic editing is concerned, many of the reformulations created by editors introduced a hapax, but without any overt change in the degree of conventionalisation of the lexical choices (though it may have effects on, for example, explicitness or simplicity, and indirect effects on conventionalisation). This category accounts for 29% of the hapax-producing corrections. Examples (9) and (10) provide some illustration of such changes (with the hapax marked in bold):

(9a) *Tax avoidance is the act where one does not register for tax, compared to tax evasion, where one* ***omits*** *tax registration to plan tax aggressively in one's advantage* [...] (A-031-O)

(9b) *"Tax avoidance" refers to individuals who do not register for tax, whereas "tax evasion" refers to individuals who omit to register for tax and plan their tax aggressively to benefit themselves.* (A-031-E)

(10a) They were then informed that they were required to upload the documentation from the desktop of the given computer. (A-031-O)

(10b) They were then informed that they had to upload documents from their computer **desktops**. (A-031-E).

Of course changes such as these may have indirect effects on the conventionalisation of the text; however, the kind of fine-grained analyses required to investigate this is viewed as a further research possibility, and this category was not investigated in more detail.

The last two subcategories deal specifically with changes at lexical and collocational level that clearly result in either a greater or a lesser degree of conventionalism in the edited text. Of the two subcategories, the conventionalising tendency is by far more frequent, accounting for 26% of hapax-producing corrections. The conventionalising tendency is particularly visible in the replacement of less conventional formulations or groupings of words with more conventional ones (see example [11]) and the replacement of coinages and idiosyncractic lexis with more conventional lexis (see example [12]).

(11a) *Current statistics on the recent status of learner enrolment in History from Grades 10 to 12 provides a* ***meek*** *reflection of growth* [...] (A-079-O)

(11b) *Current statistics on the recent status of learner enrolment in History from Grades 10 to 12 reflect reduced growth* [...] (A-079-E)

(12a) *From Adam's rib, He* ***architectured*** *a female compatible with Adam in every way.* (P-007-O)

(12b) *From Adam's rib, He created a female compatible with Adam in every way.* (P-007-E)

Sometimes conventionalisation also takes the form of an increase in formality at lexical and collocational level. While this leads to the use of less frequent lexical choices (contrary to an intuitive interpretation of conventionalisation), this tendency may be seen as a form of conventionalisation in adjusting texts to the norms of a perceived written standard, as in examples (13) and (14).

(13a) *The apartheid had the motto of seemingly innocuous "separate development" of different ethnic groups.* (A-007-O)

(13b) *The* ***raison d'être*** *of apartheid was the seemingly innocuous "separate development" of different ethnic groups.* (A-007-E)

(14a) *Other media organs and newspapers followed suit with some reasonable success in creating awareness* [...] (A-041-O)

(14b) *Other media organs and newspapers followed suit with a* ***modicum*** *of success in creating awareness* [...] (A-041-E)

Editors also prefer more established spelling and morphological variants to less established ones (even when both are acceptable), as when *finical* (A-039-O) is replaced with *finicky* (A-039-E). Furthermore, more formal plural forms are preferred to more informal ones (for example, *matrixes* (A-067-O) become *matrices* (A-067-E), and *appendixes* (A-106-O) is changed to *appendices* (A-067-E)).

The opposite tendency – the introduction of less conventionalised, more innovative forms – is far less frequent, accounting for only 1% of hapax-producing corrections, and is also less overt. There are a few examples of less typical collocations introduced through the process of editing, as in examples (15) and (16), and a handful of instances of coinage, as in example (17):

(15a) *Despite these achievements worldwide, the agricultural production in D R Congo on the contrary has shifted* [...] (I-040-O)

(15b) *Despite these achievements worldwide, the agricultural production in the DRC has* ***contrarily*** *shifted* [...] (I-040-E)

(16a) *Though the research outcomes regarding both studies don't provide a "moonshine and roses" scenario* [...] (A-012-O)

(16b) *Although the research outcomes for both studies do not present a “**moonlight** and roses” picture* [. . .] (A-012-E)

(17a) *However, the **system** is not limited to the interface design* [. . .] (I-031-O)

(17b) *However, the **system-to-be-created** is not limited to the interface design* [. . .] (I-031-E)

While a more detailed analysis of this dataset is clearly necessary, it appears that the increased frequency of hapaxes in the edited subcorpus is primarily the consequence of a tendency towards conventionalisation and increased formality in editing.

This suggestion is supported by the statistically significant effect for the interaction between corpus and register (F(3, 408) = 3,18, p < 0,05). Figure 7 demonstrates that it is, once again, in the popular register that edited and unedited texts differ most self-evidently, indicating that the tendency towards conventionalisation through editorial intervention becomes most evident in less formal registers.

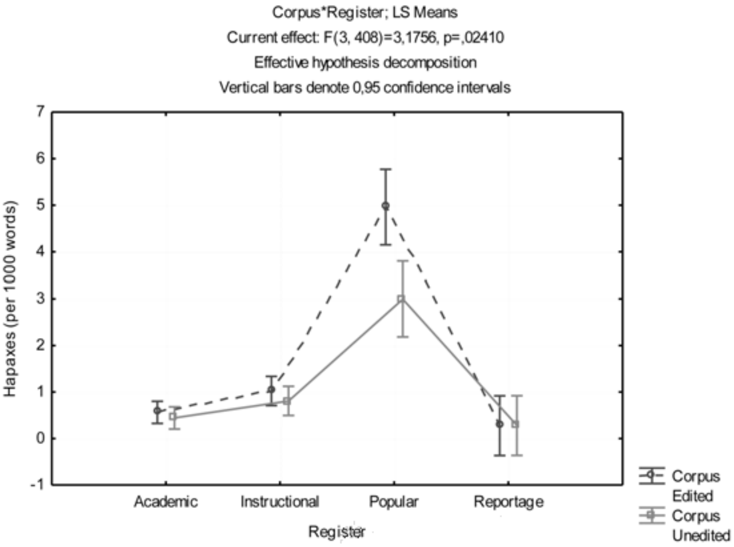


Figure 7: Frequency of hapaxes (with hapaxes caused by obligatory changes removed) per 1,000 words, in the two subcorpora and four registers

6 Discussion

The analysis reveals a statistically significant difference between the two subcorpora for seven of the eight features investigated, with only word length not demonstrating any difference between the two corpora. Editors prefer explicit, non-redundant, analytical constructions, which also tend to be associated with formal writing. Editors in all likelihood change contractions to their non-contracted forms, and add the optional complementiser *that* where it is omitted not only because these forms are more explicit and less ambiguous, but also because these forms are associated with normative conventions for formal writing. This inclination is most evident in the popular register, where editors' conventionalising impulses override the register preference for more informal usage in the form of contractions, in favour of the universal imposition of the full form – leading to an increase in the formality of the register.

Editorial preferences for explicitness are evident not only at the formal level, but also at the propositional level, with editors adding linking adverbials (particularly to mark contrast and concession) to make the relationships between propositions more overt.

There is also considerable further support for the hypothesis that editors demonstrate a tendency towards conventionalisation or normalisation. The edited subcorpus not only contains more trigram types, but also more tokens of these types, particularly in the lower-frequency ranges of trigrams. Editors do, however, also appear to have some inclination to reduce conventional lexical patterning in the most-frequent range of trigrams – where there are slightly *fewer* instances of the most overused trigrams in the edited than in the unedited subcorpus (see Table 6 and example 8). This pattern may account for the findings in Kruger (2012a), who finds that in her comparable corpus the most frequent shared trigrams in the unedited and edited subcorpora occur significantly less frequently in the edited than unedited subcorpus. The findings of this study demonstrate that, overall, editors tend to increase the degree of conventionalised lexical patterning in texts, but with a tentative indication that there may be also be an impulse to introduce some stylistic variation in editing that accompanies the preference for greater conventionalisation. The avoidance of repetition is regarded as a hallmark of good writing (as evidenced in the advice of numerous writing guides; see for example Yagoda 2013). However, as Pinker (2014: 156) points out, many style experts warn against gratuitous stylistic variation (disapprovingly termed *elegant variation* by Fowler (1908) in his chapter entitled *Airs and Graces* in *The King's English*), leading to competing injunctions to both introduce variety – and avoid it.

Hapaxes occur significantly more frequently in the edited than in the unedited subcorpus, and here, too, the qualitative analysis suggests that the editorial changes leading to hapaxes are mostly the consequence of a conventionalising tendency at the levels of punctuation, lexis and collocation – which also, at times, leads to a higher degree of formality. In this respect, the fact that the popular register has a significantly higher proportion of such single-use forms indicates that the normalising effects and the increase in formality become particularly perceptible in registers where editors' propensity to impose the norms of conventionalised written language is at odds with the innate informality of the register. However, to further investigate these findings, more fine-grained analysis of the data is necessary.

The study furthermore yields evidence that editors simplify texts at the syntactic level (as measured by sentence length), and potentially at the level of vocabulary range (as measured by TTR) – although the latter finding is somewhat ambiguous. However, editorial intervention does not appear to have an effect on lexical complexity, or lexical specificity (as indicated by word length).

Overall, this study provides support for the hypothesis that editing may affect texts in terms of formal and propositional explicitness, the degree of normalisation or conventionalisation, and relative complexity, suggesting that studies of translated language should exercise care in attributing these features to translation only. In other words, studies of the features of translated language need to consider that editorial intervention has taken place for both the original texts and the translations in a corpus. Accounting for the exact effects of this intervention is no easy task, since editing work usually remains invisible, and is highly variable – not only for translated and non-translated texts, as groups, but also for text types, and individual texts.

Understanding how editing and revision may affect translated texts (also in comparison to non-translated texts) will require at least three distinct types of studies. In the first instance, more corpus-based research is required to understand the effects that translators' self-revision has on texts. For this, draft translations extracted from various stages of the translation process are required. While process studies on translation have investigated translators' self-revision (see Dragsted 2012; Dragsted and Carl 2013; Jakobsen 2002; Mossop 2007), this needs to be extended to a more comprehensive corpus-based approach. Secondly, corpus-based research is also needed to investigate how revision and editing during the publication process affects translated texts. For this kind of investigation, corpora consisting of different versions of texts as they proceed through the editorial process prior to publication are necessary. These two types of studies will allow for the disentanglement of editorial changes made by translators themselves, and those made by others, such as revisers and editors. Lastly, corpora that allow for the comparison of the editorial processes of original versus translated texts are required, in order to investigate the similarities and differences between the processes for the two types of texts.

The findings of the study provide evidence that the potential effects of editing are a real factor requiring consideration. However, there are a number of important concerns and caveats to keep in mind. In the first instance, the findings of this study cannot be regarded as definitive. There is some disagreement with the findings of Kruger (2012a), who proposes that editors may introduce variety in editing, rather than reducing it. Some reasons for the differences in findings have already been raised above; however, it is likely that some of the other differences (such as for *that* omission, which Kruger (2012a) finds occurs more frequently in edited than unedited texts) are the consequence of corpus-design issues. In the first instance, the comparable corpus design used in Kruger (2012a) may have introduced undesirable variability, as she acknowledges (and her study was not, in the first instance, designed to investigate the differences between edited and unedited language). In this respect, the parallel corpus design used in this study provides more reliable findings. However, the findings of the study should be read against the dominance of the academic register in this corpus, which provides less scope for variability in some features as a consequence of its innate formality. Furthermore, academic editing is typified by a lower degree of editorial intervention, because of ethical concerns (see Kruger and Bevan-Dye 2010, 2013). This leads to another complicating factor that needs to be taken into consideration: the fact that different sectors of the publishing industry impose different expectations and constraints on editing. For example, in newspaper reports, succinctness (because of, amongst others, space constraints) may lead to the more frequent omission of redundant elements, and shorter sentences – not a constraint present in book editing, or academic editing.

A last matter to be taken into consideration in interpreting the findings is that a significant number of the texts in the corpus have been produced by users of different varieties of English, who use English in various configurations with their other languages (see discussion in section 4.1).[5] While this does not mean that these writers should necessarily be seen as second-language writers, it is the case that the varieties of English spoken in South Africa demonstrate distinct features. Non-native indiginised varieties, or New Englishes (such as Black South

[5] This is very typical for the South African context, where a minority of South Africans speak English as a home language – but the majority of academic output, media and print publications are in English.

African English), demonstrate a general tendency towards increased simplification and regularisation, as well as a more transparent and regular mapping of form and function (Kortmann and Szmrecsányi 2009; Szmrecsányi and Kortmann 2009). In slightly different terms, these findings can be interpreted as indicating that New Englishes are characterised by simplification and explicitation. The trend towards increased explicitness in New Englishes is well documented (see Van Rooy et al. 2010; Mesthrie 2006; Williams 1987). New Englishes are characterised in other ways by conservative, more standard-like choices (Mesthrie and Bhatt, 2008: 162) and the transfer of formal styles to less formal contexts (Van Rooy et al. 2010: 334). These features have clear implications for a study such as this one, which require further investigation.[6]

7 Conclusion

Despite limitations, the parallel corpus analysis presented in this chapter provides evidence that editorial intervention effects overall changes to texts, in terms of greater explicitness, reduced complexity, increased conventionalisation, and normalisation to a formal written standard. This implies that the effects of editing cannot be discounted in studies of translated language, and that some of the features of translated language may, potentially, be the consequence of editorial intervention rather than (only) translation. However, much further research is required to determine more definitively how the constraints of different contexts affect editorial intervention, and how editorial intervention in published original and translated texts are similar, or different – particularly given the high degree of variability in translation revision processes.

In terms of the broader framework of editing and translation as different types of mediated, or constrained, language, this chapter provides evidence that translation and editing may share certain aspects that are the consequences of constrainedness. However, there are clearly other dimensions in which the two types of constrained language are qualitatively different, not necessarily captured in the current research design. These differences are related to the parameters of constrainedness along which the two different activities differ: bilingual versus monolingual production, the degree of constraint imposed by a pre-existing

[6] The similarities between features of L2-writing, translated language and writing in the New Englishes also offer a fertile ground for further investigation in the context of the model of constrained communication set out in section 2 – all these varieties instantiate language production constrained by bilingual activation. See Gaspari and Bernardini (2010) and Kruger and Van Rooy (2016).

text, and the degree to which the activity itself is automatised. Research on revision and editing emphasises that while some aspects of revision are automatic, there is a strong component of deliberate, strategic reflection in much editing (Hayes 2004: 14; Bisaillon 2007: 306). Corrections to surface errors (i.e. at the formal level of the text) are much more frequent, because they are less effortful and have well-defined cognitively represented solutions (Largy, Chanquoy, and Dédéyan 2004: 41), while corrections that deal with the substance or structure of a text are much less frequent, since they have "ill-defined representations [...] that require [...] activating high-demanding reflection processes" (Piolat et al. 2004: 23). While translation also no doubt has both a proceduralised and declarative component, the active text production of translation means that it in all likelihood has a far greater proceduralised component than editing – and, of course, translation involves bilingual language processing.

Ultimately, as Lanstyák and Heltai (2012: 117) point out, understanding the variable role that various types of constraints may play in different kinds of constrained language will involve much more large-scale rigorous comparisons of various forms of communication that range along different parameters of constrainedness. In addition, it will be necessary to extend the investigation in various ways. Some of these have already been discussed, in terms of different types of corpora that are needed. However, the examples cited in this chapter should also indicate the widespread presence of different types of explicitating, conventionalising and simplifying alterations not indexed by the measures used in this study. Alternative methods of analysis are therefore required. Possibilities include the point-by-point, inductive analysis of aligned texts to arrive at a functional typology of editorial changes based on actual editorial practice, or comprehensive multidimensional analyses. Furthermore, the investigation needs to move beyond product-oriented methods, to also include controlled, process-oriented experimental research designs.

Acknowledgements

This work is based on research supported in part by the National Research Foundation of South Africa (Grant specific unique reference number (UID) 85291). The Grantholder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF supported research are that of the author(s), and that the NRF accepts no liability whatsoever in this regard.

I would like to express my gratitude to the language agencies and editors who contributed texts for the corpus, as well as to Deirdre Duvenage and Karien Redelinghuys for their work on the processing of the texts. Thank you also to Bertus van Rooy, who read and commented on drafts of this chapter, and to the editors and anonymous reviewers for their careful criticisms.

References

Allal, Linda & Lucile Chanquoy. 2004. Introduction: Revision revisited. In Linda Allal, Lucile Chanquoy & Pierre Largy (eds.), *Revision: Cognitive and Instructional Processes*, 1–7. New York: Springer.

Antunović, Goranka & Nataša Pavlović. 2011. Moving on, moving back or changing it here and now: Self-revision in student translation processes from L2 and L3. *Across Languages and Cultures* 12(2). 213–234.

Baker, Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In Harold Somers (ed.), *Terminology, LSP and Translation: Studies in Language Engineering In Honour of Juan C. Sager*, 175–186. Amsterdam: John Benjamins.

Bernardini, Silvia. 2011. Monolingual comparable corpora and parallel corpora in the search for features of translated language. *SYNAPS: A Journal of Professional Communication* 26. 2–13.

Bernardini, Silvia & Adriano Ferraresi. 2011. Practice, description and theory come together: Normalization or interference in Italian technical translation. *Meta* 56(2). 226–246.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-linguistic Study.* Cambridge: Cambridge University Press.

Biber, Douglas & Bethany Gray. 2012. The competing demands of popularization vs. economy: Written language in the age of mass literacy. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), *The Oxford Handbook of the History of English*, 314–328. Oxford: Oxford University Press.

Biber, Douglas & Bethany Gray. 2013. Being specific about historical change: The influence of sub-register. *Journal of English Linguistics* 41(2). 104–134.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Pearson Education.

Bisaillon, Jocelyne. 2007. Professional editing strategies used by six editors. *Written Communication* 24(4). 295–322.

Butcher, Judith, Caroline Drake & Maureen Leach. 2006. *Butcher's Copy-editing: The Cambridge Handbook for Editors, Copy-editors and Proofreaders*, 4th edn. Cambridge: Cambridge University Press.

Chesterman, Andrew. 2004. Hypotheses about translation universals. In Gyde Hanse, Kirsten Malmkjær & Daniel Gile (eds.), *Claims, Changes and Challenges in Translation Studies: Selected Contributions from the EST Congress, Copenhagen, 2001*, 1–13. Amsterdam: John Benjamins.

Chesterman, Andrew. 2014. Translation Studies Forum: Universalism in Translation Studies. *Translation Studies* 7(1). 82–90.

Dayrell, Carmen. 2008. Investigating the preference of translators for recurrent lexical patterns: A corpus-based study. *trans-kom* 1(1). 36–57.

Delaere, Isabelle, Gert De Sutter & Koen Plevoets. 2012. Is translated language more standardised than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target. International Journal of Translation Studies* 24(2). 203–224.

Dragsted, Barbara. 2012. Indicators of difficulty in translation – correlating product and process data. *Across Languages and Cultures* 13(1). 81–98.

Dragsted, Barbara and Michael Carl. 2013. Towards a classification of translation styles based on eye-tracking and keylogging data. *Journal of Writing Research* 5(1). 133–158.

Einsohn, Amy. 2006. *The Copyeditor's Handbook: A Guide for Book Publishing and Corporate Communications*. Berkeley: University of California Press.

Englund Dimitrova, Birgitta. 2005. *Expertise and Explicitation in the Translation Process*. Amsterdam: John Benjamins.

Fawcett, Peter. 1995. Translation and power play. *The Translator* 1(2). 177–192.

Fowler, Henry Watson. 1908. *The King's English*. Oxford: Clarendon Press. Bartleby.com, 1999. www.bartleby.com/116/.

Gaspari, Federico and Silvia Bernardini. 2010. Comparing non-native and translated language: monolingual comparable corpora with a twist. In Richard Xiao (ed.), *Using Corpora in Contrastive and Translation Studies*, 215–234. Newcastle: Cambridge Scholars Publishing.

Hacker, Douglas J., Carolyn Plumb, Earl C. Butterfield, Daniel Quathamer & Edgar Heineken. 1994. Text revision: Detection and correction of errors. *Journal of Educational Psychology* 86(1). 65–78.

Hawkins, John. 2003. Why are zero-marked phrases close to their heads? In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of Grammatical Variation in English*, 175–204. Berlin: Mouton De Gruyter.

Hayes, John R. 2004. What triggers revision? In Linda Allal, Lucile Chanquoy & Pierre Largy (eds.), *Revision: Cognitive and Instructional Processes*, 9–20. New York: Springer.

Hayes, John R. 2012. Modelling and remodelling writing. *Written Communication* 29(3). 369–388.

Hayes, John R. & N. Ann Chenoweth. 2006. Is working memory involved in the transcribing and editing of texts? *Written Communication* 23(2). 135–149.

Horváth, P. 2009. A lektori kompetencia [Revising competence]. Budapest: ELTE University doctoral dissertation.

Jakobsen, Arnt Lykke. 2002. Translation drafting by professional translators and by translation students. In Gyde Hansen (ed.), *Empirical Translation Studies: Process and Product*, 191–204. Copenhagen: Samfundslitteratur.

Jakobsen, Arnt Lykke. 2003. Effects of think aloud on translation speed, revision and segmentation. In Fabio Alves (ed.), *Triangulating Translation: Perspectives in Process-oriented Research*, 69–95. Amsterdam: John Benjamins.

Kenny, Dorothy. 1998. Creatures of habit? What translators usually do with words. *Meta* 43(4). 515–523.

Kortmann, Bernd & Benedikt Szmrecsányi. 2009. World Englishes between simplification and complexification. In Lucia Siebers & Thomas Hoffmann (eds.), *World Englishes – Problems, Properties and Prospects: Selected Papers from the 13th IAWE Conference*, 265–285.

Kruger, Haidee. 2012a. A corpus-based study of the mediation effect in translated and edited language. *Target. International Journal of Translation Studies* 24(2). 355–388.

Kruger, Haidee. 2012b. *Postcolonial Polysystems: The Production and Reception of Translated Children's Literature in South Africa*. Amsterdam: John Benjamins.

Kruger, Haidee & Bevan-Dye, Ayesha. 2010. Guidelines for the editing of dissertations and theses: A survey of editors' perceptions. *Southern African Linguistics and Applied Language Studies* 29(3). 293–311.

Kruger, Haidee & Bevan-Dye, Ayesha. 2013. The language editor's role in postgraduate research: A survey of supervisors' perceptions. *South African Journal of Higher Education* 27(4). 875–899.

Kruger, Haidee & Van Rooy, Bertus. 2012. Register and the features of translated language. *Across Languages and Cultures* 13(1). 33–65.

Kruger, Haidee & Van Rooy. 2016. Constrained language: a multidimensional analysis of translated English and non-native indigenised varieties of English. *English World-Wide* 37(1). 26–57.

Künzli, Alexander. 2005. What principles guide translation revision? A combined product and process study. In Ian Kemble (ed.), *Translation Norms: What Is Normal in the Translation Profession? Proceedings of the Conference Held on 13th November 2004 in Portsmouth*, 31–43. Portsmouth: University of Portsmouth.

Künzli, Alexander. 2006. Translation revision: A study of the performance of ten professional translators revising a technical text. In Maurizio Gotti & Susan Šarčevič (eds.), *Insights Into Specialized Translation*, 195–214. Bern: Peter Lang.

Künzli, Alexander. 2007. Translation revision: A study of the performance of ten professional translators revising a legal text. In Yves Gambier, Miriam Shlesinger & Radegundis Stolze (eds.), *Doubts and Directions in Translation Studies: Selected Contributions From the EST congress, Lisbon 2004*, 115–126. Amsterdam: John Benjamins.

Lanstyák, István & Pál Heltai. 2012. Universals in language contact and translation. *Across Languages and Cultures* 13(1). 99–121.

Largy, Pierre, Lucile Chanquoy & Alexandra Dédéyan. 2004. Orthographic revision: The case of subject-verb agreement in French. In Linda Allal, Lucile Chanquoy & Pierre Largy (eds.), *Revision: Cognitive and Instructional Processes*, 39–62. New York: Springer.

Leech, Geoffrey, Nicholas Smith & Paul Rayson. 2012. English style on the move: Variation and change in stylistic norms in the twentieth century. *Language & Computers* 76. 69–98.

Lutz, Jean A. 1987. A study of professional and experienced writers revising and editing at the computer and with pen and paper. *Research in the Teaching of English* 21(4). 398–421.

Mackenzie, Janet. 2004. *The Editor's Companion*. Cambridge: Cambridge University Press.

Malkiel, Brenda. 2009. From Ántonia to My Ántonia: Tracking self-corrections with Translog. In Susanne Göpferich, Arnt Lykke Jakobsen & Inger M. Mees (eds.), *Behind the Mind: Methods, Models and Results in Translation Process Research*, 149–166. Copenhagen: Samfundslitteratur.

Mesthrie, Rajend. 2006. Anti-deletions in an L2 grammar: A study of Black South African English mesolect. *English World-Wide* 27(2). 111–145.

Mesthrie, Rajend. 2010. New Englishes and the native speaker debate. *Language Sciences* 32. 594–601.

Mesthrie, Rajend & Rakesh M. Bhatt. 2008. *World Englishes: The Study of New Linguistic Varieties*. Cambridge: Cambridge University Press.

Milton, John & Paul F. Bandia (eds.). 2009. *Agents of Translation*. Amsterdam: John Benjamins.

Mondorf, Britta. 2014. (Apparently) competing motivations in morpho-syntactic variation. In Brian MacWhinney, Andrej Malchukov & Edith Moravcsik (eds.), *Competing Motivations in Grammar and Usage*, 209–228. Oxford: Oxford University Press.

Mossop, Brian. 2007. Empirical studies of revision: What we know and need to know. *JoSTrans* 8. 5–20.

Mossop, Brian. 2014. *Revising and Editing for Translators* (3rd edition). London: Routledge.

Munday, Jeremy. 2008. *Style and Ideology in Translation: Latin American Writing In English*. New York: Routledge.

Olohan, Maeve & Mona Baker. 2000. Reporting *that* in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures* 1(2). 141–158.

Olohan, Maeve. 2003. How frequent are the contractions? A study of contracted forms in the Translational English Corpus. *Target. International Journal of Translation Studies* 15(2). 59–89.

Peters, Pam. 2004. *The Cambridge Guide to English Usage*. Cambridge: Cambridge University Press.

Peters, Pam. 2006. English usage: Prescription and description. In Bas Arts & April McMahon (eds.), *The Handbook of English Linguistics*. London: Blackwell.

Pinker, Steven. 2014. *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century*. London: Penguin.

Piolat, Annie, Jean-Yves Roussey, Thierry Olive & Murielle Amada. 2004. Processing time and cognitive effort in revision: Effects of error type and of working memory capacity. In Linda Allal, Lucile Chanquoy & Pierre Largy (eds.), *Revision: Cognitive and Instructional Processes*, 21–38. New York: Springer.

Redelinghuys, Karien & Haidee Kruger. 2015. Using the features of translated language to investigate translation expertise: A corpus-based study. *International Journal of Corpus Linguistics* 20(3). 293–325.

Rijlaarsdam, Gert, Michel Couzijn & Huub van den Bergh. 2004. The study of revision as a writing process and as a learning-to-write process: Two prospective research agendas. In Linda Allal, Lucile Chanquoy & Pierre Largy (eds.), *Revision: Cognitive and Instructional Processes*, 189–207. New York: Springer.

Ritter, R. M. 2003. *The Oxford Style Manual*. Oxford: University Press.

Robert, Isabelle S. & Luuk Van Waes. 2014. Selecting a translation revision procedure: Do common sense and statistics agree? *Perspectives: Studies in Translatology*.

Robert, Isabelle S. 2008. Translation revision procedures: An explorative study. In Pieter Boulogne (ed.), *Translation and Its Others: Selected Papers of the CETRA Research Seminar in Translation Studies 2007*. http://www.kuleuven.be/cetra/papers/files/robert.pdf (accessed 30 March 2015).

Robert, Isabelle S. 2013. Translation revision: Does the revision procedure matter? In Catherine Way, Sonia Vandepitte, Reine Meylaerts & Magdalena Bartłomiejczyk (eds.), *Treks and Tracks In Translation Studies*, 87–102. Amsterdam: John Benjamins.

Schneider, Edgar. 2007. *Post-colonial Englishes: Varieties Around the World*. Cambridge: Cambridge University Press.

Scott, Mike. 2013. *WordSmith Tools 6. Liverpool: Lexical Analysis Software*. http://www.lexically.net/wordsmith/version6/index.html (accessed 30 March 2015).

Szmrecsányi, Benedikt & Bernd Kortmann. 2009. Vernacular universals and Angloversals in a typological perspective. In Markku Filppula, Juhani Klemola & Heli Paulasto (eds.), *Vernacular Universals and Language Contact: Evidence from Varieties of English and Beyond*, 33–53. New York: Routledge.

Szmrecsányi, Benedikt M. 2004. On operationalizing syntactic complexity. In *7es Journées Internationales d'Analyse Statistique des Donnés Textuelles*, 1031–1038.

Statsoft Inc. 2013. *Statistica (data analysis software system)*, version 12. http://www.statsoft.com.

Torres Cacoullos, Rena & James A. Walker. 2009. On the persistence of grammar in discourse formulas: a variationist study of *that*. *Linguistics* 47(1). 1–43.

Ulrych, Margherita & Amanda Murphy. 2008. Descriptive translation studies and the use of corpora: Investigating mediation universals. In Carol Taylor Torsello, Katherine Ackerley & Erik Castello (eds.), *Corpora for University Language Teachers*, 141–166. Bern: Peter Lang.

Van Rooy, Bertus, Lize Terblanche, Christoph Haase & Joseph Schmied. 2010. Register differentiation in East African English: A multidimensional study. *English World-Wide* 31(3). 311–349.

Webb, Vic. 2002. *Language in South Africa: The Role of Language in National Transformation, Reconstruction and Development.* Impact: Studies in Language and Society, 14. Amsterdam: John Benjamins.

Westin, Ingrid. 2002. *Language Change in English Newspaper Editorials*. Amsterdam: Rodopi.

Williams, Jessica. 1987. Non-native varieties of English: A special case of language acquisition. *English World-Wide* 8(2). 161–199.

Xiao, Richard. 2010. Idioms, word clusters & reformulation markers in translational Chinese: Can "translation universals" survive in Mandarin? *Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies 2010 Conference (UCCTS 2010).* http://www.lans.ac.uk/fass/projects/corpus/UCCTS2010Proceedings/papers/xiao.pdf (accessed 30 March 2015).

Yagoda, Ben. 2013. *How to Not Write Bad: The Most Common Writing Problems and the Best Ways to Avoid Them*. London: Penguin.

Zanettin, Federico. 2012. *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies.* Manchester: St Jerome.

Adriano Ferraresi and Maja Miličević

5 Phraseological patterns in interpreting and translation. Similar or different?

Abstract: Research in corpus-based studies of translation and interpreting has typically focused on monolingual comparable and/or interlingual parallel perspectives; only more recently intermodal comparisons have been proposed as a new paradigm, aiming to shed light on the traits that distinguish one form of language mediation from the other. Pursuing this line of research, the present contribution draws on EPTIC, a newly created intermodal corpus, to compare phraseological patterns in Italian texts translated and interpreted from English. We investigate whether translations and interpretations differ in terms of use of different types of word pairs (infrequent, highly frequent and strongly associated sequences), and further check whether differences, if any, also apply to oral vs. written non-mediated texts, and/or to mediated vs. non-mediated texts. Results indicate that translations are more phraseologically conventional than interpretations in terms of the majority of the parameters considered, and that these two forms of mediated output are more dissimilar to each other than they are to comparable non-mediated texts. We hypothesize that the observed differences are related to cognitive and task-related constraints characterizing the translation and interpreting processes.

1 Introduction

Research in corpus-based translation and interpreting studies (CBTS and CBIS) has typically been based on two prevailing approaches, the interlingual parallel and the monolingual comparable one. Within the former, translated/interpreted texts are compared to their sources, e.g. in research on shifts and translators/interpreters' strategies; within the latter, they are contrasted to comparable original written or oral production, most notably in a search for regularities (patterns, universals, . . .) characterizing translated/interpreted texts as such. Whichever the approach adopted, translation and interpreting have largely been investigated independently of one another, within the boundaries of the respective disciplines.

It is only more recently that intermodal comparisons between translated and interpreted outputs have been proposed in the literature. First introduced

DOI 10.1515/9783110459586-006

within the corpus-based paradigm by Shlesinger (1998), such an approach can be viewed as part of a more general tendency to lay the grounds for closer collaboration between translation and interpreting studies at large (see e.g. the contributions in Schäffner 2004a). The underlying idea, as remarked by Gile (2004: 23), is that "since translation and interpreting share so much, the differences between them can help shed light on each, so that besides the autonomous investigation of their respective features, each step in the investigation of one can contribute valuable input towards investigation of the other". In a wider perspective, due to sharing features with, for example, edited language and second language production, translation and interpreting are increasingly seen as instances of *constrained varieties*, i.e. types of language use that are subject to a number of cognitive and/or social restrictions, which vary along the dimensions of language activation (monolingual vs. bilingual), modality (spoken vs. written) and text production (mediated vs. non-mediated) (see Kruger 2014).

Despite the evident interest, corpus-based studies on the differences and similarities between these two forms of language mediation are still scarce. Apart from the relative novelty of the approach, one possible reason is that building appropriate corpus resources allowing this type of investigation has proved a major hurdle (Shlesinger and Ordan 2012).

Drawing on the newly created bidirectional (English<>Italian), intermodal, comparable and parallel European Parliament Translation and Interpreting Corpus (EPTIC; Bernardini et al. 2016), the present contribution aims to take steps towards filling this gap by comparing translated and interpreted Italian texts. Specifically, the focus is on phraseological patterns. Phraseology has traditionally occupied centre stage in investigations of translated vs. non-translated language under the hypothesis that the former displays untypical patterning at the lexical level compared to the latter, e.g. in terms of the word combinations used and/or their frequency, and that this may be the result of the translation task itself (Mauranen 2000). Recently, it has been suggested that such untypical patterning may also characterize other forms of constrained communication, including bilingual communication (Lanstyák and Heltai 2012: 106). In this framework, phraseology seems a particularly rewarding area of enquiry to compare translation and interpreting and assess whether and to what extent any differences or similiarities can be explained in terms of the constraints governing the two tasks.

Extending the method proposed by Durrant and Schmitt (2009), and exploiting EPTIC's multi-faceted setup, we investigate whether translated and interpreted texts differ in terms of use of different types of phraseological items. To gain a full picture, we further check whether differences, if any, also apply to oral vs. written texts in general and/or to mediated vs. non-mediated texts, thus encompassing different dimensions of variation, and novel and more traditional approaches to the analysis of translated and interpreted language.

The remainder of the chapter is structured as follows. Section 2 reviews previous work carried out in the areas of intermodal corpora and phraseology in translation and interpreting studies. In section 3 we present our research setup, describing EPTIC and the method used to identify and compare phraseological patterns across its sub-corpora. Section 4 presents the results of the analysis, and section 5 discusses them, highlighting their relevance for translation and interpreting research. Finally, section 6 offers concluding remarks and points at ways in which the present work can be extended.

2 Previous work

2.1 Intermodal corpora and the search for features of translated vs. interpreted texts

In her outline of possible ways in which translation and interpreting studies have moved, or may move, towards greater integration, Schäffner (2004b: 4) mentions the development of "models, frameworks, and research methods that can equally be applied to study the two modes". Crucially, corpus linguistics is mentioned pervasively in the literature as one such framework, alongside other linguistic and cognitively-oriented approaches (see e.g. Gile 2004: 20; Pöchhacker 2004: 115; Snell-Hornby 2006: Chapter 4.1).

The work of Miriam Shlesinger has been pivotal in shaping the research agenda of corpus-based explorations at the interface between translation and interpreting. In particular, she first put forward the idea of extending monolingual comparable corpora of interpreted and original speeches by adding an intermodal component, i.e. mediated texts produced in the written modality (i.e. translations), arguing that this would allow one to discern "the characteristics of interpreting qua interpreting" (Shlesinger 1998: 488). However, in later work she acknowledged that this would also allow one to observe "differences between the oral and written modalities of translation, [and] to observe the effects of the ontology variable (original vs. translated) as well" (Shlesinger and Ordan 2012: 47), hence redressing the balance between interpreting- and translation-centred perspectives.[1]

Focusing on this line of research, Shlesinger (2009) built a small monolingual intermodal corpus containing the translational and interpretational output produced in an experimental setting by six Hebrew translators/interpreters, who first rendered an English source text (henceforth ST) orally, and then, three years later, rendered the same ST in writing. In a follow-up study, Shlesinger and Ordan (2012) collected a larger intermodal corpus reflecting more closely the corpus setup initially envisaged by Shlesinger (1998), i.e. one containing interpreted and translated target texts (from English into Hebrew; henceforth TT) and original (Hebrew) speeches produced in authentic professional conditions, specifically in the academic domain. In the two studies, the author(s) investigate a large set of features, including measures of lexical variety, part-of-speech distributions and lexical aspects such as the use of formal/colloquial terms. Overall, their findings suggest that interpreting "exhibits far more similarities to original speech than to written translation" (Shlesinger and Ordan 2012: 47), i.e. that the modality variable (being oral vs. written) has a stronger effect than the ontology variable (being mediated vs. non-mediated).

An intermodal comparable corpus was also built by Kajzer-Wietrzny (2012). This corpus is based on the European Parliament plenary sessions, and contains texts interpreted and translated into English from different languages (French, Spanish, German and Dutch), as well as original oral texts in English. However, even though the corpus setup would have allowed intermodal comparisons, the author limited her analysis to the oral sub-corpus. The results are therefore of limited relevance in this context.

Finally, Bernardini et al. (2016) introduced EPTIC, which, unlike the corpora mentioned so far, is a bidirectional (English<>Italian), intermodal comparable and parallel corpus, including translation and interpreting outputs of pseudo-parallel STs, as well as the written and oral STs themselves (see section 3.2 for a more detailed description). As a first attempt at unearthing the potential of the corpus, the authors carried out a study of lexical simplification both at the monolingual comparable level, finding that the mediation process reduces complexity in both interpreting and translation and in both language directions, as well as the intermodal level, suggesting that interpreters simplify more than translators do. The experiment presented in sections 3–5 can be seen as a follow up to that study, focusing on a different area of linguistic enquiry, which has

[1] Unlike Shlesinger and Ordan (2012), in this chapter we use the term *translation* to refer to the written modality of language mediation (as opposed to the oral modality of mediation, i.e. *interpreting*), and not as an overarching term that encompasses both translation and interpreting.

received considerable attention within CBTS (and to a much more limited extent in CBIS): phraseology.

2.2 Phraseology in corpus-based translation and interpreting studies

In CBTS, the investigation of lexical issues at the multi-word level (phraseology, broadly conceived) has occupied a prominent position within the discipline. In general terms, units of language variously identified as collocations, set phrases, idioms, etc. are compared in corpora of translated and non-translated language, in a search for systematic differences at the quantitative and/or qualitative level; viewing phraseological units as signals of conventionalized language, the differences are then interpreted against the background of purported translation norms or universals, such as interference or standardization/normalization (for an overview, see e.g. Marco 2009: 844–847).

Despite the prominence in its neighbouring discipline, phraseology has not received the same attention in CBIS. References to the notion of collocation can be found in process- and cognitively-oriented studies: examples include Zanetti (1999) and Vandepitte (2001), who discuss collocations in the context of interpreters' anticipation strategies, as well as Shlesinger (2003), who uses complex noun phrases composed of a noun and a series of collocating adjectives as experimental items in a study on interpreters' working memory. From a more product-oriented perspective, collocations are also mentioned in passing by Setton (2011) as a salient feature for investigations of *interpretese*. However, to the best of our knowledge, no in-depth corpus-based study of phraseology in interpreted language has been carried out so far.

Going back to CBTS, since the full range of theories and methods proposed in the literature is more vast than it would be possible to chart here, only a few selected studies will be presented in what follows. These were chosen as exemplars of the two main approaches outlined in section 1, i.e. the interlingual parallel and the monolingual comparable one. We mainly report on studies that have focused on one specific type of phraseological unit, i.e. collocations, loosely defined in this context as two-word lexical sequences occurring with a higher than chance frequency (cf. also section 3.1).

Kenny (2001) and Marco (2009) exemplify the use of parallel corpora to search for and compare phraseological units in STs and the corresponding TTs. Drawing on the German-English Parallel Corpus of Literary Texts (GEPCOLT), Kenny (2001) isolates the creative/unusual word combinations involving the noun *Auge* ('eye') in the German ST sub-corpus, and looks at their translations into English; her results suggest a slight tendency of translators to normalize non-standard uses, i.e. to replace them with forms that are more lexicalized in the target language (e.g. the ST *als habe er Augen im Nacken*, a non-lexicalized combination in German meaning literally 'as if he had eyes in the back of his neck' is rendered in English by the fully conventional *as if he had eyes in the back of his head*; cf. Kenny 2001: 193); reported figures for this kind of shift are in the 22–44% range. Marco (2009) carries out an analogous study on the English > Catalan component of the Corpus Valencien de Textes Littéraires Traduits (COVALT). Focusing on the words *eye*, *foot* and *hand* in English STs, and on their translational equivalents *ull*, *peu* and *mà* in Catalan TTs, he classifies all the cases in which they occur within sentence- and phrase-length phraseological units that he terms *utterances* and *idioms* (see also Marco 2009: 845). His results challenge those of Kenny (2001) insofar as they hint at translated texts being "less phraseological than their corresponding English source texts", even though evidence emerges of "some effort on the part of translators to retain or recreate a noticeable degree of phraseological activity in translated texts" (Marco 2009: 853).

In both of the studies just mentioned, a selection of phraseological items is carried out, which has two major drawbacks: first, despite being based on explicit parameters, it involves a certain amount of subjectivity, as the authors themselves acknowledge (cf. Marco 2009: 849; Kenny 2001: 210–211); and, second, it concentrates on a very narrow subset of the phraseological units occurring in a corpus. Clearly such selection is justified by the limits imposed by manual pruning and categorization of shifts, but it may ultimately hinder generalizations of results.

The same methodological limitation is also frequent in studies adopting the monolingual comparable approach. Within this approach, phraseological patterns involving specific, more or less arbitrarily selected node words are identified in translated (sub-)corpora and then compared with those found in non-translated ones, e.g. in terms of collocation types (cf. Jantunen 2004 on the lexico-grammatical patterning of three synonymous Finnish degree modifiers meaning 'very'), or literal vs. non-literal uses (cf. Baker 2007 on the idioms *off the hook* and *out of order*).

To overcome the limitations faced by previous studies, Dayrell (2007) and Bernardini (2007) propose objective, replicable methods to select node words and their collocates for subsequent in-depth scrutiny. Specifically, Dayrell (2007) suggests that nodes should be chosen among high-frequency words having similar frequencies in the translated and non-translated sub-corpora, and that the collocates of such words should then be chosen relying on Mutual Information (MI), a statistical association measure (AM). Bernardini (2007) goes a step further, and proposes a method whereby *all* word pairs are taken into account, provided that: a) they occur in pre-determined part-of-speech patterns (e.g. Adjective-Noun, Noun-Noun, etc.); and b) they are above a certain threshold of frequency and/or MI, as measured in an external reference corpus; the implementation of the latter criterion also allows one to overcome the data sparseness problem often faced by phraseological studies in CBTS (cf. also Dayrell 2007: 381). The two authors apply their method respectively to a monolingual comparable corpus of literary texts in Brazilian Portuguese (Dayrell 2007), and to a bilingual parallel and comparable corpus of literary texts in Italian and English (Bernardini 2007). The conclusions they reach diverge in important ways: Dayrell reckons that translated texts show *less* variety in the use of collocations than comparable non-translated texts, suggesting that the former do not draw on the full range of phraseological items available in the language. Bernardini, on the contrary, concludes that translated texts make more use of collocations than their non-translated counterparts, insofar as they feature both a larger variety of, and stronger/more lexically associated word combinations.

Summing up, a blurred picture emerges from previous work on phraseology in CBTS, with some studies suggesting that translated language is less phraseologically patterned than non-translated language, while others, on the contrary, show that the translation process leads to increased language conventionality. Such divergences are deeply related to the multitude of approaches in defining the notion of phraseology itself. If anything, this might be taken as an incentive to adopt increasingly objective and replicable methods to identify and compare phraseological patterns. As concerns CBIS, very few studies of phraseology have been carried out, and the same applies to studies comparing phraseological patterns across translated and interpreted language. The investigation presented in the next sections represents our attempt to start filling these gaps.

3 Research questions, corpus and method

3.1 Research questions

The main question addressed by the present study is whether texts that are interpreted vs. translated from English into Italian differ in terms of phraseological patterns. Specifically, we focus on the use of word combinations characterized by different collocational strengths. In choosing this approach we follow the neo-Firthian frequency-based tradition in the study of English collocations, and apply (to Italian) its view of collocation as a combination of words that occur together more often than predicted by chance (see e.g. Jones and Sinclair 1974/1996).

To check whether the detected differences or similarities are indeed characteristic of interpreting vs. translation, rather than reflecting more general modality-specific language properties, we also investigate the use of phraseology in non-mediated oral and written texts. Lastly, taking full advantage of EPTIC's structure, we look at the impact of the ontology variable (in Shlesinger and Ordan's 2012 terms; cf. section 2.1), comparing phraseological patterns across mediated and non-mediated texts.

3.2 Corpus description

EPTIC, the European Parliament Translation and Interpreting Corpus (Bernardini et al. 2016) is an intermodal corpus that builds on the well-known EPIC (European Parliament Interpreting Corpus; Sandrelli and Bendazzoli 2005; Bendazzoli 2010), created by transcribing a number of original European Parliament speeches and their interpretations into selected languages (the combinations represented in EPIC are English<>Italian<>Spanish). During the creation of EPTIC, EPIC's transcripts of interpreted speeches and their STs were paired with the corresponding translated versions and respective STs. This was made possible by the fact that, for each plenary session, the European Parliament publishes so-called verbatim reports of proceedings consisting of transcripts of the speeches and their translations into all EU official languages. Crucially, despite being called verbatim, the reports are edited, sometimes considerably, starting from the addition of punctuation and the removal of context-related comments, to the correction of mistakes such as false starts, unfinished sentences or mispronunciations (see the example in Table 1). The translations of the proceedings are the result of an independently performed translation process based on the verbatim reports, without any reference to the interpreters' outputs (as confirmed by several EU officials consulted on this issue).

The language combination in EPTIC is English-Italian, including translations/ interpretations in both directions.[2] Considering all its sub-corpora, comprising simultaneous interpretations paired with their STs, plus corresponding translations and STs (a total of eight components), EPTIC can be classified as an intermodal, comparable and parallel corpus. Its structure is shown in Figure 1 (the *st-* and *tt-* prefixes indicate source and target texts, the *-in-* and *-tr-* affixes interpretations and translations, and the *-en* and *-it* suffixes the language the texts are in – English and Italian respectively).

[2] The Spanish component of EPIC has been left out, and a French component is currently being added.

Table 1: First lines of a transcribed speech and the corresponding verbatim report

Transcript of the original speech	Verbatim report
thank you very much for slotting me in the speakers' list // I'm very sorry that I was late for the debate // ehm on as we're talking about and addressing the issue of the European economy ehm in the debate on these reports ehm the point I'd like to emphasise is the vital importance of turning the Lisbon agenda from rhetoric into reality //	Mr President, as we are addressing the issue of the European economy in the debate on these reports, the point I would like to emphasise is the vital importance of turning the Lisbon agenda from rhetoric into reality.

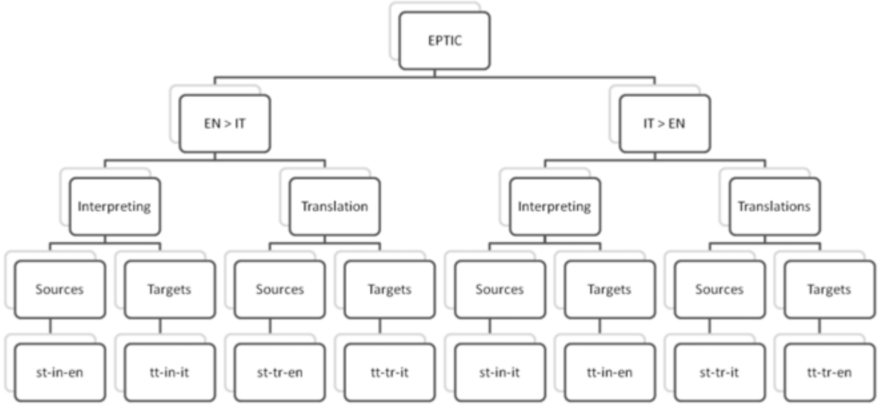


Figure 1: EPTIC structure

The construction of EPTIC is an ongoing project; at the moment the corpus as a whole contains 568 individual texts, for a total of 250,093 words (disregarding truncated words in interpreted texts). The bigger, English > Italian portion contains four versions of 81 texts, while the smaller Italian > English one has four versions of 61 texts. The majority of these texts are directly based on EPIC, i.e. they derive from the European Parliament speeches delivered at the Parliament's part-session held in February 2004. A more recently added portion, comprising 44 new texts interpreted and translated from Italian into English, comes from the March, April and July sessions of the same year.

Only the Italian component of EPTIC was used in the present study, amounting to a total of 119,548 words; the sizes of the individual sub-corpora are shown in Table 2.

Table 2: Sizes of Italian sub-corpora

Sub-corpus	N. of texts	Word count
tt-in-it	81	33,675
tt-tr-it	81	36,876
Total	162	70,551
st-in-it	61	24,866
st-tr-it	61	24,131
Total	122	48,997
Total	**284**	**119,548**

EPTIC is part-of-speech tagged and lemmatized using the TreeTagger,[3] and indexed with the Corpus WorkBench.[4] Moreover, each text within the corpus is aligned at sentence level with its ST/TT and with the corresponding text in the other modality (oral/written). Rich metadata are also encoded in the corpus and can be used to perform complex queries based on specific characteristics of texts and/or the speakers who delivered them. Available metadata, part of which were inherited from EPIC, include speech duration, delivery speed, delivery type (read, impromptu or mixed), text topic and length, as well as speaker and interpreter details (e.g. their gender and native language status).

3.3 Method

The method adopted in this study for investigating phraseological patterns in interpreted vs. translated Italian is (loosely) based on the work of Durrant and Schmitt (2009), who compared native and non-native use of English phraseology. The main merits of the chosen approach are that a) it offers a principled way of classifying collocations for the purpose of quantitative between-corpus comparisons; b) it overcomes the data sparseness problem often encountered when working with relatively small corpora by identifying different types of collocations based on a large external reference corpus (cf. also section 2.2); and c) it takes into account within-corpus variability by looking at individual texts rather than at corpora as wholes. The novelty of our study lies in applying the method to a new domain (translation/interpreting) and a new language (Italian), as well as in adjusting some of the data analysis procedures.

[3] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[4] http://cwb.sourceforge.net/

Similarly to Durrant and Schmitt, we focused on two syntactic patterns, namely modifier + noun (e.g. *precedenti osservazioni* 'previous observations') and noun + modifier (e.g. *comunità internazionale* 'international community'). All modifier + noun pairs were directly adjacent and formed of an adjective and a noun; noun + modifier combinations were also adjacent when comprising a noun followed by an adjective, whereas they were separated by a preposition (mostly *di* 'of', as in *attività (di) ricerca* 'research activities') when both elements were nouns; the latter possibility was allowed since it constitutes a counterpart of noun + noun combinations in English (such as *research activities*). Noun + modifier patterns are more frequent in Italian overall, due not only to combinations involving two nouns, but also to some qualifying adjectives obligatorily appearing after the noun (e.g. geographical and ethnic adjectives such as *europeo* 'European'). Given the complex nature of adjective placement in Italian, in which many adjectives can appear in either position (cf. *iniziale mancanza* 'initial lack' vs. *portata iniziale* 'initial reach'), and to reduce the overall complexity of the study, we do not carry out separate analyses for the two types of syntactic patterns.

Combinations involving nouns and their modifiers were chosen for reasons of methodological comparability with Durrant and Schmitt's study, and because they are among the patterns which are cited in the literature as being more collocationally productive, both in English and Italian (see e.g. Bernardini 2007 and references therein). The choice of a single pattern is of course an arbitrary decision, which limits the possibility of generalizing results to other types of collocations. Nonetheless, it provides a starting point for future studies looking at other collocationally productive patterns.

We extracted relevant word pairs from EPTIC's interpreted and translated sub-corpora (tt-in-it, tt-tr-it), and from the non-mediated spoken and written ones (st-in-it, st-tr-it) relying on available part-of-speech information. Thanks to their separate tags,[5] combinations containing proper nouns, numbers, pronouns, possessives and demonstratives were automatically excluded (cf. Durrant and Schmitt 2009: 166). On the other hand, the automatic tagging procedure inevitably led to some errors: a cursory manual inspection of the extracted lists detected sequences such as *Presidente Barroso* 'President Barroso' or *tale processo* 'such process'. Similar cases, however, were not frequent and were similarly dispersed across all sub-corpora: as such, it was assumed that they would not systematically skew the results. In order to keep subjective interventions to a minimum, it was therefore decided not to follow up on the automatic extraction of word sequences with manual pruning.

[5] The full Italian tagset is available at http://sslmit.unibo.it/~baroni/collocazioni/itwac.tagset.txt.

For each extracted sequence, three measures were computed relying on frequency data obtained from an external reference corpus of Italian: raw frequency, *t*-score and Mutual Information values.[6] The reference corpus we used was a 200-million-word random subset of the web-derived itWaC, one of the largest available corpora of Italian (Baroni et al. 2009). Using a general language corpus like itWaC has the advantage of allowing a classification of bigrams based on their frequency in Italian in general, rather than in the specific text type/genre under scrutiny; while, in principle, large specialized corpora could also be used for this task, they may provide unreliable frequency information for bigrams that are not typical of the domain they represent (Durrant and Shcmitt 2009: 160). We classified each EPTIC word sequence based on each of the three measures in the following way: infrequent/unattested vs. frequent sequences ($fq < 2$ vs. $fq \geq 2$ in itWaC), and strong vs. weak collocations based on two AMs, namely *t*-score ($t \geq 10$ vs. $t < 10$ in itWaC), and MI (MI ≥ 7 vs. MI < 7 in itWaC).[7] The infrequent/unattested category is meant to account for extremely uncommon and potentially non-standard usages in the Italian language; unlike Durrant and Schmitt (2009), who set a rather conservative frequency threshold of 5, we include in this category pairs with 0 or 1 occurrences in our reference corpus, i.e. pairs which are either not attested or would not count as collocations within a frequency-based paradigm (since a minimum frequency of 2 is required for a word pair to count as a collocation, i.e. a recurrent sequence of words; Jones and Sinclair 1974/1996). Two different AMs were taken into account to emphasize different types of word combinations: *t*-score is expected to highlight "very frequent collocations" (Durrant and Schmitt 2009: 167; see also Stubbs 1995; e.g. *diritti umani* 'human rights'), while MI gives prominence to "word pairs which may be less common, but whose component words are not often found apart" (Durrant and Schmitt 2009: 167; e.g. *partenariato strategico* 'strategic partnership'). Both measures have been abundantly used in the study of English collocations, and in British lexicography in general (see Evert 2005: 200), while their uptake in Italian linguistics and lexicography is much more limited (but see e.g. Masini 2012; Nissim et al. 2014). The cut-off points between weak and strong collocations were taken from Durrant and Schmitt, who found the values

[6] *T*-score values were calculated using the UCS toolkit (http://www.collocations.de/software.html), and MI values using an ad-hoc script implementing the formula by Church and Hanks (1990).

[7] Like Durrant and Schmitt (2009), we impose a $fq \geq 5$ criterion (in addition to $t \geq 10$ or MI ≥ 7) for sequences to be classified as high-*t*-score or high-MI, and we automatically classify all other sequences as low-*t*-score and low-MI. This approach prevents that unreliably high values of *t*-score/MI deriving from low frequency counts are treated as instances of high *t*-score/MI.

of 10 and 7 for *t*-score and MI respectively to mark the lower boundary past which native and non-native speakers differ. The adoption of cut-off points suggested in the literature, like in this case, ensures methodological comparability and is a rather common practice in phraseology studies, even when comparability comes at the expense of identifying cut-off points which are potentially better tailored to the new domain of investigation. This point is raised, among others, by Granger and Bestgen (2014: 240), who also adopt Durrant and Schmitt's parameters in their own experiment.

The number of word combination tokens belonging to infrequent/unattested, high-*t*-score and high-MI sequences was then calculated for each text in each sub-corpus and expressed as a percentage, e.g. of high-MI combinations out of the total number of word combinations found in a text, irrespective of their frequency and collocational strength; percentages were used in the analysis rather than raw frequencies due to between-corpus differences in the total number of combinations (see section 4). In order to check for a possible effect of repetitions, the same procedure was repeated for word combination types (cf. Durrant and Schmitt 2009: 171–172).

Due to some data being non-normally distributed, differences in median percentages of each type of word combinations in translated vs. interpreted, oral vs. written and mediated vs. non-mediated texts were tested for significance using non-parametric Wilcoxon rank sum tests in R.[8] After visual inspection of the data, we decided to run analyses both on complete and on cleaned data sets, i.e. with outliers removed; in the results section we only report the latter in the statistical tables, but we show the original data in the graphs. For each significant test we also calculated the effect size in the form of Pearson's correlation coefficient (r), using the function provided by Field et al. (2012: 665); coefficient values close to .10 (or lower) indicate a small effect, those around .30 point to a medium effect, while those around .50 or more are indicative of a large effect (Field et al. 2012: 58).

By applying the described procedure, we combined intermodal (tt-in-it vs. tt-tr-it and st-in-it vs. st-tr-it) and comparable (tt-in-it vs. st-in-it and tt-tr-it vs. st-tr-it) perspectives. As per our research questions, we were primarily interested in comparing the two modalities of mediation by examining the TTs resulting from the processes of interpreting and translation. The original (non-mediated) texts were looked at as a control comparison: while it should be remarked that these texts are comparable, and not parallel, with respect to the translated/ interpreted texts, any difference emerging from this comparison could be indicative of a more general oral vs. written (rather than interpreted vs. translated)

[8] http://www.r-project.org/

distinction. Lastly, we also conducted traditional monolingual comparable comparisons between mediated and non-meditated texts in the same modality, in order to verify whether and to what extent the mediation variable itself exerts an effect on phraseological patterning.

4 Results

Following the procedure outlined above, a total of 14,000 modifier + noun and noun + modifier combinations were retrieved from the Italian component of EPTIC; their distribution across the four sub-corpora is shown in Table 3.

Table 3: Summary of word combinations by sub-corpora

Sub-corpus	Total combinations retrieved	Combinations per 10,000 words	Median combinations per text
tt-in-it	3638	1080.33	30
tt-tr-it	4945	1340.98	46
st-in-it	2716	1092.25	34
st-tr-it	2701	1119.31	35

Even a superficial inspection of the data in Table 3 points to a between-sub-corpus difference in the total number of word combinations. Indeed, the statistical analysis conducted on the normalized figures revealed that the Italian sub-corpora of EPTIC do differ in the total frequencies of the word combinations under scrutiny, as shown in more detail in Table 4 and Figure 2. In particular, the sub-corpus comprised of translated texts (tt-tr-it) stands out as the richest in terms of relevant word combinations; the difference between this sub-corpus and the one of interpreted texts (tt-in-it) is highly statistically significant, as is the difference between the sub-corpora of translation TTs and comparable STs (st-tr-it); no significant difference is found between the oral and written ST sub-corpora, nor between the interpreting TT and ST sub-corpora. Note also that the interpreted and translated texts occupy the opposite ends of the scale, having respectively the lowest[9] and the highest incidence of noun + modifier and modifier + noun sequences.

[9] A possible explanation for interpreted data displaying the lowest number of noun + modifier and modifier + noun word combinations might be the tendency, hypothesized by Shlesinger (2003), for interpreters to omit adjectives as a means of reducing cognitive load. While we cannot pursue the matter further here, seeking confirmation of this hypothesis by looking at single syntactic patterns (so as to tell apart adjectival and nominal modifiers), and perusing parallel concordances might constitute a worthwhile subject for future work.

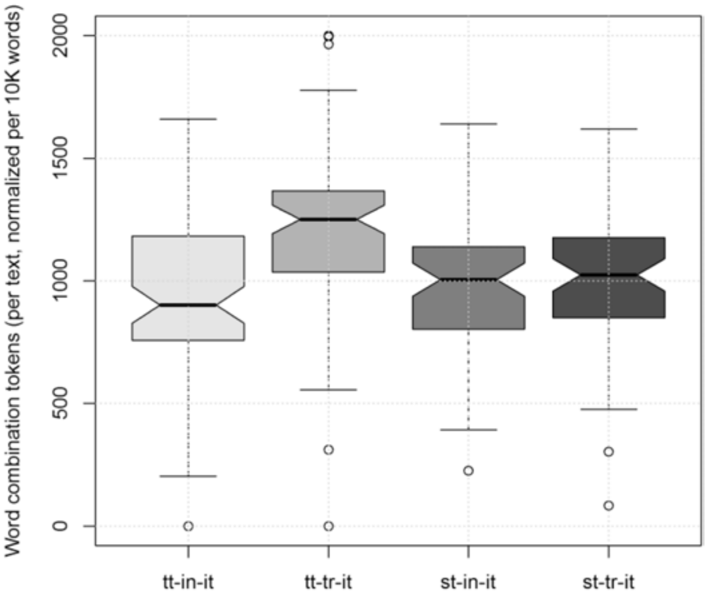


Figure 2: Total frequency of word combinations

These findings were decisive in motivating the choice to conduct further between-sub-corpus comparisons based on percentages. As also pointed out by Durrant and Schmitt (2009: 169–170), if significant differences are observed in the total number of word combinations across sub-corpora, using raw counts in further analyses might confound results: any difference observed in terms of use of different types of combinations (unattested, high-*t*-score, high-MI) could be due to a greater use of modifier + noun and noun + modifier constructions overall, rather than a greater degree of reliance on the specific type of combinations under scrutiny.

Table 4: Total frequency of word combinations

Intermodal		**Comparable**	
tt-in-it	**tt-tr-it**	**tt-in-it**	**st-in-it**
M = 920.3	M = 1250	M = 920.3	M = 1011.1
W = 1497.5, p = .000* (r = −.415)		W = 2094, p = .350	
st-in-it	**st-tr-it**	**tt-tr-it**	**st-tr-it**
M = 1011.1	M = 1030.7	M = 1250	M = 1030.7
W = 1637.5, p = .483		W = 3031.5, p = .000* (r = −.317)	

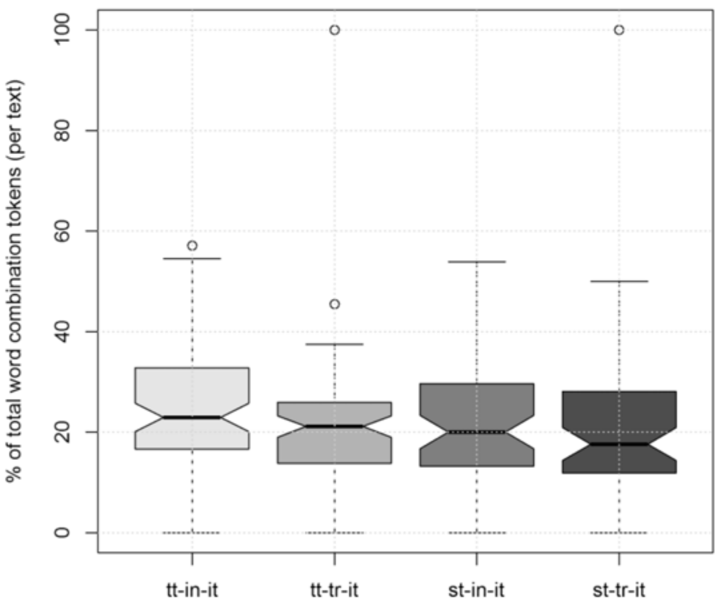


Figure 3: Infrequent/unattested word combinations – tokens

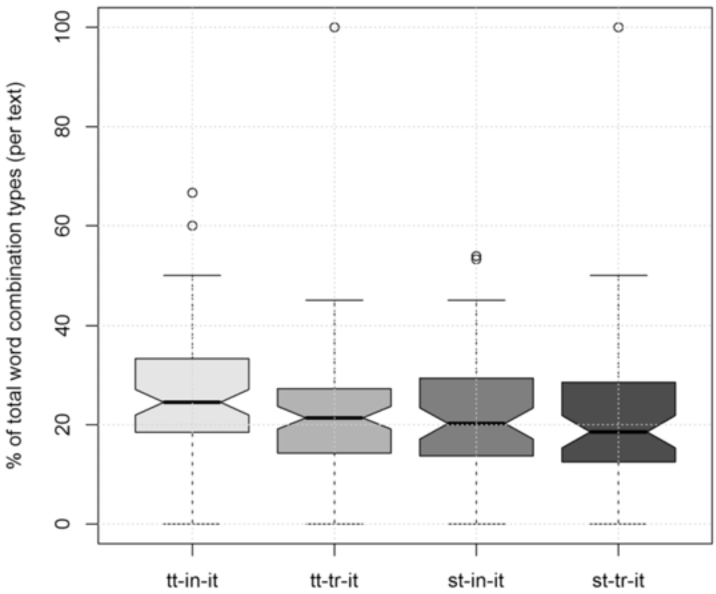


Figure 4: Infrequent/unattested word combinations – types

Table 5: Infrequent/unattested word combinations (%) – tokens and types

	Intermodal		**Comparable**	
Tokens	**tt-in-it**	**tt-tr-it**	**tt-in-it**	**st-in-it**
	M = 22.63	M = 20.64	M = 22.63	M = 20
	W = 3835.5, p = .013* (r = −.198)		W = 2764, p = .178	
	st-in-it	**st-tr-it**	**tt-tr-it**	**st-tr-it**
	M = 20	M = 17.65	M = 20.64	M = 17.65
	W = 2004, p = .368		W = 2387.5, p = .840	
Types	**tt-in-it**	**tt-tr-it**	**tt-in-it**	**st-in-it**
	M = 24	M = 20.90	M = 24	M = 20.90
	W = 3813.5, p = .016* (r = −.192)		W = 2801, p = .043* (r = −.171)	
	st-in-it	**st-tr-it**	**tt-tr-it**	**st-tr-it**
	M = 20	M = 18.39	M = 20	M = 18.39
	W = 1888.5, p = .531		W = 2522, p = .519	

Table 6: High-t-score combinations (%) – tokens and types

	Intermodal		**Comparable**	
Tokens	**tt-in-it**	**tt-tr-it**	**tt-in-it**	**st-in-it**
	M = 29.68	M = 31.82	M = 29.68	M = 29.03
	W = 2588, p = .566		W = 2384, p = .364	
	st-in-it	**st-tr-it**	**tt-tr-it**	**st-tr-it**
	M = 29.03	M = 30.63	M = 31.82	M = 30.63
	W = 1515.5, p = .288		W = 2212.5, p = .762	
Types	**tt-in-it**	**tt-tr-it**	**tt-in-it**	**st-in-it**
	M = 26.27	M = 27.87	M = 26.27	M = 28.13
	W = 2949, p = .267		W = 2239, p = .341	
	st-in-it	**st-tr-it**	**tt-tr-it**	**st-tr-it**
	M = 28.13	M = 29.22	M = 27.87	M = 29.22
	W = 1679.5, p = .355		W = 2223, p = .308	

Moving on to the central analyses, the percentages of infrequent/unattested word combinations in each of the sub-corpora are shown in Figures 3 and 4, for tokens and types respectively. Two examples of such combinations are *esimio ospite* ‘distinguished guest’ and *esseri seri* ‘serious beings’ (cf. high-frequency combinations such as *anno scorso* ‘last year’ or *esseri umani* ‘human beings’). Table 5 shows the cleaned median values and the results of the statistical tests, which detected a significant difference between the interpreted and translated

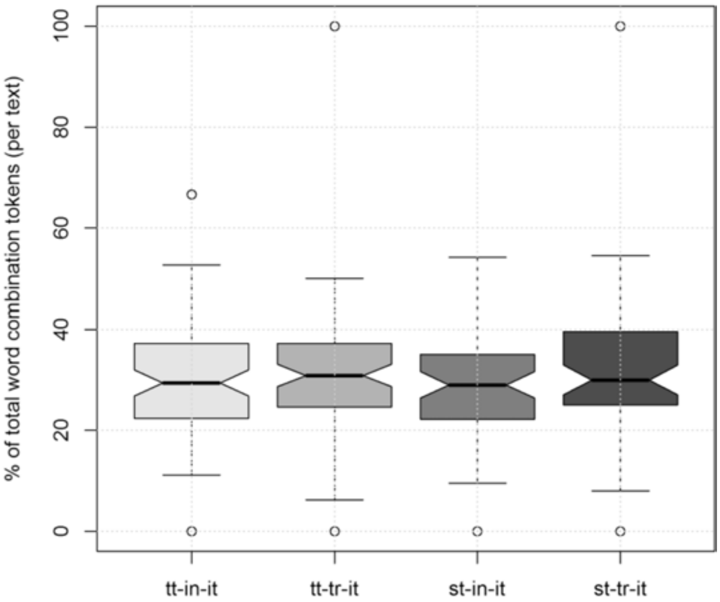


Figure 5: High-t-score combinations – tokens

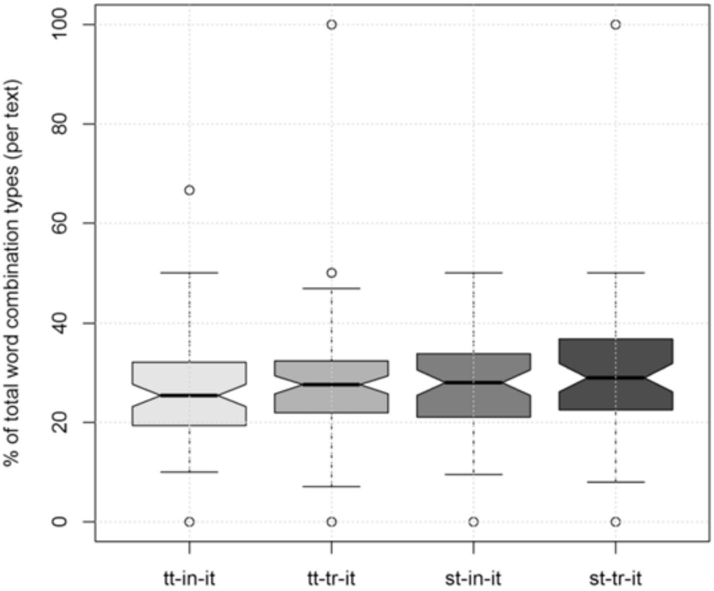


Figure 6: High-t-score combinations – types

Table 7: High-MI combinations (%) – tokens and types

	Intermodal		**Comparable**	
Tokens	**tt-in-it**	**tt-tr-it**	**tt-in-it**	**st-in-it**
	M = 25.76	M = 28.93	M = 25.76	M = 28.59
	W = 2320.5, p = .021* (r = −.186)		W = 2215.5, p = .438	
	st-in-it	**st-tr-it**	**tt-tr-it**	**st-tr-it**
	M = 28.59	M = 27.27	M = 28.93	M = 27.27
	W = 1733.5, p = .619		W = 2435.5, p = .431	
Types	**tt-in-it**	**tt-tr-it**	**tt-in-it**	**st-in-it**
	M = 22.58	M = 26.02	M = 22.58	M = 26.47
	W = 2217, p = .010* (r = −.208)		W = 1863, p = .034* (r = −.178)	
	st-in-it	**st-tr-it**	**tt-tr-it**	**st-tr-it**
	M = 26.47	M = 26.32	M = 26.02	M = 26.32
	W = 1614, p = .711		W = 2025, p = .796	

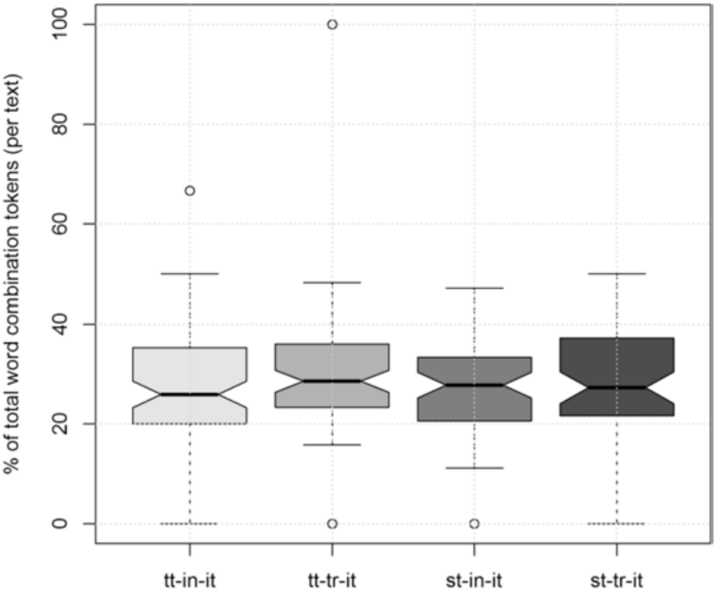


Figure 7: High-MI combinations – tokens

texts. Interpreted texts were found to contain significantly more infrequent/ unattested combinations than translated texts in both token-based and type-based analyses, which was not the case with oral vs. written STs. In addition, in the analysis of tokens neither translations nor interpretations differ from comparable STs in the same modality; however, the difference is closer to significance

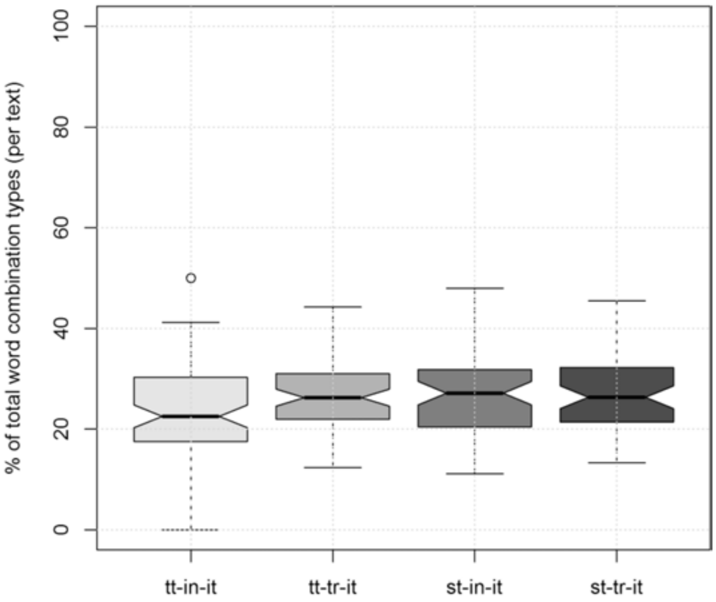


Figure 8: High-MI combinations – types

for interpretations, for which a significant difference is reached in the analysis of types.

Turning to the analysis of high-*t*-score combinations, similar percentages are found in all four sub-corpora (see Table 6, and Figures 5 and 6). That is, regardless of either modality or ontology, the Italian sub-corpora of EPTIC contain similar amounts of highly frequent modifier + noun and noun + modifier combinations. Results are fully parallel for collocation tokens and types. High-*t*-score combinations are exemplified by e.g. *sistema economico* 'economic system' and *vita pubblica* 'public life' (with low-*t*-score counterparts such as *progressi concreti* 'concrete progress' and *popolazione avicola* 'bird population').

Further, as can be seen in Table 7 and Figure 7, a single difference emerges as significant in the token-based results related to MI scores, namely that between the translated and interpreted texts: interpreted texts contain fewer highly idiomatic modifier + noun and noun + modifier sequences than translated texts (e.g. *malattie trasmissibili* 'transmissible deseases' or *protocollo aggiuntivo* 'additional protocol', as opposed to sequences with lower idiomaticity such as *politica sociale* 'social policy' or *mercato unico* 'single market'). On the contrary, no difference is observed between the oral and written non-mediated sub-corpora, nor in any of the monolingual comparable comparisons. As was the case for the infrequent/unattested pairs, however, the type-based analysis finds additional significant differences, showing that interpreted texts differ not only from their translated counterparts, but also from comparable non-mediated oral texts (see Figure 8).

5 Discussion of results

To summarize, the results presented in section 4 show that: a) at the intermodal level, translations display a significantly lower percentage of infrequent/unattested pairs and a higher percentage of high-MI collocations compared to interpretations, both in terms of word combination tokens and types; no difference is found regarding high-*t*-score collocations; b) at the monolingual comparable level, interpreted texts are found to differ from non-interpreted ones in terms of the same features and following the same trends that set them apart from translations (more infrequent/unattested combinations and fewer high-MI collocations), but differences are only detected in terms of types; no significant difference is found between translated and non-translated texts; c) no difference emerges as significant when the comparisons are conducted on the original written vs. oral texts.

Focusing first on the intermodal comparison between TTs, our analyses indicate that, limited to the collocational pattern considered, translated texts appear as more phraseologically conventional than the interpreted ones. In particular, translations use to a significantly greater extent the highly idiomatic, strongly associated expressions characterized by a high MI value, and rely significantly less on combinations that would not be regarded as phraseological within a frequency-based paradigm. While cursory inspection of infrequent sequences in EPTIC reveals that at least part of them are intuitively acceptable, well-formed expressions (e.g. *decisione quotidiana* ‘daily decision’; *associazione rivoluzionaria* ‘revolutionary association’), their overuse in interpreted texts points at interpreters also relying to a greater extent than translators on less conventional/acceptable forms in Italian (e.g. *considerazione annuale* ‘annual consideration’, meaning ‘annual scrutiny’; *necessaria legislazione* ‘necessary legislation’, which should have been rendered as *legislazione necessaria*). The only result that would seem to contradict the pattern of significantly greater collocationality of translated texts is the one concerning the high-*t*-score pairs, i.e. the common, high-frequency collocations, which are used to a similar extent by interpreters and translators.

It seems reasonable to hypothesize that these apparently contradictory tendencies related to different types of lexical sequences are compatible with the constraints characterizing the translation and interpreting tasks. As rightly pointed out by Marco (2009: 853–854), interpreting corpus-based observations in the light of underlying cognitive mechanisms requires caution. However, it would seem that process- and cognitively-oriented studies of translation/ interpreting, and off-line vs. on-line language production in general, are particularly relevant in explaining our results. Specifically, it has been suggested, e.g. by Tremblay and Tucker (2011), that frequency plays a crucial role in determining the mental availability of lexical sequences: lexically cohesive but low-frequency expressions (i.e. those featuring high MI values) are less easily retrievable from memory than high-frequency ones (i.e. those featuring a high *t*-score), especially in on-line tasks. On the other hand, production of non-standard forms (such as the unattested/infrequent sequences) has been found to also be common in off-line translation tasks, and to correlate with the time allowed for revision of the output (Jiménez-Crespo 2012).

Against this background, cognitive and task-related constraints seem to play a major role in shaping and explaining phraseological similarities and differences between translated and interpreted texts: interpreters' output displays the same degree of phraseological conventionality as translators' output in terms of the high-frequency items, which are more readily available to memory during the task. Translated texts are instead more conventional in those areas of phraseology that require more time for processing: this might explain both their greater reliance on highly idiomatic but lower frequency expressions, whose production requires some cognitive effort, and their avoidance of non-standard forms, i.e. those forms that can be omitted or substituted by more standard ones provided enough time is available for the task. It can be hypothesized that source language interference also plays a role: given the contextual constraints, interpreters are more likely than translators to activate what Dam (2000) calls form-based (vs. meaning-based) strategies, a form of direct transfer from the source language which can contribute to explaining the lower degree of phraseological conventionality of their output.[10] A follow-up study including an analysis of source texts might contribute to lending support to this hypothesis.

The significant results obtained in the monolingual comparable comparisons help us to put these findings in perspective. Specifically, we observed that the differences between interpreted and translated texts also apply to interpreted vs. original oral texts: a relatively ample use of non-standard word combinations and more limited reliance on idiomatic expressions characterized by high MI values thus emerge as features setting apart interpretations from other kinds of language production – both on- and off-line –, and this is possibly due to the unique constraints under which the interpreting task is carried out. Differences are however only detected in terms of word combination types, indicating that

[10] We thank the volume's editors for this observation.

interpreters differ from original speakers in terms of the variety, but not the overall number, of non-standard and idiomatic expressions produced. In this respect, the interpreted vs. non-interpreted opposition seems less clear-cut than the interpreted vs. translated one, where the two forms of language mediation are found to differ both in terms of tokens and types. The effect sizes obtained in the respective analyses also point in the same direction, revealing (slightly) stronger effects in the intermodal than in the comparable comparisons.

Finally, no significant difference is observed between original written and oral texts (in Shlesinger and Ordan's (2012) terms, no modality effect was found), nor between translated vs. non-translated texts (no ontology effect, Shlesinger and Ordan 2012). One might conlude that the differences observed are indeed characteristic of the translated vs. interpreted distinction, rather than applying more generally to written vs. oral production, and that translations and interpretations are more similar to non-mediated texts in the same modality than they are to each other, even though the claim is more strongly supported for translations.

Thus, our findings extend and refine those of Shlesinger and Ordan (2012: 43): based on a three way comparison that did not include written non-mediated production, they concluded that "modality may exert a stronger effect than ontology – i.e. that being oral (vs. written) is a more powerful influence than being translated (vs. original)". By also factoring in written non-mediated production, the picture that emerges is one in which the significant differences opposing interpreting and translation at the phraseological level are neither due to modality nor to ontology only, but rather to their combined effect.

6 Conclusion

In this chapter we presented a study on phraseological patterns in the EPTIC corpus, a bilingual (English<>Italian) intermodal, comparable and parallel resource comprising translated and interpreted texts and their respective sources. Focusing on several types of word combinations, i.e. infrequent expressions and attested collocations, we conducted quantitative analyses across the different components of the corpus, both from an intermodal perspective (contrasting translated vs. interpreted texts and original written vs. oral texts), and from a more traditional monolingual comparable angle (contrasting mediated vs. non-mediated texts in the same modality).

Our results indicate that translations are more phraseologically conventional than interpretations in terms of most of the parameters considered. We hypothesized that these differences may be related to the cognitive and task-related constraints characterizing the two processes. It was also found that the observed differences apply specifically to translated vs. interpreted texts, rather than to written vs. oral production in general, or to mediated vs. non-mediated texts, highlighting the need for increasingly multi-faceted corpus and research setups in the analysis of translated and interpreted data.

In terms of ways forward, a multifactorial and/or multivariate statistical analysis capturing both modality and ontology, as well as all three measures of collocational status, might be useful for detecting interaction effects that could not be examined in the separate tests that were carried out here. We also intend to tap the potential of EPTIC to carry out a more qualitative investigation of phraseological shifts at the parallel level. In particular, we are planning to scrutinize single texts where interpreter and translator behaviours are most divergent, e.g. by focusing on the texts where translators produce the highest number of high-MI collocations compared to interpreters. By relying on metadata to explore variables that have been suggested to influence interpreters' and translators' performance (e.g. the speed of delivery of the original speech or its delivery as a scripted vs. impromptu speech) we might be able to shed light on specific contextual and task-based effects, thereby deepening our understanding of the processes leading to the regularities discussed in this work.

Acknowledgement

We would like to thank Silvia Bernardini, with whom EPTIC was jointly developed, for her careful reading of this paper, and Niccolò Morselli for his work on the enlargement of EPTIC. We also thank the anonymous reviewers and the volume's editors for their valuable comments and suggestions.

References

Baker, Mona. 2007. Patterns of idiomaticity in translated vs. non-translated text. *Belgian Journal of Linguistics* 21(1). 11–21.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3). 209–226.

Bendazzoli, Claudio. 2010. *Corpora e interpretazione simultanea* [Corpora and simultaneous interpreting]. Bologna: Asterisco.

Bernardini, Silvia. 2007. *Collocations in translated text. A corpus-based study.* Middlesex University doctoral dissertation.

Bernardini, Silvia, Adriano Ferraresi & Maja Miličević. 2016. From EPIC to EPTIC. Exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies* 28(1). 58–83.

Church, Kenneth Ward & Peter Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–29.

Dam, Helle V. 2001. On the option between form-based and meaning-based interpreting: The effect of source text difficulty on lexical target text form in simultaneous interpreting. *The Interpreters' Newsletter* 11. 27–55.

Dayrell, Carmen. 2007. A quantitative approach to compare collocational patterns in translated and non-translated texts. *International Journal of Corpus Linguistics* 12(3). 375–414.

Durrant, Philip & Norbert Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics* 47(2). 157–177.

Evert, Stefan. 2005. *The statistics of word cooccurrences: Word pairs and collocations*. Stuttgart: University of Stuttgart doctoral dissertation.

Field, Andy, Jeremy Miles & Zoe Field. 2012. *Discovering statistics using R*. London: Sage Publications.

Gile, Daniel. 2004. Translation research versus interpreting research: Kinship, differences and prospects for partnership. In C. Schäffner (ed.), *Translation research and interpreting research: Traditions, gaps and synergies*, 10–34. Clevedon: Multilingual Matters.

Granger, Sylviane & Yves Bestgen. 2014. The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics* 52(3). 229–252.

Jantunen, Jarmo Harri. 2004. Untypical patterns in translations. In Anna Mauranen & Pekka Kujamäki (eds.), *Translation universals: Do they exist?*, 101–128. Amsterdam/Philadelphia: John Benjamins.

Jiménez-Crespo, Miguel Angel. 2012. Translation under pressure and the web: A parallel corpus-study of Obama's inaugural speech in the online media. *Translation and Interpreting* 4(1). 56–76.

Jones, Susan & John M. Sinclair. 1974/1996. English lexical collocations: A study in computational linguistics. In J. A. Foley (ed.), *Sinclair on lexis and lexicography*, 21–54. Singapore: UniPress.

Kajzer-Wietrzny, Marta. 2012. *Interpreting universals and interpreting style*. Poznań: Adam Mickiewicz University doctoral dissertation.

Kenny, Dorothy. 2001. *Lexis and creativity in translation. A corpus-based approach*. Manchester: St. Jerome.

Kruger, Haidee. 2014. Language change, photoshopped language and constrained communication: Some new ways of thinking about translation through corpora. *EST Newsletter* 45. 8–10.

Lanstyák, István & Pál Heltai. 2012. Universals in language contact and translation. *Across Languages and Cultures* 13(1). 99–121.

Marco, Josep. 2009. Normalisation and the translation of phraseology in the COVALT corpus. *Meta* 54(4). 842–856.

Masini, Francesca. 2012. *Parole sintagmatiche in italiano* [Collocations in Italian]. Roma: Caissa Italia.

Mauranen, Anna. 2000. Strange strings in translated language. A study on corpora. In M. Olohan (ed.), *Intercultural faultlines*, 119–142. Manchester: St. Jerome.

Nissim, Malvina, Sara Castagnoli & Francesca Masini. 2014. Extracting MWEs from Italian corpora: A case study for refining the POS-pattern methodology. *The 10th Workshop on Multiword Expressions (EACL 2014)*. Goteborg, Sweden.

Pöchhacker, Franz. 2004. I in TS: On partnership in Translation Studies. In C. Schäffner (ed.), *Translation research and interpreting research: Traditions, gaps and synergies*, 104–115. Clevedon: Multilingual Matters.

Sandrelli, Annalisa & Claudio Bendazzoli. 2005. Lexical patterns in simultaneous interpreting: A preliminary investigation of EPIC (European Parliament Interpreting Corpus). *The Corpus Linguistics Conference Series 1*. http://goo.gl/53glY5 (accessed 15 March 2015).

Schäffner, Christina (ed.). 2004a. *Translation research and interpreting research: Traditions, gaps and synergies*. Clevedon: Multilingual Matters.

Schäffner, Christina. 2004b. Researching translation and interpreting. In C. Schäffner (ed.), *Translation research and interpreting research: Traditions, gaps and synergies*, 1–9. Clevedon: Multilingual Matters.

Setton, Robin. 2011. Corpus-Based Interpretation Studies: Reflections and prospects. In A. Kruger, K. Wallmach & J. Munday (eds.), *Corpus-based Translation Studies: Research and applications*, 33–75. London: Continuum.

Shlesinger, Miriam. 1998. Corpus-based Interpreting Studies as an offshoot of Corpus-based Translation Studies. *Meta* 43(4). 486–493.

Shlesinger, Miriam. 2003. Effects of presentation rate on working memory in simultaneous interpreting. *The Interpreter's Newsletter* 12. 37–49.

Shlesinger, Miriam. 2009. Towards a definition of *Interpretese*: An intermodal, corpus-based study. In G. Hansen, A. Chesterman & H. Gerzymisch-Arbogast (eds.), *Efforts and models in interpreting and translation research: A tribute to Daniel Gile*, 237–253. Amsterdam: John Benjamins.

Shlesinger, Miriam & Noam Ordan. 2012. More spoken or more translated? Exploring a known unknown of simultaneous interpreting. *Target. International Journal of Translation Studies* 24(1). 43–60.

Snell-Hornby, Mary. 2006. *The turns of Translation Studies. New paradigms or shifting viewpoints?* Amsterdam/Philadelphia: John Benjamins.

Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative methods. *Functions of language* 2(1). 1–33.

Tremblay, Antoine & Benjamin V. Tucker. 2011. The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon* 6(2). 302–324.

Vandepitte, Sonia. 2001. Anticipation in conference interpreting: A cognitive process. *Revista Alicantina de Estudios Ingleses* 14. 323–335.

Zanetti, Roberta. 1999. Relevance of anticipation and possible strategies in the simultaneous interpretation from English into Italian. *The Interpreters' Newsletter* 9. 79–98.

Oliver Čulo, Silvia Hansen-Schirra and Jean Nitzke

6 Contrasting terminological variation in post-editing and human translation of texts from the technical and medical domain

Abstract: Post-editing is a rather new mode of translation production increasingly being studied from various angles. In this chapter, we contrast post-editing and human translation along the dimension of term translation within the domain of Languages for Specific Purposes. We make use of the perplexity coefficient to measure terminological variation in term translation from English into German. Our findings reveal levels of variation on the terminological level in the post-edited texts close, but not identical, to those of the machine translation outcomes. They thus indicate a shining through of the machine translations in the post-editing products, motivating further research into the properties of post-edited texts within corpus-based translation studies. On the basis of our observations, we discuss potential effects of this shining through, such as diminished quality of machine translation if post-edited texts are used for re-training, and we critically examine the applicability of the perplexity coefficient as a quality measure for term translation.

1 Introduction

Post-editing (PE) has become a more and more pervasive phenomenon in the translator's world. It has been studied in various settings and for various language pairs. While PE is mostly associated with gains in productivity, and while questions like the quality of post-edited texts are a consistent topic in research, there is still a lot to be learned about the differences between PE and human translation (HT).

Some observations of the influence of PE on the process and the product of translations have already been made. Mesa-Lao (2014) reports that phases like an orientation phase or a final revision phase, characteristic of HT, are often missing in PE. Čulo and colleagues (2014) observe an influence of machine translation (MT) on PE products on the lexical level. For instance, what the authors classify as unidiomatic renderings of the phrase *In a gesture*, namely the translation *In einer Geste*, was observed in a number of fully post-edited

DOI 10.1515/9783110459586-007

versions of the German versions. Similarly, lexical inconsistencies introduced by MT, such as a missing distinction between *Krankenpfleger* 'male nurse' and *Krankenschwester* 'female nurse' which were both used to refer to one person, was left uncorrected by half of the post-editors in their experiment. These observations point to a certain influence of MT on PE on the lexical level.

One question thus to be dealt with is that of how PE products differ in terms of linguistic properties from HT and how much these properties are influenced by MT. This kind of research goes along the lines of Lapshinova-Koltunski's work on register differences between various types of MT and HT products (Lapshinova-Koltunski 2013; Lapshinova-Koltunski 2015).

In the present chapter, based on the observations of Čulo and colleagues (2014), we assume that post-edited texts will differ in their lexical profile from human-translated texts, due to influence from MT. In order to test for this, we contrast technical and medical texts from English to German that were post-edited and translated. As proposed by Carl and Schaeffer (in press), we make use of the perplexity coefficient (cf. section 3.3) to measure consistency of term translation. As we are dealing with texts from the domain of Languages for Specific Purposes (LSP), we will focus on the terminological level and will be looking at variation in translation of terms. In addition, by looking at medical and technical texts, we intend to remedy a shortcoming of a significant body of PE studies conducted so far, namely the reliance on newspaper texts and so-called general language. Most of professional translation takes place in the LSP domain, which is why we believe that these text types deserve further attention in PE research.

In the following, we will briefly provide some background information on LSP in translation studies and professional translation as well as on post-editing. Section 3 focuses on the methodology including a description of the dataset, the criteria for the analysis, and the explanation of the parameters *perplexity coefficient* and *term variation*. The data will be analysed in section 4 and discussed in section 5. Finally, we will draw conclusions and give an outlook on future work.

2 Background

2.1 LSP in translation

LSP represent domain-specific discourse in which experts communicate in a highly optimized, condensed and impersonal manner. They are constrained by formal and linguistic conventions aiming at effectively transmitting the kind of complex information which needs to be conveyed given the high degree of sophistication and specialization. For instance, they are characterized by an extensive use of technical and specialist vocabulary as well as condensed syntactic structures (Halliday and Martin 1993). Typical LSP register features comprise among others term density, nominal style, grammatical complexity and passive. The qualitative and quantitative realization of these features may differ according to the degree of expertise on the vertical layer and interdisciplinary diversity on the horizontal layer (cf. Biber 1995).

In LSP translation, typological differences and contrastive gaps come into play which might cause translation difficulties. Furthermore, LSP translation currently faces additional challenges: the speed of progress in the sciences and an increasing specialization widen the gap between the current state of research and the amount of knowledge an average translator can have acquired in the process of his/her education or professional career. In addition, the global market as well as political forces (like the European Union) increase the demand for translations in such a way that they exhaust human resources. Finally, web-based publishing and dissemination of texts increase the speed of publication workflows, which again increases the pressure on the translator.

However, the translation industry has developed strategies and technologies in order to adapt its working environment to the challenging demand. In the area of technical documentation, controlled languages are used to improve the source text quality. This optimizes the consistency of terminology and domain-specific collocations as well as syntactic structures typical of the respective LSP (e.g. imperatives or passives), which facilitates the translation process and improves translation quality (cf. Hartley and Paris 2001). Additionally, online databases and translation corpora can serve as reference to solve typological or LSP specific translation problems (cf. Hansen-Schirra 2008; Pearson 2003; Bowker 1999). From a technical perspective, computer-aided translation (CAT) tools help to cope with lexical and grammatical LSP translation problems.

Concerning the lexicon of LSP, translators use terminology management tools in order to cope with equivalence and consistency problems (e.g. Mayer, Schmitz, and Zeumer 2002; Schmitt 2003). Within this context, it is a commonly accepted technique to extract terms on the basis of LSP corpora (e.g. Heid et al. 1996 for monolingual lexicography; Bernhard 2006 for multilingual work). Even the extraction of bilingual terminology from comparable as well as parallel corpora has proven to be successful (e.g. Vintar 1999; Carl, Rascu, and Haller 2004). The output can then be used as lexical input for MT systems or as term banks for human translators.

In terms of LSP grammar, translation memory (TM) systems facilitate the translation workflow since they offer pre-translations and consistency management for redundant phrases and sentences. Moreover, predefined or pre-translated phrases (e.g. typical realizations of imperatives or passives) or collocations can be looked up (cf. Seewald-Heeg 2005). State-of-the-art IT solutions integrate terminology management tools into translation memory systems and software localization tools (cf. Reinke 1999). They also combine MT and TM for the translation of fuzzy matches. However, the quality, consistency and equivalence of the automatically translated segment have to be approved by a human translator or post-editor; this is where PE comes into play.

2.2 Post-Editing

Due to the growing demand for translational services in a globalized market, the use of so-called PE has moved into focus as a more efficient and cost-effective method of translation. In order to facilitate and support high-quality human translation, PE has been integrated in existing TM systems and CAT tools (cf. Folaron 2010; O'Brien 2012). This integration into translational environments optimizes the translation workflow and minimizes the PE effort.

However, the productivity and efficiency of PE significantly depends on the quality of the MT output, which in turn is particularly suited for closely related language pairs and text domains with a considerable amount of redundancy (cf. Fiederer and O'Brien 2009; O'Brien 2010). Typical text types include product manuals and technical documentation, where controlled languages are used in order to improve the consistency and processability of the source texts (Aymerich 2005; Kirchhoff et al. 2011).

Within this realm customers' needs can be satisfied with respect to time and quality by offering several levels of post-editing. The so-called light or fast post-editing delivers the main content in a comprehensible and accurate form with only essential corrections (O'Brien, Roturier, and de Almeida 2009; O'Brien 2010). By contrast, the result of full or conventional post-editing should be almost indistinguishable from a translation from scratch by a human translator (Wagner 1985). Text intention and quality expectations are relevant for deciding what type of PE will be the best choice for the purpose (cf. Specia 2011a).

The question concerning how to measure the effort needed to produce the final product plays a key role in the whole PE process. Apart from saving time, translators behave differently when post-editing, especially regarding their methods to detect and plan corrections (Koponen 2012; Koponen et al. 2012; Krings 2001; Specia 2011b). From a more technical perspective, exploiting keylogging and eyetracking data, for instance by calculating editing distances, can shed light on the quality of the MT output. For instance, pauses during keylogging as well as fixation durations, saccade lengths and regressions reveal problems and difficulties in the MT output (Doherty and O'Brien 2009; Doherty, O'Brien and Carl 2010). Compared to statistical MT evaluation techniques (e.g. BLUE: Papineni et al. 2002, METEOR: Banerjee and Lavie 2005) which are based on the comparison of MT produced segments with segments from a reference translation, the triangulation of user-activity data provides metrics to measure MT quality which are more oriented towards standards in professional human translation (Plitt and Masselot 2010).

The integrated PE workflow requires additional skills different from those a classical translator training generally provides, therefore it has to be acknowledged as a competency in its own right in educational curricula. Next to domain-specific knowledge, additional skills are required with regard to MT knowledge, integrated CAT solutions, terminology management, pre-editing, controlled languages, programming as well as a positive attitude towards the technical paradigm (O'Brien 2002; Wagner 1985). So, PE has been an increasingly important topic not only within the translation industry, but also within research.

Due to these special characteristics of the PE task, several web-based projects have been developed, partly in joint efforts, like CAITRA[1], PET[2], CASMACAT[3], TransType[4] and MateCat[5]. They aim to simplify the PE process for the translator and integrate the PE task into a TM environment. As opposed to existing TM systems like SDL Trados that integrate MT systems into the existing TM tool, the mentioned web-based projects were developed specifically for the PE process. While CAITRA is a test suite for improving automatic translation technologies (Koehn 2009), features like self-tuning, user-adaptive and informative MT are integrated in MateCat (Federico, Cattelan, and Trombetti 2012). PET is mainly designed for the evaluation of machine translation through post-editing (Aziz, Sousa, and Specia 2012), whereas TransType offers text prediction on the basis of the translator's input (Langlais, Foster, and Lapalme 2000). CASMACAT focuses on visualization and enhanced user-friendly input methods and aims at "cognitive analysis that provides insight into the human translation process to guide our development of a new workbench for translators" (Ortiz-Martínez et al. 2012).

All these efforts head towards improving MT output and data-driven human-machine interaction in translation. However, from a translational perspective it is rather critical that the existing projects do not use market-relevant text types. MT systems often use the Europarl Corpus as training material – a text type which is translated within the European Union, but which does not meet

[1] CAITRA: an experimental Web-based interactive computer-aided translation tool.

[2] PET: post-editing tool.

[3] CASMACAT: cognitive analysis and statistical methods for advanced computer-aided translation.

[4] TransType: an interactive text prediction tool.

[5] MateCat: machine translation enhanced computer-assisted translation.

the needs for postediting technical documentation, which constitutes another important market segment for LSP translators. Systematic and empirical research integrating market relevant LSP data into CAT solutions including PEMT can still be considered as a desideratum – a research gap which this study tries to bridge. A similar problem can be observed in process-oriented research into PE: especially with earlier versions of the CRITT TPR database containing eye-tracking and key-logging data of translation as well as post-editing sessions (Carl 2012), a significant proportion of it and of the studies conducted on the basis of it was based on newspaper articles – a text type typically not considered for translation. While data from other domains such as these described in the following section have recently been included, including further data from other LSP domains would be a welcome addition from the viewpoint of translation studies.

As discussed in section 2, it is LSP translation which mostly relies on integrated CAT solutions including PE. However, systematic and empirical research based on market-relevant LSP data can still be considered a desideratum – a research gap which this study aims to help bridge.

3 Method

3.1 Data collection

The data used for the analysis in section 4 was collected in an experiment in which participants were asked to translate (HT), fully post-edit (FPE) and lightly post-edit (LPE) texts from either the technical or the medical domain in the translation direction English to German. The data collection is a generic collection in the sense that we did not aim at studying one specific phenomenon (e.g. term translation), it is rather aimed at contrasting the PE and HT processes and products on a broad scale.

The participants were all students enrolled in a translation studies degree. They had at least two years of training and had passed at least one course on translating in the domain they would translate and post-edit for in the experiment. Some had minor post-editing experience, but not all, as post-editing is not established as a mandatory course. Therefore, it can be assumed that all participants were better trained in HT than in the PE tasks.

The LSP texts were of low formality level and originated from texts freely available on the internet. The three technical texts selected for the experiment are comparable parts taken from a dish washer manual, the three medical texts were taken from package leaflets ranging from a vaccine against measles to human insulin for diabetes patients and medication for treatment of cancer. All texts were about 150 words long. The texts can be found in the Appendix to this chapter.

Each translator was asked to translate, rapidly post-edit and fully post-edit a text. The texts were automatically pre-translated by Google Translate for the PE tasks. A permutation scheme was set for the three sessions for each domain so that each text would be translated, fully and rapidly post-edited equally often, but by different participants (cf. Table 1). The participant groups consisted of 12 advanced translation students for the technical and 9 for the medical texts. Thus, the technical texts were translated and fully post-edited four times each, the medical texts three times. The participants used Translog-II for all three tasks and eye movements as well as key strokes were logged. The eye-tracking and keylogging data were not used for the present study, however.

Table 1: Permutation scheme for text translation, full, and light post-editing

Participant	Text 1	Text 2	Text 3
P01	HT	LPE	FPE
P02	FPE	HT	LPE
(…)			

As we could not assume any formal training in PE, participants were given instructions on how to post-edit. Part of the instructions for the full PE was to ensure terminological consistency of the post-edited texts. In general, only one term is used to describe one concept in LSP texts. As all participants had had at least basic training in LSP translation, it was assumed for HT that translators would follow the general norm to use terms consistently in the target text. In one case two term naming variants were used synonymously in the original text in a package leaflet: the Latin-derived term *varicella* and the equivalent English term *chickenpox*. But even in this case, participants opted for one variant in human translation, either the more formal variant *Varizellen* or the more colloquial variant *Windpocken* (see Table 4).

As ensuring terminological consistency was not part of the assignment for light PE, we ruled out these data for the analysis presented below.

3.2 Analysis criteria

For the medical source texts, we manually identified 58 term candidates, for the technical source texts we found 51 term candidates. Term candidates were grouped by concept, i.e. while *varicella* and *chickenpox* are naming variants, we classified them as variants of one term. Only ca. 10% of the term candidates appeared three or more times per text. We ruled out terms appearing only once or twice per text for the purposes of the present pilot study. Term candidates appearing only once cannot be tested for variation within a text. For term candidates only appearing twice within a text, we expected too strong a distortion: if each time a different translation is used for a term which only appears twice in the text, we have variation in 50% of the cases, which seems too strong a statement for a sample so small. We confirmed the term status of the remaining terms by looking them up in the online European Union term database IATE[6]. Term candidates based on a term listed in IATE received term status as well (e.g. *upper filter assembly* which is a subterm of the verified term *filter assembly*).

In a next step, we had to decide what to classify as variation in terms of term translation. A clear case of variation is the use of a synonym. For instance, if *dishwasher* was translated once as *Spülmaschine* and as *Geschirrspüler* another time within the same text, we counted this as variation (cf. 3.3 for the calculation schema).

However, there were also special cases, of which the following three are of particular interest.

First, we ruled out typing errors as types of variation. They would, of course, be relevant when assessing HT or PE quality, but in our context they do not indicate lexical variation, as these are not cases of use of a synonym. For instance, if a *filter* was translated as *Filter* twice, but once as *Filöter*, we did not count this as variation, as there is no such thing as a *Filöter*, let alone as synonym to *Filter*.

Then, there were borderline cases such as the translation of the phrase *where measles is common*. In our text set, this was rendered as *Masernvorkommen* 'measles occurrence' in one session and as *Masernprävalenz* 'measles prevalence' in another session. The part *where … is common* was thus nominalized during the translation into German and this nominalization was then compounded with *Masern*. In a strict sense, as *Masern* and *Masernprävalenz* are not naming variants of one term, this would be an instance of variation. However, nominalization is a common process for translations into German (cf. e.g. Čulo et al. 2008), and compounding is a typical feature of the German language. This would thus be a well-motivated case of variation. Taking into account, then, that the first part of the compound, *Masern*, is a consistent rendering of the English term *measles*, we decided not to count this specific case as variation.

What we did classify as variation, however, were non-translations, i.e. where a term was left out in the translated texts (e.g., because it was inferable through context). Thus, if three occurrences were translated twice by the same term and once not at all, this was an instance of variation. While omitting a term e.g. in an elliptical structure in the target language might be well motivated, an omission

[6] http://iate.europa.eu

of a term in the same spot both in the machine translation output and then also in the fully post-edited text would indicate an influence of MT on PE.

3.3 Perplexity coefficient and term variation

The perplexity coefficient (cf. Carl and Schaeffer (in press) for a more in-depth explanation) is used in MT to measure how difficult it is for a translation model to opt for a translation. If there are many equally likely translation choices, perplexity increases, or in other words: the lower the perplexity, the more reliably a machine can opt for a translation.

The perplexity coefficient *PP* is an exponential function defined by the entropy value *H*:

$$PP = 2^H$$

Entropy is an expectation measure borrowed from information theory which states with what level of certainty we can expect a certain event. In fact, entropy expresses the level of uncertainty: the value increases the more uncertain an event is. We will conceptualize a translation decision (i.e. the translation of a term) as an "event" here. The entropy value is calculated by the following formula

$$H = \sum p_i I(p_i)$$

where $I(p_i)$ is calculated on a logarithmical base:

$$\log_2\left(\frac{1}{p_i}\right)$$

The application of the perplexity coefficient to variation in segment translation as in Carl and Schaeffer (in press) is rather straightforward. For every segment, the number of translation variants throughout the translations is counted and the probability for each variant to occur is calculated. On the basis of this, the perplexity coefficient is calculated (see example in Table 2 and Table 3).

However, this simple method is not applicable in our experimental setting. One of the properties of terms is that they may have synonyms, and a translator (or the institution deciding on translation norms) may opt for one or the other variant (then called the *preferred term*). As long as this variant is used consistently, the norm is not violated. In our setting, we did not define preferred terms beforehand. Thus, for two translation sessions A and B, we may get a session A in which, e.g. for *dish washer*, the term *Geschirrspüler* is used five times, but *Spülmaschine* was used five times in session B. However, this is not variation in our sense, as both translators stuck with what they decided to be the preferred term.

In order to model variation in term translation, we thus need to classify the translation choices or "events" differently. Here we follow the conceptualisation of term realisation "events" in terminology theory: for each session, we defined a *preferred term* a posteriori, variations of this term were categorized as *synonymous terms*. We thus count two main types of events: *translation-by-preferred-term* (pref.t.trans) and *translation-by-synonymous-term* (syn.trans), where each different synonymous term used counts as a different subtype of event (simply numbered, i.e. syn.1.trans, syn.2.trans, etc.).

Table 2: Mapping of term variant to translation event type

Session	term translation	term frequency	event type
P21_FPE	Geschirrspülmaschine	4	pref.t.trans
	Spülmaschine	2	syn1.trans
P10_FPE	Spülmaschine	2	pref.t.trans
	Geschirrspülmaschine	2	syn.1.trans
	Geschirrspüler	1	syn.2.trans

Table 2 exemplifies this with the term *dishwasher*, as it was translated in two different FPE sessions. For participant P21, defining the preferred term was simple, as one term, namely *Geschirrspülmaschine*, was used more often than the other. The event type syn1.trans covers all translations by means of the synonym *Spülmaschine*. For participant P10, assigning the preferred term status to one of the two synonyms *Spülmaschine* or *Geschirrspülmaschine* is arbitrary, as both appear equally frequently; as *Spülmaschine* appeared first in the translation, we chose this as the preferred term. For P10, we also have two synonym-translation subtypes, which are simply numbered: syn.1.trans covers all translations of *dish washer* by means of the synonym *Geschirrspülmaschine*, syn.2.trans is the category for translation by means of *Geschirrspüler*.

Table 3: Aggregated translation event type frequency for FPE of participants P10 and P21

event type	frequ.	prob.
pref.t.trans	6	0.55
syn.1.trans	4	0.36
syn.2.trans	1	0.09

To sum up, even though we have different preferred terms and synonymous terms for each translation session, we can aggregate the numbers by means of assigning them to the more abstract classes pref.t.trans, syn.1.trans and syn.2.

trans (and, of course, potentially syn.3.trans etc.). When covering variation for all FPE sessions, this data then is distilled into a table as in Table 3, here based on the two sessions listed in Table 2. We are using an aggregate of variation data over several sessions for one specific term, as we are interested in contrasting different translation settings. If we were interested in looking at individual variation of the participants, we would look at all term translations by each single participant and would not require a scheme for aggregating the variation data.

While the probability values for each translation type already give a first impression of the variation in term translation, the perplexity value enables us to express the variation by means of one single value. It is important to note here that this comparison is indeed possible between terms in different sessions, but not necessarily between different terms. The reason for this is that the values the perplexity coefficient can take depend on the number of translation instances we take into account. While the lower bound is always 1 (indicating no variation), the upper bound always is the number of instances we take into consideration (indicating maximum variation, i.e. a different synonym for each translation instance).[7] Also, the intermediate values are not directly comparable. For a term with four occurrences in the source and with one instance of variation in the translation, we get a perplexity value PP ≈ 1.75, for a term with five occurrences and one instance of variation, we get PP ≈ 1.65.

The perplexity values were calculated using R[8], with the entropy function as provided by the package agrmt. Aggregate perplexity for the example in Table 3 is 2.5.

4 Analysis

Table 4 lists the perplexity values for MT and the aggregate perplexity values for FPE and for HT per term. A first look already reveals that the perplexity values for MT and FPE are usually quite similar.

In order to test whether this visual impression can be confirmed, we used the non-parametrical Kendall’s Tau correlation test. If FPE and HT showed a correlation, this might mean that certain norms are adhered very strongly by human translators and post-editors, but are not followed by MT. If MT and FPE showed a correlation, this would indicate that either the terms proposed by the

[7] We have no mathematical proof for this as yet, but this proved true for all terms we analyzed in this chapter.

[8] http://www.r-project.org/

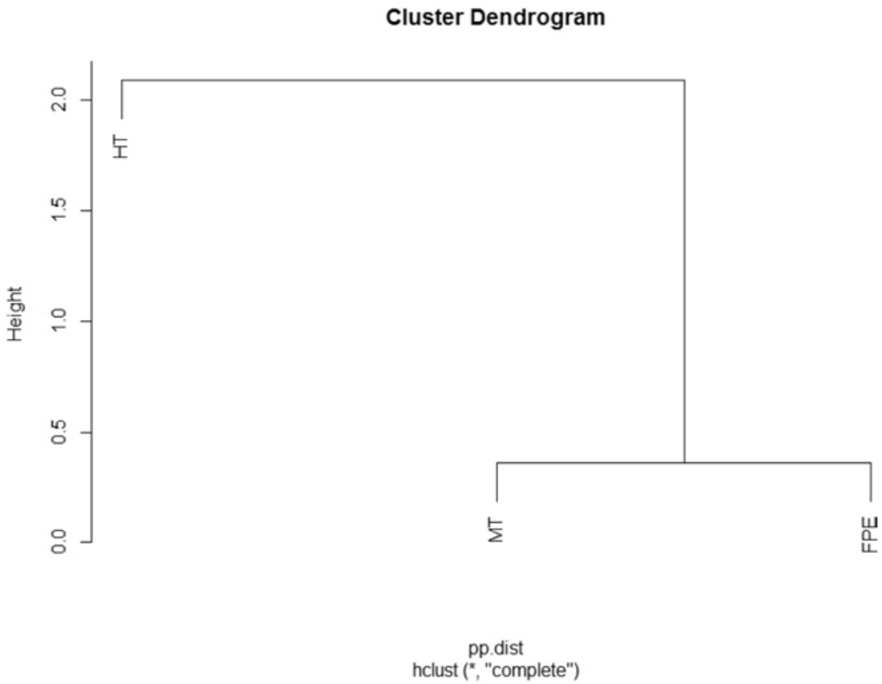
MT system were acceptable for the target text, or, even though the task description for FPE specifically stated that the MT output should be adapted to reach terminological consistency, post-editors did not arrive at this goal. Indeed, Kendall's Tau correlation test reveals a correlation 0.76 ($p < 0.01$) between MT and FPE, but only a weak negative correlation for HT and FPE ($r = -0.29$) which did not prove significant ($p \approx 0.29$).

One might have expected the MT and FPE data to deviate from each other more strongly in terms of terminological consistency. Post-editors had both the source text and the MT product available and could thus choose to override the translation decision made by the machine. Furthermore, the task specification for term translation specifically required to deviate from the MT output where necessary. However, a terminology list was not part of the translation job. We suggest, though, that the correlation detected by Kendall's Tau indicates an influence of MT on the terminological choices in the FPE task, but at this point we can only speculate about the causes for this. Potential causes for and implications of this will be discussed in the following section.

Table 4: Perplexity value (MT) and aggregated perplexity value (FPE and HT) per translation of ST term

ST term (frequency in ST)	MT	FPE	HT
vaccine (3)	1	1	1.57
measles (4)	1	1	1
varicella (1) / chickenpox (3)	1.75	1.82	1
Disease	1	1	1
Protaphane (6)	1	1	1
anticancer medicine (3)	1	1	1.57
dishwasher (text 1: 5)	2.81	2.07	1.23
dishwasher (text 2: 5)	1.96	1.48	1
rinse aid (4)	1	1	1.33
filter (3)	1	1	1.42
upper filter assembly (3)	1.89	1	1.75

We also applied clustering in order to see which vectors, formed by the perplexity values for each mode of translation from Table 4, would group best. Figure 1 shows the dendrogram for the agglomerative hierarchical clustering of the MT, FPE and HT vectors. While the FPE and MT vectors merge into one cluster very early, the HT vector forms a distant cluster of its own. This supports our assumption that there is a close relation between MT and FPE.



vectors

Figure 1: Dendrogram for hierarchical clustering of MT, FPE and HT vectors

5 Discussion

Our analysis has focused on measuring the variation of term translation and contrasting it according to the mode of translation (i.e. which technology was or was not used). In order to express the degree of variation, we used the perplexity coefficient, borrowing it from machine translation. The strong correlation between MT and FPE in our results suggests that the outcomes in FPE are influenced by the outcomes in MT.

Our study nevertheless suffers from two potential drawbacks. First, the sample which we evaluated is small. The main focus of our study is, from our viewpoint, on the method of evaluation and on identifying lexical dimensions along which the different translation modes can be contrasted. Of course, the exact nature of the results will depend heavily on the characteristics of the MT systems used. Second, we did not check for potential spill-over effects, i.e. whether the order of tasks influenced the results.

Given these drawbacks, we will discuss our findings from two perspectives. First, in which direction does our observation of a close correlation between MT and FPE in term variation point and which further questions might follow? Second, while we were successful in measuring variation by means of the perplexity coefficient, what exactly do the numbers say – in our view – in the context of our experiment setup and other potential setups?

The FPE products exhibit similar properties in terms of variation in term translation as the MT products, which is not that surprising given that the FPE products are in based on the MT products. We would have expected the post-editors to act more independently of the MT outcome, correcting the terminological inconsistencies produced by MT. They were, however, more or less left unchanged. While overall, there was variation for fewer terms in both MT and FPE than in HT, it is this non-independence from the MT output which in our view deserves more attention.

This non-independence does not only extend to cases of variation, but also to cases of non-variation: As Table 4 shows, post-editors will also not vary where MT is already very consistent. While we do not have compelling examples for this in our data sets, it is in theory possible that a certain variation or cases of non-translation would be a good or even necessary option. We will exemplify this by means of cognates. Cognates are pairs of words which share many formal and semantic aspects, such as the English *system* and German *System*. They have to be distinguished, though, from false friends such as the English *actual* and German *aktuell* ‘current’. Also, we assume that only words of the same word class can be seen as cognates in a strict sense, and that changes of word class involve more (and potentially different) processes than a cognate translation. The following example is taken from an evaluation of cognate translation in MT and HT (Čulo and Nitzke 2016):

> Source: “[...] political stability rested on the acceptance in all classes of the legitimacy [...]”
>
> Target: “[...] beruhte ihre Stabilität darauf, daß alle Klassen die Legitimität [...] akzeptierten”
>
> Lit.: [...] *rested their stability on that all classes the legitimacy [...] accepted*

The English word *acceptance* has a cognate in German, *Akzeptanz*, which, however, we would argue does not convey the exact same meaning: intuitively, the German word *Akzeptanz* is used in cases where an increasing or decreasing degree of acceptance is meant; for the process or fact of agreeing to a fact, the German verb *akzeptieren* seems a more apt translation. The variation – i.e. the shift from nominal to verbal form and thus a non-cognate translation – is desired in this case.

One potential cause we hypothesize for the close similarities in (non-)variation between MT and FPE – which needs to be confirmed by deeper inspection of the keylogging and eyetracking data – is that post-editors focus more on the micro-level of the text than on the macro-level/the overall text. Correcting MT-specific errors in single segments *and* ensuring consistency over the stretch of a text may thus be a task translators need to be trained for specifically, as they require heightened levels of effort. The lexical inconsistencies reported by Čulo and colleagues (2014) did not occur in monolingual editing of MT output which we believe to require overall less cognitive effort as there is only one text to focus on. Another potential cause might be that post-editors rely more on the MT output than they notice or are willing to admit, despite general scepticism towards MT being repeatedly reported. This would thus not be a problem of too narrow focus, but a lack of critical attitude towards the text being revised (or, as we would say in this setting, post-edited). Indeed, we may be looking at a mix of these two causes.

In effect, we could, in a sense, say that the MT product is *shining through* in the FPE products. Shining through is a phenomenon well entrenched in translation studies. The notion was introduced by Teich (2003) in her study on linguistic properties of popular scientific English and German texts and their translation into the other language. For passives, she found that, e.g., translations into German exhibit more passives than German original texts, probably due to the influence of English originals which, too, exhibit them more often than German originals. We can see a similar phenomenon in our study for the variation in term translation: The PE products tend to exhibit variation where the MT products exhibit variation, they show similar peaks of perplexity values for the same term, and there is no variation when there is none in MT. Whether a correlation between MT and FPE can be found on the syntactic level for choice of syntactic constructions in our samples should be verified in an additional step.

A question to be asked from the viewpoint of empirical corpus-based translation studies is whether the factor post-editing is a new register-defining factor, i.e. whether all post-edited texts will exhibit similar linguistic properties. We propose that post-editing products should further shift into the focus of corpus-based translation studies to better understand their – potentially hybrid between MT and HT – nature. However, it is too early to jump to conclusions about the overall linguistic behaviour of post-edited texts: Lapshinova-Koltunski's findings (this volume) show that in terms of linguistic properties, some modes of human production may cluster better with MT modes than with other human production modes (e.g. professional translations seem to be closer to rule-based MT than to student translations).

A question to be asked from the viewpoint of MT research is how this will influence the future development of MT systems and their output. If PE continues its surge as a method in the translation world, we will get more and more post-edited translation products. If our observation that post-edited texts exhibit properties close to those of machine translated texts on term level (and potentially other levels) holds and if these texts are fed back into MT systems for training purposes, in theory the features and translation choices that MT systems can opt for would be narrowed down as they are constantly being reproduced, and MT products would become linguistically flatter (in a sense). One might even ask whether an LSP in which a lot of MT and PE is used would be subject to any kind of language change in the long run due to frequent exposure of producers and recipients to the linguistic specifics of machine translated and/or post-edited texts.

The results also indicate that teaching PE to translators will require more than just background knowledge of the technologies used in MT. As has been stated before also by others, machine translated texts pose specific problems that translators need to be trained to recognize (see e.g. O'Brien 2002). On that note, we would like to stress that we believe that PE should be done by translators as opposed to mere monolingual or bilingually informed text correctors. Concepts such as terminological consistency, but also the necessary variation that may come with translation, are not easily grasped without dedicated translation training.

The second question that we want to turn to is that of how well applying the perplexity coefficient, borrowed from MT, aligns with translation settings in which humans are involved. Is it any sort of quality measure, as it is in MT, and if so, is this true for all translation settings?

First, the name of the coefficient in itself is somewhat problematic for the application to translations. *Perplexity* clearly is a term that was meant for deterministic machines: the easier a decision is made, the better. When applied to measuring variation in translation, the term perplexity suggests a strongly normative stance. With respect to terminology, this is not actually all wrong: in the realm of LSP translation discussed in this chapter, there are clear norms and outlines as to how to deal with terms. However, this is not true in other settings. In a popular scientific article in which LSP terms may be used, we would, in fact, expect a certain degree of variation on the terminological level due to the norms associated with this text type.

However, even in LSP translation we may observe well motivated variation. Recall, for instance, the example of the phrase *where measles are common* which was translated by compound nouns into German (cf. section 3.2). And, as soon as we go beyond terminological considerations, from the stance of a more descriptively oriented translation research, there is little that we can say about how things should be translated. With these considerations in mind, we would put forward that while the perplexity coefficient is a good means to measure variation, it cannot be understood as a quality measure, in contrast to its use in MT.

All in all, the underlying idea of the perplexity measure as it is used in machine translation and as we applied it here are different ones. As for the naming of the coefficient itself, it is probably best to leave it at that for the time being, with all reservations kept in mind.

Another reason why we believe that the perplexity coefficient does not lend itself to the use as a quality measure is that we see no way to pin down absolute thresholds. In an idealized setting for the kinds of LSP translation discussed in this chapter, the perplexity value would be its lower bound, i.e. 1, indicating no variation at all. For our most frequent term which appears six times, the upper bound of perplexity is 6, which would indicate that each occurrence of the term was translated differently (cf. section 3.3). For our small experiment set, a perplexity value below 1.5 indicates only minimal variation, i.e. for all translation instances in all given texts, a synonym was used only once. All of this is, however, only true for a data set of the size as in the experiment described here. The authors are currently not aware of a method to sensibly normalize the perplexity value across different settings – something that would be a prerequisite for making it a quality measure comparable across different data sets.

Our evaluation setup aggregated all post-edited, machine-translated and human-translated texts into one "bag of words" per translation mode, as we were aiming to contrast variation levels between different translation settings and not individual variation. We suggest, however, that the alternative approach, i.e. calculating perplexity for a term per text, may allow to identify motivated term variation: if the perplexity value for a term is consistently above 1 (and ideally always the same for each text), this might be a case of motivated variation, but only if a term variant (or another process like compounding, recall the *measles*-example from section 3.2) is used consistently in the same position throughout the texts. A perfect word alignment for all translated texts with the source text would, of course, basically allow for the same trick, but is not available without major human correction effort for the automatic alignment.

To sum up, when applied to term translation, the perplexity coefficient can point to certain directions and thus be a good means of orientation, when e.g. identifying an unexpectedly high level of variation in a German LSP translation, but an exact interpretation is still subject to deeper scrutiny of the expected as well as unexpected cases of variation.

6 Conclusions and outlook

We applied the perplexity coefficient borrowed from MT to measure terminological variation in translation. We presented a schema of how to classify translation variants across different translation settings and sessions as different types of events in order to make them comparable in terms of consistency, and usable for perplexity calculation.

We found that the perplexity coefficient is adequate for the task of measuring consistency in term translation as it takes a perspective on the data which is guided by a norm of consistency. For various reasons, though, we would oppose any interpretation of the perplexity coefficient as quality measure.

The method we used could be developed in various ways. For instance, a per-text instead of an aggregate calculation of perplexity will not only model individual variation. When overlaid with calculations from other texts, similar levels of variation for a certain term may indicate motivated variation. It remains to be investigated for which other types of variation besides variation in term translation the perplexity coefficient can be used.

We also found that FPE products were similar to MT products along the dimension of variation in term translation, indicating a shining through of the MT in the FPE product. The exact level or pattern of terminological variation may, of course, change according to such factors as the training data of the MT system or on methods like terminology injection. Still, this observation, in our view, motivates a deeper contrastive investigation of the macro translation process types PE and HT into how they differ in their linguistic properties.

References

Aymerich, Julia. 2005. Using machine translation for fast, inexpensive, and accurate health information assimilation and dissemination: Experiences at the Pan American Health Organization. Paper presented at the *Ninth World Congress on Health Information and Libraries*, Salvador – Bahia, Brazil.

Aziz, Wilker, Sheila C. M. Sousa & Lucia Specia. 2012. PET: A tool for post-editing and assessing machine translation. Paper presented at the *Eighth International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Banerjee, Satanjeev & Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (ACL)*, 65–72. Michigan, USA.

Bernhard, Delphine. 2006. Multilingual term extraction from domain-specific corpora using morphological structure. *The Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters and Demonstrations (EACL)*, 171–174. Trento, Italy: Association for Computational Linguistics.

Biber, Douglas. 1995. *Dimensions of register variation*. Cambridge: Cambridge University Press.

Bowker, Lynne. 1999. Exploring the potential of corpora for raising language awareness in student translators. *Language Awareness* 8(3/4). 160–173.

Carl, Michael. 2012. The CRITT TPR-DB 1.0: A database for empirical human translation process research. In Sharon O'Brien, Michel Simard & Lucia Specia (eds.), *Workshop on Post-editing Technology and Practice (AMTA)*, 9–18. San Diego, USA.

Carl, Michael, Ecaterina Rascu and Johann Haller. 2004. Using weighted abduction to align term variant translations in bilingual texts. *The Fourth International Conference on Language Resources and Evaluation (LREC)*, 1973–1976. Lisbon, Portugal.

Carl, Michael & Moritz Schaeffer. In press. Measuring translation literality. In Arnt-Lykke Jakobson & Barto Mesa-Lao (eds.), *Translation in transition: Between cognition, computing and technology*. Amsterdam: John Benjamins.

Čulo, Oliver, Silke Gutermuth, Silvia Hansen-Schirra & Jean Nitzke. 2014. The influence of post-editing on translation strategies. In Laura Winther Balling, Michael Carl, Sharon O'Brien, Michel Simard & Lucia Specia (eds.), *Post-editing of Machine Translation: Processes and Applications*, 200–218. Cambridge: Cambridge Scholars Publishing.

Čulo, Oliver, Silvia Hansen-Schirra, Stella Neumann & Mihaela Vela. 2008. Empirical studies on language contrast using the English-German comparable and parallel CroCo corpus. Paper presented at the *Sixth International Conference on Language Resources and Evaluation workshop on building and using comparable corpora (LREC 2008)*. Marrakech, Morocco.

Doherty, Stephen & Sharon O'Brien. 2009. Can MT output be evaluated through eye tracking? *The twelfth Machine Translation Summit*, 214–221. Ottawa, Canada.

Doherty, Stephen, Sharon O'Brien & Michael Carl. 2010. Eye tracking as an MT evaluation technique. *Machine Translation* 1(24). 1–13.

Federico, Marcello, Alessandro Cattelan & Marco Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. Paper presented at the *Conference of the Association for Machine Translation in the Americas (AMTA 2012)*. San Diego, USA.

Fiederer, Rebecca & Sharon O'Brien. 2009. Quality and machine translation – A realistic objective? *Journal Of Specialised Translation* 11. 52–74.

Folaron, Deborah. 2010. Translation tools. *Handbook of Translation Studies*, vol. 1, 429–436. Amsterdam: John Benjamins.

Halliday, M. A. K. & James R. Martin. 1993. *Writing science: Literacy and discursive power*. London/Washington, D.C.: Falmer Press.

Hansen-Schirra, Silvia. 2008. Interactive reference grammars: Exploiting parallel and comparable treebanks for translation. In Elia Yuste Rodrigo (ed.), *Topics in language resources for translation and localisation*. 23–37.

Hartley, Anthony & Cécile Paris. 2001. Translation, controlled languages, generation. In Erich Steiner & Colin Yallop (eds.), *Exploring translation and multilingual text production. Beyond content*, 307–326. Berlin/New York: Mouton de Gruyter.

Heid, Ulrich, Susanne Jauss, Katja Krüger & Andrea Hohmann. 1996. Term extraction with standard tools for corpus exploration: Experience from German. Paper presented at the *4th International Congress on Terminology and Knowledge Engineering*. Frankfurt, Germany.

Kirchhoff, Katrin, Anne M. Turner, Amittai Axelrod & Francisco Saavedra. 2011. Application of statistical machine translation to public health information: A feasibility study. *Journal of the American Medical Information Association* 18(4). 473–478.

Koehn, Philipp. 2009. A web-based interactive computer aided translation tool. Paper presented at the *Software Demonstrations (ACL-IJCNLP 2009)*. Suntec, Singapore.

Koponen, Maarit. 2012. Comparing human perceptions of post-editing effort with post-editing operations. *The Seventh Workshop on Statistical Machine Translation (ACL)*, 181–190. Montreal, Canada.

Koponen, Maarit, Wilker Aziz, Luciana Ramos & Lucia Specia. 2012. Post-editing time as a measure of cognitive effort. *The Workshop on Post-editing Technology and Practice (AMTA 2012)*. San Diego, USA.

Krings, Hans Peter. 2001. Repairing texts: Empirical investigations of machine translation post-editing processes. In Geoffrey S. Koby, Gregory M. Shreve, K. Mischerikow & S. Litzer (eds.), *Translation Studies Series*. Kent, Ohio: Kent State University Press.

Langlais, Philippe, George Foster & Guy Lapalme. 2000. Unit completion for a computer-aided translation typing system. *Machine Translation* 15(4). 267–294.

Lapshinova-Koltunski, Ekaterina. 2013. VARTRA: A comparable corpus for the analysis of translation variation. *The 6th Workshop on Building and Using Comparable Corpora*, 77–86. Sofia, Bulgaria.

Mayer, Felix, K.-D. Schmitz & J. Zeumer (eds.). 2002. eTerminology. Professionelle Terminologiearbeit im Zeitalter des Internet [eTerminology. Working in terminology in the internet era]. Paper presented at *Symposium des Deutschen Terminologie-Tags e.V.* Köln: DTT.

Mesa-Lao, Bartolomé. 2014. "Gaze Behaviour on Source Texts: An Exploratory Study Comparing Translation and Post-Editing." In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard & Lucia Specia (eds.), *Post-Editing of Machine Translation*. Cambridge Scholars Publishing.

O'Brien, Sharon. 2002. Teaching post-editing: A proposal for course content. *Sixth Conference of the European Association for Machine Translation (EAMT Workshop)*, 99–106. Manchester, U.K.

O'Brien, Sharon. 2010. Introduction to Post-Editing: Who, What, How and Where to Next? Paper presented at the conference of the *Association for Machine Translation in the Americas (AMTA 2010)*. Denver, Colorado.

O'Brien, Sharon. 2012. Translation as human–computer interaction. *Translation Spaces* 1(1). 101–122.

O'Brien, Sharon, Johann Roturier & Giselle de Almeida. 2009. Post-Editing MT output views from the researcher, trainer, publisher and practitioner. Paper presented at the *Machine Translation Summit XII (MTS 2009)*. Ottawa, Ontario, Canada.

Ortiz-Martínez, Daniel, Germán Sanchis-Trilles, Francisco Casacuberta Nolla, Vicent Alabau, Enrique Vidal, José-Miguel Benedí, Jesús González-Rubio, Alberto Sanchís & Jorge González. 2012. *The CASMACAT Project: The next generation translator's workbench*. Madrid, Spain.

Papineni, K., S. Roukos, T. Ward & W. J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. *The 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318. Philadelphia, USA.

Pearson, Jennifer. 2003. Using parallel texts in the translator training environment. In Federico Zanettin, Silvia Bernardini & Dominic Stewart (eds.), *Corpora in Translator Education*, 15–24. Manchester: St. Jerome.

Plitt, Mirko & François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics* 9(3). 7–16.

Reinke, Uwe. 1999. Überlegungen zu einer engeren Verzahnung von Terminologiedatenbanken [Reflections on connecting terminological databases more closely]. *Translation Memories und Textkorpora* 16. 64–80.

Schmitt, Peter A. 2003. Fachlexikographie in der Internet-Ära: Vom PC zum polytechnischen Großwörterbuch [Specific lexicography in the internet era: from pc towards a polytechnical dictionary]. *Lebende Sprachen* 3. 97–113.

Seewald-Heeg, Uta. 2005. Der Einsatz von Translation-Memory-Systemen am Übersetzerarbeitsplatz [The use of translation memory systems in the workplace]. *MDÜ–Mitteilungen für Dolmetscher und Übersetzer* 52(4–5). 8–38.

Specia, Lucia. 2011a. *Quality estimation of machine translation*. Dublin: Dublin City University.

Specia, Lucia. 2011b. Exploiting objective annotations for measuring translation post-editing effort. In Mikel L. Forcada, Heidi Depraetere & Vincent Vandeghinste (eds.), *The 15th conference of the European Association for Machine Translation (EAMT)*, 73–80. Leuven, Belgium.

Teich, Elke. 2003. *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. (Text, Translation, Computational Processing. Vol. 5). Berlin/New York: Mouton de Gruyter.

Vintar, Špela. 1999. A parallel corpus as a translation aid: Exploring EU terminology in the ELAN Slovene-English parallel corpus. Paper presented at the *34th Colloquium of Linguistics*. Germersheim, Germany.

Wagner, Emma. 1985. Post-editing Systran – A Challenge for Commission Translators. *Terminologie et Traduction* 3.

Appendix

Medical Text 1

What is ProQuad?

ProQuad is a vaccine against measles, mumps, rubella, and varicella (chickenpox). ProQuad is available as a powder and solvent that are made up into a suspension for injection. The active substances are attenuated (weakened) viruses for the diseases.

What is ProQuad used for?

ProQuad is given to children from 12 months of age to help protect them against the four diseases: measles, mumps, rubella, and chickenpox. ProQuad may also be given to children from nine months of age in certain situations, for example as part of a national vaccination programme, during an outbreak or for travel to a region where measles is common. The medicine can only be obtained with a prescription.

How does ProQuad work?

ProQuad is a vaccine. Vaccines work by ‘teaching’ the immune system (the body’s natural defences) how to defend itself against a disease. ProQuad contains weakened forms of the viruses that cause measles, mumps, rubella and chickenpox.

Medical Text 2

1. What PROTAPHANE IS AND WHAT IT IS USED FOR

Protaphane is human insulin to treat diabetes. Protaphane is a long-acting insulin. This means that it will start to lower your blood sugar about 1 ½ hours after you take it, and the effect will last for approximately 24 hours. Protaphane is often given in combination with fast-acting insulin products.

2. Before YOU USE PROTAPHANE

Do not use Protaphane

- If you are allergic (hypersensitive) to this insulin product, metacresol or any of the other ingredients (see 7 Further information). [...]
- If you have trouble with your kidneys or liver, or with your adrenal, pituitary or thyroid glands.
- If you are drinking alcohol: Watch for signs of a Hypo and never drink alcohol on an empty stomach.
- If you are exercising more than usual or if you want to change your usual diet.
- If you are ill: Carry on taking your insulin.

Medical Text 3

How does Hycamtin work?

The active substance in Hycamtin, topotecan, is an anticancer medicine that belongs to the group “topoisomerase inhibitors.” It blocks an enzyme called topoisomerase I, which is involved in the division of DNA. When the enzyme is blocked, the DNA strands break. This prevents the cancer cells from dividing and they eventually die. Hycamtin also affects non-cancer cells, which causes side effects.

How has Hycamtin been studied?

Hycamtin as an infusion has been studied in more than 480 women with ovarian cancer who had failed one treatment with platinum-containing anticancer medicines. Three studies were “open,” meaning that the medicine was not compared to any other treatment and the patients knew that they were receiving Hycamtin. The fourth study involved 226 women, and compared Hycamtin with paclitaxel (another anticancer medicine). The main measure of effectiveness was the number of patients whose tumours responded to treatment.

Technical Text 1

IMPORTANT SAFETY INSTRUCTIONS

WARNING: When using the dishwasher, follow basic precautions, including the following:

- Read all instructions before using the dishwasher.
- Use the dishwasher only for its intended function.
- Use only detergents or rinse agents recommended for use in a dishwasher, and keep them out of the reach of children.
- When loading items to be washed:
  - Locate sharp items so that they are not likely to damage the door seal; and
  - Load sharp knives with the handles up to reduce the risk of cut-type injuries.
- Do not wash plastic items unless they are marked "dishwasher safe" or the equivalent. For plastic items not so marked, check the manufacturer's recommendations.
- Do not abuse, sit on, or stand on the door, lid, or dish racks of the dishwasher.
- Under certain conditions, hydrogen gas may be produced in a hot water system that has not been used for two weeks or more. HYDROGEN GAS IS EXPLOSIVE [. . .]

Technical Text 2

What's New in Your Dishwasher

Energy

Congratulations on purchasing your water and energy efficient dishwasher! This dishwasher cleans by spraying the dishes with water and pauses to allow the detergent to soak into and release the soils on the dishes. The cycles are longer due to the soak and pauses for exceptional cleaning. Several models contain an optical water sensor. The optical water sensor is used to determine the optimum water and energy consumption for great cleaning performance. The first cycle using the sensor will run longer to calibrate the optical sensor.

Performance

Rinse Aid

Using rinse aid will optimize your drying and wash performance. This dishwasher is specifically designed to be used with rinse aid for improved drying performance and controlling buildup of hard water deposits. Energy efficient dishwashers use less water and energy, so they depend on the water "sheeting" action of rinse aid for total optimal performance.

Technical Text 3

Filtration System

Your dishwasher has the latest technology in dishwasher filtration. This triple filtration system minimizes sound and optimizes water and energy conservation while providing superior cleaning performance. Throughout the life of your dishwasher, the filter will require maintenance to sustain peak cleaning performance.

The triple filter system consists of 2 parts, an upper filter assembly and a lower filter.

- The upper filter assembly keeps oversized items and foreign objects, along with very fine food particles, out of the pump.
- The lower filter keeps food from being recirculated onto your dishware.

The filters may need to be cleaned when:

- Visible objects or soils are on the Upper Filter Assembly.
- There is degradation in cleaning performance (that is, soils still present on dishes).
- Dishes feel gritty to the touch.

It is very easy to remove and maintain the filters. The chart below shows the recommended cleaning frequency.

Ekaterina Lapshinova-Koltunski

7 Exploratory analysis of dimensions influencing variation in translation. The case of text register and translation method

Abstract: The present study investigates the interplay between two dimensions influencing translation: text register and translation method. This is achieved by a corpus-based analysis which involves the extraction of specific linguistic features occurring in multiple translations of the same texts. These translations differ, on the one hand, in registers the texts belong to, and on the other hand, in the translation method applied (human vs. machine translation). Our analysis is based on two frameworks – register theory and corpus-based translation studies – which also serve as sources for the definition of the features under analysis. Our quantitative analysis is supported with statistical methods. Unsupervised techniques are used to trace the degree of variation caused by the two dimensions, and also to identify the dimension having a greater impact on the translations. The results of our analysis shed light on the main factors influencing translation, and also deliver explanations for translation errors. In addition, further factors affecting linguistic features of translations, e.g. the experience involved, are traced in the present analysis. In this way, the study contributes to a better understanding of both translation product and translation process, and provides information which is useful for both the improvement and evaluation of translation.

1 Research goals and motivation

In the present study, we analyse the interplay between two dimensions influencing variation in translation: translation methods (human and machine translation) and text registers (e.g. fiction, political speeches). Our starting assumption is that the interplay between translation method and text register is reflected in the lexico-grammar of translated texts. As shown by Neumann (2013), translations are influenced by both language and context of situation (i.e. register a text belongs to). Linguistic features of translations vary according to these different dimensions. We believe, however, that there are more dimensions at

DOI 10.1515/9783110459586-008

play in translation than has been presumed so far. For example, due to recent developments in translation-oriented language technologies, translations are increasingly produced not only by human translators, but also by machine translation systems. There are also mixed forms of translations, such as computer-aided human translation or post-edited machine translation, with more classes within each subtype (e.g. human translations assisted by different tools, such as translation memories or terminology databases, or produced by experienced vs. inexperienced translators, with rule-based or statistical machine translation systems). Our assumption here is that translation types produced by/resulting from these different translation methods constitute another possible context of variation for translation, alongside language and register. We call these translation subtypes *translation varieties*. To date, variation along the third parameter (translation method) has not received much attention in studies devoted to translation. Some studies that address both human and machine translations (cf. Babych and Hartley 2004; Papineni et al. 2002; Popović 2011; White 1994) focus solely on translation error analysis, using human translation as a reference in the evaluation of machine translation outputs. Only a few of them operate with linguistically-motivated categories (e.g. Fishel et al. 2012; Popović 2011), but once again, these categories are used to detect errors in translations. To our knowledge, the only study dealing with the differentiation between human and machine translation is Volansky et al. (2011). They derive the features they analyse from corpus-based translation studies and try to detect those which are common to both translation varieties (e.g. contextual function words, part-of-speech patterns) and those that differentiate them (average sentence length, passive verb ratio, etc.). Their dataset contains newspaper articles only, and hence, cannot be used to reveal variation along the parameter of register. In Zampieri and Lapshinova-Koltunski (2015), automatic text classification techniques were applied to differentiate between different registers and methods of translated texts, using data-driven and not linguistically motivated features.

Our primary interest is in linguistic features of translation varieties, such as active vs. passive verb constructions, preferences for certain functional verb classes, modality meanings, proportion of nominal vs. verbal phrases. Our assumption is that they reflect the interplay of the dimensions described above: language-specific (both source and target language), register-specific, as well as translation method-specific. In this study, we focus on the variation affected by register and translation method only. This decision is due to the nature of the dataset we have at our disposal: it consists of English-to-German translations only. As shown in various studies (see Teich 2003 or Neumann 2013 among others), variation in translation is realised by different linguistic phenomena which are situated at various linguistic levels (morpho-syntax, lexis and text). This study will, therefore, investigate the quantitative distributions of linguistic features reflected in the lexico-grammar of texts. The comparison of the distribution of these features across translation varieties will provide insights into the interplay of the two dimensions under analysis.

We believe that this type of analysis will allow us to formulate general statements based on larger quantities of data rather than individual texts, as in Neumann (2013). Quantitative analysis requires corpus-based methods and involves the classification and counting of features, as well as their statistical validation, see McEnery and Wilson (2001) and Biber et al. (1998). Our objective is to shed light on linguistic properties of translated texts produced by both humans and machines with a view to arriving at a better understanding of the translation product, and, to some extent, of the translation process. We also aim at showing the usefulness of the results obtained in this study for the fields of translation evaluation and quality estimation. Although registers have been taken into account in several studies on translation quality evaluation in human translation, they have remained under-researched in the field of machine translation.

2 Related work

As already mentioned in section 1, translations may vary according to different parameters, e.g. language, register and translation method. The present section examines studies that deal with variation in translation along all three parameters. These studies serve as important sources for the aggregation of linguistic features required for a corpus-based analysis.

2.1 Language variation and translation

Studies on translation variation along the language dimension concentrate on differences between translations and non-translated texts in both source and target languages (Hansen 2003; House 2014; Matthiessen 2001; Steiner 2004; Teich 2003). Most of them are related to the notion of *translationese*, invoked by Gellerstam (1986) to describe differences between original and translated texts. Baker (1993, 1995) elaborates specific features of translations that are believed to be universal, irrespective of source and target language. These universals include explicitation, a tendency to spell things out rather than leave them implicit; simplification, a tendency to simplify the language used in translation; normalisation, a tendency to exaggerate features of the target language and to conform to its typical patterns; and levelling out, the similarity of individual texts among each other in a set of translations if compared to individual texts in a set of non-translations. For the latter, we prefer to use the term *convergence*, which means a relatively higher level of homogeneity of translated texts with regard to their scores of lexical density, sentence length, etc. Another translation feature which is ignored by Baker is Toury's law of interference (Toury 1995). We prefer to use the term *shining through*, i.e. features of the source texts observed in translations, see Teich (2003).

In some recent approaches, machine learning techniques have been applied to the analysis of translationese. Text classification is used to identify translation features or to differentiate between translations and non-translations. For instance, Baroni and Bernardini (2006) analyse the features of Italian translations on the basis of monolingual comparable corpora (translations into Italian from a number of languages, including English, Arabic, French, Spanish and Russian, as well as their comparable non-translated originals) using machine learning techniques for text categorisation. Their work shows that it is possible to distinguish translations from non-translations automatically. Ilisei et al. (2010) differentiate between Spanish translated from English and non-translated texts by means of a simplification measure based on average sentence length, proportion of simple and complex sentences, lexical richness, etc. Another example is Kurokawa et al. (2009) who investigate the possibility of automatically determining whether a piece of text is an original text or a translation using word and part-of-speech n-grams. They are especially interested in the implication of their work for machine translation performance, but do not consider the identification of human- vs. machine-translated texts (see also Lembersky et al. 2012).

However, none of the above-mentioned studies compares registers, although Baroni and Bernardini (2006) do point out the importance of text registers. They acknowledge that the use of a comparable corpus representing one specific register is a drawback of their experiment.

2.2 Register variation and translation

Studies within register and genre theory, e.g. Quirk et al. (1985), Halliday and Hasan (1989), Biber (1995), analyse contextual variation in languages. In their terms, languages vary with respect to usage context. These contexts influence the distribution of particular lexico-grammatical patterns which manifest language registers. The canonical view is that situations can be characterised by the parameters of field, tenor and mode of discourse. Field of discourse relates to processes and participants (e.g. Actor, Goal, Medium), as well as circumstantials (Time, Place, Manner, etc.) and is realised in lexico-grammar in lexis and colligation (e.g. argument structure). Tenor of discourse relates to roles and attitudes of participants, author-reader relationship, and is reflected in stance expressions or modality. Mode of discourse relates to the role of the language in the interaction and is linguistically reflected at the grammatical level in Theme-Rheme constellations, as well as cohesive relations at the textual level. In other words, the contextual parameters of registers correspond to sets of specific lexico-grammatical features, and variation across different registers can be seen in the distribution of these features which are expressed in lexico-grammatical patterns.

Multilingual register studies concern linguistic variation, i.e. the distribution of lexico-grammatical features, not only across registers but also across languages, comparing the settings specific for the languages under analysis, e.g. Biber (1995) on English, Nukulaelae Tuvaluan, Korean and Somali, Hansen-Schirra et al. (2012) and Neumann (2013) on English and German. In addition, the latter two also consider register analysis in translations. Some other scholars, e.g. House (1997) and Steiner (1996, 1998, 2004), also integrate register analysis in translation studies. However, they either do not account for distributions of these features, or analyse individual texts only. Likewise, De Sutter et al. (2012) and Delaere and De Sutter (2013), in their analysis of translated Dutch, pay attention to register variation, but concentrate on lexical features only.

Applying a quantitative approach, Neumann (2013) analyses an extensive set of features and shows the degree to which translations are adapted to the requirements of different registers, thereby detecting a further dimension to the study of translation properties and showing how both register and language typology are at work.

2.3 Variation in translation method

As previously mentioned (see section 1), studies involving both human and machine translation mostly focus on translation error analysis, i.e. automatic error detection. Human translation usually serves as a reference for the comparison of different machine translation outputs. However, some of them do consider linguistic properties, or linguistically-motivated errors, see for instance Popović (2011), Popović and Burchardt (2011) or Fishel et al. (2012). The latter operate with such features as missing words (content vs. grammatical), incorrect words (incorrect disambiguation, wrong lexical choice, etc.), wrong word order, etc. This classification also includes style errors, which are partly related to register. Yet, this error type is analysed on the level of words only, and thus, further register settings involved, e.g. part-of-speech classes, passive vs. active verb construction are not taken into account. In other words, none of these studies provides a comprehensive descriptive analysis of specific features of different translation methods.

El-Haj et al. (2014) compare translation style and consistency in human and machine translations of Camus's novel *The Stranger*, which was translated from French into English and Arabic. They use readability as a proxy for style, and then measure how it varies within and between translations. Their work, however, is of an evaluative nature, as they aim at quality evaluation of both human and machine translation.

To our knowledge, the only study describing linguistic differences between human and machine translation is Volansky et al. (2011). The authors analyse human and machine translations, as well as comparable non-translated texts. They use a range of features based on the theory of translationese (simplification, explicitation, normalisation and shining through), and find out that the features specific to human translations can also be used to identify machine translation. The authors show that there are features of human translations which coincide with those of machine-translated texts, whereas other features differ across these two translation methods (see section 1 above). The combination of their methodology and translationese features can reveal similarities and dissimilarities between human and machine translations.

3 Linguistic features, data and methodology

3.1 Features under analysis

For our analysis, we selected a set of features derived from the studies on language, register and translation method variation described in section 2. These features represent lexico-grammatical patterns of more abstract concepts, e.g. textual cohesion expressed via pronominal or nominal reference, evaluative patterns expressed via certain syntactic constructions. The selected features were chosen because they reflect linguistic characteristics of all texts under analysis, are content-independent (do not contain terminology or keywords), and are easy to interpret, thereby yielding insights into the differences between the variables under analysis (cf. Volansky et al. 2011). As a result, some of the features analysed in the studies mentioned above were excluded from our analysis. For instance, no token n-grams were used here, as they are rather content-dependent, and reflect domains. In addition, as we have multiple translations of the same texts in our data, this kind of feature is not suitable for our analysis. We also used groupings of nominal and verbal phrases (based on chunk annotations) rather than part-of-speech n-grams, as they are easier to interpret than n-grams.

The set of selected features used in our analysis is outlined in Table 1. The first column shows the lexico-grammatical patterns that we extracted for the quantitative analysis. The second column presents the correspondence of these patterns to the context parameters (field, tenor and mode of discourse) (see

Table 1: Features under analysis

	lexico-grammatical patterns	register analysis	translationese studies
1	content vs. total words	mode	simplification
2	nominal vs. verbal parts-of-speech and phrases (np.chunk, vp.chunk)	field	shining through/normalisation
3	*ung*-nominalisation (ungnom)	field	shining through/normalisation
4	nominal (all.np) vs. pronominal (pronnp) and demonstrative vs. personal reference (perspron, dempron)	mode	explicitation, shining through/ normalisation
5	abstract or general nouns (gen.nouns) vs. all other nouns	field	explicitation
6	logico-semantic relations: additive, adversative, causal, temporal, modal	mode	explicitation
7	modality: obligation, permission, volition	tenor	shining through/normalisation
8	evaluation patterns	tenor	shining through/normalisation

section 2.2). The third column links the lexico-grammatical patterns with the translation features outlined in section 2.1.

Content words and their proportion to the total number of words in a text (row 1) represent lexical density, which is related to informational density in a text. This corresponds to the mode parameter in register theory and simplification in translationese studies (lexical richness of translations). The number of nominal and verbal parts-of-speech, as well as their groupings into nominal and verbal phrases or chunks (row 2) reflect participants in the field parameter, shining through and normalisation (as languages use different grammatical structures, which is reflected in translations, and English tends to be more verbal than German; see Steiner 2012). For the same reasons, field and shining through / normalisation can also be analysed via the distribution of nominalisations (*ung*-nominalisations in row 3). Reference expressed either in nominal phrases or in pronouns (row 4) reflects textual cohesion in the parameter of mode. From the point of view of translationese studies, this feature can point to explicitation, as pronouns are less explicit than nouns or nominal phrases. Moreover, preferences for personal or demonstrative pronouns in different languages (in our case English and German) can be reflected in shining through and normalisation. The distribution of abstract or general nouns and their comparison to other nouns (row 5) gives information about lexical choices (parameter of field) and preferences for more concrete or abstract words in translations (explicitation). Conjunctions (including both grammatical conjuncts such as *und – and*, *aber – but* and multiword expressions like *aus diesem Grund – that is why*), for which we analyse distributions of logico-semantic relations (row 6), belong to the parameter of mode as they express cohesion, and at the same time to the explicitation feature, as they explicitly mark relations in discourse.

Modal verbs, e.g. *können – can, müssen – must* (row 7), express modality, i.e. the parameter of tenor. They are grouped according to different meanings, and also reveal cross-linguistic contrasts, as described by Teich (2003) and König and Gast (2012) for differences between English and German. That is why their distribution in translation also reflects normalisation and shining through. Similarly, these phenomena (tenor and shining through/normalisation) are reflected in evaluation patterns (e.g. *es ist interessant/wichtig zu wissen… – it is interesting/important to know*, row 8).

Information on the structural properties of the features, including examples illustrating subcategories, is presented in section 3.3 below.

3.2 Corpus data

To our knowledge, the only corpus resource suitable for our research agenda is VARTRA-SMALL (cf. Lapshinova-Koltunski 2013). This corpus contains different translation varieties, the texts of which are translated from English into German. The translations included in VARTRA-SMALL were produced with the following methods: by 1) professional humans (PHT, Professional Human Translation) and 2) student translators (SHT, Student Human Translation); as well as with MT systems: 3) a rule-based MT system (RBMT, Rule-Based Machine Translation) and two statistical MT systems – 4) Google Statistical Machine Translation (GSMT) and 5) Moses Statistical Machine Translation (MSMT). The dataset contains multiple translations of the same texts, which vary both in translation method and text register, and thereby represent different translation varieties (as defined above).

Translations by professionals (PHT) were exported from the already existing CroCo corpus (Hansen-Schirra et al. 2012). The SHT variant was produced by student translators with at least a BA degree, who have little or no experience in translating. Translators in the SHT production were assisted with different translation memories (available in the OPUS collection at http://opus.lingfil.uu.se/), used with the help of Across[1], a computer-aided translation tool which can be integrated into the usual work environment of a translator. We did not time the

[1] http://www.my-across.net/

translation tasks, and did not take into account other settings of the translation process, e.g. time spent on terminology search. The translation task was not part of any examination, and the students could freely make decisions on the use of additional reference resources.

The rule-based machine translation variant was produced with SYSTRAN, whereas for statistical machine translation we used two systems – Google Translate (GSMT), and the in-house Moses-based system (MSMT). MSMT was trained with EUROPARL, a parallel corpus containing texts from the proceedings of the European parliament, cf. Koehn (2005). This training set is one of the largest existing datasets which are freely available. Moreover, this corpus is used in most studies on machine translation. The decision to include two SMT systems is justified by the fact that the first one (GSMT) is trained with enormous data Google has at its disposal, whereas MSMT is trained with a parallel corpus, which is smaller than the data available at Google, and has a register restriction (as it contains transcripts of political speeches only). In this way, we have two translation varieties displaying shortages in experience or data: inexperienced translators in SHT and insufficient training data in MSMT.

Each translation subcorpus contains translations of the same texts which cover seven registers of written language: political essays (ESSAY), fictional texts (FICTION), manuals (INSTR), popular-scientific articles (POPSCI), letters to share-holders (SHARE), prepared political speeches (SPEECH), and tourism leaflets (TOU). It should be noted that some of these registers represent a continuum between written and spoken dimensions, i.e. SPEECH, which is written-to-be-spoken, and FICTION, which, because of the dialogues it contains, is on the border between spoken and written texts. The dataset contains both frequently machine-translated texts, e.g. political speeches, and those which are usually not translated with MT systems, such as fiction. All translation variants in VARTRA-SMALL comprise ca. 600,000 tokens. All subcorpora are tokenised, lemmatised, and tagged with part-of-speech information, segmented into syntactic chunks and sentences. The annotations in VARTRA-SMALL were obtained with Tree Tagger (Schmid 1994).

The subcorpora are encoded in the CWB format (CWB 2010) and can be queried with the help of CQP regular expressions, the syntax of which is described in Evert (2005).

3.3 Feature extraction

As mentioned in 3.2, VARTRA can be queried with CQP, which allows the definition of language patterns in the form of regular expressions based on string, part-of-speech and chunk tags, as well as further constraints. Table 2 outlines a number of examples for the queries used in our corpus. For instance, query 1 is used to differentiate between personal and demonstrative reference. Here, we simply search for personal or demonstrative pronouns and make use of the part-of-speech annotation in our corpus. In the second query, we also add lexical information, reducing our search to items which are tagged as nouns and end with *-ung*. This query is used to extract German *ung*-nominalisations. The third query is even more restricted, as we reduce our search to certain lexical items, such as modal verbs. We utilise this restriction to classify different modal meanings. The classification of different logico-semantic relations expressed via conjunctions is achieved with the help of query 4. Here, we do not use part-of-speech annotation, as we are interested not only in grammatical conjunctions, like *und (and), oder (or), deswegen (therefore)*, but also in multiword expressions such as *darüber hinaus (in addition)* and *aufgrund dessen (that is why).* Manually compiled lexical lists are used to extract these items (ranging from ca. 80 to ca. 100 types per list). The lists of conjunctions expressing logico-semantic relations were derived from Lapshinova and Kunz (2014), who describe a procedure to semi-automatically annotate these relations in a multilingual corpus.

Table 2: Queries for feature extraction

<table>
<tr><th></th><th>feature category</th><th>CQP query</th></tr>
<tr><td>1</td><td>personal</td><td>[pos="PP.*"]</td></tr>
<tr><td></td><td>demonstrative</td><td>[pos="PD.*"]</td></tr>
<tr><td>2</td><td><i>ung</i>-nominalisation</td><td>[pos="NN.*"&lemma=".*ung.*"]</td></tr>
<tr><td>3</td><td>obligation</td><td>[pos="VM.*"&lemma="müssen|sollen"]</td></tr>
<tr><td></td><td>permission</td><td>[pos="VM.*"&lemma="können|dürfen"]</td></tr>
<tr><td></td><td>volition</td><td>[pos="VM.*"&lemma="wollen|mögen"]</td></tr>
<tr><td>4</td><td>additive</td><td>$additive-conjunction</td></tr>
<tr><td></td><td>adversative</td><td>$adversative-conjunction</td></tr>
<tr><td></td><td>causal</td><td>$causal-conjunction</td></tr>
<tr><td></td><td>temporal</td><td>$temporal-conjunction</td></tr>
<tr><td></td><td>modal</td><td>$modal-conjunction</td></tr>
<tr><td>5</td><td>abstract nouns</td><td>[pos="NN.*"&lemma=$abstract_nouns”]</td></tr>
<tr><td>6</td><td>evaluation patterns</td><td>"es|Es"[pos="VAFIN"][pos!="$.|$,"]{0,3}
[pos="AD.*"][]?"da(ss|ß|\"s)"
"es|Es"[pos="VAFIN"][pos!="$.|$,"]{0,3}
[pos="ADJ.*"] []? "zu|wenn|f(ü|\"u)r"
"(A|a)m" [pos="AD.*"&word=".*ste.*"]</td></tr>
</table>

In some cases, e.g. for the extraction of abstract nouns, we also add a part-of-speech restriction, as shown in query 5. For evaluation patterns (derived from pattern grammar by Francis and Hunston 2000), we need more morpho-syntactic restrictions. These include sequences of parts-of-speech and lexical elements to extract seven evaluation patterns (see Table 3). For the time being, we do not classify them according to their evaluative meaning, simply considering the amount of evaluation in translation varieties.

Table 3: Evaluation patterns

evaluation pattern	German example	English translation
es+Verb *BE*+ADJ+,+*dass*	*es ist erforderlich, dass*	it is required that
Verb *BE*+*es*+ADJ+,+*dass*	*ist es erforderlich, dass*	is it required that
es+Verb *BE*+ADJ+,+*zu/wenn/für*	*es ist besser, wenn*	it is better if
Verb *BE*+*es*+ADJ+,+*zu/wenn/für*	*ist es besser, wenn*	is it better if
Verb *machen*+*es*+ADJ	*machen es schwerer*	to make it more difficult
es+ADJ+Verb *machen*	*es schwerer machen*	to make it more difficult
am+ADJsuperlative	*am wichtigsten ist dabei*	the most important thing here

CQP facilities allow us to count the extraction results and sort them along the texts, registers and varieties they occur in. The extracted frequencies of these features are saved in a matrix for further validation in R (version 3.0.2; R Core Team 2013).

3.4 Statistical analysis

For our analysis, we use an unsupervised technique – hierarchical cluster analysis (HCA), see Baayen (2008) and Everitt et al. (2001). With the help of this technique, we can discover differences and similarities between the translation varieties (subcorpora) under analysis.

Unsupervised data analysis, which is sometimes called knowledge discovery (cf. Murphy 2012), allows us to discover ‘interesting structures’ in the data. In our case, we are looking for structures in the form of translation clusters which are formed according to different dimensions, i.e. translation method and register. Whereas in a supervised case (used in Lapshinova-Koltunski and Vela 2015; Zampieri and Lapshinova-Koltunski 2015), we would define the clusters and the features specific to them, in an unsupervised case, we are free to choose the number of clusters as we like. Applying this technique, we hope to trace the interplay of variation dimensions in our data, to see which of them has a greater impact on the clustering of translations, and also to discover further structures in our data.

In hierarchical cluster analysis, a set of dissimilarities for the *n* objects being clustered is used. These dissimilarities are calculated by the Euclidean distance, i.e. distance between datasets, in our case between the subcorpora under analysis. Euclidean distance is one of the most straightforward and generally accepted ways of computing distances between objects in a multidimensional space.

We employed the complete-linkage method to perform clustering, as it yields compact clusters of approximately equal diameter. According to this method, the similarity between two clusters is the “worst-case” similarity between any pair of subcorpora in the two clusters. This means that at each step, the two clusters separated by the shortest distance are combined. The ’shortest distance’ between two clusters is based on all pair-wise distances between the elements of both clusters, so that the distance between clusters equals the distance between those two elements (one in each cluster) that are farthest away from each other. The shortest of these links that remains at any step causes the fusion of the two clusters whose elements are involved.

Complete-linkage clustering can avoid chaining (clusters may be forced together due to single elements being close to each other, even though many of the elements in each cluster may be very far apart from each other) (cf. Everitt et al. 2001).

The results of hierarchical clusters are represented graphically in a dendrogram, which is a branching diagram that represents the relationships of similarity among a group of entities. The variables are shown as leaves (subcorpora), and the branches or clades – the clusters. The arrangement of the clades tells us which leaves are most similar to each other. The height of the branch points indicates how similar or different they are from each other: the greater the height, the greater the difference. Highly correlated clusters are nearer the bottom (also leftmost or outermost, depending on the representation) of the dendrogram.

4 Analyses, results and interpretation

This section presents the results of the analysis which we performed in several steps. First, we analyse intra-dimensional variation, i.e. variation along the dimension of translation method and along the dimension of register (4.1). Second, we combine both dimensions and analyse variation influenced by both dimensions (4.2). In section 4.3, we analyse the features contributing to the resulting variation.

4.1 Intra-dimensional variation

a) Variation across translation methods

First, the variation across translation methods is analysed, for which we have five dependent variables: PHT, SHT, RBMT, GSMT and MSMT. Distance measures are calculated for each variable on the basis of the total number of occurrences of the features. The subcorpora, each representing the lowest (most left) leaves of the dendrogram (see Figure 1) are joined up to the top (on the right side of the graph). The dendrogram is then read from the bottom up (from left to right in our case), identifying in this way, which subcorpora, and hence which translation varieties, are most similar to each other.

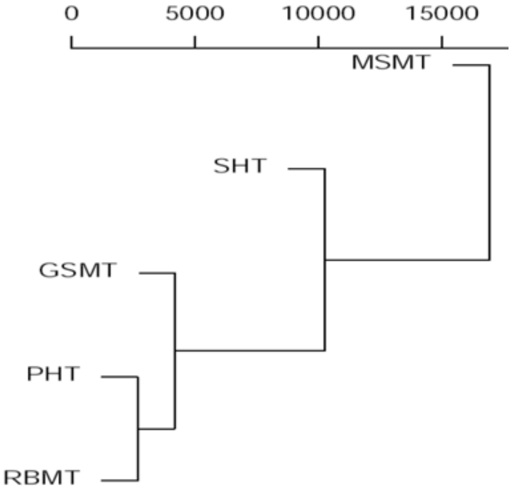


Figure 1: Variation across translation methods

The first node to join together is that of PHT and RBMT. The connection between them is the closest link to the bottom of the diagram, which means that the distance between them is the smallest. Together, they join with GSMT. This indicates that every node within this cluster is more similar to every other node within this cluster than to any nodes that join at a higher level (SHT and MSMT in this case). We move further and join this node with SHT, then, moving further to MSMT. The length of the horizontal lines indicates the degree of difference between branches.

The greatest differences (the longest lines between the last two clusters) are observed for SHT and MSMT, which vary strongly from other subcorpora. We assume that the variance of SHT and MSMT from the rest can be explained by another dimension of variation that comes into play – *experience* (which includes both degree of experience of human translators and the amount of training data in a statistical machine translation system) involved in the translation process. Differences in this dimension were pointed out in several studies, e.g. by Göpferich and Jääskeläinen (2009) and Carl and Buch-Kromann (2010) for human translation or Estrella et al. (2007) and Koehn and Haddow (2012) for machine translation. Although none of them combine human and machine translation in their analyses, some parallels can be observed in both translation varieties. For instance, Göpferich and Jääskeläinen (2009) state that with increasing translation competence, translators focus on larger translation units, just as in (phrase-based) machine translation, large training data sets make it possible to learn longer phrases (see Koehn 2010).

In our data, both SHT and MSMT reveal a certain lack of experience: the former is produced by novice translators, whereas the latter is trained with a restricted set of data. We assume that human translators and statistical MT systems behave in a related fashion in terms of their performance. That is the degree of experience in human translation and the amount of data in machine translation contribute to similar outcomes in both translation varieties. Translations by novice translators are similar to those produced with a statistical system trained with a small amount of data, whereas texts translated by professional translators may be comparable to those produced with SMT systems trained with larger data. However, this trend will need to be further tested in future experiments.

b) Variation across registers

We also analyse the variation across registers of all translated texts (not differentiating between methods). In this case, we have seven dependent variables: ESSAY, FICTION, INSTR, POPSCI, SHARE, SPEECH and TOU. The dendrogram representing distances and clusters of these variables is shown in Figure 2.

Moving from left to right in Figure 2, we observe the following groups of registers: 1) ESSAY and TOU; 2) FICTION; 3) INSTR and POPSCI and 4) SHARE and SPEECH. The next two nodes join 1) with 2), and 3) with 4), although FICTION varies more strongly from ESSAY and TOU than INSTR and POPSCI from SHARE and SPEECH. The greatest difference lies between two bigger groups of registers, which we can observe at the very right.

Similarities between ESSAY and TOU in German non-translated texts have been observed by Neumann (2013) for almost all sub-dimensions of field, tenor and mode (see Neumann 2013: 313, Table 60). Similarly, Diwersy et al. (2014) found a clear tendency of translations to normalise their features in order to adapt to target language conventions. The authors used the same dataset as that employed by Neumann (2013). Our results show that the same tendency can be observed not only for professional human translation (as in the case of the two studies mentioned here), but also for machine translation. Neumann (2013) also demonstrates the individuality of FICTION in both English and German. She points out that translations in the FICTION register do not exhibit any deviation from the non-translations (both English and German), which can probably

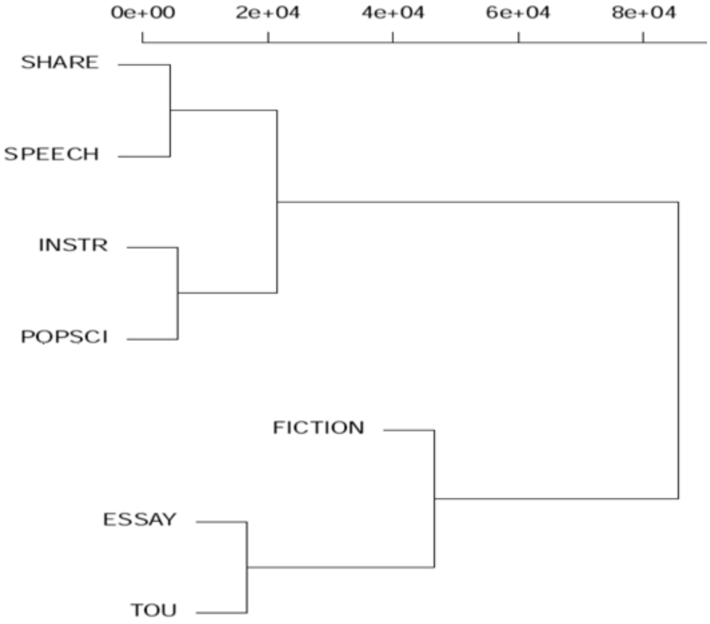


Figure 2: Variation across registers

be explained by the fact that the register features of fictional texts coincide in both languages. As a result, this register is distinctive from the other registers under analysis. The features of SHARE and SPEECH in our translations coincide, and we suggest that the phenomena of shining through and normalisation are at play here, which means that the features of both source and target languages interact in translations of these registers. In Neumann's description of register profiles, SHARE and SPEECH would coincide if we take into account both languages, see the profiles of the English and German original registers in Neumann (2013: 309, 313) for details. At the same time, the similarities between INSTR and POPSCI in our data rather contradict Neumann's definition of register profiles. In her description, these two registers coincide only in mode of discourse, if both languages are considered. If compared monolingually, that is, within each language, INSTR and POPSCI have similarities in English only, and only in the parameter of tenor. However, the author also states that register profiles of translations might deviate from those of non-translations. We suspect that in the case of INSTR and POPSCI we observe this kind of deviation, which explains the unexpected clustering of these registers in our data.

Comparing the distance indication on the scale in the two figures, we find that the difference between register clusters is greater than that between translation methods. On this basis, it can be concluded that variation along the register dimension is greater in our translation data than variation along the translation method dimension. To test this, we need to combine both dimensions in one single analysis.

4.2 Inter-dimensional variation in translation varieties

To analyse the interplay between translation methods and registers, and to find out which of these two dimensions causes greater variation in our data, we calculated the distances between registers of all translations that underlie the hierarchical clustering. In this case, we have 35 dependent variables, for which distances are calculated. The results are represented in Figure 3.

The dendrogram clearly reveals two very distinct groups: the bottom group seems to consist of two more distinct clusters, while the clustering of most classes in the upper group is more levelled out.

If we start to generate the tree from the outermost nodes (or leaves), we observe a clear predominance of register features for the clustering of fictional texts, tourism leaflets and political essays. The only exception here is the set of tourism texts translated by students, which varies more from other translations of TOU, though ultimately it joins the cluster formed by TOU and ESSAY.

It is worth nothing that SHT-FICTION stands out in the corresponding class. These results are in line with our observations regarding groups of registers in section 4.1 above, where we show that FICTION builds a class of its own. The tendency to group according to experience (which we observed within the intra-dimensional variation across translation methods in 4.1) is detected on the smallest nodes of fictional and tourism texts: PHT shows more similarities with RBMT and GSMT. However, in the register of political essays, both varieties of human translations are clustered together before they are clustered with the varieties of the MT systems. On the whole, the results for this group confirm the observations that variation along the register dimension is more prominent in translations than along the translation method dimension.

The clusters comprising TOU, ESSAY and FICTION on the one hand, and INSTR, POPSCI, SHARE and SPEECH on the other, reveal different variation behaviour. The latter does not demonstrate any prominence of either register or method, as the two bigger classes within this group are built up by a mixture of registers: one of instruction manuals and popular-scientific articles, and the other made up by SHARE and SPEECH texts. The dimension of experience is not present in this group of translations. However, we observe a tendency for

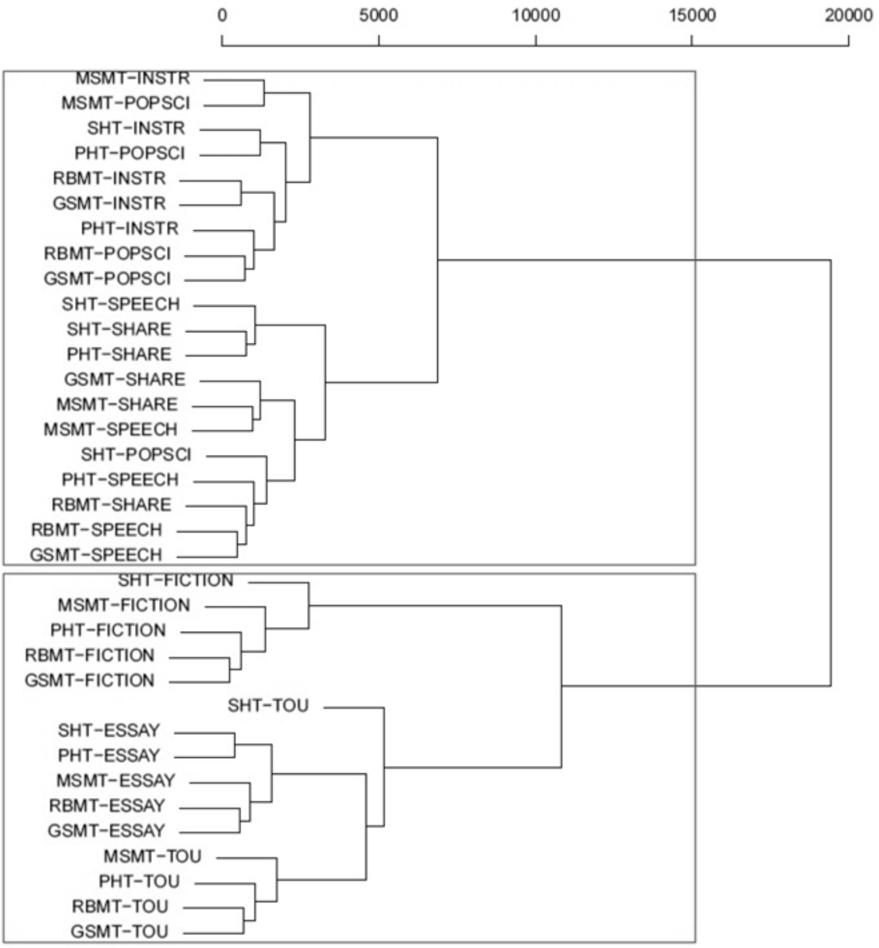


Figure 3: Variation across registers and translation methods

translations to form human vs. machine clusters. In some cases, we even observe a more fine-grained method-driven clustering: Moses-generated translations of INSTR and POSPCI, as well as SHARE and SPEECH are grouped, which means that these translations do not diversify in terms of register, and at the same time, the resulting translation varieties deviate more strongly from other translations.

In general, groups in the INSTR-POPSCI-SHARE-SPEECH cluster are more heterogeneous and diverse, and contrary to the TOU-ESSAY-FICTION cluster, it is rather difficult to identify which of the two dimensions predominates in this dataset. Both registers and methods are less pronounced and distinct in the second cluster, and the new dimension of variation which we discovered in section 4.1 is not observed here at all.

As we are not able to define which of the two dimensions under analysis is more prominent in the case of the second group, we tested how many clusters we need to describe this dataset, excluding ESSAY, FICTION and TOU from the analysis. The underlying assumption here is that if a resulting cluster contains more variables representing translation method, it shows the prominence of the register dimension, while the prominence of the method dimension is demonstrated by a greater number of register variables.

The maximum number of clusters that we can have should not exceed five (as we have four registers and five translation methods), and accordingly, we cut the tree into five clusters (see Table 4).

Table 4: Cluster membership in a five-cluster tree

	SHT	PHT	RBMT	GSMT	MSMT
INSTR	1	1	1	1	5
POPSCI	2	1	1	1	5
SHARE	3	3	2	4	4
SPEECH	3	2	2	2	4

Cluster 1 contains two registers and four translation methods, and thereby reflects the register dimension. Cluster 2 is represented by three registers and four translation methods, being rather on the borderline of the two dimensions. Clusters 3 and 4 are both built up by two registers and two translation methods. Interestingly, they contain the same register variables (SHARE and SPEECH), opposing in this way human translation vs. statistical machine translation (difference in the dimension of translation method). The last cluster (5) includes MSMT only, and also reflects translation method.

Summarising these observations, we can say that the greatest differences are still observed for the register dimension, and not that of translation method, as variation along the latter is observed rather for smaller groups of variables. However, we admit that register groupings are different across human vs. machine translation. In general, machine translations show a lower number of distinctive registers than human translations. This finding can be explained by the fact that register features have not been taken into account in machine translation to date. In studies on machine translation, authors mostly operate with the notion of domain, which covers the lexical level only, cf. Lapshinova and Pal (2014). Therefore, on the level of register settings, the translations produced by MT systems do not vary as much as those produced by humans who are aware of register constraints. This observation does not tie in with the findings in section 4.1 above, in which the clustering of registers in all translations together were taken into account. This means that if a translation is analysed in terms of register settings, we need to take into account the method with which this translation was produced. For instance, in order to judge the quality of a translation regarding its correspondence to the register standards in the target language, the production method involved needs to be considered.

4.3 Features involved in the cluster formation

To analyse the influencing factors, i.e. the features contributing to the classification of the translation varieties in our data, we need to consider the numeric data. Table 6 presents the original feature set for each of the seven clusters. We chose this number of clusters according to the number of registers in our data as they seem to be more prominent than methods, although some of them are mixed, see Table 5.

Table 5: Main clusters in the analysed data

clusters	translation varieties
1	political essays
2	fictional texts
3	instruction manuals and popular-scientific texts except SHT-POPSCI
4	letters-to-shareholders and political speeches except SHT- and PHT-SHARE and SHT-SPEECH
5	human translations of letters-to-shareholders and SHT-SPEECH
6	student translations of tourism texts
7	tourism leaflets except SHT-TOU

The figures presented in Table 6 represent median values for the variables we have used in the cluster analysis, which are broken down by cluster groups. These values change if a different number of clusters is selected[2]. Comparing the figures across clusters, we can characterise each cluster according to the set of features specific to it. For instance, cluster 1 (political essays) is characterised by an average distribution of nominal and verbal parts-of-speech and phrases, a relatively low number of conjunctive relations. The amount of pronominal reference here is also lower than average. General nouns, *ung*-nominalisations,

[2] We have used the aggregate() function in the R clustering package.

Table 6: Features contributing to cluster definition

cluster	1	2	3	4	5	6	7
content.words	7037	4407	9469	10352	11098	6633	8569
np.chunk	4651	3384	6274	7039	7332	4115	5506
vp.chunk	1712	1570	2344	2594	2908	1381	1495
nominal	8885	5679	11483	13239	13772	7993	10536
verbal	4197	3485	5857	6397	7124	3361	4177
additive	726	622	717	1042	1169	752	880
adversative	313	250	502	414	453	232	274
causal	115	169	271	133	179	65	79
temporal	384	332	614	566	628	213	308
modal	96	146	248	131	159	57	89
pronnp	39	154	82	67	62	35	21
all.np	3399	2717	4602	5121	5140	2799	3969
gen.nouns	121	52	107	139	138	37	49
all nouns	3464	1852	4816	5292	5701	3394	4417
obligation	72	32	46	72	76	8	13
permission	76	38	174	96	162	81	65
volition	18	35	32	22	30	5	6
evaluation	11	4	8	12	12	5	7
ungnom	685	94	422	786	756	134	211
perspron	521	1127	781	1141	1084	288	408
dempron	122	102	175	263	252	43	68

as well as modal verbs expressing obligation are more frequent here than on average. Modality, especially that with the meaning of obligation, is one of the indicators of argumentative goal orientation (a part of the context parameter of field in register theory, see section 2.1). Argumentative texts contain significantly more modality than texts pursuing other goals.

Cluster 1 has a high number of modal verbs of obligation, when compared to other clusters (cluster 5 only has a higher number of them). The verbs are used in their meaning of personal obligation, rather than logical necessity, see example (1).

(1) *Erstens* ***müssen*** *wir die ansteigende Produktion ausgleichen, indem wir einen sauberen und sparsamen Energieverbrauch in den Vordergrund stellen. Zweitens* ***müssen*** *wir unsere internationalen Beziehungen mit den Verbraucher- und Erzeugerländern ausbauen. Drittens* ***müssen*** *wir unsere Energiequellen erweitern und diversifizieren […]* (PHT-ESSAY).
‘First, we **must** balance the rising production by providing a clean and efficient energy consumption in the foreground. Second, we **must** expand our international relationships with the consumer and producer countries. Third, we **must** expand our energy sources and diversify […]’.

At the same time, the translations of cluster 1 also demonstrate features of goal type exposition, which are reflected in lexico-grammar by the low number of personal pronouns and a higher number of nouns, including general nouns and *ung*-nominalisations. These characteristics certainly correspond with the observation made by Neumann (2013: 184).

The main distinctive features of cluster 2 include first a prevalence of pronominal reference, which is expressed with personal pronouns in most cases, see example (2). This usually characterises spoken language, and as we know, FICTION is on the borderline between written and spoken language as it contains conversations.

(2) ***Er*** *glaubte* ***es*** *nicht einmal ansatzweise, aber* ***er*** *wollte, dass* ***ich es*** *glaubte oder mir Gedanken darüber machen würde. Und* ***er*** *könnte sich dann über* ***mich*** *lustig machen* (SHT-FICTION).
‘**He** did not even start to believe **it**, but **he** wanted **me** to believe **it** or think about it. And then **he** could laugh about **me** [...]’.

Moreover, pronouns also correspond to one of the characteristics of narrative texts. Further characteristics include description of events in clausal structures rather than in nominal structures, which leads to a lower amount of nominalisations in this cluster. Again, the features characterising cluster 2 correspond to those defined for FICTION, see Steiner (2012) and Neumann (2013).

Cluster 7 contains all translations of tourism texts except the one produced by inexperienced translators. We assume that the main difference is in the proportion of nominal and pronominal reference in translation variants of this register. SHT-TOU seems to contain more pro-forms than the other translation varieties, see examples in (3).

(3a) *Nordirlands Bevölkerung liebte schon immer die Natur. Die Menschen aus Ulster sind keine Stubenhocker! Manche trödeln an den ’Loughs’ (Seen oder Meeresarmen) herum, andere verbringen* ***ihre*** *Freizeit mit Angeln oder Bootfahren, wieder andere machen Ausflüge mit der Familie in die Berge oder in die Waldparks speziell am Wochenende* (PHT-TOU).
‘Northern Ireland’s population has always loved nature. The people of Ulster are no couch potato! Some dawdle on the ’Loughs’ (lakes or estuaries) around, others spend **their** leisure time with fishing or boating, others make trips with the family in the mountains or in the forest park especially on weekends.’

(3b) *Das kulturelle Erbe Nordirlands besteht größtenteils aus* ***seiner*** *Landschaft. Die Menschen in Ulster halten sich gerne in der freien Natur auf. In* ***ihrer***

Freizeit schlendern ***sie*** *die Küste entlang, und an den Wochenenden unternehmen* ***sie*** *einen Familienausflug in die Berge* (SHT-TOU).
'The cultural heritage of Northern Ireland consists largely of **its** landscape. The people of Ulster like to stay outdoors. In **their** spare time **they** stroll along the coast, and on the weekends **they** take a family trip to the mountains.'

(3c) *Das Erbe von Nordirland ist in großem Maße ländlich. Ulster-Leute sind Leute im Freien.* ***Sie*** *verbringen* ***ihre*** *Freizeit, die um die Küste oder das Gehen auf Familienexpeditionen zu den Bergen an den Wochenenden pottering ist* (RBMT-TOU).
'The heritage of Northern Ireland is largely rural. Ulster people are people in outdoors. **They** spend **their** leisure time, which is pottering around the coast or going on family expeditions to the mountains on weekends.'

(3d) *Das Erbe von Nordirland ist weitgehend ländlich geprägt. Ulster Menschen sind Outdoor-Menschen.* ***Sie*** *verbringen* ***ihre*** *Freizeit werkeln rund um die Küste oder gehen auf Familie Expeditionen zu den Bergen am Wochenende* (GSMT-TOU).
'The heritage of Northern Ireland is characterized largely as rural. Ulster people are outdoor people. **They** spend **their** free time pottering around the coast or go on family expeditions to the mountains on weekends.'

(3e) *Die Erbe von Nordirland ist weitgehend rural. Ulster Menschen sind im Freien people.* ***Sie ihre*** *Freizeit pottering etwa der Küste oder an Familie entstand zu den Bergen auf weekends* (MSMT-TOU).
'The heritage of Northern Ireland is largely rural. Ulster people are people of outdoors. **They their** free time pottering around the coast or family built up to the mountains on weekends.'

SHT-TOU forms its own cluster (nr. 6), which means that it is distinctive from the other translation varieties. However, we are not able to claim that the relation between nominal and pronominal reference is the only indicator of this variation, and we need to test the interplay of further features in this subcorpus to be able to test the differences of SHT-TOU from other translations of tourism texts.

The heterogeneous cluster 3 (containing instruction manuals and popular-scientific texts) is characterised by conjunctive relations and modal verbs, which indicate the argumentative goal. Interestingly, previous studies did not reveal commonalities between these two registers. In Neumann (2013), instruction manuals appear to be very distinct from the other registers under analysis.

SHARE and SPEECH, which compose the fourth cluster, have a number of commonalities. According to Neumann (2013), these two registers seem to be closer in English than in German, which might indicate the influence of the source texts on our translations.

Cluster 5 is the second smallest cluster and contains the human-translated SHARE texts and political speeches translated by novice translators. The figures in Table 6 show that this cluster is characterised by a large amount of both nominal and verbal classes, general nouns and nominalisations, as well as additive and temporal conjunctive relations. These are the same features which also characterise cluster 4, and thus the other translations of the same two registers. However, translations in cluster 5 reveal a greater amount of conjunctive relations and modality meanings than those in 4.

Table 7: Significance of differences between translation methods of SHARE measured with the Chi-squared test

	SHT	RBMT	GSMT	MSMT
PHT	0.03748	4.343e-12	2.541e-11	3.059e-12
SHT		< 2.2e-16	< 2.2e-16	< 2.2e-16
RBMT			0.023	0.2177
GSMT				0.6856

Table 8: Significance of differences between translation methods of SPEECH measured with the Chi-squared test

	SHT	RBMT	GSMT	MSMT
PHT	0.003122	0.01938	0.3268	0.2729
SHT		0.0001679	7.313e-05	3.41e-05
RBMT			0.3378	0.2591
GSMT				0.9885

We calculate p-values (calculated with the Chi-squared test) to see the differences between translation varieties in terms of modal verb distribution. Considering the p-values for modal verbs expressing permission in SHARE and SPEECH (Tables 7 and 8), we clearly see that SHT and PHT differ from machine translation in SHARE (also differing from each other), whereas SHT differs from all other translation varieties in SPEECH (PHT differs from SHT only). It is interesting to note that whereas in human translations of SHARE modal verbs express both possibility and obligation (expressed with *können* and *dürfen*) with possibility/ability prevailing, machine-translated SHARE contains possibility verbs only.

Overall, the features contributing to our cluster formation correspond to the features specific to the registers involved, as described by Neumann (2013) or Steiner (2012), with some exceptions. From this, it follows that the dimension of register has more influence on variation in translation than that of translation method, at least in the dataset analysed here. The features contributing to the generation of the resulting clusters can shed light on the linguistic properties of different translations. At the same time, the observed exceptions, i.e. emerging mixed classes, point to further issues that we need to analyse more closely. This requires an investigation of the features contributing to such mixed class formation. We believe that these features can point to translation problems caused by textual differences between the source and the target language. To test this, we need to include original data (of both source and target language) into our analysis.

Moreover, the observed differences between human and machine translations in register classification show that the linguistic features of translations are influenced by the methods involved in the translation process. These differences result partly from the fact that in the development of machine translation systems, no attention is paid to register settings.

5 Conclusion and future work

The present study analysed the interplay between two dimensions influencing the linguistic settings of translations: register and translation method. We applied unsupervised analysis techniques with the aim of discovering new structures in our translation data. Our assumption was that translations in our dataset would cluster according to either registers the texts belong to or translation methods involved in the translation process, indicating in this way the prominence of one of the two dimensions.

Our results show that both dimensions are present in the groupings of our subcorpora. For some of them, e.g. all translations of fictional and tourism texts, as well as political essays, we clearly see the prevalence of the register dimension. Interestingly, variation along translation method can be detected within register-specific clusters. Moreover, we observe another dimension of variation – that of translation experience. By contrast, translations of the other texts (instruction manuals, popular-scientific texts, political speeches and letters to shareholders) are classified into more fine-grained clusters, for which individual tendencies are observed, influenced by either one of the two dimensions. In this case, smaller clusters of human vs. machine translations, as well as mixed register clusters are observed. The predominance of one or the other dimension is observed on lower nodes. The existence of mixed classes in our data shows that a more detailed analysis of their linguistic properties is needed. The emergence of these classes might indicate the existence of phenomena causing translation problems, e.g. ambiguities. Differences in their resolution result in the independent variation of translations that we observe in our data. This kind of analysis, however, needs to take non-translated texts into consideration, alongside translated ones.

We also think that we need to perform a more detailed analysis of the dimension of translation experience, as the amount of experience of a translator and the amount of data involved in the training of an SMT system seem to have a similar influence on the outcome of the translation process.

Overall, the results of our analysis deepen our knowledge of the linguistic properties of translated texts, which, in turn, furthers our understanding of variation processes in translation. Knowledge of the variation caused by translation methods also provides us with information on method-specific features, which can facilitate their improvement, especially in the area of machine translation. In this way, the results of our analysis can be used for the enhancement of MT systems, as well as the evaluation and quality estimation of both human and machine translation.

In our future work, we plan to include supervised techniques of analysis, which will allow us to detect distinctive features of each translation variety. We believe that a combination of unsupervised and supervised techniques, as elaborated in Diwersy et al. (2014), is needed to detect linguistic properties on a more fine-grained level, as hierarchical clustering allows us to observe tendencies, as well as differences and similarities between certain subcorpora. We also need to include an intra-lingual analysis of the same features in English and German non-translated texts, as we believe that they also have a major impact on translation varieties. We will then be able trace the variation along the additional dimension, that of language, which was not taken into account in the present analysis. In this case, we expect to observe different degrees of shining through and normalisation in the translation varieties under analysis.

References

Baayen, Harald. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Babych, Bogdan & Anthony Hartley. 2004. Modelling legitimate translation variation for automatic evaluation of MT quality. Linguistic Resources and Evaluation Conference (LREC2004), Lisbon, Portugal, 833–836. http://www.lrec-conf.org/proceedings/lrec2004/pdf/707.pdf.

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Gill Francis, Mona Baker & Elena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 233–250. Amsterdam: John Benjamins.

Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies* 7(2). 223–243.

Baroni, Marco & Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21(3). 259–274.

Biber, Douglas. 1995. *Dimensions of register variation. A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press.

Carl, Michael & Matthias Buch-Kromann. 2010. Correlating translation product and translation process data of professional and student translators. Annual Conference of the European Association for Machine Translation (EAMT) http://www.mt-archive.info/EAMT-2010-Carl.pdf.

CWB v3.0. 2010. The IMS Open Corpus Workbench. http://www.cwb.sourceforge.net.

Delaere, Isabelle & Gert De Sutter. 2013. Applying a multidimensional, register-sensitive approach to visualize normalization in translated and non-translated Dutch. *Belgian Journal of Linguistics* 27. 43–60.

De Sutter Gert, Isabelle Delaere & Koen Plevoets. 2012. Lexical lectometry in corpus-based translation studies: Combining profile-based correspondence analysis and logistic regression modeling. In Michael P. Oakes & Ji Meng (eds.), *Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research*, 325–345. Amsterdam: John Benjamins.

Diwersy, Sascha, Stefan Evert & Stella Neumann. 2014. A semi-supervised multivariate approach to the study of language variation. In: Benedikt Szmrecsanyi & Bernhard Wälchli (eds.), *Linguistic variation in text and speech, within and across languages*, 174–204. Berlin: Mouton de Gruyter.

El-Haj, Mahmoud, Paul Rayson & David Hall. 2014. Language independent evaluation of translation style and consistency: Comparing human and machine translations of Camus' novel "The Stranger". In Petr Sojka, Ales Horák, Ivan Kopeček & Karel Pala (eds.), *The 17th International Conference TSD 2014. Lecture Notes in Computer Science*, vol. 8655, XVI. Springer.

Estrella, Paula, Olivier Hamon & Andrei Popescu-Belis. 2007. How much data is needed for reliable MT evaluation? Using bootstrapping to study human and automatic metrics. *Machine Translation Summit XI*, 167–174.

Everitt, Brian S., Sabine Landau, Morven Leese & Daniel Stahl. 2001. *Cluster Analysis* (4th edn.). London: Arnold.

Evert, Stefan. 2005. *The CQP Query Language Tutorial*. IMS: Universität Stuttgart.

Fishel, Mark, Rico Sennrich, Maja Popovic & Ondrej Bojar. 2012. TerrorCAT: a Translation Error categorization-based MT Quality Metric. *The Seventh Workshop on Statistical Machine Translation*, 64–70, Montreal, QC, Canada, Association for Computational Linguistics.

Francis, Gill & Susan Hunston. 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English* (Studies in Corpus Linguistics, vol. 4). Amsterdam/Philadelphia: John Benjamins.

Gellerstam, Martin. 1986. Translationese in Swedish novels translated from English. In Lars Wollin & Hans Lindquist (eds.), *Translation Studies in Scandinavia*, 88–95. CWK Gleerup, Lund.

Göpferich, Susanne & Riitta Jääskeläinen. 2009. Process research into the development of translation competence: Where are we, and where do we need to go. In Susanne Göpferich & Riitta Jääskeläinen (eds.), *Process Research into Translation Competence. Special Issue of Across Languages and Cultures* 10(2). 169–191.

Halliday, M. A. K. & Ruqaiya Hasan. 1989. *Language, context and text: Aspects of language in a social semiotic perspective*. Oxford University Press.

Hansen, Silvia. 2003. *The nature of translated text: An interdisciplinary methodology for the investigation of the specific properties of translations.* Saarland University: German Research Center for Artificial Intelligence doctoral dissertation.

Hansen-Schirra, Silvia, Stella Neumann & Erich Steiner (eds.). 2012. *Cross-linguistic corpora for the study of translations. Insights from the language pair English – German* (Text, Translation, Computational Processing). Berlin/New York: Mouton de Gruyter.

House, Juliane. 1997. *Translation quality assessment. A model revisited.* Tübingen: Günther Narr.

House, Juliane. 2014. *Translation quality assessment. Past and present*. Routledge.

Ilisei, Iustina, Diana Inkpen, Gloria Corpas Pastor & Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, 503–511. Berlin/Heidelberg: Springer.

Koehn, Philipp & Barry Haddow. 2012. Towards effective use of training data in statistical machine translation. *The Seventh Workshop on Statistical Machine Translation (WMT'12)*, 317–321. Stroudsburg, PA, USA: Association for Computational Linguistics.

Koehn, Philipp. 2010. *Statistical machine translation* (1st edn.). New York: Cambridge University Press.

König, Ekkehard & Volker Gast. 2012. *Understanding English – German Contrasts* (3rd edn.), Berlin: Erich Schmidt.

Kurokawa, David, Cyril Goutte & Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. Machine Translation Summit XII. http://www.mt-archive.info/MTS-2009-Kurokawa.pdf.

Lapshinova-Koltunski, Ekaterina. 2013. VARTRA: A comparable corpus for analysis of translation variation. *The Sixth Workshop on Building and Using Comparable Corpora*, 77–86. Sofia: Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-2500.

Lapshinova-Koltunski, Ekaterina & Kerstin Kunz. 2014. Conjunctions across languages, registers and modes: Semi-automatic extraction and annotation. In Ana Diaz Negrillo & Francisco Javier Díaz-Pérez (eds.), *Specialisation and variation in language corpora. Linguistic insights*, 77–104. Peter Lang.

Lapshinova-Koltunski, Ekaterina & Santanu Pal. 2014. Comparability of corpora in human and machine translation. 7th Workshop on Building and Using Comparable Corpora (BUCC). *Building Resources for Machine Translation Research, Linguistic Resources and Evaluation Conference (LREC-2014)*, Reykjavik, 42–48. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-Bucc2014%20Proceedings.pdf.

Lapshinova-Koltunski, Ekaterina & Mihaela Vela. 2015. Measuring 'Registerness' in Human and Machine Translation: A Text Classification Approach. Workshop on *Discourse in Machine Translation* at EMNLP-2015. Association for Computational Linguistics, 122–131. http://aclweb.org/anthology/W15-2517.

Lembersky, Gennadi, Noam Ordan & Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics* 38(4). 799–825.

Matthiessen, Christian. 2001. The environments of translation. In Erich Steiner & Colin Yallop (eds.), *Exploring translation and multilingual text Production: Beyond content*, 41–124. Berlin/New York: Mouton de Gruyter.

McEnery, Tony & Andrew Wilson. 2001. *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.

Murphy, Kevin P. 2012. *Machine learning: A probabilistic perspective*. Cambridge Massachusetts/London: MIT Press.

Neumann, Stella. 2013. *Contrastive register variation. A quantitative approach to the comparison of English and German*. Berlin, Boston: Mouton de Gruyter.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. BLEU: A Method for automatic evaluation of machine translation. *The 40th Annual Meeting on Association for Computational Linguistics*, 311–318.

Popović, Maja. 2011. Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics* 96. 59–68.

Popović, Maja & Aljoscha Burchardt. 2011. From human to automatic error classification for machine translation output. 15th *International Conference of the European Association for Machine Translation (EAMT-2011)*, Leuven, Belgium, European Association for Machine Translation, 265–272. http://www.ccl.kuleuven.be/EAMT2011/proceedings/pdf/eamt2011-proceedings.pdf.

R Core Team. 2013. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna. http://www.R-project.org/.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees *International Conference on New Methods in Language Processing*, 44–49, Manchester, UK. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.

Steiner, Erich. 1997. An extended register analysis as a form of text analysis for translation. In Gerd Wotjak & Heide Schmidt (eds.), *Modelle der Translation – Models of translation*, 235–253. Frankfurt/M.: Vervuert.

Steiner, Erich. 1998. A register-based translation evaluation. *Target. International Journal of Translation Studies* 10(2). 291–318.

Steiner, Erich. 2004. *Translated Texts. Properties, variants, evaluations*. Frankfurt/M.: Peter Lang Verlag.

Steiner, Erich. 2012. A characterization of the resource based on shallow statistics. In Silvia Hansen-Schirra, Stella Neumann & Erich Steiner (eds.), *Cross-linguistic corpora for the study of translation. Insights from the language pair English-German*, 71–90. Berlin: Mouton de Gruyter.

Teich, Elke. 2003. C*ross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. Berlin: Mouton de Gruyter.

Toury, Gideon. 1995. *Descriptive translation studies and beyond*. Amsterdam/Philadelphia: John Benjamins.

Volansky, Vered, Noam Ordan & Shuly Wintner. 2011. More human or more translated? Original texts vs. human and machine Translations. *11th Bar-Ilan Symposium on the Foundations of Artificial Intelligence (BISFAI)*. https://www.cs.bgu.ac.il/~adlerm/iscol11/bisfai11_submission_38.pdf.

White, John S. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and further approaches. *The 1994 Conference of the Association for Machine Translation in the Americas*, 193–205.

Zampieri, Marcos & Ekaterina Lapshinova-Koltunski. 2015. Investigating Genre and Method Variation in Translation Using Text Classification. In P. Král & V. Matoušek (eds.), *Text, Speech, and Dialogue: 18th International Conference (TSD-2015)*, Lecture Notes in Computer Science, vol 9302, 41–40, Springer, Cham. http://link.springer.com/content/pdf/10.1007%2F978-3-319-24033-6_5.pdf.

Bert Cappelle and Rudy Loock

8 Typological differences shining through. The case of phrasal verbs in translated English

Abstract: Are phrasal verbs less numerous in English translations if the source language is a Romance language than if the source language is a Germanic one? This chapter sets out to answer that question. In a subcorpus of English fictional texts translated from Romance languages, *up*, *out* and *down*, which represent phrasal verb use rather well, are indeed underused when compared with non-translated English fiction from the British National Corpus, while no significant difference is to be found for this set of items between non-translated English and English translated from Germanic languages. This finding is strong evidence for source-language interference, as Romance languages on the whole do not have close equivalents to phrasal verbs, while Germanic languages do. This effect appears stronger than any source-language-independent translation universal that could in principle have played a role, such as normalization (exaggerated use of phrasal verbs, which are typical of the English language) or levelling-out (avoidance of phrasal verbs, which are generally felt to be rather colloquial). A comparison of French prefixed verbs with morphologically simplex ones in *Le Petit Prince* further shows that the former are more likely to be translated by phrasal verbs than the latter, again supporting source-language influence, as phrasal verbs resemble prefixed verbs in being composed of a verb and an added element. Our study thus stresses the relevance of taking into account typological differences (and similarities) between source and target language in translation studies.

1 Introduction

1.1 Phrasal verbs

English would be a rather different language without combinations such as *give up*, *cool down*, *run away*, *throw up*, *figure out*, *rub off*, and so on. Phrasal verbs, also known as verb-particle combinations, are among the most frequent constructions in English (Biber et al. 1999, Gardner and Davies 2007), which is why there are countless books, websites and apps on the English language learning

DOI 10.1515/9783110459586-009

market that encourage learners to develop their ‘phrasal verb skills’. In a recent monograph on the diachronic tracing of the construction back to Germanic and Proto-Indo-European preverbs, Thim (2012: 244) states that “throughout the history of English, phrasal verbs have always had a place in the ‘common core’ of the language”.

Not only are they undeniably central to the English language but many of them are also clearly situated towards the colloquial end of the formality spectrum: consider for example *bum around*, *freak out*, *piss off*, not to mention combinations that might be rather unfit for print. On the basis of corpus-evidence, Biber et al. (1999: 409) conclude that “[o]verall, conversation and fiction show much greater use of the most frequent phrasal verbs than news and academic prose.” One can easily confirm Biber et al.’s findings based on the Longman Spoken and Written English Corpus with data from the British National Corpus (Davies 2004-), which reveal that particles in conversation are more than ten times as frequent as in medicine-related academic texts and still more than five times as frequent as in administrative writing. Marks (2005: s.p.) sums up “widespread popular wisdom about phrasal verbs among learners and teachers”, namely “that they are colloquial, casual, informal, characteristic of speech rather than writing and perhaps even a bit sloppy or slovenly, uneducated and not quite proper” [bullets from the original article removed – B.C. and R.L.].[1] Such characterizations are not new. They are echoes from style commentators in the eighteenth century. For example, the famous lexicographer Samuel Johnson, in his *Dictionary of the English Language* (1755), already provided labels such as “improper”, “low”, “common”, “barbarous”, “familiar”, “vulgar” or “less elegant” to some of the phrasal verbs he attested (cf. Wild 2010: 207).

1.2 Intra-language differences, translation universals and source-language interference

The two properties of phrasal verbs as a class – their high frequency and their informality – accords to them the status of ideal test object to evaluate the validity of certain alleged ‘translation universals’ (for introductions of which, see Baker 1993, Halverson 2003, Mauranen and Kujamäki 2004), which generally come in

[1] It should be noted that the informality of phrasal verbs is a tendency. As Marks (2005: s.p.) points out, “individual phrasal verbs can have distributions that go against the grain of this generalisation. For example, *carry out* is equally common in newspapers and academic writing, but rare in conversation and fiction, and *point out* is more common in academic writing than in the other three genres.”

contradiction with the idea that the source language interferes with the linguistic characteristics of target texts. Indeed, since the advent of corpus-based translation studies in the 1990s, researchers have agreed on the existence of the third code, that is, differences between original and translated language (intra-language differences) but have been divided over the way these should be accounted for. While some argue in favor of the existence of translation universals, those “features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems” (Baker 1993: 243), others consider intra-language differences as resulting from source-language interference. In the literature, most researchers seem to support the translation universals hypothesis (e.g. Baker and Olohan 2000, Olohan 2003, Jiménez-Crespo 2011, Laviosa 1998, Øverås 1998) but other researchers have claimed that source-language interference, also known as ‘shining through’ (Teich 2003), must play (at least) a role (e.g. Koppel and Ordan 2011, Cappelle and Loock 2013). Of course, one might wonder whether interference is not itself a kind of translation universal, albeit one of a very different nature from that of all the others.[2]

Although initial studies provided evidence for the influence of translation universals, more recent studies have clearly questioned their role in the differences that can be observed between original and translated language (see e.g. Becher 2010, Lind 2007 or Kruger and Van Rooy 2012). Their very existence is controversial and many suggested universals have been severely criticized these last ten years (e.g. Becher 2010, House 2008, Corpas Pastor et al. 2008, Mauranen and Kujamäki 2004), with the term ‘universal’ itself being questioned (see Lind 2007). Among the criticism against translation universals is the fact that source languages are not taken into consideration and that many of the studies based on intra-language differences to prove their existence focus on only one genre, very often literary texts (see e.g. Becher 2010 for a criticism of corpus studies based on the Translational English Corpus) and with (British) English as either source or target language.

Recent studies, based on genre-controlled corpora (see Lefer and Vogeleer 2013 for a special issue on this question for the normalization universal), do show that results are genre-sensitive, and that so-called translation universals might not be so universal. For instance, Delaere et al. (2012) have investigated

[2] Source-language interference *has* been listed as a potential translation universal (cf. Toury’s 1995 “law of interference”), but translation universals *à la Baker* clearly exclude it (see definition above) and most studies on translation universals do not mention interference as one of them (see Mauranen 2004) for a discussion on the unclear status of interference in relation to translation universals).

the use of standard and non-standard Belgian Dutch in translated vs. original texts for several registers (fiction, non-fiction, press texts, administrative texts, external communication) and have shown that intra-language differences are "text type dependent, as some of the results (partially) confirm the general trend of translations being more standardized than non-translation (fiction, external communication, administrative texts) but other results do not (journalistic texts and non-fiction)" (Delaere et al. 2012: 220). Also, Delaere and De Sutter (2013) have investigated whether the normalization universal (which they relate to "risk-aversion behavior") is source-language- and/or register-dependent: their results show that the tendency of translators to normalize Dutch translated texts is not only the result of the translation process. They show that several other factors interact: register, source language, and target audience. For instance, their results show that translators resort more to Standard Dutch in journalistic texts (register effect), but in administrative texts, it is the source language (English vs. French) that has an impact. With a special emphasis on another geographical variety of English (South African English) and using a multi-register corpus, Kruger and Van Rooy (2012) have shown that when tested together and not in isolation, linguistic features associated with translation universals fail to discriminate between original and translated texts. They investigated (i) explicitation, associated with the frequency of optional complementizer *that*, contractions and linking adverbials; (ii) normalization, associated with the frequencies of lexical bundles, coinages and loanwords; and (iii) simplification, associated with lexical diversity and mean word length. What their results, based on a comparable corpus of original and translated South African English, show is that with the exception of two of them (optional *that* and lexical diversity), the linguistic features that were investigated do not show significant differences between original and translated texts. More importantly, their study shows that cross-register differences are also present in their corpus of translated English, suggesting that register variation *is* present in translated language, *contra* what the leveling out universal might suggest.

As far as the influence of the source language is concerned, some studies have shown that the linguistic characteristics of translated texts differ depending on the source language. For instance, Cappelle (2012) shows that fewer manner-of-motion verbs are to be found in English translated from French than in English translated from German, a result that is interpreted as resulting from the typological difference between the two source languages, that is, the fact that French, like most Romance languages, belongs to the typological group of verb-framed languages, while English and German, like the other Germanic languages, are satellite-framed languages (see below). Among the other studies that have argued in favor of source-language interference instead of translation universals to account for observed differences between translated and original language is Dai and Xiao (2011), who have shown that Chinese translated from English contains many more occurrences of the passive voice than original Chinese. Given the fact that the passive voice is much more frequent in translated than in original Chinese, Dai and Xiao conclude, through the observation of their corpus, that more than 8 out of 10 occurrences of the passive voice in translated Chinese result from the interference of the source language. In a similar way, Cappelle and Loock (2013) have shown that English translated from French contains fewer occurrences of existential *there*-constructions (e.g. *there is a dog in the garden*) than original English while French translated from English contains more occurrences of existential *il y a*-constructions (e.g. *il y a un chien dans le jardin*), although the two constructions are translationally equivalent. Once again, in the light of cross-linguistic results obtained for original English and original French (existential constructions are much more frequent in original English), Cappelle and Loock suggest a strong case of interference, as results for translated English and translated French point to opposite directions.

1.3 Aim of the chapter

Our chapter aims to adopt the same kind of approach as the one to be found in Cappelle (2012), for phrasal verbs this time, so as to check whether their frequency in English translated texts differs depending on the source language, specifically the source language family (Romance vs. Germanic), or whether results are homogeneous, paving the way for an interpretation related to the influence of translation universals. Depending on whether one selects, say, *normalization* (“the tendency to exaggerate the features of the target language and to conform to its typical patterns” (Baker 1996: 183)) or, rather, *levelling out* (“the tendency of translated text to gravitate towards the centre of a continuum” (Baker 1996: 184)) as one’s translation universal of choice, the frequency of occurrence of phrasal verbs in translated (i.e., non-original) English may be hypothesized to be either enhanced or reduced. Indeed, on the one hand, normalization manifests itself in a boost of common target language features, and one could therefore predict translated texts to display an over-use of phrasal verbs, which are very common items in English. On the other hand, levelling out is the tendency of translated texts to share similar characteristics, thus eradicating any register/genre-related differences, and since phrasal verbs are generally considered to belong to more informal registers, we should expect fewer phrasal verbs in translated texts. That is to say, if a translator aims to make a text sound like a genuinely English one, s/he might use many phrasal verbs, leading to an over-representation of phrasal verbs in translated English compared to original

English – the effect, as Lind (2007) puts it, that "translated texts, like converts, are more normal than normal". Alternatively, if a translator is wary of using expressions that are felt to be too colloquial, s/he might use more neutral, morphologically simple verbs over the generally more informal phrasal verbs, leading to an underrepresentation of the latter in translated English compared to original English. We perhaps make it appear here as if the translator makes these decisions consciously, but this need not be the case. Translators are not necessarily aware of selecting options that could be analysed as particularly target-like, including items that sound informal or even slangy or, conversely, of producing a lexically more 'cautious' text, whose edges have been smoothed, so to speak.

We strongly believe that contrastive issues should be taken into account in comparing translated and non-translated variants of a single language (Cappelle 2012, Cappelle and Loock 2013, Loock, De Sutter and Plevoets 2013). In the present chapter, therefore, we distance ourselves from hypothesized universal laws of translational behaviour (whether consciously applied or not) which do not take the relation between the source language and the target language into consideration.[3] To the extent that normalization and levelling out can be formulated, respectively, as "make the translation sound like a typical, authentic text in the target language" and "use lexical and grammatical items that fall within the conventional core of the target language's lexico-grammar", they are what Chesterman (2004) calls T[arget]-universals, which are concerned with the comparison of translations in a language to other texts in the same language. In other words, they do not make reference to differences between the target language and the source language. Yet, if we want to find out whether translated English uses more or fewer phrasal verbs than non-translated English, we should not ignore this important question: How readily does a phrasal verb present itself as a translation *of what is in the source text*?

[3] We leave aside another criticism that we could level at such translation universals, namely that some of them come in pairs that are mutually incompatible. In the case of normalization and levelling out, these two reasonable-sounding candidates for universality in translation behaviour make contradictory predictions about the frequency of phrasal verbs. For another example, apart from the tendency of *explicitation*, translated language has also been claimed to display *implicitation* (Klaudy and Károly 2005). Needless to say, both these hypothesized generalizations cannot hold universally, at least not if *universally* is taken to mean what it is supposed to mean, namely 'for all translators, for all language pairs, in all texts, in every place'. The problem in weakening the universality constraint and allowing these proposed universals to be mere tendencies, as is customary in translation studies, is that these universals lose their predictive potential and effectively become unfalsifiable, thus having little or no scientific value. See Chesterman (1997) for discussion of these concerns.

We will now refine this question against the background of Cappelle's (2012) study, which linked the underuse of a closely related target-language feature in English, namely manner-of-motion verbs, to a typological difference between the source language and the target language (section 2). We will then present two corpus-based studies, one in which translated English from two typologically different source language families (Romance vs. Germanic) is compared to non-translated English (section 3) and another in which we look at which kinds of verbs in a French text are likely to tempt the translator to use a phrasal verb (section 4). In our conclusion, we will sum up the main findings and present some methodological reflections (section 5).

2 Framing typology and *translationese*

In Cappelle (2012), it was shown that translated English contains fewer manner-of-motion verbs when the source language is French than when the source language is German. The most likely reason for this, it was suggested, is that French, like most Romance languages, belongs to the typological group of verb-framed languages, while English and German, like the other Germanic languages, are satellite-framed languages (Talmy 2000). This typological difference pertains to whether the most central semantic aspect in an event of change of position or state, among other kinds of events, is preferentially expressed by the verb root or by a sister to it – a prefix or a particle, for instance. For the expression of a motion event, the core semantic element is the so-called path, which refers to direction, source or goal. Thus, in the English sentence *A UFO whizzed by*, it is the particle *by* that encodes the path, just like the separable verb prefix *vorbei* in the German equivalent sentence *Ein UFO sauste vorbei*. Since the verb root itself does not have to express this aspect of meaning, it is 'freed up' to encode a more secondary semantic aspect: in this case, the high speed of motion and the accompanying sound this produces. In French, by contrast, as in Romance languages more generally, the direction of motion is typically expressed by the main verb, as in *Un OVNI passa* ('A UFO passed', 'A UFO went by'). The verb, having been assigned with the task of encoding this aspect of meaning, can then no longer express manner of motion, which would then be expressed, if at all, in a constituent functioning as adjunct, such as *à grande vitesse* 'with high speed' or *dans un sifflement* 'with a whistling sound'.

What emerged from that corpus investigation is that the difference in overall framing preferences appears to leave its trace in a corpus of English translations from French. Specifically, this corpus exhibited a higher proportion of path verbs (*leave*, *rise*, etc.) to manner-of-motion verbs (*crawl*, *leap*, etc.) than both a reference corpus of non-translated English texts and a corpus of English translations from German. Though this difference may not be noticeable in a single individual text, the cumulative evidence appearing from corpora of translated and non-translated English cannot be ignored: the underuse of manner-of-motion verbs in translations from French is an example of source-language interference.

In what follows, we will attempt to answer the following questions:

(1) a. Do English translations from Romance languages contain as many phrasal verbs as texts originally written in English?

b. Do English translations from other Germanic languages contain as many phrasal verbs as texts originally written in English?

In light of the typological differences between Germanic and Romance languages briefly reviewed above, we predict that the first question will be answered negatively and the second positively. Such findings from translation studies would complement what we know from second language acquisition studies, where it appears that phrasal verbs are underproduced by undergraduate learners whose L1 lacks a similar category (e.g. Hebrew, Chinese) but not by undergraduate learners whose L1 has a close equivalent (e.g. Dutch, Swedish) (Laufer and Eliasson 1993, Liao and Fukuya 2004). In addition, if our expectation for the first sub-question (1a) is borne out, we will address the following question:

(2) Assuming that English translations from Romance languages are not completely devoid of any phrasal verbs, which source language expressions are they the translations of?

For this second question, which we will answer with reference to French as a source language, we hypothesize that morphologically complex verbs in the source text (e.g. *re-venir* ‘go back’) are more likely to be translated by a phrasal verb than simplex verbs (e.g. sortir ‘leave’, ‘go out’). Such a result would again be in line with our assumption that what is in the source text may be ‘shining through’ in the translation, to use Teich’s (2003) concept, as the presence of a bound derivational morpheme in the French source text could be seen as a trigger for a translation into a particle to form a phrasal verb in the English translation. For English as a source language, previous studies demonstrated such a structure-preserving effect. In a study on the translation of English phrasal verbs into German, Claridge (2002) found that translators typically used structures resembling the source forms, translating the verb by a literal equivalent and translating the particle by a separable or inseparable prefix. In English translations into Russian, too, phrasal verbs are often translated by prefixed verbs (Mudraya et al. 2005). German and Russian are typologically similar to

English in that they make extensive use of particles or prefixes. In other words, they have a frequently used equivalent to phrasal verbs. While French is not considered to be a satellite-framed language, it does contain prefixed verbs and these may account to a large extent for the use of phrasal verbs in the translation. It has been shown in previous research (Paillard and Videau 2008) that the most frequent translation strategy for French verbs starting with *dé-* or *de-* is a phrasal verb in English. Such a finding lends support to the correctness of our hypothesis, but to have full certainty, we should check whether prefixed verbs in French more often lead to a phrasal verb in the target text than non-prefixed ones do.

Throughout this chapter, we assume that phrasal verbs are, by and large, satellite-framed structures. This assumption is justified insofar as particles frequently express the path in a motion event (e.g. *walk in*) or a more metaphorical or abstract change of state in other events (e.g. *cool down*, *wake up*). Even in aspectual combinations like *play along*, the particle encodes an aspect of meaning which in verb-framed languages would more typically be expressed in the main verb, for instance as something close to *accompany someone playing*. Only in highly idiomatic combinations (e.g. *make out* ‘kiss in a sexual way’) is it less clear that the particle expresses the core semantic part of a complex event. Even so, it can hardly be denied that phrasal verbs are compatible with, and indicative of, a language’s satellite-framed typological nature.

3 Phrasal verbs in translated English from typologically different languages: A large-scale quantitative corpus-based approach

3.1 Method

3.1.1 Overall design

The methodology adopted here involves three corpora: (i) a reference corpus of English texts, (ii) a comparable corpus of English texts translated from source languages A1, A2, A3, etc. and (iii) another comparable corpus of English texts translated from source languages B1, B2, B3, etc. Here, as opposed to Cappelle (2012), we do not focus on the proportion of manner-of-motion verbs to path verbs but on the frequency of occurrence of particles relative to corpus size. Another difference is that we here extend the comparable corpora considerably:

we do not just look at translations into English from French and German but at translations into English from Romance languages and Germanic languages.

3.1.2 Corpora

Our corpus study rests on an analysis of fictional texts in original and translated English. The reason why we have restricted our analysis to fictional texts is (i) their availability in different corpora (as opposed to press or technical texts for instance) and (ii) the fact that phrasal verbs are frequent in this register (see above). Our reference corpus was the 100 million word British National Corpus (BNC), which we searched via the Brigham Young University web interface (Davies 2004-). We restricted our searches to the subcorpus of fictional texts, whose total size is 15,909,312 words. This component consists of 476 texts, whose average length is 34,656 words (information obtained via http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides). For translations into English we used the Translational English Corpus (TEC), available at http://www.llc.manchester.ac.uk/ctis/research/english-corpus/. From this corpus, we selected two subcorpora with English fictional texts published between 1980 and 1993, which is also the time range from which the texts in the BNC date. For one of the TEC subcorpora, we selected as source languages all available Romance languages (Brazilian Portuguese, Catalan, European Spanish, French, Italian and Latin American Spanish). This subcorpus, which we can refer to as TEC < Rom, has a total number of 1,952,690 words and consists of 32 texts, with an average length of 111,669 words (shortest: 25,915 words; longest: 197,422 words). For the other subcorpus, we aimed (somewhat overoptimistically, as we will point out shortly) at covering all available Germanic languages (Danish, Dutch, German, Icelandic, Norwegian and Swedish) in the entire TEC corpus. This subcorpus (TEC < Ger) totals 1,146,785 words and comprises 14 texts, whose average length is 81,913 words (shortest: 14,288 words; longest: 166,973 words). Because of our preselection of fiction from a specific period and because of the make-up of the TEC corpus itself, our subcorpora are not balanced for individual languages, though. TEC < Rom is made up of mainly French (61%), with Brazilian Portuguese (12%), European Spanish (11%), European Portuguese (6%), Italian (4%), Latin American Spanish (3%) and Catalan (3%) only being minor parts of this corpus. TEC < Ger has an even less equal spread over individual languages, as it predominantly consists of German texts (90%), complemented by only one other Germanic language, namely Swedish (10%). Because of this scarcity of data for some Romance and especially Germanic languages, we will not report findings for individual languages. Future research should take a closer look at possible differences among the languages within the two families studied here.

3.1.3 Search items

As it would be quite unfeasible to search and check all verb-particle combinations in these corpora by hand, we carefully selected search items that could serve as proxy for the entire class of phrasal verbs. Since verbs form an open lexical class and particles a closed one, we restricted our searches to the latter. However, especially as the TEC is not tagged, we had to be careful to choose words that do indeed often occur as particles, in order to maximize precision (i.e., the fraction of retrieved items that are relevant), while also making sure that we used items that are representative of the class of particles as a whole, in order to maximize recall (i.e., the fraction of relevant items that are retrieved). The choice of suitable items was important, since many words that can be used as particles occur in fact more frequently as prepositions – as is the case for *in* and *on*, for instance. For our selection of search items, we relied on data in Gardner and Davies (2007: 346), according to which the words *out* and *up*, of all particle candidates, are not only used most frequently as particles (in 97.3% and 87.4% of their total occurrences, respectively) but are also the two most frequently used particles, together accounting for almost half (46%) of all particle occurrences in the BNC. In short, these two words have high precision and reasonably good recall values as representatives of the class of verb particles. Given the unclear grammatical status of *out of* (see Cappelle 2001), we removed this combination from our selection. In order to have a more representative sample of particles, we decided to add one more word to our set of search items, namely *down*. It appears that *down* is another frequent particle, with 79.2% of all occurrences of this item being tagged as a particle, as reported by Gardner and Davies (2007). This is still a reasonably good precision value, and with *down* added to *up* and *out*, we thus obtained a small set of items which jointly form a representative subset of the class of particles, making up 57.3% of all particle occurrences, to be precise (still according to Gardner and Davies 2007). So, the three search items, *up*, *out* (minus *out of*) and *down* are all predominantly used as particles and together represent well over half of the occurrences of all particles.

3.1.4 Statistical analysis

Results of the searches of the three items in the three different corpora were analysed statistically using the Log-likelihood test, a test which is similar to the Chi-square test but is not subject to certain assumptions about how the data are distributed (cf. McEnery, Xiao and Tono 2006: 55). We provide the results for the three individual particles, but as we are not interested in how the relative frequencies of individual particles may differ across corpora, we summed over the occurrences of all three test items in each of the three corpora. The results obtained in the BNC were compared with those obtained in the corpus of translations from Romance languages and separately with those obtained in the other translational corpus.

3.2 Results

In the BNC fictional subcorpus, the three search words yielded 124,529 occurrences, or 7827 occurrences per million words. In the corpus of translations into English from Romance languages (TEC < Rom), we retrieved 11,880 occurrences of these words, or only 6084 per million words. By contrast, the corpus of translations into English from (other) Germanic languages yielded 8951 occurrences of the three search items, which amounts to 7805 per million words, virtually the same frequency as in the BNC reference corpus. The details of the findings are given in Table 1 and graphically represented in Figure 1.

Table 1: Number of occurrences of *up*, *out* and *down* (raw and normalized per million words (pmw)) in the fiction component of the BNC and in two comparable corpora of translated texts into English from Romance and Germanic languages

	BNC		**TEC < Rom**		**TEC < Ger**	
	raw	**pmw**	**raw**	**pmw**	**raw**	**pmw**
up	54850	3448	5595	2865	4169	3635
out	38507	2420	3747	1919	2753	2401
down	31172	1959	2538	1300	2029	1769
total	124,529	7827	11,880	6084	8951	7805

The conspicuous underrepresentation of the three search items in the corpus of English fiction translated from Romance languages, compared to their frequency in the BNC fiction component, is statistically extremely significant (Log-likelihood = 738.32; p < .0001). Given the large sample sizes, the small difference between the BNC results and the results from TEC < Ger were in danger of having proved statistically significant as well, but (summarized over the three particles) they are not (Log-likelihood = 0.067; p = 0.8).

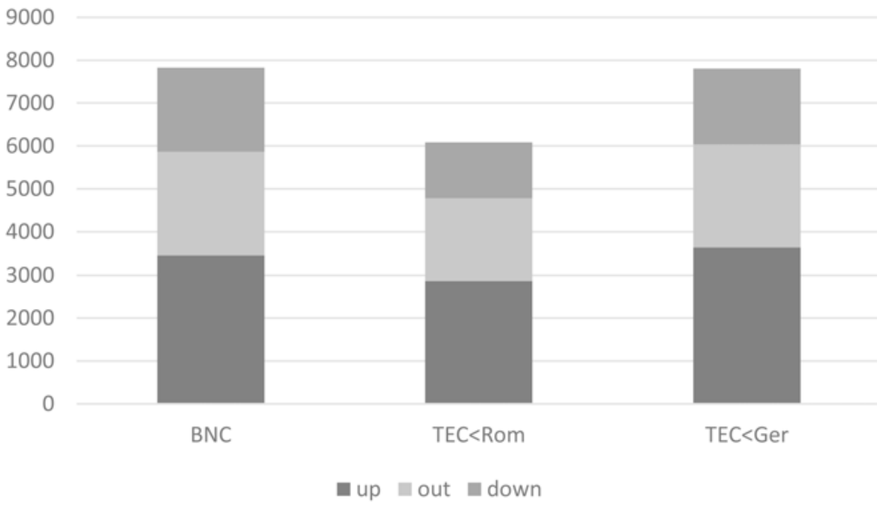


Figure 1: Number of occurrences of *up*, *out* and *down* per million words in the fiction component of the BNC and in two comparable corpora of translated texts into English from Romance and Germanic languages

3.3 Interim discussion

The present findings strongly suggest that an account of the differences between original and translated English which is based on translation universals, in particular normalization or levelling-out, should be treated with utmost circumspection. The results clearly differ depending on the language family to which the source language belongs. Although we have not studied the influence of each individual source language on the linguistic characteristics of the English translated texts, it is possible to consider 'source language family interference' as a significant effect in translation, *contra* any translation universal that is claimed to hold irrespective of the source and target language involved. Of course, one might argue that it is only to be expected that texts from the same genetic and typological family as English lead to translations that are closer to non-translated English than texts from genetically more distant and typologically rather different languages do. We fully agree that our results are in line with what was to be expected. Obviously, a finer-grained approach involving a comparison with each individual source language would be desirable, especially as not all languages within a single family may be equally strong representatives of a typological profile (see Iacobini and Masini 2003) for the occurrence of phrasal verbs in Italian, in particular). This is beyond the scope of this chapter, which used the TEC to investigate translated English, as the corpus does not contain samples that are large enough for each individual source language.

Still, we would like to stress that our results allow us to cast very serious doubts on the validity of each of the two translation universals discussed in section 1.3. Source-language (family) interference has a stronger influence than these and any other translation universals that could be considered, since our results are source-language-family dependent. Quite clearly, translations into English from Romance languages fail to be fully normalized with respect to the presence of phrasal verbs. While such normalization, if that is what it is, seems to be successful for translations from Germanic languages, it falls short for translations from Romance languages. In other words, it does not obtain across-the-board and, by consequence, it would seem wrong to consider normalization a translation *universal*.

It could be objected that the sheer fact that phrasal verbs are used in translations from languages that do not use them (but see section 4 for an important qualification of this 'absence' in the source language) can only be analyzed as the result of a normalization tendency. Such an objection can easily be countered. While it cannot be denied that the presence of phrasal verbs in English translated from Romance languages could be interpreted as translators' tendency to 'normalize' the target language by using a common feature of it, this does not lead to the translation sounding "more normal than normal", to use a formulation we quoted earlier. In other words, what we dismiss here is normalization as a T(arget language)-universal, that is, a source-language-independent tendency manifesting itself as a stable difference between translations in the target language and originally produced texts in that target language (Baker 1996). It would be very strange, of course, if an English translation from a Romance language contained no, or hardly any, phrasal verbs: after all, the job of the translator is to produce an English text, and phrasal verbs are an integral part of that language's lexicon. The point, though, is that normalization, as a T-universal, predicts that there will be more of these very normal lexical items in translations into English than in original English. This prediction is clearly not borne out. Note also that even with Germanic languages as source languages, there is no overrepresentation of phrasal verbs, which we should have found if normalization leads to exaggeration of target-language features.

Likewise, we can quite safely dismiss levelling-out as a translation universal (cf. again section 1.3). If we only considered translations into English from Romance languages, this tendency could have served as a suitable explanation for the underrepresentation of phrasal verbs, but we are then left to explain why in translations from other languages, this presumed levelling out does not obtain. So, levelling out, too, loses its appeal as a source-language-independent translation universal. Since our results are source-language-family dependent, *any* such translation universal that could be at play is undermined by interference from the source language.

There just might be a way of salvaging normalization and levelling-out. While, as we saw above, they lead to opposite predictions in the case of phrasal verbs (overrepresentation and underrepresentation, respectively), we could say that, when they are both at work (as they should be, given that they are universals), they cancel each-other out. That is, using phrasal verbs in an over-indulgent way (i.e., normalization/exaggeration) may be tempered by the simultaneously operative avoidance of lexico-grammatical structures, including phrasal verbs, that have a too informal ring about them (i.e., levelling-out), and vice versa. The results that we obtained above (under-representation in translations from Romance languages; no noteworthy difference in translations from Germanic languages) may be compatible with such neutralisation. They could then be considered, crucially, as the added effect of just one more ‘universal’ thrown in, namely Toury’s (1995: 275) “law of interference”, or Tirkkonen-Condit’s (2004) “Unique Items Hypothesis”, according to which linguistic features that are typical of, or even unique to, the target language, when compared to the source language, have a tendency to be under-represented in translations. Even if there is no neutralization at work of any other universals, if they exist at all, the law of interference or the Unique Items Hypothesis can definitely *not* be rejected, in the light of our findings. The law of interference covers “phenomena pertaining to the make-up of the source text [that] tend to be transferred to the target text” (Toury 1995: 275). Toury’s definition could – with the help of some mental gymnastics, admittedly – be applied to cases where the phenomena (here: phrasal verbs) are actually absent from the source text. Indeed, one could argue that the absence of phrasal verbs is part of the make-up of source texts written in (most) Romance languages. Because of this absence in the source language, a more useful concept here is “Unique Item” in the target language. After all, particles, with the exception of Italian, are unique to English as a target language when the source language belongs to the Romance family. If we now adhere to the strictly scientific principle of not multiplying hypotheses beyond those that are needed to explain the observable facts, we can let Occam’s razor do its work and simply dismiss normalization and levelling-out – prime examples of translation universals – as having a major impact on the occurrence of phrasal verbs in translated English.

4 Phrasal verbs in translated English from French: Comparing *Le Petit Prince* and its English translation

In a follow-up study, we tried to answer the following question: If there is a phrasal verb in a text translated into English from a Romance language, what corresponds to this item in the source text? The motivation for asking this question is that Romance languages, with the exception of Italian, do not have any phrasal verbs. If such structures do end up in the translation, this could be evidence that there still is some degree of normalization at work, albeit not to the extent that target-language features are exaggerated in the translation, as the data reported above clearly show.

Although French does not have anything that could be called ‘phrasal verbs’, it does make use of morphologically complex verbs (e.g. *re-venir* ‘come back’, *sur-voler* ‘fly over’) which, just like phrasal verbs in English, belong to the set of satellite-framing structures (Kopecka 2006, Pourcel and Kopecka 2006), as these prefixed verbs can be decomposed into an element referring to a path (the prefix) and an element referring to (manner of) motion as such. In other words, such verbs have two semantic components, in much the same way that English motional phrasal verbs do. The hypothesis we can therefore formulate is that such complex morphological items in the source text are more likely to act as a trigger for the use of a phrasal verb in the target text than morphologically simplex items in the source text.

4.1 Method

4.1.1 Overall design

For this part of our study, we needed to take a look at what is actually there in the source text, something we did not do in the corpus study reported on in section 3, where we merely knew that the source text belonged to either a Romance language or a Germanic language. We opted to use a single relatively short text in French and its English translation, the idea being that this research leans rather more to qualitative research, possibly at the expense of full representativeness. The verbs used in the source text were classified as either morphologically simplex (non-derived) or morphologically complex (containing a prefix). We identified the phrasal verbs in the translation with the aim of finding out whether, as we expect, there are more of them whose source expression is a morphologically complex verb than can be expected on the basis of the overall frequency of complex verbs, relative to simplex verbs, in the entire source text.

Thus, suppose there are 900 simplex verbs and 100 complex verbs in the source text and suppose, furthermore that there are 100 phrasal verbs in the translation, then in the still imaginary case that all of these 100 phrasal verbs had as their source expression the 100 complex source verbs, there would be an extreme case of source-language interference. By contrast, if 90 of the 100 phrasal verbs had a simplex verb as their source expression and the other 10 had a complex verb as their source expression, there would be no association whatsoever between the morphological complexity of source verbs and their translations' membership to the class of phrasal verbs.

4.1.2 Corpus

Within a pilot study, the text we took a closer look at is the well-known children's book *Le Petit Prince* by Antoine de Saint-Exupéry and its English translation *The Little Prince* by Katherine Woods, both of which appeared in 1943. The reasons why we selected this text and its translation, which we realize are not very modern ones, are that (i) the original text is widely available and has been translated into many languages, such that our results for the English translation could be compared with those for other target languages and (ii) this is a children's book, for which one can expect there to be a large number of dynamic scenes, including motion, rather than merely reflective passages. Easily obtainable electronic versions of these texts were sentence-aligned automatically with the help of PlusTools, a Windows Word plugin. The aligned output was then corrected manually. The original text is 14,952 words long and its translation 17,066 words long.

4.1.3 Coding

A master's student from the University of Lille 3, whose mother language is French, was instructed to classify all lexical verbs in the source text as either morphologically simple (disregarding inflectional endings) or morphologically complex (specifically, having a prefix). Deciding whether a verb is morphologically complex or not is not always easy. The student was asked not only to consult an etymological dictionary (namely, the online database Ortolang, http://www.cnrtl.fr/etymologie/) but also to use her best judgement in classifying a verb as multi- or monomorphemic and be consistent in her choices. For instance, the verb *comprendre* ‘understand’ was morphologically complex in Latin but is no longer felt to be so for contemporary speakers of French and therefore was classified as simplex. By contrast, verbs like *a-néantir* ‘destroy’ *dé-ranger* ‘disturb’, *r-assurer* ‘reassure’ were classified as complex. The analysability of the verbs for present-day speakers of French was used as a criterion to determine the simplex or complex nature of the verbs. A further instruction was that extremely light verbs, such as linking verbs and auxiliaries (e.g. *être* ‘to be’, *pouvoir* ‘can’), should be removed from the analysis.

Semi-automatic identification of phrasal verbs in the target text was carried out by ourselves. We did not do this fully automatically, because we did not have a tagged version of the text and we wanted to make sure that we only retrieved occurrences of *on*, *off*, *through*, etc. in their use as particles, not as prepositions. Note that the particles *up*, *down* and *out* taken together have 93 occurrences in the translation of *Le Petit Prince*; 93/17,066 equals a normalized frequency of 5499 per million words. This compares rather well to the normalized frequency of these lexical items grouped together in TEC<Rom (which, as we have seen, is 6084 per million words) and confirms the finding from the preceding section that English translations from Romance languages underuse phrasal verbs. We also determined the category of these phrasal verbs’ source expressions (simplex or complex). We realize that the analysability of prefixed lexemes in French is an extremely complex issue, and that there is no such thing as a homogeneous French audience; therefore, the MA student’s and our assessment of prefixed verbs as being synchronically polymorphemic or not may not generalize perfectly to all speakers of French. According to standard morphological analysis, word parts should be endowed with meaning in order to be identified as morphemes in the language. In the present analysis, this was thought to be the case for *é-* in *s’écrier* ‘cry out’, for instance, but it is not clear that all speakers of French perceive this lexeme’s complexity. In Appendix A, we list all contexts in the source text containing French verbs which were considered to be prefixed and which have a phrasal verb in the English translation. In Appendix B, we list all further verb forms in the source text that were considered to be prefixed, but whose translation was not a phrasal verb.

4.1.4 Statistical analysis

All the simplex and complex source verbs were counted by the same master’s student who did the coding, while we counted the phrasal verbs which had a simplex verb and those which had a complex verb as a source expression. In the English translation of *Le Petit Prince*, there are 149 phrasal verbs, which were classified depending on their source language structure: ‘simplex’ (e.g. *find out* < *savoir*), ‘(morphologically) complex’ (e.g. *go on* < *reprendre*), ‘syntactically complex’ (e.g. *put off* < *remettre à plus tard*), ‘nothing’ in cases of explicitation (e.g. *go on* < Ø) and ‘other’ (e.g. *a month has gone by* < *ça fait déjà un mois*). We disregarded the last three cases, in other words, we retained only those phrasal verbs that corresponded with a simplex or a morphologically complex verb in the source text. We converted the results into a 2x2 contingency table, on which we performed the Chi-Square Test of Association (with Yates correction) using the web calculator developed by Richard Lowry available at http://vassarstats.net/tab2x2.html. In addition, the Fischer Exact Probability Test was performed via GraphPad’s online application available at http://graphpad.com/quickcalcs/contingency2/.

4.2 Results

Out of the 149 phrasal verbs in the translation, 98 have a morphologically simplex verb and 29 have a morphologically complex verb in the French source text (with 12 phrasal verbs being the translation of a syntactically complex structure, 5 having nothing directly corresponding with them in the source text and another 5 corresponding to other structures). In the entire source text, there were another 1592 simplex verbs and another 123 complex verbs whose translation did not correspond with a phrasal verb in the target text. These results are shown in Table 2 (with added percentages based on raw totals) and represented graphically by means of two pie charts in Figure 2 below:

Table 2: Distribution of phrasal verbs versus other translations of simplex versus complex verbs in Le Petit Prince

	Phrasal verb as translation	**Other translation**
Simplex source verb	98 (6%)	1592 (94%)
Complex source verb	29 (19%)	123 (81%)

Clearly, when the verb is morphologically complex in the French source text, it is much more likely to be translated with a phrasal verb in the English translation than when that source verb is simplex. The observed difference is statistically extremely significant ($\chi^2 = 36.27$; $p < .0001$). Fisher’s Exact Probability Test also produces a two-tailed p-value smaller than 0.0001, confirming that association between kinds of source verbs (simplex/complex) and translations (phrasal verbs/other) can be considered to be extremely statistically significant.

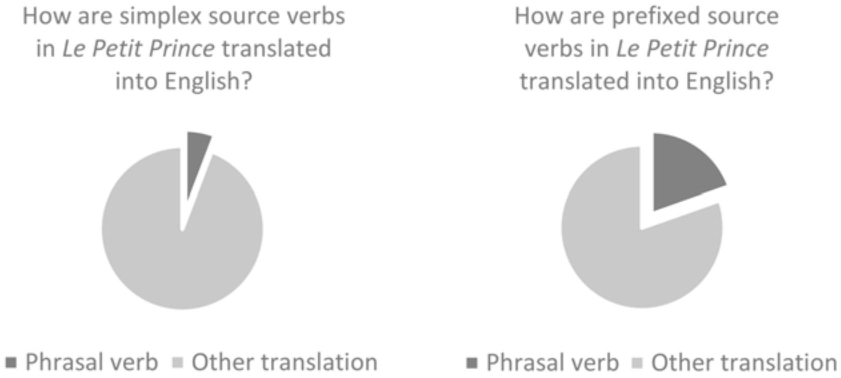


Figure 2: Morphologically simplex versus morphologically complex verbs in *Le Petit Prince* and the share of phrasal verbs corresponding to them in the English translation

4.3 Discussion

The English translation of *Le Petit Prince* contains over a hundred phrasal verbs, even though there were no phrasal verbs in the source text, French being a language that lacks such a structure. This could be taken as evidence that there was some degree of normalization at work. Yet, such a conclusion is trivial. After all, it is nothing more than a translator's job to render a text into the target language and if that target language makes frequent use of phrasal verbs, then the occurrence of such structures is not at all surprising. Moreover, as we saw before, translations into English from Romance languages do not contain more phrasal verbs than non-translated English, but to the contrary.

If phrasal verbs are used in the translation when there is no phrasal verb in the source text, what prompts the translator to introduce these 'unique items'? We surmised that, again, there may be some source-language influence at work. When the source text contains an expression that displays a satellite-framed encoding strategy, whether for motion or a semantically more abstract (e.g. resultative) event, then the distance to a phrasal verb as translation is somehow smaller than when there is no satellite-framed item in the source text. That is, a prefixed verb in French leads more easily to a phrasal verb in English, since at an underlying cognitive level, these structures show crucial similarities. Table 3 below provides a sample of prefixed verbs in *Le Petit Prince* and their phrasal verb translations in *The Little Prince*.

Table 3: Sample of morphologically complex verbs in *Le Petit Prince* and their corresponding translations involving a phrasal verb. All prefixes are underlined, as are English particles that are semantically equivalent to them

s'écrier → cry out	*s'enhardir → pluck up courage*
refaire → do over	*s'enfermer → shut oneself up*
rapporter → bring back	*reprendre → go on*
revenir → come back	*poursuivre → go on*

As Table 3 shows, however, only in some cases does the prefix in French actually correspond with a semantically equivalent particle in English. Often, the particle is not at all any close translation of what the prefix expresses (cf. also Paillard and Videau 2008). For instance, in *s'enfermer*, *en-* basically means *in* but this reflexive verb is translated by *shut (oneself) up*, not by *shut (oneself) in*. Thus, it may be that the translator noticed something semantically complex in the source language and was then, at some unconscious level, encouraged to render this complex item by a phrasal verb in the target language, without necessarily feeling the need to preserve the internal semantic makeup of the source item in the translation.

5 Conclusion

In this chapter, we have shown that translations into English from Romance languages contain fewer phrasal verbs than translations into English from Germanic languages, which do not differ noticeably from non-translated English in the frequency of phrasal verbs. The most likely explanation for this is that all Germanic languages are satellite-framed and thus frequently use structures that are very similar to phrasal verbs in English, while Romance language are verb-framed and do not use such structures as frequently, if at all. This result casts doubt on the validity of normalization as a translation universal, as the frequency of phrasal verbs is apparently not similar to that in non-translated English, or even higher, *independently of the source language used*. For the same reason, this result also undermines the validity of levelling-out, as phrasal verbs are not underused (because of their general informality), *independently of the source language used* – indeed, they are not in the translations from Germanic source languages.

Furthermore, as we also expected, a French simplex verb is less frequently translated into English by a phrasal verb than a French prefixed verb is. The most plausible explanation of this is that French prefixed verbs are satellite-framed structures and thus structurally and especially conceptually close to

English phrasal verbs, which then present themselves more easily as translations than in the case of simplex source verbs, even if the prefix is not necessarily rendered by a semantically equivalent particle.

Methodologically, we would like to stress three points that may be useful for similar studies. Firstly, we have demonstrated that when facing a linguistic phenomenon that would require time-consuming manual filtering, it is possible to select a few items which can be retrieved automatically by relying on independent research which shows that these items, when used as search queries, do not yield much noise. As a consequence, especially with quite large corpora (ranging here from over a million to almost sixteen million words), one can use a few well-chosen search items that do not necessitate much, if any, filtering of noisy examples afterwards, as representatives of the phenomenon under research. In our example, corpus frequencies for the words *up*, *out* (at least after extraction of occurrences of *out of*) and *down* can be 'trusted' as indicators of the use of phrasal verbs.

Secondly, we have shown how a large-scale, quantitative corpus-based approach can and sometimes has to be combined with a smaller-scale, more qualitative study, where data have to be coded manually and with considerable deliberation, as was done in this study for *Le Petit Prince* and its English translation. Results obtained by such more painstaking research still allow – and should undergo – a minimum of statistical validation, if the data are sufficiently large. We hope to have shown that the combination of different methods is especially fruitful when one method drives hypotheses that can be tested with another.

Thirdly, and most importantly, we believe that the application of a certain method can never be a goal in itself. Any methodology should remain subservient to answering a question or supporting a hypothesis that helps to advance, by however small a step, our current state of knowledge. In the present chapter, we have used a rather simple method to discredit normalization and levelling-out as translation universals and to show that the typological nature of the source language 'shines through' in the frequency of use of phrasal verbs in English translations, thus supporting the high importance of source-language influence. We have used an equally simple method to support the linguistically relevant idea that French, in its use of prefixed verbs, is not a *perfect* verb-framed language. While these prefixes do not necessarily carry information on 'path' but, rather, are remnants of previous stages in which the language was satellite-framed (Latin and perhaps Old French), these vestiges of satellite-framing appear to be synchronically relevant, at least to L2 users of French translating into English, as prefixed verbs are translated more often than non-prefixed ones by phrasal verbs in English, which we have regarded in this chapter as satellite-framed structures. In short, our methods have been used to some advantage to further our current knowledge in the fields of both translation studies and language typology.

Acknowledgements

We would like to thank the organizers and participants of the workshop "New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies" held as part of the 46th Annual Meeting of the Societas Linguistica Europaea at Split, Croatia (18–21 September 2013) for valuable comments and encouragement. We are grateful to Lindsay Leichnig, a master's student at the University of Lille 3, for classifying all the verbs used in Saint-Exupéry's *Le Petit Prince* and performing the counts on them. We also thank the editors and two anonymous reviewers for their thoughtful comments, occasional criticisms, and detailed suggestions, which helped us improve our text. Of course, any remaining errors are ours.

References

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In M. Baker et al. (eds.), *Text and Technology*, 233–250. Amsterdam / Philadelphia: John Benjamins.

Baker, Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In Harold Somers (ed.), *Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager*, 175–186. Amsterdam / Philadelphia: Benjamins.

Baker, Mona & Maeve Olohan. 2000. Reporting *that* in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures* 1(2). 141–158.

Becher, Viktor. 2010. Abandoning the notion of "translation-inherent" explicitation: Against a dogma of translation studies. *Across Languages and Cultures* 11(1). 1–28.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson.

Cappelle, Bert. 2001. Is *out of* always a preposition? *Journal of English Linguistics* 29(4). 315–328.

Cappelle, Bert. 2012. English is less rich in manner-of-motion verbs when translated from French. *Across Languages and Cultures* 13(2). 173–195.

Cappelle, Bert & Rudy Loock. 2013. Is there interference of usage constraints? A frequency study of existential *there is* and its French equivalent *il y a* in translated vs. non-translated texts. *Target. International Journal of Translation Studies* 25(2). 252–275.

Corpas Pastor, Gloria, Ruslan Mitkov, Naveed Afzal & Viktor Pekar. 2008. Translation universals: do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA-08)*, Waikiki, Hawaii, 21–25.

Chesterman, Andrew. 1997. Explanatory adequacy and falsifiability in translation theory. In Kinga Klaudy & János Kohn (eds.), *Transferre Necesse Est: Proceedings of the 2nd International Conference on Current Trends in Studies of Translation and Interpreting*, 219–224. Budapest: Scholastica.

Chesterman, Andrew. 2004. Beyond the particular. In Anna Mauranen & Pekka Kuyamaki (eds.), *Translation Universals: Do they exist?*, 33–49. Amsterdam/Philadelphia: John Benjamins.

Claridge, Claudia. 2002. Translating phrasal verbs. In Bernhard Kettemann & Georg Marko (eds.), *Teaching and Learning by Doing Corpus Analysis: Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz, 19–24 July, 2000*, 361–373. Amsterdam: Rodopi.

Dai, Guangrong & Richard Xiao. 2011. SL 'shining-through' in translational language: A corpus-based study of Chinese translation of English passive. *Translation Quarterly* 62. 85–108.

Davies, Mark. 2004. *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). http://corpus.byu.edu/bnc/.

Delaere, Isabelle, Gert De Sutter & Koen Plevoets. 2012. Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target. International Journal of Translation Studies* 24(2). 203–224.

Delaere, Isabelle & Gert De Sutter. 2013. Applying a multidimensional, register-sensitive approach to visualize normalization in translated and non-translated Dutch. *Belgian Journal of Linguistics* 27, Marie-Aude Lefer & Svetlana Vogeleer (eds.), 43–60. Amsterdam: John Benjamins.

Gardner, Dee & Mark Davies. 2007. Pointing out frequent phrasal verbs: A corpus-based approach. *Tesol Quarterly* 42(2). 339–359.

Halverson, Sandra L. 2003. The cognitive basis of translation universals. *Target. International Journal of Translation Studies* 15(2). 197–241.

House, Juliane. 2008. Beyond intervention: universals in translation, *Trans-kom* 1. 6–19.

Iacobini, Claudio & Francesca Masini. 2003. The emergence of verb-particle constructions in Italian: locative and actional meanings. *Morphology* 16(2). 155–188.

Jiménez-Crespo, Miguel A. 2011. The future of "universal" tendencies: A review of papers using localized websites. Proceedings of UCCTS 2010, *Using Corpora in Contrastive and Translation Studies*, Edge Hill University, UK. http://www.lancs.ac.uk/fass/projects/corpus/UCCTS2010Proceedings/.

Klaudy, Kinga & Krisztina Károly. 2005. Implicitation in translation: An empirical justification of operational asymmetry in translation. *Across Languages and Cultures* 6(1). 13–28.

Koppel, Moshe & Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1318–1326. Portland, Oregon: Association for Computational Linguistics.

Kruger, Haidee & Bertus Van Rooy. 2012. Register and the features of translated language. *Across Languages and Cultures* 13(1). 33–65.

Kopecka, Anetta. 2006. The semantic structure of motion verbs in French: Typological perspectives. In Maya Hickmann & Stéphane Robert (eds.), *Space in Languages: Linguistics Systems and Cognitive Categories*, 83–101. Amsterdam/Philadelphia: John Benjamins.

Laviosa, Sara. 1998. Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43(4). 557–570.

Laufer, Batia & Stig Eliasson. 1993. What causes avoidance in L2 learning: L1-L2 difference, L1-L2 similarity, or L2 complexity? *Studies in Second Language Acquisition* 15(1). 35–48.

Lefer, Marie-Aude & Svetlana Vogeleer (eds.). 2013. *Interference and Normalization in Genre-Controlled Multilingual Corpora*. Belgian Journal of Linguistics 27. Amsterdam: John Benjamins.

Liao, Yan & Yoshinori J Fukuya. 2004. Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning* 54(2). 193–226.

Lind, Sarah. 2007. Translation Universals (or laws, tendencies, or…?) *TIC Talk 63 (Newsletter of the United Bible Societies Translation Information Clearinghouse)*. http://www.ubs-translations.org/tt/past_issues/tic_talk_63_2007/.

Loock, Rudy, Gert De Sutter & Koen Plevoets. 2013. Teasing apart Translation Universals and Source-Language Interference: A case study on derived adverbs in English and French. Talk presented at *ICLC 7 – UCCTS 3*, Ghent, 11–13 July 2013.

Marks, Jonathan. 2005. The truth revealed: phrasal verbs in writing and speech. *MED Magazine* 34. http://www.macmillandictionaries.com/MED-Magazine/October2005/34-Feature-PV-Spoken-Written.htm.

Mauranen, Anna. 2004. Corpora, universals and interference. In Anna Mauranen & Pekka Kujamäki (eds.), *Translation Universals: Do they Exist?*, 65–82. Amsterdam/Philadelphia: John Benjamins.

Mauranen, Anna & Pekka Kujamäki. 2004. *Translation Universals: Do They Exist?* Amsterdam/Philadelphia: John Benjamins.

McEnery, Tony, Richard Xiao & YuKio Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London and New York: Routledge.

Mudraya, Olga, Scott S. L. Piao, Laura Löfberg, Paul Rayson & Dawn Archer. 2005. English-Russian-Finnish cross-language comparison of phrasal verb translation equivalents. In C. Cosme, C. Gouverneur, F. Meunier & M. Paquot (eds.), *Proceedings of Phraseology 2005: An Interdisciplinary Conference*, *13–15 October 2005, Louvain-la-Neuve, Belgium*, 277–281.

Olohan, Maeve. 2003. How frequent are the contractions? *Target. International Journal of Translation Studies* 15(1). 59–89.

Øverås, Linn. 1998. In search of the third code: An investigation of norms in literary translation. *Meta* 43(4). 557–570.

Paillard Michel & Nicole Videau. 2008. Les verbes français préfixés en *dé-* et leurs traductions en anglais. In M. Paillard (ed.), *Préfixation, prépositions, postpositions*, 75–91. Rennes: Presses Universitaires de Rennes.

Pourcel, Stephanie & Anetta Kopecka. 2006. Motion events in French: Typological intricacies. Unpublished ms., Brighton: University of Sussex and Nijmegen: Max Planck Institute for Psycholinguistics.

Talmy, Leonard. 2000. *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.

Teich, Elke. 2003. *Cross-Linguistic Variation in System and Text*. Berlin: Mouton de Gruyter.

Thim, Stefan. 2012. *Phrasal Verbs: The English Verb-Particle Construction and its History*. Berlin: Mouton de Gruyter.

Tirkkonen-Condit, Sonja. 2004. Unique items: over- or under-represented in translated language? In Anna Mauranen & Pekka Kujamäki (eds.), *Translation Universals: Do They Exist?*, 177–184. Amsterdam/Philadelphia: John Benjamins.

Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond*. Amsterdam/Philadelphia: John Benjamins.

Wild, Catherine. 2010. *Attitudes towards English usage in the late modern period: The case of phrasal verbs*. Glasgow: University of Glasgow doctoral dissertation.

Appendix A: Prefixed verb contexts in *Le Petit Prince* with phrasal verbs in the English translation

Below is a list of sentences or sentence fragments with prefixed verbs in *Le Petit Prince* (with the prefix given in boldface) and corresponding passages containing a phrasal verb in the English translation (with the particle in boldface). Note that there is not always a direct semantic correspondence between the French prefix and the English particle.

*Je **re**fis donc encore mon dessin*	*So then I did my drawing **over** once more.*
*Alors il s’**é**cria: […]*	*He cried **out**, then: […]*
*L’astronome **re**fit sa démonstration en 1920*	*So in 1920 the astronomer gave his demonstration all **over** again*
*Et comme il se sentait un peu triste à cause du souvenir de sa petite planète abandonnée, il s’**en**hardit à solliciter une grâce du roi: […]*	*And because he felt a bit sad as he remembered his little planet which he had forsaken, he plucked **up** his courage to ask the king a favor: […] [No direct semantic correspondence]*
*Il faut exiger de chacun ce que chacun peut donner, **re**prit le roi.*	*“One much require from each one the duty which each one can perform,” the king went **on**.* [On *refers to continuation after an interruption, and so has some of the resumptive meaning expressed by* re-]
*Je vais **re**partir !*	*So I shall set **out** on my way again. [Note that it is the adverb* again *here that captures the meaning of the prefix in the source text.]*
*acheva le buveur qui s’**en**ferma définitivement dans le silence.*	*The tippler brought his speech to an end, and shut himself **up** in an impregnable silence. [No direct semantic correspondence]*

*Le petit prince **pour**suivit: […]*	*The little prince went **on** with his explanation: […] [No direct semantic correspondence]*
*S’il s’agit par exemple de la découverte d’une grosse montagne, on exige qu’il en **r**apporte de grosses pierres.*	*For example, if the discovery in question is that of a large mountain, one requires that large stones be brought **back** from it.*
*On pourrait **en**tasser l’humanité sur le moindre petit îlot du Pacifique.*	*All humanity could be piled **up** on a small Pacific islet. [No direct semantic correspondence]*
*Où sont les hommes ? **re**prit enfin le petit prince.*	*“Where are the men?” the little prince at last took **up** the conversation again. [Note that it is the adverb* again *here that captures the meaning of the prefix in the source text.]*
*Bonjour, répondit poliment le petit prince, qui se **re**tourna mais ne vit rien.*	*“Good morning,” the little prince responded politely, although when he turned **around** he saw nothing.*
*Mais le renard **re**vint à son idée:*	*But he came **back** to his idea.*
*Les autres pas me font **r**entrer sous terre.*	*Other steps send me hurrying **back** underneath the ground.*
*Le lendemain **re**vint le petit prince.*	*The next day the little prince came **back**.*
*Il eût mieux valu **re**venir à la même heure, dit le renard.*	*“It would have been better to come **back** at the same hour,” said the fox.*
*Tu **re**viendras me dire adieu, et je te ferai cadeau d’un secret.*	*Then come **back** to say goodbye to me, and I will make you a present of a secret.*
*Et il **re**vint vers le renard:*	*And he went **back** to meet the fox.*
*Ils **re**viennent déjà ? demanda le petit prince…*	*“Are they coming **back** already?” demanded the little prince.*
*Comme le petit prince s’**en**dormait, je le pris dans mes bras, et me remis en route.*	*As the little prince dropped **off** to sleep, I took him in my arms and set out walking once more. [No direct correspondence]*

*Les hommes, dit le petit prince, ils s’**en**fournent dans les rapides, mais ils ne savent plus ce qu’ils cherchent.*	*“Men,” said the little prince, “set **out** on their way in express trains, but they do not know what they are looking for.” [No direct semantic correspondence]*
***Re**viens demain soir…*	*Come **back** tomorrow evening…*
*Lorsque je **re**vins de mon travail, le lendemain soir, j’aperçus de loin mon petit prince assis là-haut, les jambes pendantes.*	*When I came **back** from my work, the next evening, I saw from some distance away my little price sitting on top of a wall, with his feet dangling.*
*je veux **re**descendre !*	*I want to get **down** from the wall. [No direct semantic correspondence]*
*Tu vas pouvoir **r**entrer chez toi…*	*Now you can go **back** home…*
*Moi aussi, aujourd’hui, je **r**entre chez moi…*	*I, too, am going **back** home today…*
*Quand je réussis à le **re**joindre il marchait décidé, d’un pas rapide.*	*When I succeeded in catching **up** with him he was walking along with a quick and resolute step. [No direct semantic correspondence]*
*Il hésita encore un peu, puis il se **re**leva.*	*He still hesitated a little; then he got **up**. [No direct semantic correspondence]*
*Mais je sais bien qu’il est **re**venu à sa planète, car, au lever du jour, je n’ai pas retrouvé son corps.*	*But I know that he did go **back** to his planet, because I did not find his body at daybreak.*
*écrivez-moi vite qu’il est **re**venu*	*Send me word that he has come **back**.*

Appendix B: Other verbs coded as prefixed in *Le Petit Prince*

Below is a list of all other verb forms coded as prefixed by the MA student; the reader may beg to differ about the prefixed nature of some of these forms, for example in the case of *ajouter*. Note, however, that eliminating tokens from this list would only make our findings stronger, not weaker, as these are the forms for which the translation does not contain a phrasal verb.

ajouta
ajouta
ajouta
ajouta
ajouta
ajouta
ajouta
ajouta
ajouta
ajouta
ajoutais
ajouté
ajouter
ajustait
anéantir
apaisent
apercevais
apercevrai
aperçu
aperçus
aperçus
aperçut
aperçut
aperçut
aperçut
aperçut
décrire
défait
démodent
dépendra
dérangé
dérangea
dérangeaient
dévisser
disparu
écrasent
efforçai
égaré
égaré
embaumait
embaumait
embellit
émerveilla
emportait
emportent
emporter
emporter
emporter
emporter
emporter
encombre
enferme
enferme
enfonça
enfuir
enlève
enroula
éprouvai
éprouve
éprouve
épuisait
intimide
ralluma
ralluma
rallumer
rallumer
rassuré
rassurent
réchauffait
recommença
recommença
recompte
reconnaître
redevint
refis
rejoindre
réjouir
remua
répandait
repartir
repose
reprit
reprit
ressemblaient
ressemblait
ressemblait
ressemblait
ressemble
ressemble
ressemblent
ressemblent
ressemblent
ressemblent
ressemblent
ressembleraient
retournais
retrouvé
réveille
réveillé
réveiller
réveillons
revoir
revoir
revoir
revu
s'attendrir
s'écria
s'écria
s'écria
s'écrient
s'étire
se découragea
se réveiller
surprendre
surveille
surveillé

Kerstin Kunz, Stefania Degaetano-Ortlieb, Ekaterina Lapshinova-Koltunski, Katrin Menzel and Erich Steiner

9 English-German contrasts in cohesion and implications for translation[1]

Abstract: This study discusses findings from a corpus-based comparison of cohesive features in English and German written and spoken registers with a view to translation studies. We use several multivariate techniques to empirically analyse our corpus data and to interpret it with respect to four research questions. These concern contrastive differences in the overall degree of cohesion, in the strength of cohesive relations, the meaning relations established and the breadth of inter- and intralingual register variation. We hereby add a focus on semantic relations across grammatical domains to the available lexicogrammatical accounts of language contrast, which provides a background for making suggestions for translation strategies.

1 Cohesion and contrastive linguistics: the added value for the study of translation

The present study discusses findings on cohesion in an English-German comparable corpus as a step towards deriving potential translation strategies for this language pair. We start from the claim that there are systemic and textual contrasts between English and German on the level of cohesion. Systemic contrasts in cohesion concern differences in the linguistic resources of the two languages to establish relations of meaning across grammatical domains, i.e. above the phrase level, between different clauses, sentences or larger stretches of text. The question here is which (cohesive) devices are available in each language to explicitly indicate particular cohesive relations to other linguistic expressions (called *antecedent* in the literature, especially in the case of coreference). We have discussed these differences in Kunz and Steiner (2012) for coreference, Kunz and Steiner (2013) for substitution, Kunz and Lapshinova-Koltunski (2014) for conjunction and Menzel (2014) for ellipsis. A summary of these studies is provided in section 2.

[1] The GECCo project was funded by the Deutsche Forschungsgemeinschaft under STE 840/6-2 and KU 3129/1-2. http://www.gecco.uni-saarland.de/GECCo/index.html

DOI 10.1515/9783110459586-010

It is especially relevant for translation studies to see which of these resources are used and how they create cohesive relations in naturally occurring texts of English and German. Therefore, our focus is on identifying contrasts in the textual realizations of cohesion. This includes not only the investigation of the explicit signal, the cohesive device, but also the cohesive relation triggered. Depending on the type of cohesion, several cohesive devices may be used to create a *cohesive chain*, containing more than two linguistic elements and stretching over longer textual passages than two adjacent sentences. Knowledge about these textual contrasts between English and German original texts should impact on the conscious use of particular cohesive strategies when translating or interpreting within this language-pair.

1.1 Motivation and main research objectives

Our research objective is not to draw any conclusions about properties of translation, or translationese, such as explicitation, standardization or interference (see e.g. Baker 1993; Toury 1995). We rather aim at complementing contrastive works on differences between English and German such as Hawkins (1986) or König and Gast (2012). While these approaches mainly focus on systemic features in lexicogrammar, our corpus-based work examines instantiations of textual relations across grammatical domains. Furthermore, the findings from our contrastive study are a point of departure for making suggestions about systematic and adequate translation strategies on the level of text/discourse. The importance of cohesion in general and of coreference resolution in particular has variously been addressed in the literature on translation (cf. among others Baker 1992: 180; Becher 2011: 55; Blum-Kulka 1986; Doherty 2002: 160; Fabricius-Hansen 1996; Hatim and Mason 1990: 192; House 2004; Königs 2011: 72), though usually in a programmatic or at best example-based way. Our study aims at a more comprehensive account and at one that has improved empirical grounding.

For this purpose, we compare corpora of English and German original texts in the first instance, rather than translations and originals. A contrastive study aiming at wide coverage requires accounting for the textual variation as a signal of variation in contextual configurations. Our corpus resource therefore comprises comparable subcorpora in 10 different English and German registers. In this way, we obtain findings about the cohesive norms informing general strategies for translations between the two languages. Additionally, the corpus constellation yields data on variation in different written[2] and spoken registers and allows deriving register-specific translation (and interpreting) strategies.

[2] Written-spoken refers to a mode distinction in terms of register theory.

Our approach goes beyond the investigation of individual cohesive phenomena. Following the classification by Halliday and Hasan (1976), our analyses cover various features of coreference, substitution, ellipsis, cohesive conjunction and lexical cohesion. Applying several statistical methods we capture language- and register-specific preferences as to the meaning relations that are established by cohesion, as to the forms to encode these relations, and as to the interaction of cohesive types.

1.2 Research questions and methodology

Our research design addresses four research questions about English-German contrasts which are especially relevant for translation studies. Research questions (1), (2) and (4) are based on assumptions that have already been discussed in contrastive works, however mostly with a view to lexicogrammar, while to our knowledge, question (3) has not been addressed so far:

(1) How cohesive are the texts in our corpus?

(2) How strong are the cohesive relations?

(3) Which semantic relations are generally expressed and which relations are preferred over others?

(4) How much cohesive variation is there in one language as compared to the other? How much difference is there between (written and spoken) registers?

Several statistical methods are applied to evaluate our corpus data in terms of questions (1) to (4) and to obtain insights about a) English-German language contrast, b) register variation and, within register variation, c) variation between written and spoken modes.

Question (1) is a very general one and concerns contrasts in the overall degree of cohesion. The term implies the number of cohesive devices per linguistic unit occurring per text, i.e. tokens, not types. Some strands of contrastive pragmatics (e.g. House 1997: 84) suggest that German and English differ along the explicitness – implicitness dimension, with German showing a preference for expressing meaning more explicitly by linguistic signals than English. This assumption is supported for lexicogrammar, information packaging and grammatical metaphor (cf. Fabricius-Hansen 1996; Steiner 2005; Hansen-Schirra et al. 2012). We expect that these differences also manifest themselves in the choice of cohesive devices for explicitly indicating relations of coherence, i.e. at a deeper conceptual level of the text (cf. Beaugrande and Dressler 1981). Another assumption formulated in the literature is that variation in terms of cohesive vs. non-cohesive expressions per text depends not only on language, but also on register (following on from Hansen-Schirra et al. 2007: 249; Kunz 2010: 395). We particularly expect variation between written and spoken registers. The degree of cohesion should be higher in spoken than in written registers, along with reduced information density, less metaphorical semantics to grammar mapping and a looser grammatical structure, which cannot be dealt with in the frame of this study. All these features are said to reflect the particular constraints of oral communication such as reduced working memory capacity, speaker interaction or a noisy environment (cf. Halliday 1989: 41; Levy and Jaeger 2007).

The degree of cohesion is measured in our study by relating the total number of cohesive devices to the total number of tokens per corpus or subcorpus. Yet, a lower degree of cohesion in one language/register relative to another does not necessarily imply a lower degree of explicitness in general as encoding could be provided on a different linguistic level such as lexicogrammar.

Research question (2) considers differences in the strength of the cohesive relation[3]. The term *strength of relation* is chosen in this study to cover not only coreference but also other types of cohesion. Our focus is on the following two aspects:

(I) How explicitly is the relation indicated by linguistic signals in the cohesive device? This has to do with parameters such as semantic reduction vs. semantic specification; multifunctionality vs. single precise function of the cohesive device. Here again, we draw on literature on contrastive pragmatics to suggest that German may use rather specific/explicit cohesive devices to strengthen cohesive relations while in English cohesive relations may be vaguer, created by more underspecified/less explicit cohesive devices. Take, for instance, the coordinating conjunction *and* and its counterpart *und* in German. Both are semantically vague and may be applied for a variety of logico-semantic relations or even pragmatic functions (see e.g. Carston 2002). We expect a higher amount of more specific conjunctive devices in German to express relations of addition (e.g. *darüber hinaus* or *außerdem*). Additionally, we expect spoken registers to employ more cohesive devices

[3] Studies on coreference use terms such as *accessibility* or *salience* of referents, which are related to our notion of *strength*. Different degrees of accessibility are indicated by varying degrees of specification or explicitness in anaphors (Ariel 1990, 2001; Prince 1981; Gundel et al. 1993). Studies with an information-theoretic perspective look into the predictability of upcoming referents in relation to the choice of the coreferring expressions (antecedents and anaphors), suggesting that the degree of accessibility / predictability of a referent at a particular point in the text is indicated by structural and semantic features of coreferring expressions. These interact with information structure and syntax of coreferring expressions, number of previously mentioned coreferring expressions as well as properties of chains (distance, size and length; see e.g. Grosz et al. 1995; Kaiser 2003; Lambrecht 1994; Sanders and Spooren 2001; Strube and Hahn 1999).

than written registers overall (see question (1) above) but at the same time these devices may be shorter and semantically more underspecified. Structural indicators for measuring cohesive devices along the dimension of implicitness/explicitness or underspecification/specification in this study are elliptical constructions vs. other cohesive types, pronouns as nominal heads vs. determiners/modifiers (within coreference and substitution), connects vs. conjunctive adverbials (within cohesive conjunction). A cohesive relation is "strong" in the sense of (2) if it is specific in its encoding and functionally unambiguous.

(II) How much does the cohesive relation contribute to the overall textual or thematic coherence? Two more specific aspects can be addressed in the frame of this study: chain size (number of elements in one chain) and number of (different) chains: while higher numbers of elements in one chain contribute to thematic continuity, higher frequencies of different chains per text reflect thematic progression or variation. There are no assumptions in the literature suggesting contrasts in terms of cohesive relations. A cohesive relation is "strong" in the sense of question (2) if it occurs in a chain of many elements and/ or if a text has many chains of its type.

The degree of cohesion (research question (1)) may or may not impact on the strength of relation: whenever high frequencies of cohesive devices are employed to establish the same type of relation – e.g. a succession of conjunctive devices to establish a temporal sequence or long coreference chains – the sequence or the coreference chain is strengthened. Yet, extensive use of cohesive devices may also translate into a high number of short chains, which are only of local relevance for the text.

Question (3) deals with differences in the type of meaning relations that are signalled by cohesion. Differences in the meaning relations expressed would point to different cultural preferences for expressing four main types of semantic relations which can be indicated on the basis of cohesion in English and German by the following cohesive types:

- Coreference: Identity between instantiated referents, mainly expressed by subtypes of personal and demonstrative coreference;
- Type reference: involving comparison between instantiated or generic referents belonging to the same referential type, expressed by subtypes of comparative reference, ellipsis and substitution;
- Similarity: Sense relations between different types of referents, expressed by lexical cohesion (general nouns);
- Logico-semantic relations: relations such addition, cause, time, contrast and manner as expressed by subtypes of cohesive conjunction (additive, causal, adversative, temporal and modal conjunctions).

Section 2 presents a discussion and exemplification of the cohesive types/devices realizing these meaning relations. To our knowledge, no studies exist that deal with contrasts along the dimension of meaning relations expressed by cohesive devices based on empirical data of the type offered here.

The research question can be approached from different angles for which we employ different statistical methods (see in more detail below): if we compare the two languages with respect to the distribution in percentage of the main cohesive types, we can see which type is more important for the overall cohesion in one language (or mode) than the other. Correspondence analysis shows how certain cohesive features cluster together and where the biggest differences and similarities lie: between languages or between registers (including spoken and written modes). On the other hand, text classification reveals those cohesive features which are (strongly) distinctive, i.e. which mainly contribute to the differences.

Question (4) finally examines differences in the breadth of variation between written and spoken registers of English and German. From a general perspective, this research question dives into language contrast by relating it to intralingual register variation. There are several works concerned with lexicogrammar suggesting that distinctions along the register dimensions of written vs. spoken, and formal vs. colloquial may be weaker in English than in German (Collins 2012; Leech et al. 2009: 20, 239; Mair 2006: 183). So far, this has not been seriously investigated for patterns of cohesion at all. We here attempt a first assessment for cohesion by examining differences in variation with respect to the research questions (1) to (3) raised above. The statistical methods will be correspondence analysis and text classification.

Knowledge about contrasts in terms of these four research questions by translators will impact on the local and possibly also on the global translation strategies that are consciously chosen for establishing textuality or coherence as well as thematic progression in texts. It may direct translators towards decisions such as the following in order to adhere to target language conventions on the level of cohesion:

- Generally explicitate textual relations which are left implicit in the source text by inserting cohesive devices in the target text.
- Move a textual relation that is preferred in the source text into the background in the target text by implicitating particular meaning relations and explicitating others.
- Make a vague cohesive relation stronger by using a more specific cohesive device or by using more cohesive devices in one cohesive chain.
- Use more different cohesive devices to mark one conceptual type of cohesive relation.

In addition, our study may shed light on whether any of these decisions generally holds for translating into a particular translation direction or whether it strongly depends on the register. Any recommendations made here are, of course, limited by the quantitative size and registerial spread of our corpus (cf. section 3.1 below). The fact that all of the written texts (approx. 1 million tokens) are published texts should at least provide a bottom line of quality assurance.

1.3 Overview of sections

The chapter is structured as follows: section 2 will be concerned with a definition of cohesive types and an overview of contrasts between the English and German language systems as to the resources available for establishing cohesion. Section 3 will describe the corpus and outline the methodologies for annotating and querying cohesive features as well as the multivariate techniques for analysing our corpus data. Section 4 will describe the findings obtained by the multivariate techniques, which will then be interpreted with respect to our four research questions in section 5.

2 Cohesion in English and German

2.1 Cohesive types and semantic distinctions

The types and subtypes of cohesion discussed below are classified on the basis of English (following Halliday and Hasan 1976) and occur in our presentations and discussions of data in sections 3 and 4 of this chapter (cf. Table 1 later on). In order to be able to relate these types to each other in a cross-linguistic comparison, we need to specify the semantic distinctions which they encode. Research question 1 in section 1 (the overall degree of cohesion) requires a maximally comprehensive view of cohesive types/devices. For an investigation of research question 2 (the strength of the cohesive relation), we need the structural realization of the types. The differences in the type of meaning relations signalled by cohesion (research question 3) is accounted for by the semantic distinctions encoded in the different types of cohesive devices. Finally, differences in the breadth of variation between written and spoken registers of English and German (question 4) can be stated either in terms of devices or in terms of semantic distinctions – hence both cohesive types/devices and their semantic distinctions need to be part of our model.

The relationship of cohesive coreference encodes identity between textually instantiated entities, including events (cf. examples (1) and (2)).

(1) *We work for prosperity and opportunity because **they**'re right. **It**'s the right thing to do.* [EO_ESSAY]

(2) *Wir arbeiten für Wohlstand und Chancen, weil **das** richtig ist. Wir tun **damit** das Richtige.* [GTRANS_ESSAY]

'We work for prosperity and opportunity because **that** is right. We do **thereby** the right'[4]

The English example (1) uses the personal reference pronouns *they* and *It* to corefer to the entities *prosperity and opportunity* in the first case, and to the event *working for prosperity and opportunity* in the second. The German corpus translation uses the demonstrative *das* and the demonstrative deictic *damit*, referring to the event *working for prosperity and opportunity* in both cases, but encoding an additional adverbial relation of *instrument* in the second. This is one of the typical cases of translation between English and German where the coreference relation as such is preserved, but it is not coreference exactly to the same entities in both cases, and it is semantically enriched by the instrumental relation in the second.

For coreference, we distinguish personal and demonstrative types. Their subtypes are based on their functions, e.g. acting as heads or modifiers in a text, or expressing local and temporal relations. Here, we also include demonstrative deictics and definite articles.

We separately analyse the category of coreference chains in terms of the number of referents and referring expressions in chains (antecedents and anaphors), as well as the number of chains and chain length. Coreference chain is thus not a separate cohesive category, but the chain aspect of coreference. These data are included in our statistical analyses in sections 3 and 4, but are not extensively discussed in this chapter.

Comparative reference is semantically distinct from personal and demonstrative reference. It does not create identity of reference (coreference) but rather evokes a relation of comparison between referents, events or propositions of the same type (see e.g. Halliday and Matthiessen 2013: 632; Schubert 2008: 35). We here include two subtypes, which express general (see 3) or particular comparison (see 4).

[4] The English translation in this example is provided as a gloss translation for the sake of comprehensibility. The other translations are drawn from our translation corpus ETRANS/GTRANS, as indicated. Note that the TRANS-examples in this study are outside the data of original texts used in section 4.

(3) *Although we believe privatization is beneficial, it is not part of the trade negotiations. Countries will continue to make* ***such*** *decisions for themselves.* [EO_ESSAY]

(4) *Ich habe diese einfache Art der Verschlüsselung nur deshalb gewählt, weil wir damit bei kleinen Zahlen bleiben, bei denen sich das Verfahren leichter nachvollziehen lässt. Eine* ***bessere*** *Möglichkeit bietet sich Herrn Weiss, wenn […]* [GO_POPSCI]

I chose this form of encryption only so we could keep to small numbers, where the operation is easier to follow. Mr. White, however, can choose a **better** way by […] [ETRANS_POPSCI]

Similarly to comparative reference, cohesive substitution and ellipsis encode comparison between different entities of the same type, based on co-denotation of that presupposed type.

(5) *In future studies of adult stem cell potential, it will be crucial to rule out the possibility that stem cells are merely fusing to local* ***cells*** *rather than generating new* ***ones***. [EO_POPSCI]

(6) *Daher ist bei künftigen Studien zum wahren Potenzial von adulten Stammzellen unbedingt auszuschließen, daß die Zellen nur mit den* ***lokal vorhandenen (0) verschmelzen, statt wunschgemäß neue (0) zu erzeugen***. [GTRANS_POPSCI]

In the German translation of the English original in (6), we see first a change in the type of cohesive device (substitution *one* and lexical *cells* by German ellipses). Second, ellipsis and substitution by themselves do not encode coreference, but rather co-denotation (*neue (0)* in 6). The latter does not presuppose joint reference to the same individual entity, but rather joint denotation of the same class of entities. Another instance of this semantic relationship can be seen in the denotation of the various occurrences of *cell(s)* in (5).

For substitution and ellipsis we define categories depending on what is substituted or elided: nominal or verbal phrase, or their parts, a part of a clause or even a whole clause (nominal, verbal and clausal in Table 1).

Cohesive conjunction[5] encodes logico-semantic relations between discourse units, usually propositions, such as addition, contrast, cause. We have already

[5] The term *conjunction* is used here as in Halliday and Hasan (1976) (cf. also Halliday and Matthiessen 2013: 593ff) as a type of cohesion. This needs to be kept distinct from *conjuncts*, in the sense of Quirk et al. (1985: 631ff), which refers to one type of cohesive *adverbials*, and from *conjunction* as a word class. We are using the term *connect* here to refer to cohesive conjunctions as opposed to adverbials.

seen this in the adding of an instrumental adverbial relation in (2) above. (7) and (8) show the shift in the semantic type of conjunctive relation – in this case from adversative *dagegen* to additive *moreover.*

We again have the preservation of the general type of cohesive device (conjunctive relation), but a change in subtype, and hence in the encoded semantic relation.

(7) ***Dagegen*** *ist das Gewicht der Bauwirtschaft mit 15 Prozent gegenüber Westdeutschland (4 Prozent) noch entschieden zu hoch.* [GO_ESSAY]

(8) ***Moreover****, the significance of the construction industry (15 %) remains much too high when compared with western Germany (4 %).* [ETRANS_ESSAY]

Conjunctive relations are analysed in terms of the logico-semantic relations they explicitate between discourse segments in section 3 below: additive (relation of addition), adversative (relation of contrast or alternative), causal (relation of causality), temporal (time-relation between events) and modal (relation between events connected by an evaluation of the speaker). We also consider the restrictions in their syntactic function (coordinating conjunctions and conjunctive adverbials), and include them into our definition of types under analysis (additive connects, additive adverbials, etc. in Table 1). Subordinating conjunctions are excluded from our analysis as their encoding is grammatical, rather than cohesive only.

Lexical cohesion involves sense relations between lexical items (e.g. hyperonymy, part-whole relations), yet also semantically weaker relationships of collocation (Halliday and Hasan 1976: 284).

(9) *Sweetheart. That's what that weather was called. Sweetheart weather, the prettiest day of the year. And that's when it started. On a day so pure and steady trees* ***preened****. Standing in the middle of a concrete slab, scared for their lives, they* ***preened****. Silly, yes, but it was that kind of day [...]* [EO_FICTION].

(10) *HERZBLATT. So wurde das Wetter genannt. Herzblattwetter, der schönste Tag des Jahres. Und da hat es angefangen. An einem Tag so rein und beständig, daß die Bäume sich* ***herausputzten****. In der Mitte eines Stücks Beton, um ihr Leben besorgt,* ***putzten*** *sie* ***sich*** *das Gefieder. Albern, ja, aber so ein Tag war es.* [GTRANS_FICTION]

In examples (9) and (10) from our corpus, we see a case where lexical cohesion is very much preserved. Notable exceptions are the variable translation of English *preen* as *sich putzen, sich herausputzen*, establishing a hyponymic relationship in German differently from English, and the translation of the English *that kind of day* using the general noun *kind (of)* by the German *so ein Tag* using the comparative reference item *so* for the English lexical expression.

There is no one-to-one relationship between semantic distinctions and patterns expressing them: coreference as a semantic relationship, for example, may be encoded in grammatical constructions or in cohesive configurations. And even within the latter, it may involve cohesive devices other than reference.

One further example for the multiple possible mapping relationships between semantics and variously lexicogrammar or cohesion is given below, this time using logico-semantic relations as an example (cf. (11) to (15)):

(11) *The performance was* ***followed*** *by a round of applause.*

(12) ***After*** *the performance, there was a round of applause.*

(13) ***After*** *the performance ended, there was a round of applause.*

(14) *The performance ended.* ***Afterwards****, there was a round of applause.*

(15) ***After*** *the event, there was a round of applause.*

The semantic relationship of temporal precedence is variously encoded in (11)–(15) through lexical, grammatical and cohesive devices. And in (15), it is additionally encoded through a combination of the grammatical preposition *after*, the demonstrative reference item *the* and the lexically cohesive hyponymy relationship between *performance* and *event*.

Most of the semantic distinctions encoded in cohesive devices can be expressed lexicogrammatically or cohesively across and even within languages – and we are particularly interested in the cohesive encoding and its systemic and instantial (textual) differences between English and German.

As we said at the beginning of this section, our classifications of cohesive devices are initially based on the account given for English in Halliday and Hasan (1976). Although the systemic possibilities for cohesive devices in the two languages are relatively similar, at least for the more general parts of the two systems, there are contrastive differences as well. The most important of these will be mentioned in the next section. Where such non-matching systems exist, they will show up in one-sided occurrences in the data: demonstrative pronominal adverbs (deictics) or demonstrative articles are rare or non-existent in our English data, for example, whereas verbal substitution and general nouns are very infrequent in our German data. In our lists of cohesive categories for the two languages, though, we are aiming at comprehensiveness – so no categories were excluded if they are specific to only one of the two language systems. Table 1 gives an overview of the categories of cohesion and their realizational types annotated in the corpus:

Table 1: Cohesive categories

Categories of cohesion	Realizational types
coreference	personal head, personal modifier, demonstrative head, demonstrative modifier, demonstrative local, demonstrative temporal, pronominal adverbs, definite articles
coreference chain	number of antecedents, number of anaphors, number of chains, chain length
comparative reference	comparative general, comparative particular
substitution	nominal, verbal, clausal
ellipsis	nominal, verbal, clausal
conjunctive relations	additive connects, additive adverbials, adversative connects, adversative adverbials, causal connects, causal adverbials, temporal adverbials, modal adverbials
lexical cohesion	general nouns

Our categories of cohesion are thus language specific, but comprehensive for the two languages. There is no assumption that they would form the *tertium comparationis* or the basis for a definition of translational equivalence. The semantic relationships encoded by these devices would be better candidates, but even they are not necessarily preserved in translations, as our examples in this section demonstrate. However, the bases for comparison here are not the systems, but rather textual frequencies of occurrences by language, register, and by the generalized written vs. spoken modes, even where the systems are similar or identical.

2.2 Systemic differences and some associated tendencies of instantiation

In the area of personal coreference (cf. Kunz and Steiner 2012), we find marginally more systemic distinctions in German overall (encoding of social distance in forms of address). In addition to distinctions in terms of lexical base forms, the encoding in German of grammatical as opposed to natural gender influences local resolvability of antecedent-anaphor chains with 3rd person singular pronouns differently from English; cf. examples (16) and (17) below.

(16) *Denn Erhards Philosophie war nicht einfach ein singulärer Geistesblitz –* ***sie*** *stand in einer langen deutschen Tradition des Bemühens um das Glück der großen Zahl.* [GO_ESSAY]

(17) *For Erhard's philosophy was not just a singular flash of inspiration –* ***it*** *was part of a long German tradition of seeking the happiness of the majority.* [ETRANS_ESSAY]

The English *it* in (17) may have either the subject or the subject complement as antecedent of the preceding clause, although the syntactic parallelism strongly suggests the former. The German *sie* in (16), by contrast, does not have any potential ambiguity.

Even where the systemic options for cohesive reference coincide, we find frequent alternative use of demonstrative reference in German as in examples (1, 2) above. Finally, there is frequent use of cohesive substitution, ellipsis, or lexical cohesion combined with reference devices (articles, demonstratives) for the encoding of coreference as a semantic relation.

In the area of demonstrative coreference, German has a diversified system of demonstratives for all the major referents and relations, including 'demonstrative deictics' (*darüber, damit, dabei* etc.). For many of these, it also has *asserting vs. questioning (darüber* vs. *worüber)* and *near vs. far (darüber* vs. *hierüber)* variants. There are also semantic and/or registerial differences between analytic and synthetic variants in particular *(darüber vs. über das vs. da rüber)*. In terms of instantiation, German demonstratives serving as anaphors to complex antecedents may be preferred to personal *it* in English and *es* in German (as in examples 1 and 2 above). Finally, the English proximity-distinction between *this/that* may be more systematic and frequent than its German counterparts.

Comparing substitution in English and German (Kunz and Steiner 2013), we find that nominal and verbal forms are less grammaticalized in German, i.e. retaining more of their lexical meanings (*ein(e,r,s), tun, so*) than those of English. Clausal substitution relies on etymologically related forms in the two languages, but with differing meaning relations. Generally in the area of substitution there are more forms in German at the borderline to other cohesive types (comparative reference, conjunction) than in English, and they may exhibit multifunctionality in both languages, though in different ways, as in the case of *so* (König 2015). This is illustrated in (18) and (19) below:

(18) *He thought he recognised the twisted thorn trees, and might indeed have* ***done so***. [EO_FICTION]

(19) *Es wollte ihm scheinen, als erkenne er die krummen Weißdornbäume wieder, und* ***das*** *mochte sich durchaus* ***so verhalten*** [...] [GTRANS_FICTION]

The English verbal substitution in (18) is translated as a combination of demonstrative reference *(das)*, comparative reference *(so)*, and a general verb *(verhalten)*. The German *so*, classified here as comparative reference because of its remaining "manner"-meaning, is on the borderline between reference and substitution. The cohesive effect of (19) is similar to the one in (18), but the cohesive devices are different.

Comparing conjunction in English and German (cf. Kunz and Lapshinova-Koltunski 2014), we find a richer inventory in German (pronominal adverbs/ demonstrative deictics) encoding fine-grained distinctions of meanings. For these and other distinctions, we find more multi-word constructions in English, for example use of English *that is why* for German *deshalb*. Because of the generally freer word order of German, there is more positional flexibility in German for the encoding devices, such as *deshalb*, which may occur in clause initial or clause-internal position. Additionally, more of the systemically available forms in German are at the borderline to reference and substitution, a possible overall indicator of a high multifunctionality of cohesive devices in German. *Deshalb* is a combination of demonstrative *des-* and logically conjunctive *halb* (like *therefore* in English), and German *dabei* in its adversative meaning is a combination demonstrative *da* with a preposition *bei*, fusing into an overall adversative conjunctive relation expression.

Moving onwards to ellipsis (cf. Menzel 2014), we first note the systemic possibility of ellipsis remnants with morphological agreement suffix to license elided nouns in German in the domain of NPs (21). In English, this possibility is very rare, hence the classification of *'one/s'* as substitution in (20).

(20) *People who need this science, I would make an effort to tell them we have real sciences, hard sciences, we don't need imaginary* ***ones***. [EO_FICTION]

(21) *Den Leuten, die diese Wissenschaft brauchen, also, ich würde mir extra Mühe geben, ihnen zu erzählen, daß wir richtige Wissenschaften haben, hieb- und stichfeste Wissenschaften, wir brauchen keine imaginäre* ***(0)***. [GTRANS_FICTION]

German cases of ellipsis are on the borderline to substitution, if the ellipsis remnant can be analysed either as a pronoun replacing the noun or as a determiner in an incomplete phrase. This is the case if its inflectional paradigm allows the insertion of a supposedly elided head noun only after certain agreement suffixes (*eine, keine* (0) vs. *eines,r keines,r (/) (one, none)*). In the domain of VPs, we find more possibilities for omission of lexical verbs after operator/ modal verbs in English. In the domain of the clause and its constituents, there are more possibilities for fragment clauses in German. There are probably more possibilities for *true fragments* in German altogether because of less ambiguity due to morphological markers. Those true fragments do not necessarily have underlying sentential structures that were subject to deletion/omission; therefore they will not be subsumed under the category of cohesive ellipsis. Exophoric or situational ellipses refer to the extralinguistic context and do not fall under the category of cohesive ellipsis in our annotation either.

The area of lexical cohesion is currently under investigation and receives little coverage in our chapter here. We do give a preliminary analysis of the distribution of general nouns, yet in conjunction with referential modification (*these, such* etc.) where necessary.

3 Methodology

3.1 Corpus Resources

For the corpus-based analysis of instantiations of cohesive categories we envisage here, we use GECCo, a German-English corpus containing written and spoken texts (cf. Lapshinova-Koltunski et al. 2012). The whole corpus contains ca. 1.3 million tokens and six subcorpora: English and German originals and their translations (extracted from CroCo; Hansen-Schirra et al. 2012), as well as two spoken subcorpora: German written originals (GO), English written originals (EO), English spoken originals (EO-SPOKEN) and German spoken originals (GO-SPOKEN), translations of German written originals into English (ETRANS) and translations of English written originals into German (GTRANS).[6] The two written subcorpora (EO and GO) consist of texts from eight registers: popular-scientific texts (POPSCI), tourism leaflets (TOU), prepared speeches (SPEECH), political essays (ESSAYS), fictional texts (FICTION), corporate communication (SHARE), instruction manuals (INSTR) and corporate websites (WEB). The two spoken subcorpora contain academic speeches (ACADEMIC) and interviews (INTERVIEW). This text collection provides 20 subcorpora (the size of the subcorpora are given in Table 3, see section 4.1 below) which serve as variables for our corpus-based statistical analysis.

To extract frequency information on the occurrence of cohesive categories under analysis in these subcorpora, we deploy annotations available in GECCo.

[6] The translation corpora ETRANS and GTRANS are not analysed empirically in this study but are used for the illustration of examples. Corpus size for our empirical investigation proper is thus only between 700,000 and 800,000, and only 30,000–40,000 tokens for the individual registers. This does raise issues of representativeness, but is a usual size for small richly annotated corpora.

They include information on tokens, lemmas, morpho-syntactic features (e.g. case, number), parts-of-speech, grammatical chunks along with their syntactic functions, clauses, and sentence boundaries. The annotation of the written subcorpora was partly imported from CroCo, whereas for the spoken part, we use the Stanford POS Tagger (Toutanova et al. 2003) and the Stanford Parser (Klein and Manning 2003). The corpus is encoded in the CWB format (CWB 2010) and can be queried with Corpus Query Processor (CQP; Evert 2005). The described annotation levels provide us with additional information on cohesive types, i.e. for coreference or conjunctive relations: morpho-syntactic preferences of antecedents and anaphors, position of coordinating conjunctions and conjunctive adverbials in a clause, etc. Information on cohesive devices and their categories is also annotated in the corpus, and includes both functional and structural subtypes of coreference, conjunction, substitution, ellipsis and lexical cohesion, as outlined in Table 1 in section 2. For reference chains, our annotations provide us with the information on the number of antecedents, anaphors, chains, as well as chain length.

For the annotation of cohesive categories, semi-automatic procedures were applied, which include a rule-based tagging of cohesive candidates and their manual post-correction by humans. The procedures involve an iterative extraction-annotation process based on the method derived from the system used for the YAC chunker (see Kermes and Evert 2002; Kermes 2003). The system is based on the option of the CWB tools to incrementally enhance corpus annotations, as query results deliver not only concordances of the searched structures but also information on their corpus positions. This permits the importation of information on queried data back into the corpus. In this way, we annotate candidates for cohesive categories, which are then corrected manually by human annotators with the help of MMAX2 (Müller and Strube 2006), as visualization options of this tool allow annotators to decide whether the candidates tagged by the automatic procedures have a cohesive function and belong to the given category. Moreover, in instantiations of cohesive relations, the borderlines between their categories may be blurred, e.g. the same realizational form (e.g. English or German *so*) may serve as different cohesive devices, depending on the context in which it is realized, see examples (18) and (19). These ambiguities can be resolved in the course of manual correction.

Manual procedures are also used for annotation of coreference chains, as human annotators manually identify antecedents and link them to the cohesive referring expressions (anaphors) which were automatically tagged by our system. A detailed description of the semi-automatic procedures of coreference, substitution and conjunctive relations is given in Lapshinova-Koltunski and Kunz (2014).

The instantiations of the categories given in Table 1 above can be easily extracted from the corpus, as relevant information is annotated and can be queried with CQP. Table 2 contains examples of queries used for data extraction.

Table 2: Query examples used to extract the categories under analysis

	Query	Explanation
1	[_.mention_chain_id="set.*"& .mention_antecedent="none"]	element in a reference chain & does not have an antecedent, an antecedent itself
2	[_.mention_chain_id="set.*"& .mention_antecedent!="none"]	element in a reference chain & an anaphor has an antecedent
3	[_.mention_func="poss.*"]	personal reference with a modifying function (pers_mod)
4	[_.mention_func="temporal"]	temporal demonstrative reference (dem_temporal)
5	[_.mention_chain_id="set.*"] & post-processing	all reference chains (nr_of_chains)
6	[_.conj_func="additive" & _.conj_type="connect"]	additive coordinating conjunctions (additive_connect)
7	[_.conj_func="additive" & _.conj_type="adverbial"]	additive adverbials (additive_adverbial)
8	[_.substitution_type="verbal"]	all cases of verbal substitution
9	[_.ellipsis_type="clausal"]	all cases of clausal ellipsis
10	[_.noun_type="general"]	general nouns

For instance, query 1 is used to extract information on the number of antecedents in cohesive chains, query 2 to extract the number of referring expressions in cohesive chains. With the help of query 3, we can identify how many referring expressions function as personal modifiers, whereas query 4 is used to identify all cases of cohesive demonstrative reference with a temporal function. Query 5 is enhanced with a post-processing procedure to count chains which have the same ID (chain_id) per text. Queries 6 and 7 are built to differentiate between coordinating conjunctions and adverbials expressing additive relations. We apply queries like in 8 and 9 to extract different types of substitution and ellipsis, and query 10 is used to extract occurrences of general nouns. The extracted numeric results are saved in tables for statistical validation which is described in the analysis presented in section 4 below.

3.2 Statistical methods applied

As already mentioned in section 1, we aim to analyse contrasts between languages and registers. This will help us to identify those phenomena that are relevant for translation analysis in terms of the four research questions raised in the introduction.

To answer these questions, we apply different types of quantitative analysis. We use descriptive data analysis to obtain information on the frequency of cohesive devices and in terms of distributions of main cohesive types. The findings provide a basis for interpretations in terms of research question (1) and also partially for question (3). In order to interpret our results with respect to questions (2), (3) and (4), we use explorative techniques, viz. correspondence analysis (CA; Venables and Smith 2010; Baayen 2008; Greenacre 2010), and supervised techniques, i.e. classification with support vector machines (SVM; Vapnik and Chervonenkis 1974; Joachims 1998).

3.2.1 Descriptive data analysis

Descriptive data analyses are employed for two purposes: First, for the investigation of general frequencies, we relate the total number of cohesive devices to the total number of tokens per corpus and subcorpus. The results are tested for significance using the Pearson's chi-squared test. Second, relating the frequencies of main cohesive features to the total number of cohesive features per corpus and subcorpus, we obtain insight into distributions of main cohesive types.

3.2.2 Correspondence analysis (CA)

Correspondence analysis allows us to see which variables (e.g. languages or registers) have similarities and which differ from each other. Moreover, we are able to trace the interplay of categories of the cohesive devices under analysis. Some of the independent variables defined in our corpus (e.g. English and German) lead to clearly distinguished classes in terms of the exploratory technique of CA, while others (e.g. registers) are less well distinguishable than with text classification (with support vector machines, see below).

An input for CA is a table of numeric data, in our case frequencies of the categories under analysis across registers and languages (subcorpora). First, distances (differences) between rows, and distances between columns are calculated. In a second step, these distances are represented in a low-dimensional map (we use a two-dimensional map for the representation). The larger the differences between subcorpora, the further apart these subcorpora are on the map. Likewise, dissimilar categories of cohesive devices are further apart. Proximity between columns and rows (subcorpora and cohesive devices) in the merged map is as good an approximation as possible of the correlation between them.

In computing this low-dimensional approximation, correspondence analysis transforms the correlations between rows and columns of our table into a set of uncorrelated variables, called principal axes or dimensions. These dimensions are computed in such a way that any subset of k dimensions accounts for as much variation as possible in one dimension, the first two principal axes account for as much variation as possible in two dimensions, and so on. In this way, we can identify new meaningful underlying variables, which ideally correlate with such variables as language or register, indicating the reasons for the similarities or differences between these subcorpora. The degree of the contribution of a certain dimension to the plot will show where the greatest differences lie.

The *ca* package (cf. Nenadic and Greenacre 2007) is used to perform correspondence analysis in the R environment (cf. Venables and Smith 2010). The output of the correspondence analysis is plotted into a two dimensional graph. The length of the arrows indicates how pronounced a cohesive device is for the overall analysis, see Jenset and McGillivray (2012) for details. The position of the points in relation to the arrows indicates the relative importance of a cohesive device for a subcorpus. The arrows pointing in the direction of an axis indicate a high correlation with the respective dimension, and thus, a high contribution of the feature to this dimension, see Figure 2 in section 4.2 below.

3.2.3 Support vector machines (SVM)

We use text classification with SVM to observe more fine-grained differences between the two languages (English and German) and the registers with respect to the features analysed. The major difference to CA is that with SVM (a supervised method) we impose the variables (language, register) on the data (rather than getting possible variables represented by the dimensions from the CA, an unsupervised method). Thus, while CA helps to get an overview of whether the features under analysis really reflect the variables that one wants to consider, with SVM, as we impose the variables, we can inspect in detail the whole range of features that make the variables distinct from one another. With SVM texts are classified according to the respective variables which are represented as classes. The classes defined in our corpus and used for classification are: (1) the languages, i.e. English and German, and (2) the registers shown in Table 3, again for each language, with particular consideration of the mode-dimension in some cases. The most distinctive features are drawn by the observation of how well they contribute to the distinction of specific classes.

Text classification tasks have been widely used, mostly based on bag-of-words representations (see e.g. Fox et al. 2012; Joachims 1998), where documents are represented by the words occurring in them. Other studies have used linguistic features generated out of linguistic theories (e.g. Argamon et al. 2008; Degaetano-Ortlieb et al. 2014), where the documents are represented by the occurrences of linguistic features rather than the occurrences of all words. In our approach, we use linguistic features based on cohesion.

For classification, we use support vector machines (SVM; Vapnik and Chervonenkis 1974), as they are known to obtain very good results on many relevant features (see Joachims 1998; Manning et al. 2008). In principle, SVM performs a binary classification trying to separate two classes from each other. As we have to solve a multi-class problem, we use a pairwise classification, i.e. one-versus-one classifiers are built. Considering the register distinction, for example, a classifier is built for each register pair (EO_ACADEMIC vs. EO_INTERVIEW, EO_ACADEMIC vs. EO_ESSAY, etc.). Classification is performed with the data mining platform Weka (Witten et al. 2011). As a dataset, we use a matrix of linguistic features per text for each class. For interpretation, we consider the classification accuracy (overall and for each class) as well as the F-Measure, i.e. the harmonic mean or weighted average of Precision and Recall (Powers 2011; Van Rijsbergen 1979). Additionally, we inspect the SVM weights. The higher the weight of a feature, the more distinctive it is for a particular class, regardless of its positive or negative sign, which only indicates the class it belongs to.

In the analysis, we perform two classifications according to the distinction of language and register, analysing which cohesive categories and features contribute most to the distinctions.

4 Quantitative analyses of English and German

4.1 Descriptive data analysis

We start with a comparison of frequency distributions, which allow interpretations in terms of the overall degree of cohesion, our research question (1).

The total number of cohesive devices per register and per language (see TOTAL, last row) is presented in Table 3. It also provides information on the total number of tokens per text register as well as the distribution in percentage of cohesive devices in relation to the total number of tokens.

Table 3: Cohesive devices across languages and registers

	English Originals (EO)			German Originals (GO)		
	Cohesive devices		**Total tokens**	**Cohesive Devices**		**Total tokens**
	abs.	**in %**	**abs.**	**abs.**	**in %**	**abs.**
ACADEMIC	4317	10.64	40559	5284	12.09	43703
INTERVIEW	3896	10.28	37898	6130	15.25	40198
FICTION	4612	12.47	36996	3679	10.00	36778
INSTR	1575	4.35	36167	1785	4.84	36880
ESSAY	2316	6.62	34998	2432	6.82	35668
POPSCI	2306	6.56	35148	2929	8.10	36177
SHARE	2050	5.72	35824	2332	6.62	35235
SPEECH	2354	6.71	35062	2759	7.79	35399
TOU	1739	4.84	35907	1807	4.94	36574
WEB	2194	6.07	36119	2109	5.89	35779
Total	27359	7.50	364678	31246	8.39	372391

Table 4 illustrates whether the contrasts between main corpora EO and GO (see ALL in last row), and contrasts per register between languages are significant, using Pearson's chi-squared test. A p-value below 0.05 shows significant differences (+) between the languages and the registers of each language. In case the p-value is above 0.05, differences are not significant (–).

Table 4: Cohesive devices across languages and registers: results of the chi-squared tests

EO ⇔ GO	p-value	Significance
ACADEMIC	<.0001	+
INTERVIEW	<.0001	+
FICTION	<.0001	+
INSTR	<.003	+
ESSAY	>.05	–
POPSCI	<.0001	+
SHARE	<.0001	+
SPEECH	<.0001	+
TOU	>.05	–
WEB	>.05	–
All	<.0001	+

Considering the results in Table 3, there is only a slight difference in the distributions in percentage between the main corpora EO and GO (TOTAL), i.e. all registers per language taken together. However, as the results from the Pearson's chi squared test in Table 4 illustrate, the contrast between languages is significant as the p-value is below 0.05. Furthermore, we note variation if we compare the findings for each language per register: Table 3 illustrates that equally low distributions are found in both languages for the registers INSTR, WEB, TOU, ESSAY. Here the contrasts are not significant, except for INSTR, where the contrast is slightly significant (see Table 4). The distributions in the register SPEECH and SHARE are also rather low; differences between the languages are slightly significant. The contrasts for the registers POPSCI, FICTION, INTERVIEW and ACADEMIC are highly significant; while the amount of cohesive devices in POPSCI are somewhat in the middle in both languages, frequencies for FICTION and the two spoken registers are the highest of all registers in both languages. With some exceptions, language contrasts between written registers (hence translation relevant registers) are generally less pronounced than in the spoken registers. The greatest differences are attested for the register INTERVIEW, where we note a distribution of 10.28% cohesive devices in relation to all tokens in English and 15.25% in German. The register of FICTION stands out as it is the only register in which the frequencies in German (10.00%) clearly lie below those in English (12.47%). Quite interestingly, there is considerable variation across registers, language internally. The differences between registers in German are more pronounced than in English, ranging from 4.84% in INSTR to 15.25% in INTERVIEW (compare English, ranging from 4.35% in INSTR to 12.47% in FICTION). More variation in the degree of cohesiveness is therefore observed between registers in German than English, and between written and spoken registers, in particular.

Figure 1 illustrates the distributions of cohesive devices signalling the main types of cohesion of all cohesive devices per register and per language. The findings are taken as a basis of interpretation (together with the analyses in sections 4.2 and 4.3 below) for research questions (2) to (4). The types shown in Figure 1 are personal and demonstrative coreference, comparative reference (comparison particular and general taken together), substitution (nominal, verbal and clausal taken together), ellipsis (nominal, verbal and clausal taken together), conjunction (additive, adversative, temporal, causal and modal) and general nouns (as the one type of lexical cohesion analysed in this study).

Figure 1 reveals general tendencies in the distribution of meaning relations (research question (3): we note a strong preference for relations of coreference and conjunctive relations in both languages. Most cross-linguistic similarities are observed for the register of FICTION, where we find the highest amount of coreference relations and lowest distributions for conjunctive relations. Apart from that, the amount of conjunctive relations is clearly higher and the amount of indicators of coreference slightly lower in German than English. All registers

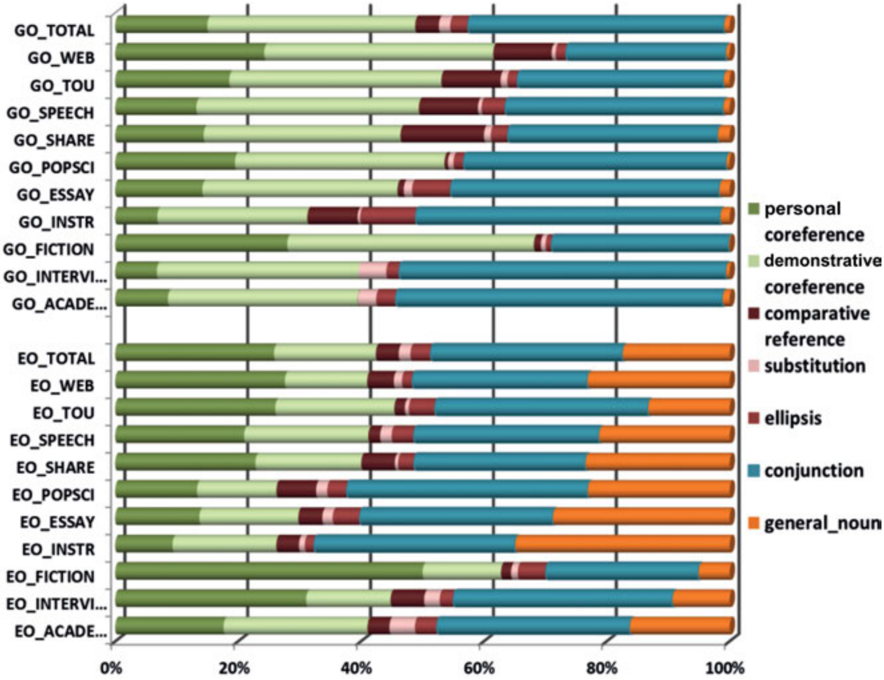


Figure 1: Distributions of main cohesive types per language and register

in German favour demonstratives for realizing coreference relations over personal reference, whereas the opposite is the case for most English registers. Comparative reference is preferred in five registers of German for realizing type-reference/ comparison over substitution and ellipsis while English exhibits more even distributions. We observe a rather high amount of general nouns in English whereas this category can almost be neglected in German. German spoken registers are characterized by very high frequencies of conjunctive relations and demonstratives and higher distributions of substitution than other German registers. The differences between English registers altogether seem to be less pronounced than between German registers. Evidence for this tendency will be further provided with the analyses in sections 4.2 and 4.3.

4.2 Exploratory analysis

We go on with correspondence analysis as presented in 3.2.2 above, which allows us to see which subcorpora have commonalities and which differ significantly from each other, so that we are able to detect where the biggest differences and similarities lie in the use of cohesive devices. Moreover, it provides us with information on the interplay of cohesive categories, and we can see which cohesive relations are generally preferred over others, as well as how much cohesive variation there is in different languages and registers.

We use frequencies of the cohesive categories in different subcorpora to calculate distances (see section 3.2.2). In the first step, we determine how many dimensions are needed to represent and interpret our data. For this, principal inertias of dimensions (eigenvalues) are analysed.

Table 5: Principal inertias of dimensions

	dim1	dim2	dim3	Further
value	0.184688	0.09516	0.070853	under 0.05
in %	40.29	20.76	15.46	under 10%
cum in %	40.29	61.05	76.51	100%

We can see from Table 5 that the first two dimensions (dim 2) can explain only 61.05% of the inertia. This means that some profiles (combinations of subcorpora) would lie more along the third or fourth axis etc., which are not visualized on a two-dimensional graph. However, we need to include the third dimension into our analysis to get a cumulative value of 76.51%, which is a satisfying coverage. So, we use two figures to represent the results in a two-dimensional map (Figures 2 and 3).

Figure 2 shows a graph representing the first two dimensions. Concerning dimension 1, we see a clear distinction between English and German subcorpora (along the x-axis on the left and on the right from zero respectively), with the exception of GO-FICTION, which is represented on the left side of the plot. However, GO-FICTION is not clearly visible on the graph, which is also seen in Table 6 illustrating χ2-distance (distances calculated on the basis of rows and columns, see section 3.2.1 above) to the centroid (dim1 and dim2) as well as the quality of display (qlt) of a variable in the graph (which amounts to only 5.4% for GO-FICTION). GO_ESSAY and GO_POPSCI have also very low values (their quality is <10%), although situated on the right (German) side of the axis. Thus, the higher the quality the better the subcorpora are represented by dimension 1 and 2, while the lower the quality the less representative these two dimensions are for the respective subcorpora.

We assume that the distinction along the first dimension (x-axis) reflects language contrasts in the use of particular cohesive features. Focusing on the features, cohesive devices on the right side are specific for German, whereas those on the left side are specific for English. Note that the three subcorpora with a lower quality of display, i.e. which are weakly represented in the

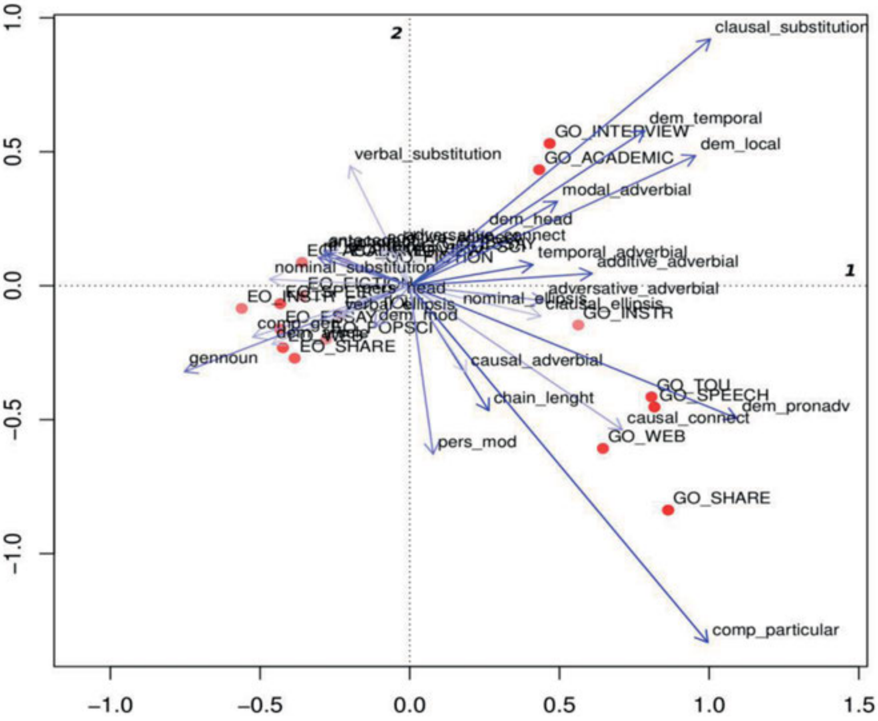


Figure 2: Data in dimensions 1 and 2

first dimension, show distinction on the basis of register specific- rather than language-specific cohesive features, and therefore, should not be considered for the analysis of language contrasts. This assumption is confirmed by the results from text classification (see section 4.1.2), which show that German political essays, fictional texts and popular-scientific articles differ more strongly from the other registers in German. These three registers should be excluded also from the interpretation along the second dimension (y-axis), where we note a separation between written and spoken registers, as both English and German INTERVIEW and ACADEMIC subcorpora (representing the spoken registers) are situated above the zero, while the written subcorpora are situated below.

The graph also provides us with information on the relevance of cohesive features contributing to the separation along both dimensions by considering the direction of the arrow. The arrows pointing to the right (such as additive, adversative, temporal conjunctive relations) and to the left (substitution and lexical cohesion), contribute more to the distinction between languages, whereas

Table 6: CA numerical results for the first two dimensions

subcorpora	qlt in %	dim1	dim2
EO_ACADEMIC	51.1	−0.837390	0.281427
EO_ESSAY	54.9	−1.010199	−0.521716
EO_FICTION	32.8	−0.831127	−0.119376
EO_INSTR	50.4	−1.307082	−0.272163
EO_INTERVIEW	13.4	−0.499686	0.283402
EO_POPSCI	38.0	−0.650140	−0.647820
EO_SHARE	67.8	−0.895342	−0.876845
EO_SPEECH	71.7	−1.006731	−0.217105
EO_TOU	22.1	−0.540643	−0.355184
EO_WEB	74.9	−0.986197	−0.749286
GO_ACADEMIC	85.5	1.006409	1.402345
GO_ESSAY	9.5	0.212667	0.363926
GO_FICTION	5.4	−0.223956	0.200083
GO_INSTR	47.0	1.310359	−0.474780
GO_INTERVIEW	87.9	1.088093	1.719372
GO_POPSCI	4.0	0.055976	0.332257
GO_SHARE	91.5	2.009601	−2.717135
GO_SPEECH	90.3	1.902325	−1.467849
GO_TOU	84.9	1.879612	−1.345706
GO_WEB	81.1	1.501930	−1.969208

those pointing up and down, such as coreference devices (especially personal subtypes) and features of coreference chains (expressed in chain length and number of chains[7]) contribute rather to the distinction between written and spoken registers. Some arrows are situated in the middle of the field, e.g. those of clausal substitution or comparative reference. This means that they contribute to the two distinctions observed.

Differences can also be seen in terms of cohesive variation in both languages with regard to both cohesive features and registers within languages. English registers seem to be more alike, as the distances between their points are shorter than those between German registers. The same tendency can be stated for the arrows indicating cohesive features – those specific for English.

To compensate the under-representation of the data (61.1%, see again Table 5), we add the third dimension, which allows us to increase the quality of subcorpora display, see Table 7. Particularly, the representation of GO-FICTION increases rapidly (from 5.4% to 88.4%).

[7] Not visible in the graph but visible in the numeric output data.

Table 7: Numeric data for the third dimension and quality in the three-dimensional representation

register	qlt	dim3
EO_ACADEMIC	58.7	−0.537025
EO_ESSAY	91.6	−1.421468
EO_FICTION	91.9	1.812896
EO_INSTR	87.2	−1.824372
EO_INTERVIEW	17.0	0.446020
EO_POPSCI	80.1	−1.360351
EO_SHARE	81.7	−0.800174
EO_SPEECH	80.2	−0.565533
EO_TOU	24.0	−0.279389
EO_WEB	82.5	−0.577310
GO_ACADEMIC	88.5	−0.433591
GO_ESSAY	12.7	0.318894
GO_FICTION	88.4	1.683101
GO_INSTR	51.3	−0.657484
GO_INTERVIEW	91.5	−0.537037
GO_POPSCI	22.0	0.845048
GO_SHARE	91.7	−0.213742
GO_SPEECH	90.9	−0.294369
GO_TOU	85.0	0.126993
GO_WEB	87.1	0.903294

In Figure 3, we plot the third dimension in a two-dimensional graph (together with dimension 1). We suggest that the separation along this dimension shows which registers show similarities (independently from the languages), and which cohesive features contribute to these similarities. For example, we can clearly see this for ACADEMIC, FICTION, SPEECH and SHARE, whereas other registers (ESSAY, INTERVIEW, POPSCI, TOU and WEB) seem to be rather language-dependent. Moreover, the plot allows us to identify cohesive features contributing to these similarities. In our case, these are especially fictional texts with (personal) reference as a feature that makes them similar.

By the results of the correspondence analysis, we see that the first dimension represents the variable of language, and thus shows us where the differences between languages lie. The second dimension represents the variable of mode, which groups registers according to written and spoken features. And finally, the third dimension is the one of registers – yet another confirmation of a widely-shared assumption (e.g. Biber 2014) that apart from language itself, the written-spoken and the narrative-non-narrative dimension are strong interlingual dimensions of register contrast.

Now, looking for an answer to the question, where the greatest differences lie, we look again at Table 5. The greatest differences lie between languages, as this dimension contributes ca. 40%, the two others in the ranking list are the dimensions related to registers (21% and ca. 16% respectively), which, to our opinion, comprise a considerable proportion.

Summing up, with the help of correspondence analysis we were able to see where the differences and similarities lie, observe the breadth of cohesive variation in both languages, as well as find out which cohesive features contribute to the distinction along these dimensions (language and register).

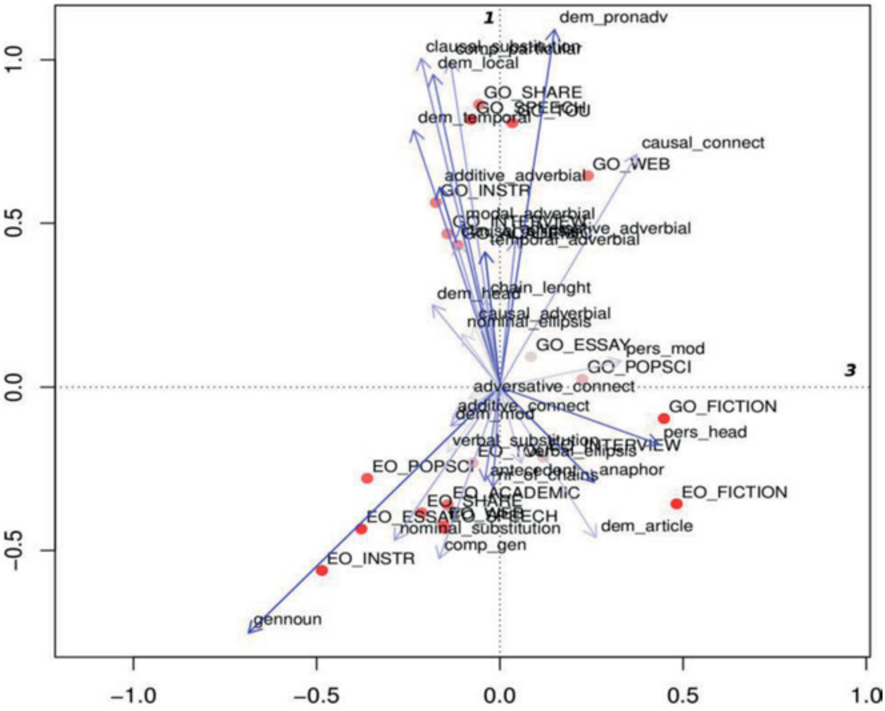


Figure 3: Data in dimensions 3 and 1

However, this analysis procedure does not allow us to precisely define which cohesive features are distinctive for the different subcorpora under analysis. Therefore, we use supervised classification techniques in the next step.

4.3 Supervised analysis

For the supervised approach, we use support vector machines (SVM) as a text classification technique (see also section 3.2.3). The classes are defined beforehand on the basis of (1) the two languages (i.e. English and German), and (2) the registers represented in the GECCo corpus. In the analysis we perform two classifications, analysing which cohesive categories and features contribute most to the distinctions of language and register.

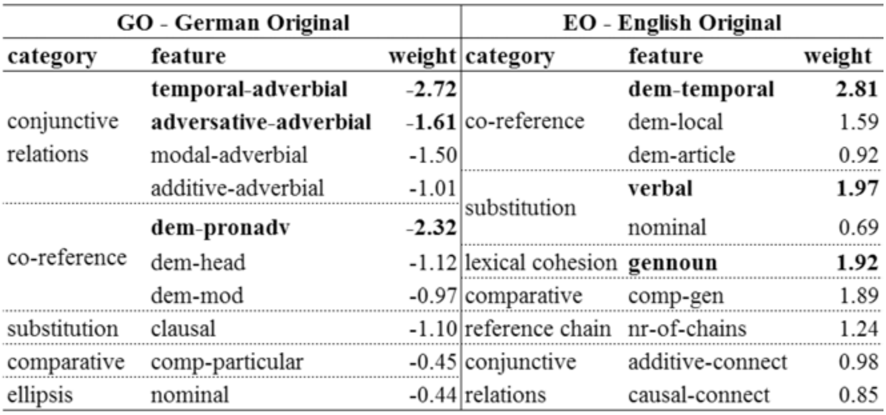
The classification by language achieves an overall accuracy of 98.96%. Table 8 shows that for both languages the accuracies as well as the F-measure are very high, i.e. the languages are quite well distinguished from one another. This matches the findings obtained by correspondence analysis (section 4.1.1).

Table 8: Accuracy and F-Measure for the language distinction

language	accuracy in %	F-Measure
GO	97.90	0.99
EO	100.00	0.99

Table 9 shows the top 10 most distinctive features for each language according to the cohesive category.[8]

Table 9: Top 10 features for German and English



GO - German Original			EO - English Original		
category	feature	weight	category	feature	weight
conjunctive relations	**temporal-adverbial**	**-2.72**	co-reference	**dem-temporal**	**2.81**
	adversative-adverbial	**-1.61**		dem-local	1.59
	modal-adverbial	-1.50		dem-article	0.92
	additive-adverbial	-1.01	substitution	**verbal**	**1.97**
co-reference	**dem-pronadv**	**-2.32**		nominal	0.69
	dem-head	-1.12	lexical cohesion	**gennoun**	**1.92**
	dem-mod	-0.97	comparative	comp-gen	1.89
substitution	clausal	-1.10	reference chain	nr-of-chains	1.24
comparative	comp-particular	-0.45	conjunctive relations	additive-connect	0.98
ellipsis	nominal	-0.44		causal-connect	0.85

[8] Rather than by accuracy (also known as true positives or recall), we ranked the registers by F-Measure as it accounts not only for recall (correctly classified texts) but also for precision (misclassified texts into that register).

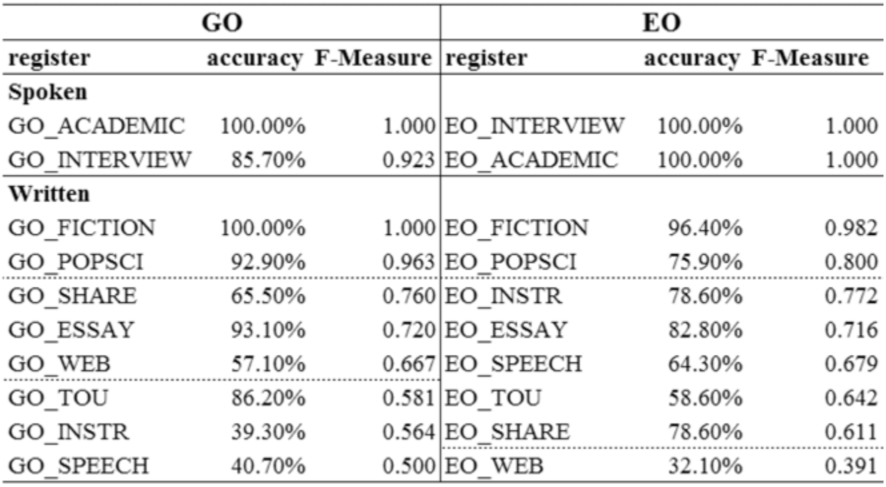
For features of conjunctive relations, German typically uses conjunctive adverbials (temporal, adversative, modal, and additive) while English uses connects (additive, causal). In terms of coreference, German and English differ in the use of demonstrative reference (GO dem-pronadv, EO dem-temporal). Additionally, while German uses comparatives of the particular type, English uses comparatives of the general type. In terms of substitution, clausal substitution is typical for German and verbal substitution for English (see section 5 for examples).

In summary, German and English are quite well distinguished by cohesive features, showing the biggest differences in the preference of (a) conjunctive adverbials and demonstrative reference for German, and (b) temporal demonstratives, in particular, as well as verbal substitution and general nouns associated with lexical cohesion for English (see features in bold in Table 9).

For the classification of registers by cohesion features, we achieve an overall accuracy of 76.46%, which is clearly lower than for the distinction by language. Thus, registers show some differences, but are less well distinguishable from one another. This is again in line with the results of the correspondence analysis in section 4.2. However, while correspondence analysis is an unsupervised method, in text classification, which is a supervised method, we impose the classes (in this case the register classes), getting more insights into the distinction of registers and which features contribute best to this distinction.

Table 10 shows the registers for each language ranked by F-Measure. We observe that the spoken registers for both languages show relatively high F-Measures (0.92-1.0), being very well distinguished by cohesion features from the

Table 10: Accuracy and F-Measure for German and English registers (ranked by F-Measure)

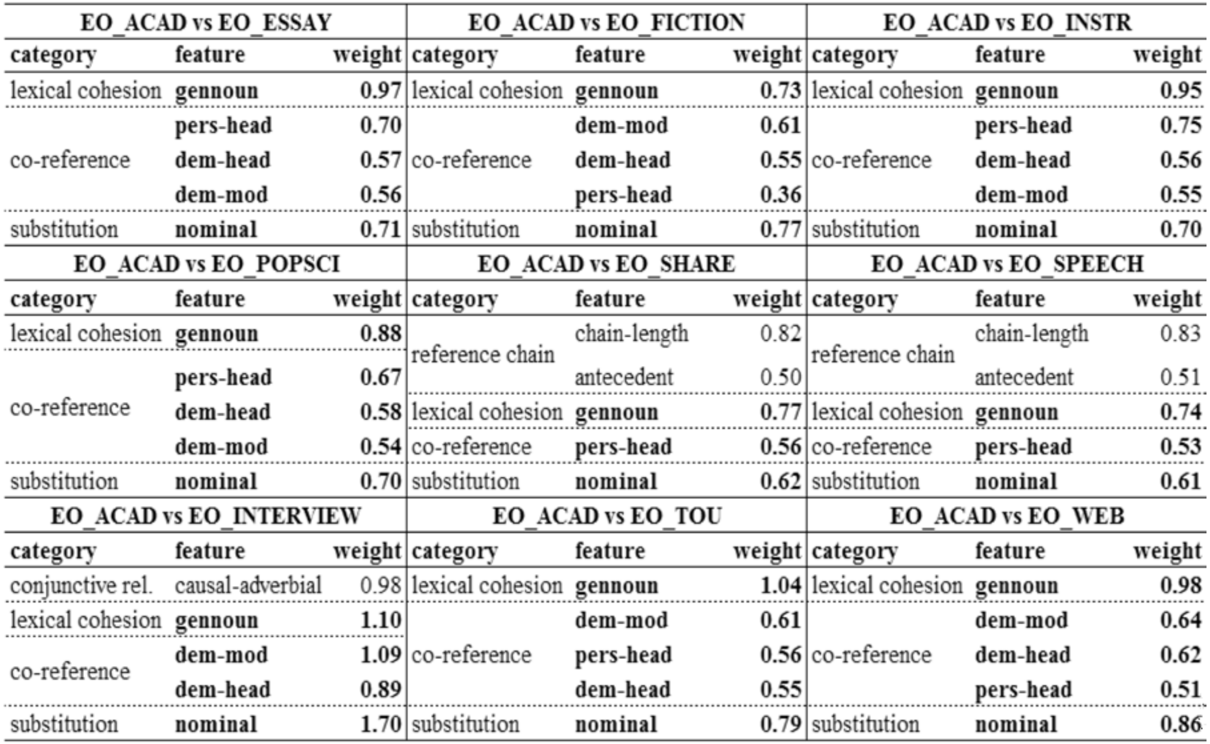


GO			EO		
register	**accuracy**	**F-Measure**	**register**	**accuracy**	**F-Measure**
Spoken					
GO_ACADEMIC	100.00%	1.000	EO_INTERVIEW	100.00%	1.000
GO_INTERVIEW	85.70%	0.923	EO_ACADEMIC	100.00%	1.000
Written					
GO_FICTION	100.00%	1.000	EO_FICTION	96.40%	0.982
GO_POPSCI	92.90%	0.963	EO_POPSCI	75.90%	0.800
GO_SHARE	65.50%	0.760	EO_INSTR	78.60%	0.772
GO_ESSAY	93.10%	0.720	EO_ESSAY	82.80%	0.716
GO_WEB	57.10%	0.667	EO_SPEECH	64.30%	0.679
GO_TOU	86.20%	0.581	EO_TOU	58.60%	0.642
GO_INSTR	39.30%	0.564	EO_SHARE	78.60%	0.611
GO_SPEECH	40.70%	0.500	EO_WEB	32.10%	0.391

Table 11: Top 5 distinctive features for all GO_ACAD binary classification pairs

GO_ACAD vs GO_ESSAY			GO_ACAD vs GO_FICTION			GO_ACAD vs GO_INSTR		
category	**feature**	**weight**	**category**	**feature**	**weight**	**category**	**feature**	**weight**
conjunctive relations	**modal-adverbial**	**1.32**	conjunctive relations	**modal-adverbial**	**1.54**	conjunctive relations	**modal-adverbial**	**1.01**
	additive-adverbial	**0.79**		**additive-adverbial**	**0.86**		adversative-adverbial	0.75
	adversative-connect	0.67	co-reference	dem-local	0.59		**additive-adverbial**	**0.73**
	temporal-adverbial	0.66	substitution	**verbal**	**0.65**		adversative-connect	0.61
co-reference	dem-local	0.51		clausal	0.55	substitution	**verbal**	**0.57**
GO_ACAD vs GO_POPSCI			**GO_ACAD vs GO_SHARE**			**GO_ACAD vs GO_SPEECH**		
category	**feature**	**weight**	**category**	**feature**	**weight**	**category**	**feature**	**weight**
conjunctive relations	**modal-adverbial**	**1.47**	conjunctive relations	**modal-adverbial**	**1.09**	conjunctive relations	**modal-adverbial**	**1.04**
	adversative-connect	0.57		adversative-connect	0.72		temporal-adverbial	0.69
co-reference	dem-local	0.50		temporal-adverbial	0.54		additive-connect	0.61
substitution	clausal	0.62		**additive-adverbial**	**0.40**		**additive-adverbial**	**0.60**
	verbal	**0.74**	substitution	**verbal**	**0.48**	substitution	**verbal**	**0.66**
GO_ACAD vs GO_INTERVIEW			**GO_ACAD vs GO_TOU**			**GO_ACAD vs GO_WEB**		
category	**feature**	**weight**	**category**	**feature**	**weight**	**category**	**feature**	**weight**
conjunctive relations	**modal-adverbial**	**1.77**	conjunctive relations	**modal-adverbial**	**0.83**	conjunctive relations	**modal-adverbial**	**0.84**
	adversative-adverbial	1.67		temporal-adverbial	0.61		adversative-connect	0.67
co-reference	dem-pronadv	1.10		additive-connect	0.53		**additive-adverbial**	**0.66**
	dem-mod	1.08	co-reference	dem-mod	0.54		additive-connect	0.52
substitution	nominal	0.64	substitution	**verbal**	**0.55**	substitution	**verbal**	**0.51**

Table 12: Top 5 distinctive features for all EO_ACAD binary classification pairs



EO_ACAD vs EO_ESSAY			EO_ACAD vs EO_FICTION			EO_ACAD vs EO_INSTR		
category	**feature**	**weight**	**category**	**feature**	**weight**	**category**	**feature**	**weight**
lexical cohesion	**gennoun**	**0.97**	lexical cohesion	**gennoun**	**0.73**	lexical cohesion	**gennoun**	**0.95**
co-reference	**pers-head**	**0.70**	co-reference	**dem-mod**	**0.61**	co-reference	**pers-head**	**0.75**
	dem-head	**0.57**		**dem-head**	**0.55**		**dem-head**	**0.56**
	dem-mod	**0.56**		**pers-head**	**0.36**		**dem-mod**	**0.55**
substitution	**nominal**	**0.71**	substitution	**nominal**	**0.77**	substitution	**nominal**	**0.70**
EO_ACAD vs EO_POPSCI			**EO_ACAD vs EO_SHARE**			**EO_ACAD vs EO_SPEECH**		
category	**feature**	**weight**	**category**	**feature**	**weight**	**category**	**feature**	**weight**
lexical cohesion	**gennoun**	**0.88**	reference chain	chain-length	0.82	reference chain	chain-length	0.83
co-reference	**pers-head**	**0.67**		antecedent	0.50		antecedent	0.51
	dem-head	**0.58**	lexical cohesion	**gennoun**	**0.77**	lexical cohesion	**gennoun**	**0.74**
	dem-mod	**0.54**	co-reference	**pers-head**	**0.56**	co-reference	**pers-head**	**0.53**
substitution	**nominal**	**0.70**	substitution	**nominal**	**0.62**	substitution	**nominal**	**0.61**
EO_ACAD vs EO_INTERVIEW			**EO_ACAD vs EO_TOU**			**EO_ACAD vs EO_WEB**		
category	**feature**	**weight**	**category**	**feature**	**weight**	**category**	**feature**	**weight**
conjunctive rel.	causal-adverbial	0.98	lexical cohesion	**gennoun**	**1.04**	lexical cohesion	**gennoun**	**0.98**
lexical cohesion	**gennoun**	**1.10**	co-reference	**dem-mod**	**0.61**	co-reference	**dem-mod**	**0.64**
co-reference	**dem-mod**	**1.09**		**pers-head**	**0.56**		**dem-head**	**0.62**
	dem-head	**0.89**		**dem-head**	**0.55**		**pers-head**	**0.51**
substitution	**nominal**	**1.70**	substitution	**nominal**	**0.79**	substitution	**nominal**	**0.86**

other registers. Considering the written registers, fiction is best distinguished from the other registers for both languages (F-Measure of 1 for German and 0.98 for English). Popular scientific texts (POPSCI) are also quite well distinguished, in particular for German (F-Measure of 0.96 for German and 0.80 for English). The other written registers are less well distinguished; most obviously when considering the F-Measure, even though they show relatively high accuracies (consider, e.g., GO_TOU with an accuracy of 86.20% but an F-Measure of 0.58).

In the following, as an example, we present in more detail the contribution of features for the German and English ACADEMIC register, as it shows the best classification results. Tables 11 and 12 show the top 5 distinctive features for GO_ACADEMIC and EO_ACADEMIC, respectively. Focusing on feature categories, we can see that features of conjunctive relations prevail for GO_ACADEMIC (see Table 11), while coreference features prevail for EO_ACADEMIC (see Table 12) in the top 5. By inspecting more closely the top 5 features across all register pairs of ACADEMIC and another register for each language, we can detect some more fine-grained differences.

Table 11 (boldfaced features) shows that GO_ACADEMIC is distinguished from the other registers by adverbial modal conjunctions (present in all pairs), verbal substitution (in 7 pairs), and adverbial additive conjunctions (in 6 pairs). Note that modal conjunctions are more distinctive than additive ones or verbal substitution by the sum of SVM weights of all pairs (compare 10.9, 5.0, and 4.9, respectively).

From Table 12, we can see that EO_ACADEMIC is distinguished by lexical cohesion (9 pairs, 8.2 weight), nominal substitution (9 pairs, 7.4 weight) as well as by demonstrative reference (dem-mod: 7 pairs, 5.5 weight; dem-head: 7 pairs, 4.6 weight) and to some extent by personal reference (8 pairs, 4.6 weight) (see section 5 for examples).

In summary, GO_ACADEMIC and EO_ACADEMIC are very well distinguished from the other registers. Moreover, they distinctively use different cohesive features, which clearly reflect the language differences, where conjunctive relations are mostly distinctive for German and coreference mostly for English.

5 Interpretation of findings and implications for translation

In this final section, we attempt an overall interpretation of the statistical results that have been discussed in section 4 with respect to the four research questions addressed in the introduction (section 1). Our focus here is on identifying contrasts in cohesion in terms of three variables: language (English vs. German), mode of production (written vs. spoken) and register. As explained in section 1, the mode of production *translation vs. original* will not be foregrounded here in terms of translationese. Rather, we will take the contrastive interpretations for each research question as a starting point for suggestions with regard to preferred translation strategies, to some extent, also with a view to interpreting.

Let us begin with the first research question, which is concerned with contrasts in the degree of cohesion. As the discussion in section 1 suggests, this question does not consider the relation between different cohesive features (as our other research questions do), but relates these features to other textual ranks. Cohesive devices serve as explicit indicators of textual relations to other linguistic expressions, beyond the level of grammar. The degree of cohesion can be calculated if we take the frequencies of all cohesive devices together, and relate them to the total number of tokens, as presented in section 4.1, Tables 3 and 4. The frequencies show that although there are only moderate differences in the distributions, language contrast between the main corpora EO and GO (all registers per language taken together) is significant. Some assumptions in the literature, assuming a preference for more explicit strategies in German as compared to English[9], which have been attested for lexicogrammar by other studies (e.g. Hawkins 1986), therefore seem to be corroborated for the level of cohesion. Even more pronounced contrasts surface if we compare language contrasts per register. While similar distributions (no significant contrasts) or slightly significant differences are found in both languages for most written registers, contrasts between English and German in the registers POPSCI, FICTION, INTERVIEW and ACADEMIC are highly significant – although the direction of the differences is not the same across the board (e.g. FICTION in Table 3). So except for the registers popular scientific and fictional texts, language contrasts between written registers (hence translation relevant registers) are less marked than in the spoken registers. Therefore, the general translation method in terms of degree of cohesion for translators would be to keep the overall number of cohesive devices at a similar level as in the source texts, or to slightly increase it when translating from English into German. Our data shows however, that language contrasts in these and all other registers exist in the features used for creating cohesion (see below). Marked differences for the popular science texts suggest that translators should use cohesive devices more extensively when translating scientific texts

[9] But note that other studies and interpretations have convincingly argued that over-generalizations in Hawkins 1986, as well as the role of word order for explicitness of encoding, ignore the counterbalance achieved by important properties of the system of English (e.g. König and Gast 2012, Fischer 2013).

from English into German. This requires verification by integrating written academic texts into the corpus in the future. FICTION is one notable exception to the overall tendency as the frequencies in German are lower than those in English. Our analyses may have to be combined with qualitative studies in the future to see whether strategies for literary translations should differ from those applied to non-fictional and specialized texts. The information that the differences between registers are more pronounced in German than in English, (mostly due to the higher frequencies in the spoken registers than in the written registers) – might be relevant for interpreting. We assume that register variation in the degree of cohesion may even be more marked in German than English, when integrating data from more spoken registers.

The interpretation of data in terms of strength of relation is partially drawn from Figure 1, permitting a comparison of general distributions of cohesive forms, from correspondence analyses, as shown in Figures 2 and 3, and text classification as shown in table 9. One relevant aspect for this parameter is whether cohesive relations are indicated by explicit and specific vs. weakly specified cohesive devices. Devices of ellipsis and substitution generally are considered to create weaker semantic ties than reference and lexical cohesion not only because of structural reduction but also because the specific semantic relation between referents is often less clear. Stronger relations of identity are created by referential modifiers as these are combined with devices of lexical cohesion, vaguer and/or ambiguous relations are established by referential pronouns, and the neuter pronoun *it*, in particular. Moreover, demonstratives serve as focus lifters and therefore are more explicit than devices of personal reference. Stronger relations of comparison are created by comparative reference, vaguer relations by substitution and ellipsis. Conjunctive adverbials indicate logico-semantic relations more explicitly than coordinating conjunctions (connects) since they usually contain more information in terms of the specific meaning relations established. General nouns establish less explicit sense relations than other types of lexical cohesion, e.g. repetition.

The data evaluated so far points to considerable differences between English and German in the explicit creation of logico-semantic relations. German prefers usage of adverbials and English mainly employs coordinating conjunctions (compare examples 22 vs. 23).

(22) *Aufgrund der veränderten Sicherheitslage konnte die Mannschaftsstärke um 40 Prozent reduziert werden.* ***Außerdem*** *wurden, […], knapp 11000 ehemalige Soldaten der Nationalen Volksarmee der DDR in die nun gesamtdeutsche Bundeswehr integriert.* [GO_ESSAY]

The new security situation made it possible to reduce personnel by 40 %. ***Furthermore****, almost 11,000 former soldiers from the GDR's National People's Army (NVA), [. . .] , were integrated into the new all-German Bundeswehr.* [ETRANS_ESSAY]

(23) *Mr. Bush has time and again demonstrated his commitment to open trade.* ***And*** *he is determined to extend the benefits of open markets to the world's poorer nations.* [EO_ESSAY]

Distributions of ellipsis are equally low in both languages. Moreover, they do not belong to the top distinctive features and therefore are no indicators of variation in strength of relation, from the perspective of language contrast. We note, however, that German generally shows a higher tendency towards comparative reference than English (especially towards comparative particular) in the written registers, which points to a more explicit marking of relations of type reference. The fact that clausal substitution is a distinctive feature of German has to be interpreted against the background that comparative reference plays a greater role for the overall creation of type reference/comparison than substitution or ellipsis in German relative to English. In addition, German shows a preference for demonstratives, hence focusing devices, for creating identity of reference (see Figure 2). For instance, consider examples (1) and (2) above, where German favours a demonstrative pronoun (e.g. *das*) or a demonstrative deictic (e.g. *damit*) over the neuter pronoun *It*, which is vaguer with respect to the (scope) of the antecedent.

While individual types of demonstratives (local and temporal) are distinctive features for coreference in English, the overall distributions for demonstratives are lower in all registers, whereas distributions for the two types of personal reference are generally higher than in German.[10] Hence, German seems to be more explicit with respect to the creation of identity of reference.

In English, we find a clear preference for general nouns. This may point to a tendency in English towards creating weak relations of similarity by lexical cohesion. Whether general nouns here are indeed an indicator of weak relations has to be seen when evaluating data about other types of cohesion such as synonymy, meronymy, etc.

So the overall assumption that German tends towards more explicit cohesive devices and thus creates semantically stronger relations seems to be confirmed along most dimensions in our data. The findings suggest that translators should

[10] Some of these differences are due to contrasts in biological gender marking in English vs. grammatical gender in German.

use more explicit devices in their target texts than in the source texts when translating from English into German. For instance, additive or adversative connects should often be transferred to adverbials in translations. Demonstrative pronouns should be used more often instead of personal pronouns (e.g. *dies/das* instead of *es/it*) and demonstrative modifiers in combination with types of lexical cohesion such as synonymy more often than the definite article in combination with a general noun. And finally, comparative particular should occur more often than comparative general or substitution. Opposite strategies should be favoured when translating from German into English.

Furthermore, we observe a stronger variation in the strength of cohesive devices between German registers than English registers (see Figures 2 and 3). We note a preference for substitution pointing to weaker relations in spoken than written German, and also a very low distribution of comparative reference in GO_ACADEMIC and GO_INTERVIEW (see Figure 1). German spoken registers are also marked in the heavy use of demonstrative pronouns and a lower distribution of demonstrative modifiers (combined with lexical cohesion). This may correlate with a preference to summarize large textual passages but needs to be tested by examining the textual scope of coreferential antecedents. The tendency towards substitution and also elliptical constructions reflects weaker relations in spoken than written English. Hence altogether, we observe that there is a tendency in both languages for spoken registers to use more cohesive devices (as shown in Table 1) but at the same time these devices are less explicit than in the written registers. This again may be relevant information for interpreters and also translators, who additionally have to be aware of the fact that different features are responsible for the variation in mode in German as compared to English.

In the introduction, we mentioned features in chains as another indicator of strength of cohesive relations. A high number of elements in one coreference chain reflects a stronger relation of identity than a chain with a low number of elements. By contrast a high number of different coreference chains points to high thematic variation. It may contribute to weakening the relation of individual coreference chains if a high number of intervening referring expressions belonging to different coreference chains are between elements of the same chain. Differences in terms of chains can only be drawn from correspondence analysis, see Figures 2 and 3, partially visible only in numeric data output, and information about number of chains as a distinctive feature also from text classification (Table 9). They have to be examined in combination with other statistical methods and related to information about distance between chain elements in the future. In English, we note a higher number of different coreference chains (see Table 9, where number of chains is a distinctive feature of

English) and a lower number of elements per coreference chain than in German (see Figure 2, where the arrow for chain length points more to the right side, hence marking German). This points to a higher variation with respect to the referents that are taken up in English, hence more different referents seem to be important in the textual world. By contrast textual continuity seems to be favoured in German as we find a lower number of different coreference chains per text and also a higher number of coreferring expressions in one reference chain. Taken together with a higher distribution of coreferential devices in general and demonstrative coreference in particular, our current data seems to point again to stronger preference for identity of reference and for marking these relations more explicitly in coreference chains in German than English. This may be counterbalanced by more variation in German in the sense relations for creating lexical cohesion. Generally speaking, translation strategies English to German for coreference chains may generally imply using a higher number of coreferring expressions while local relations, which may only consist of two elements, may drop out, e.g. because of remetaphorization. By contrast translation from German into English may contain more short reference chains, which may be instantiated because of a less metaphorical realization of meaning relations in the translations, as compared to German originals.

We now discuss our corpus data in terms of the semantic meaning relations that are preferably expressed by cohesion (question (3) in our introduction). The information is mainly drawn from the frequency distributions shown in Figure 1, section 4.2.1, and text classification analysis (Tables 9, 11 and 12, section 4.2.3). Summarizing our observations above, the most important types of cohesion – in written and spoken registers of both languages – are coreference and conjunction, and fewer relations are indicated by ellipsis, substitution and general nouns. Hence, logico-semantic relations and relations of identity between referents seem to prevail over comparisons of different entities of the same referential type (type reference). Yet, general nouns is the only subtype of lexical cohesion that is considered in this study. We expect that sense relations/relations of similarity between different types of referents will play a far greater role once we have included other types of lexical cohesion, e.g. repetition, hyperonymy and meronymy, into our data[11].

Apart from these commonalities, we observe language-specific contrasts in the meaning relations signalled by cohesion. Again the differences in German between registers are more pronounced than those within English. First of all,

[11] It would be interesting to see in future works whether these preferences also hold in other and typologically more diverse languages.

a stronger tendency is attested for German towards explicitly expressing logico-semantic relations via conjunctive relations (particularly conjunctive adverbials) on the textual level, and thus a stronger tendency towards relating propositions rather than entities. This may have to do with the types of logico-semantic relations. The results drawn from text classification (see Table 9) show that relations of time and contrast are distinctive for German (see examples 24 and 25).

(24) *Die Forschungsarbeit von Maria Reiche in der Wüste von Nazca beginnt erst 1946 nach dem Ende des Krieges.* ***Solange*** *ist es ihr verboten, als Deutsche das Stadtgebiet von Lima zu verlassen.* [GO_POPSCI]

Maria Reiche should confirm this theory. She started her research work in the dessert of Nazca in 1946, because her wasn't allowed [sic!] to leave the city Lima until the end of the Second World War. [ETRANS_POPSCI]

(25) *[...]. ich wollte was machen, was ich ins Familienleben integrieren kann und zu Hause machen kann und wo ich* ***trotzdem*** *mit Menschen zu tun hab.* [GO_INTERVIEW]

'[...] I wanted to do something I could integrate into family life and at the same time work with people'

By contrast, the most important logico-semantic relations that are overtly expressed by conjunctive devices in English are addition (extensive use of *and)* and cause (*because*, *therefore*).

Creating identity between referents seems to play a major role in both languages, but seems to be even more important for the overall cohesiveness of texts in English than German. Furthermore, we see a semantic contrast in terms of the referents that are taken up in the textual world. Relations of time reference and location (see 26 below) seem to be characteristic for English, while German favours identity between abstract entities, evoked by pronominal adverbs (see 27).

(26) *President Bush has clearly articulated that the United States will "lead by example". We have a destination. To get* ***there****, we need to turn our attention towards implementation.* [EO_SPEECH]

(27) *Andererseits müssen wir den Dialog gerade mit der islamischen Welt verstärken und intensivieren.* ***Dabei*** *geht es darum – bei allem Respekt für unterschiedliche Traditionen – die sämtlichen Weltkulturen gemeinsamen Werte sichtbar zu machen.* ***Dazu*** *gehört auch das unzweideutige Eintreten für die Menschenrechte.* [GO_ESSAY]

On the other hand, we must strengthen and intensify dialogue with the Islamic world. The important thing ***here*** *– with all due respect for different traditions – is to reveal the common values shared by all world cultures.* ***This*** *also includes unambiguous support for human rights*. [ETRANS_ESSAY]

Type reference – comparisons of individual referents of the same class – seems to be equally relevant for English and German (all register taken together), if we consider the overall distributions in Figure 1. However, the semantic relation of comparison is indicated by different lexicogrammatical patterns. As shown in Table 8, adjectives of general comparison and substitution, in particular verbal substitution, are distinctive features in English (see examples (28) and (29)).

(28) *Although we believe privatization is beneficial, it is not part of the trade negotiations. Countries will continue to make* ***such*** *decisions for themselves*. [EO_ESSAY]

(29) *They also tend to brood or ruminate more than satisficers* ***do***. [EO_POPSCI]

Adjectives of particular comparison (see example (4) above), clausal substitution, nominal ellipsis, and also clausal ellipsis (as in example 30) are preferred over other features in German:

(30) *Frankreich und Großbritannien verfügen über starke eigene sicherheitspolitische Fähigkeiten, Deutschland* ***[ ]*** *nicht* ***[ ]***. [GO_SPEECH]

France and Great Britain have strong security policy capabilities of their own. Germany does not. [ETRANS_SPEECH]

One feature we identify as a distinctive feature of English is the use of general nouns that are combined with the definite article, and less often, with demonstrative modifiers. Quite often, they take up longer stretches of text, as in (31).

(31) *If you believe, as I do strongly, that Britain must occupy a central place in Europe's decision-making, it is highly regrettable that this* ***argument*** *has been so grossly distorted in the British political debate in recent years.* [EO_SPEECH]

General nouns plus demonstrative reference therefore are an important means to create identity between referents in English but not in German data.

These contrasts between English and German in the meaning relations preferred to establish cohesion seem to be especially relevant for translation. Yet, suggestions in terms of conscious translations strategies when translating from one language into the other depend on the global translation method, and on which linguistic level translational equivalence should be attempted. An overt translation would probably be reflected on the level of cohesion by a transfer of meaning relations from source to target text. This may not necessarily involve an imitation of cohesive forms, but may imply maintaining the general distribution of semantic relations of identity of reference, type reference, logico-semantic relations and similarity. Covert translation strategies would imply adhering to the target language conventions with respect to preferences in the meaning relations expressed by cohesion. For instance these may result in:

- Expressing more relations between propositions explicitly by cohesive conjunction, in particular logico-semantic relations of time and contrast.
- Relating abstract entities by demonstratives.

when translating from English into German.

Covert translations from German into English may contain shifts towards a more pronounced marking of:

- Logico-semantic relations of addition and cause.
- Identity relations of human entities, time and place.

It has to be noted, though, that the types of shifts between source and target texts often depend on the register.

For instance, the comparison of ACADEMIC in Tables 11 and 12 suggests that the register can be classified in German by a preference for modal conjunctions (as in 32), verbal substitution (33) and additive conjunction (34).

(32) ***Also*** *Sie sehen, Terminologie ist sehr wichtig.* (conjunction modal-adverbial) [GO_ACADEMIC]

‘So you see, terminology is quite important.’

(33) *[…] ein Patient sich sein ähm noch intaktes Arbeitsgedächtnis zunutze machen kann, indem er die Buchstabenfolge einfach wiederholt, ja, und das […] vierzig Sekunden lang* ***tun*** *kann.* (verbal substitution) [GO_ACADEMIC]

‘[…] a patient can ehm use his still intact working memory by just repeating the word sequence, yes, and doing this […] for forty seconds.’

(34) *Da kriegt der Student ein reales Gebilde,* ***beispielsweise*** *dieses Glas […]. (conjunction additive adverbial)* [GO_ACADEMIC]

‘So the student gets a real object, for example, this glass […].’

This points to a tendency in German academic texts towards providing explanation and towards hedging in combination with elaboration. By contrast

English academic texts are characterized (compared to other English registers) by general nouns (35), nominal substitution (36), demonstrative (37 and personal reference 38).

(35) *Differences between IM systems are smoothed over, making it easier to communicate. We follow this* ***principle*** *in the way contacts are organized.* [EO_ACADEMIC] (lexical cohesion)

(36) *[…] this is, probably the most important* ***one*** *in distinguishing liver atrophy and muscle atrophy.* [EO_ACADEMIC] (nominal substitution)

(37) *[…] today i'd like to talk about the solution to* ***this*** *problem.* [EO_ACADEMIC] (demonstrative modifier)

(38) *[…] so* ***this*** *is what's called a tile.* [EO_ACADEMIC] (demonstrative head)

(39) *in ninety-three* ***he*** *came to the University of Michigan as an associate professor.* [EO_ACADEMIC] (personal reference)

These clustered features reflect a tendency in English towards summarizing observations or facts and towards exemplifying and personalizing research.

We now attempt a summary of the above discussions in terms of breadth of variation, thus relating our research question (1) to (3) to research question (4), breadth of variation. So far we have observed that contrasts between English and German not only concern differences in terms of the number of cohesive devices employed to create cohesion, but also in the linguistic patterns for establishing particular relations via cohesion and in terms of which semantic relations are preferably expressed. For all contrasts discussed so far, and as correspondence analyses in 4.2 clearly show (Figures 2 and 3), German registers in general are more strongly differentiated than English ones in the use and clustering of cohesive features. We also see that both languages distinguish written and spoken mode, but again German more strongly than English (see Table 10). If we compare the results obtained by the different statistical methods, we additionally see that some written registers, especially in English, show similarities, in terms of degree of cohesion, strength of relation as well as meaning relations. For instance, Table 10 shows that political speeches, touristic texts and texts from the web are not well distinguished. Hence the assumptions raised in the literature, that German shows a higher breadth of variation between registers, text types or genres, are here strikingly confirmed for patterns of cohesion. This might mean for translators that greater attention to cohesive variation should be paid generally when translating from English into German than from German into English. However, again register peculiarities have to be taken into account.

6 Summary and conclusions

Our observations about contrasts in cohesion between English and German are relevant for translation studies as they clearly show that adequate translation strategies cannot rely on knowledge about contrastive lexicogrammar alone – which is, of course, not a new insight. However, awareness of contrasts in the systemic resources for creating cohesion, though somewhat less traditional, is only one further step. Translators and interpreters additionally need awareness of more and less preferred patterns that distinguish the languages, and in particular, the registers within them.

Firstly, our corpus-based findings suggest that cohesive patterns differ in English and German in the overall frequency of cohesive devices employed as well as in the strength with which cohesive relations are indicated. These contrasts concern different aspects relating to explicitness or specification: a) the extent to which meaning relations are expressed explicitly by cohesive strategies in relation to implicit strategies or strategies on other linguistic levels and b) if expressed, the variation in explicitness of the cohesive devices and chains. German tends towards more explicitness with respect to both aspects. This leads us to suggest that German translations of English originals should explicitate meaning relations by using a higher number of cohesive devices, by using a higher number of semantically specific devices such as demonstratives and conjunctive adverbials and a higher number of elements in coreference chains. Opposite translation strategies should be used in the opposite direction.

In addition, we note language contrasts in the type of meaning relations that are expressed, for instance a preference for logico-semantic relations in German and a preference for coreference relations and general nouns in English. Depending on which global translation method is adequate for a given register or for a particular skopos of translation, the translation process may involve bleaching out particular meaning relations on the level of cohesion and enforcing others. However, more comprehensive studies are needed, relating our observations on the level of cohesion to lexicogrammar and information structure in order to see which coherence relations, if marked explicitly, are indicated by cohesion, information structure or lexicogrammar and how these different levels interact.

Furthermore, our research highlights contrasts in the degree of variation between English and German registers. Most importantly, we were able to confirm the assumption that the breadth of variation between German (written and spoken) registers is greater than between English registers, for the level of cohesion. Our study therefore supports other contrastive approaches on the level of lexicogrammar and calls for increased awareness of intralingual as well as interlingual register variation by translators and interpreters.

In future research, we hope to provide insights that have implications for research on translationese, by adding corpus-based findings about contrasts in mode, i.e. translations vs. originals. Properties of translated texts – like textual properties in general – are at the core of "textuality", and one of the most immediate manifestations of textuality are cohesive constellations permeating all texts and discourses. Cohesion is thus an immediate target for research into textuality of translated or interpreted texts, and as a large part of the GECCo-corpora, the CroCo subcorpora, have originals and their translations, the road into that kind of research is open.

References

Argamon, Shlomo, Jeff Dodick & Paul Chase. 2008. Language use reflects scientific methodology: A corpus-based study of peer-reviewed journal articles. *Scientometrics* 75(2). 203–238.

Ariel, Mira. 1990. *Accessing Noun-Phrase Antecedents*. London: Routledge.

Ariel, Mira. 2001. Accessibility theory: An overview. In Ted Sanders, Joost Schilperoord & Wilbert Spooren (eds.), *Text representation: Linguistic and psycholinguistic aspects*, 29–88. Amsterdam: John Benjamins.

Baayen, Harald. 2008. *Analyzing linguistic data. A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Baker, Mona. 1992. *In other words: A course book on translation*. London: Routledge.

Baker, Mona. 1993. Corpus linguistics and translation studies – implications and applications. In Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 233–250. Amsterdam / Philadelphia: John Benjamins.

Beaugrande, Robert-Alain de & Wolfgang Ulrich Dressler. 1981. *Einführung in die Textlinguistik.* Tübingen: Niemeyer.

Becher, Viktor. 2011. *Explicitation and implicitation in translation. A corpus-based study of English-German translations of business texts*. Hamburg: Hamburg University doctoral dissertation.

Biber, Douglas. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast* 14(1). 7–34.

Blum-Kulka, Shoshana. 1986. Shifts of cohesion and coherence in translation. In Juliane House & Shoshana Blum-Kulka (eds.), *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language acquisition Studies*, 17–35. Tübingen: Narr.

Carston, Robyn. 2002. *Thoughts and utterances: The pragmatics of explicit communication*. Oxford: Blackwell.

Collins, Peter. 2012. Grammatical variation in English worldwide: The role of colloquialization. *Linguistics and the Human Sciences* 8(3). 289–306.

CWB. 2010. The IMS Open Corpus Workbench. http://www.cwb.sourceforge.net.

Degaetano-Ortlieb, Stefania, Peter Fankhauser, Hannah Kermes, Ekaterina Lapshinova-Koltunski, Noam Ordan & Elke Teich. 2014. Data mining with shallow vs. linguistic features to study diversification of scientific registers. Paper presented at the 9th edition of the *Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, 26–31 May.

Doherty, Monika. 2002. *Language processing in discourse. A key to felicitous translation*. London: Routledge.

Doherty, Monika. 2006. *Structural propensities: Translating nominal word groups from English into German*. Amsterdam: John Benjamins.

Evert, Stefan. 2005. *The CQP query language tutorial*. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart. CWB version 2.2.b90.

Fabricius-Hansen, Cathrine. 1996. Informational density: A problem for translation theory. *Linguistics* 34. 521–565.

Fischer, Klaus. 2013. *Satzstrukturen im Deutschen und Englischen. Typologie und Textrealisierung*. Berlin: Akademie Verlag 2013.

Fox, Neal, Omran Ehmoda & Eugene Charniak. 2012. Statistical stylometrics and the Marlowe – Shakespeare authorship debate. Paper presented at *The Georgetown University Roundtable on Language and Linguistics (GURT)*, Washington D.C, USA.

Greenacre, Michael. 2010. *Correspondence analysis in practice*. CRC Press.

Grosz, Barbara J., Aravind K. Joshi & Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics* 21. 203–225.

Gundel, Jeanette, Nancy Hedberg & Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2), 274–307.

Halliday, Michael A. K. 1989. *Spoken and written language*. Oxford: Oxford University Press.

Halliday, Michael A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London/New York: Longman.

Halliday Michael A. K. & Christian Matthiessen. 2013. *Halliday's introduction to functional grammar*. London: Routledge.

Hansen-Schirra, Silvia, Stella Neumann & Erich Steiner. 2007. Cohesion and explicitation in an English-German Translation Corpus. *Languages in Contrast* 7(2). 241–265.

Hansen-Schirra, Silvia, Stella Neumann & Erich Steiner. 2012. *Cross-linguistic corpora for the study of translations. Insights from the language pair English – German*. Berlin/New York: Mouton de Gruyter.

Hatim, Basil & Ian Mason. 1990. *Discourse and the Translator*. London: Longman.

Hawkins, John A. 1986. *A comparative typology of English and German. Unifying the contrasts*. London: Croom Helm.

House, Juliane. 1997. *Translation quality assessment*. Tübingen: Narr.

House, Juliane. 2004. Explicitness in discourse across languages. In Juliane House, Werner Koller & Klaus Schubert (eds.), *Neue Perspektiven in der Übersetzungs- und Dolmetschwissenschaft*, 185–208. Bochum: AKS. 185–208.

Jenset, Gard B. & Barbara McGillivray. 2012. Multivariate analyses of affix productivity in translated English. In Michael Oakes & Meng Ji (eds.), *Quantitative methods in Corpus-Based Translation Studies*, 301–324. Amsterdam/Philadelphia: John Benjamins.

Joachims, Thorsten. 1998. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning (ECML98)*, 137–142.

Kaiser, Elsi. 2003. Word order, grammatical function, and referential form: On the patterns of anaphoric reference in Finnish. *Nordlyd* 31(1), 245–260.

Kermes, Hannah & Stefan Evert. 2002. YAC – A recursive chunker for unrestricted German text. In Manuel Gonzalez Rodriguez & Carmen Paz Suarez Araujo (eds.), *Third International Conference on Language Resources and Evaluation*, 1805–1812.

Kermes, Hannah. 2003. *Off-line (and On-line) Text analysis for computational lexicography*. Stuttgart: Universität Stuttgart doctoral dissertation.

Lapshinova-Koltunski, Ekaterina & Kerstin Kunz. 2014. Detecting cohesion: Semi-automatic annotation procedures. Paper presented at *Corpus Linguistics*, Lancaster, UK, July.

Kerstin Kunz & Marilisa Amoia. 2012. *Compiling a multilingual corpus*. In Heliana Mello & Massimo Pettorino (eds.), *VIIth GSCP International Conference: Speech and Corpora*, 29–34. Firenze: Firenze University Press.

Klein, Dan & Christopher D. Manning. 2003. Accurate unlexicalized parsing. *41st Annual Meeting on Association for Computational Linguistics*, 423–430. Stroudsburg, PA, USA: Association for Computational Linguistics.

König, Ekkehard. (2015). Manner deixis as source of grammatical markers in Indo-European languages. In Carlotta Viti (ed.). *Perspectives on historical syntax*, 33–60. John Benjamins.

König, Ekkehard & Volker Gast. 2012. *Understanding English-German contrasts. Grundlagen der Anglistik und Amerikanistik* (3rd edn.). Berlin: Erich Schmidt Verlag.

Königs, Karin. 2011. *Übersetzen Englisch – Deutsch. Ein systemischer Ansatz*. München: Oldenbourg Verlag.

Kunz, Kerstin. 2010. *Variation in English and German Coreference*. Frankfurt a. M.: Peter Lang.

Kunz, Kerstin & Erich Steiner. 2012. Towards a comparison of cohesive reference in English and German: System and text. In Maite Taboada, Susanna Doval Suárez & Elsa González Álvarez (eds.), *contrastive discourse analysis. Functional and corpus perspectives*, 208–239. London: Equinox.

Kunz, Kerstin & Erich Steiner. 2013. Cohesive substitution in English and German: A contrastive and corpus-based perspective. In Karin Aijmer & Bengt Altenberg (eds.), *Advances in Corpus-Based Contrastive Linguistics. Studies in honour of Stig Johansson*. Amsterdam: John Benjamins.

Kunz, Kerstin & Ekaterina Lapshinova-Koltunski. 2014. Cohesive conjunctions in English and German: Systemic contrasts and textual differences. In Lieven Vandelanotte, Kristin Davidse, Caroline Gentens & Ditte Kimps (eds.), *Recent advances in Corpus Linguistics: Developing and exploiting corpora* (Language and computers – studies in practical linguistics 78), 229–262. Amsterdam/New York: Rodopi.

Lambrecht, Knud. 1994. *Information structure and sentence form. Topic, focus and the mental representation of discourse referents*. Cambridge: Cambridge University Press.

Leech, Geoffrey, Marianne Hundt, Christiane Mair & Nicholas Smith. 2009. *Change in contemporary English. A grammatical study*. Cambridge: Cambridge University Press.

Levy, Roger & T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. Paper presented at the 20th *Conference on Neural Information Processing Systems* (NIPS).

Mair, Christian. 2006. *Twentieth-century English. History, variation and standardization*. Cambridge: Cambridge University Press.

Manning, Christopher D., Prabhakar Raghavan & Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

Menzel, Katrin. 2014. Ellipsen als Stil- und Kohäsionsmittel in deutschen und englischen politischen Reden. In Torsten Leuschner & Maria Koliopoulou (eds.), *Germanistische Mitteilungen. Zeitschrift für Deutsche Sprache, Literatur and Kultur* 40(1). 31–50.

Müller, Christoph & Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn & Joybrato Mukherjee (eds.), *Corpus technology and* *language pedagogy: New resources, new tools, new methods*, 197–214. Frankfurt a. M.: Peter Lang.

Nenadic, Oleg & Michael Greenacre. 2007. Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software* 20(3). 1–13.

Powers, David M. W. 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* 2(1). 37–63.

Prince, Ellen F. 1981. Towards a taxonomy of given-new information. In Peter Cole (ed.), *Radical Pragmatics*, 223–255. New York: Academic Press.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Sanders Ted & Wilbert Spooren. 2001. Text representation as an interface between language and its users. In Ted Sanders, Joost Schilperoord & Wilbert Spooren (eds.), *Text Representation*, 1–26. Amsterdam/Philadelphia: John Benjamins.

Schubert, Christoph. 2008. *Englische Textlinguistik. Eine Einführung*. Berlin: Erich Schmid Verlag.

Steiner, Erich. 2005. Some properties of texts in terms of 'information distribution across languages'. *Languages in Contrast* 5(1). 49–72.

Strube, Michael & Udo Hahn. 1999. Functional centering: Grounding referential coherence in information structure. *Computational Linguistics* 25(3). 309–344.

Toury, Gideon. 1995. The nature and role of norms in translation. In idem, *Descriptive Translation Studies and Beyond*, 53–69. Amsterdam-Philadelphia: John Benjamins.

Toutanova, Kristina, Dan Klein, Christopher D. Manning & Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. *The 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 173–180. Morristown, NJ, USA: Association for Computational Linguistics.

Van Rijsbergen & Cornelis Joost. 1979. *Information retrieval* (2nd edn.) Butterworth.

Venables, William N. & David M. Smith. 2010. *An introduction to R. Notes on R: A programming environment for data analysis and graphics*.

Vapnik, Vladimir N. & Alexey J. Chervonenkis. 1974. *Theory of pattern recognition*. Nauka, Moscow.

Witten, Ian H., Eibe Frank & Mark A. Hall. 2011. Data mining: Practical machine learning tools and techniques. *The Morgan Kaufmann Series in Data Management Systems*. Elsevier Science.

Index