Empirically modelling translation and interpreting

Edited by

Silvia Hansen-Schirra, Sascha Hofmann & Bernd Meyer

Translation and Multilingual Natural Language Processing ??

language science press

Translation and Multilingual Natural Language Processing

In this series:

1. Fantinuoli, Claudio & Federico Zanettin (eds.). New directions in corpus-based translation studies.

ISSN: 2364-8899

Empirically modelling translation and interpreting

Edited by

Silvia Hansen-Schirra, Sascha Hofmann & Bernd Meyer

language science press

Silvia Hansen-Schirra, Sascha Hofmann & Bernd Meyer (eds.). 2015. *Empirically modelling translation and interpreting* (Translation and Multilingual Natural Language Processing ??). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog

ISBN: 000-0-000000-00-0 (Digital)
000-0-000000-00-0 (Hardcover)
000-0-000000-00-0 (Softcover)
ISSN: 2364-8899

Cover and concept of design: Ulrike Harbort
Fonts: Linux Libertine, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Language Science Press
Habelschwerdter Allee 45
14195 Berlin, Germany
langsci-press.org

Storage and cataloguing done by FU Berlin

Contents

Chapter 1

Universals of editing and translation

Mario Bisiada
Universitat Pompeu Fabra

It has been claimed that translation universals are really "mediation universals" (Chesterman 2004; Ulrych & Murphy 2008), pertaining to the more general cognitive activity of mediating a text rather than specifically translating it. Among those linguistic activities that share the alleged mediation effect with translating are editing and revising. In this chapter, I critically examine the theory of "mediation universals" by comparing unedited translations with edited translations and with edited non-translations. The focus is on explicitation, normalisation/conservatism and simplification. The operationalisations are partly adopted from a similar study on English by Kruger (2012), which the present study seeks to replicate for German management and business articles. The results do not support the notion of mediation universals for the present corpus but rather show that translated texts are recognisable as such even after the editing process. Editorial influence on translated language in this genre is shown to be strongest in terms of sentence length and lexical diversity, where unedited and edited translations differ significantly from each other. Here, editors approximate the language to that of the non-translations, though the unedited translations have a greater average sentence length than the non-translations. That finding does not support the usual observation that translated texts have shorter sentences than non-translations, but highlights the importance of studying editorial influence in translation. That translations are hybrid texts, influenced by many agents other than the translator is now trivial knowledge. Yet corpus research in translation studies still relies mainly on published translations. The findings in this chapter argue for including unedited manuscripts in corpus-based studies of translated language to avoid missing phenomena of translated language that may be removed at the editing stage and to be able to differentiate which features really pertain to the translation act and which are affected by editorial influence.

1 Introduction

The notion of translation universals has been subject to debate for a long time (Baker 1993; Chesterman 2004; Mauranen & Kujamäki 2004). Its status today is

Mario Bisiada. Submitted. Universals of editing and translation. In Silvia Hansen-Schirra, Sascha Hofmann & Bernd Meyer (eds.), *Empirically modelling translation and interpreting*, 1–35. Berlin: Language Science Press.

problematic (see House 2008), though few would dispute that differences exist between translated and non-translated texts. Much of the controversy surrounding the issue is about the term "universal" (Chesterman 2014: 86), while the line of enquiry itself still seems productive and interesting because "the quest for universals is no more than the usual search for patterns and generalizations that guides empirical research in general" (Chesterman 2014: 87).

To advance translation studies as an empirical discipline, it is necessary to test existing theories with empirical methods and to suggest new models based on empirically tested (and testable) data. This process can be facilitated by conceiving studies in a replicable and rigorously transparent fashion, that is, they should enable other researchers to retrace the steps taken by the investigator, so that they can test the results in another language, genre or setting. To promote the use of statistical significance testing in our discipline, it would be useful for scholars to cite the sources where the significance tests they employ are documented, just as it is done with other tools or ideas that they use in their work. Merely stating the name of a statistical test without reference assumes that it is common knowledge, which in many disciplines of the humanities is arguably not the case.

The aim of this chapter is to draw attention to the influence of editors on the translation text, which so far has not received much attention in models of translation. Studying texts before and after editing can provide great insights into the translation process, which is here defined as "the period commencing from the moment the client contacts the translator and ending when the translation reaches the addressee" (Muñoz Martín 2010: 179).

Most analyses of translated language are based solely on corpora of published translations, and few attempts have been made to build a corpus of unedited translations (for an early such design, see Utka 2004). But published texts have usually undergone some kind of editing process involving various language users prior to their release. The study of manuscript translations informs current theories of translation by differentiating linguistic features that are present throughout the translation process from features whose frequency in the text was increased or decreased at the editing stage.

A holistic view of the translation process, obtained by studying manuscript translations alongside their published versions, will greatly increase the accuracy of the claims we make about translated language, improve the "ecological validity of experimental settings" (Muñoz Martín 2010: 179) (see also Saldanha & O'Brien 2013: 110), and allow insights into the linguistic effects of editing, an as yet underresearched aspect of language use. I discuss the influence of editors

in such a holistic view of the translation process in greater detail in Bisiada (In prep.).

This chapter investigates three proposed translation universals, explicitation, normalisation and simplification, aiming to find out how these are affected by editorial intervention. Two subcorpora that each exhibit one type of mediation (one translated but not edited, the other not translated but edited) are compared with a third subcorpus that exhibits both types of mediation, that is, the texts were translated and then edited.

If these universals were really universals of translation, then they should also be visible after editing because they would not otherwise have been noticed by the existing research using published translations. In that case, the subcorpus of translated and edited texts would share more similarities with the subcorpus of manuscript translations.

If, on the other hand, the universals were produced by any kind of "mediation effect", as some scholars argue (see Section 2), the subcorpus of translated and edited translations should exhibit differences to both the manuscript translations and the published non-translations, because the editors' intervention would have "superseded" particular features of the translation and brought the texts closer to the published texts.

The chapter is structured as follows: Section 2 discusses existing claims that translation universals are really "mediation universals". In Section 3.1, I describe the corpus and the operationalisations of the three translation universals that were tested in this study. I then explain the statistical methods used and the procedure that I took to ensure statistical significance of the findings (Section 3.2). Section 4 contains the analysis of explicitation (4.1), normalisation/conservatism (4.2) and simplification (4.3). Finally, Section 5 contains a summary of the findings and a discussion of their implications.

2 Universals of "mediated discourse"?

Translating and editing are considered to be forms of mediation. Lefevere argues that what translating has in common with "other modes of rewriting, such as editing, historiography, criticism, anthologising and the production of abridged or simplified texts" is that it "presuppose[s] a certain degree of mediation on the part of the writer/translator to adapt texts to the new audience" (Lefevere 1992: 9).

In his analysis of translation universals, Chesterman (2004) calls translation an act of "constrained communication", arguing that universals pertinent to trans-

lation may also be found

> in other kinds of constrained communication, such as communicating in a non-native language or under special channel restrictions, or any form of communication that involves relaying messages, such as reporting discourse, even journalism.
>
> (Chesterman 2004: 10–11)

Crucially, he argues that "it may be problematic eventually to differentiate factors that are pertinent to translation in particular from those that are pertinent to constrained communication in general" (Chesterman 2004: 10–11).

Ulrych & Murphy (2008) adopt the notion of "constrained communication" and the list of linguistic activities that are claimed to share particular features, so-called "mediation universals" (2008: 149). They even add to that category by arguing that "editing, copy-editing, revision or postediting" as well as ghostwriting are also types of mediated discourse (2008: 150). What unites texts of that kind, in their view, is that "they are processed, or rewritten, for particular audiences and are thus mediated for a purpose" (Ulrych & Murphy 2008: 151). Like Chesterman, they argue that "the notion of translation universals may be usefully replaced by that of *mediation universals* which may be identified in various kinds of mediated discourse" (Ulrych & Murphy 2008: 149).

By the above definition, most publicly available texts, except perhaps spontaneous online discourse such as comments, posts or tweets, could be described as "mediated discourse". Such a wide applicability not only makes the term itself less useful. It also makes the hypothesis of "mediation universals" difficult to disprove, as very few texts are available that would be considered "unmediated discourse". Internet discourse might be one possibility, but one would have to ensure that the authors are native speakers, have not reported any discourse or relayed any messages and have not revised their text. The reliability of such a corpus would seem to be rather low.

To back up their claims, Ulrych & Murphy (2008) conduct a study of mediated discourse, where they draw on the EuroCom corpus, a parallel corpus of written texts drafted by non-native speakers of English at the European Commission and the same texts edited by native speakers. The size of the corpus at the time of analysis is reported as one million words in each part of the corpus (Ulrych & Murphy 2008: 152). The object of study is to investigate "whether there are typical phraseologies within mediated discourse as such" (Ulrych & Murphy 2008: 155) by comparing the edited texts both with the non-edited texts in the

corpus and with the British National Corpus (BNC), which they call "a corpus of non-mediated native-speaker language" (Ulrych & Murphy 2008: 155).

Analysing three-word clusters, they find that *in order to* and *as well as* are used rather often in the EuroCom corpus, though more commonly in the non-edited than in the edited texts. Further, they are used less often in the BNC, from which they conclude that "they are not used frequently in speech or writing in non-mediated English" (Ulrych & Murphy 2008: 159).

However, these findings do not seem very convincing. As a "reference corpus of native-speaker, non-mediated English" (Ulrych & Murphy 2008: 159), the BNC may be problematic. It contains extracts from, among other things, national newspapers, specialist periodicals, academic books, popular fiction and university essays, the authors of which are unlikely to be native speakers in all cases. And even if that were the case, a claim that is not made anywhere in the description of the BNC (Burnard 2009), the corpus does not seem to contain much "non-mediated" English. It does contain some unpublished letters and essays that may be considered non-edited, and thus non-mediated. But for the most part, it consists of published, and thus mediated, texts, as newspapers, periodicals, journals and books have all been edited, copy-edited and revised to some extent.

Elsewhere, Ulrych (2009) claims that the boundaries between translating and editing as forms of mediation are becoming blurred. Unfortunately, it is not clear just what is meant by editing, specifically who does the editing. The research approach taken by Ulrych & Murphy (2008) outlined above suggests that the editing is done by someone other than the translator. However, the reference to "hybrid forms such as transediting" (Ulrych 2009: 219) seems to suggest that it is the authors or translators themselves who do the editing (for a valuable critique of the term "transediting", see Schäffner 2012).

The existence of "mediation universals", then, has never really been substantiated by empirical evidence. That has not kept it from being used, albeit with different understandings: to refer to non-native speaker language use (Ulrych & Anselmi 2008; Gaspari & Bernardini 2010; Rabadán & Izquierdo 2013: 79; Xiao & Hu 2015: 175), to bilingual communication (Lanstyák & Heltai 2012), to interlingual revision (Robertson 2010: 63) or to "texts produced under the constraint of linguistic or cultural contact" (Zanettin 2014: 183). Even the term "mediation" itself is used without a commonly accepted definition (for a totally unrelated use of the term "mediated discourse", see Scollon 2001; Norris & Jones 2005).

One empirical analysis of "mediation universals" and, more specifically, the mediation effect of editing, was conducted by Kruger (2012). The 1.2 million word corpus she draws on has three subcorpora: firstly, translations from Afrikaans

to English, secondly, originally English texts that were edited by professional language editors, and thirdly, those same texts in their manuscript form before editing took place (2012: 360). All texts are from the time span 1997 to 2010 and the genres are academic, instructional, popular and reportage texts (2012: 359). Her aim is to investigate whether "the universals of translated language are the consequence of a cognitive mediation effect that is shared among different kinds of mediated language" (Kruger 2012: 358). Her analysis focuses on the three suggested translation universals explicitation, normalisation/conservatism and simplification (more details on the operationalisations she uses to study these universals are given in Section 3).

Her findings do not support the hypothesis that translation universals are really mediation universals as there is a "consistent difference between the translated and edited subcorpus" in each of the three types of universals investigated (Kruger 2012: 380). Instead, she argues that the differences she finds between the two corpora can be attributed to either of the facts that they differ in processing (monolingual vs bilingual) and in production circumstances (free vs constrained) (Kruger 2012: 381). She also suggests that editing as a form of mediation does not involve explicitation or simplification, at least not as much as in translation, which she explains by the fact that editing does not involve the production of a new text (2012: 382).

Translating and editing also differ in that translating may to a larger extent be guided by the tendency of risk aversion (Pym 2005; 2008) than editing, because translators produce a text while editors work on an existing text. The linguistic mediation that translators undertake and which tends to make them "avoid misunderstandings at all costs" (Becher 2010: 20) is different to the mediation entailed by the act of editing, as it is either the translator or the author that will be blamed in case of communication problems. Universals affected by risk aversion are thus more likely to surface during the translation act than at the editing stage.

On top of that, translators are often pressed for time and paid by the hour, working on several jobs at the same time, while the editors tend to be in-house employees (that is true at least for the editors who worked on the data in my corpus). Editors have told me that the quality of the translation is an important factor affecting the time they spend on an article, though different concepts of what exactly is "quality" in translation exist (Drugan 2013; Mossop 2014; House 2015b). Thus, the different production circumstances further argue against the existence of "mediation universals".

3 Methodology

3.1 Corpus details and operationalisations

The present study draws on a 300,000 word corpus of management articles with three subcorpora (detailed in Table 1). The translated subcorpus (TR) consists of manuscript translations into German of articles that originally appeared in the *Harvard Business Review*, an American magazine for business leaders and managers. The manuscript translations were provided to me by the translation company Rheinschrift. They date from 2006 to 2011 and were commissioned by the *Harvard Business Manager*, the German sister publication of the *Harvard Business Review*. The texts are drafts that were checked for accuracy within the translation company and then sent to the publisher.

Table 1: Corpus details

Subcorpus	Translated?	Edited?	Texts (n)	Size (words)
TR	yes	no	27	106,829
TR+ED	yes	yes	27	104,448
ED	no	yes	27	88,312

The subcorpus of translated and edited texts (TR+ED) consists of the edited and published versions of the translations in the TR subcorpus. The edited (ED) subcorpus consists of articles that were written by a range of authors for the *Harvard Business Manager* and published there in 2008.

For the analysis, the three subcorpora were part-of-speech tagged and lemmatised using TreeTagger (Schmid 1995) with the Stuttgart-Tübingen tagset for German (Schiller et al. 1999).

As stated above, the setup of this corpus study is inspired by the corpus method used in Kruger (2012), which is an exemplary scientific work in that the comprehensive and detailed description of the author's methodology allows other researchers to replicate her study or adopt its methods. The present study also uses edited translations and non-translated articles, but instead of unedited, non-translated texts, it uses unedited translated texts, which means that in this study, all texts would count as mediated.

Kruger (2012) makes useful observations regarding differences between monolingual and bilingual text production and how they differ from editing, which involves no actual production of text. She states that her subcorpus of translations

contains "[p]ublished texts as well as ephemera" (Kruger 2012: 360), yet later describes it as involving only bilingual mediation (see Table 7 in 2012: 380). I would argue, though, that published translations are also mediated monolingually, because they are usually also edited before publication.

If published translations, then, have been "mediated twice", the effect of the mediation that takes place first may be obscured. Differentiating the linguistic effects of translating and editing thus requires the study of unedited translations, which is why I have chosen the present corpus structure over the one used by Kruger (2012).

The overview below lists the variables by which each translation universal was operationalised in Kruger (2012: 362) (on the left) and the variables used in this study (on the right). To replicate her study to the best possible degree, I have used her operationalisations as far as that was feasible for the analysis of German. Where this did not seem to be the case, for instance with contracted forms (German does not have this feature in written language) or inclusive language (no conventionalised forms exist in German), I have introduced other operationalisations that I consider relevant for the analysis of the given universal. A brief rationale for the applicability of each operationalisation will be given in each appropriate analysis section.

Explicitation

More complete/less economical surface realisation in translation

Frequency of use of optional complementiser *that*	Frequency of use of *dass* ('that')
Frequency of use of full forms versus contracted forms	

More explicit relations between conceptual propositions in text

Frequency of linking adverbials	Frequency of linking adverbials
	Frequency of pronominal adverbs
	Conjunction vs preposition ratio

Normalisation/conservatism

Frequency of coinages and loanwords	Degree of unconventional language use
Frequency of lexical bundles	Frequency of lexical bundles
Use of inclusive language	Passive alternatives

Simplification

Lexical diversity	Lexical diversity
Mean word length	Mean word and sentence length

3.2 Statistical significance

As we need to test the difference among the means of three corpora for statistical significance, a one-way analysis of variance (ANOVA) will be used. This test requires the data to be normally distributed and have approximately equal variances, though it is "fairly tolerant of all but gross departures from normality and homogeneity of variance" (Butler 1985: 132; see also Lowry 2012: ch. 14.1). As the data is not always normally distributed, I have chosen an equal sample size of 27 texts for each corpus to increase the robustness of the test.

Where the p-value yielded by the ANOVA is close to the significance threshold, I have also conducted a Kruskal-Wallis test, which is a distribution-free alternative to the ANOVA (Lowry 2012: ch. 14a; Cantos Gómez 2013: 45), to ensure the accuracy of the reported significance. The confidence level of $\alpha = 0.05$ is considered to be statistically significant and the confidence level of $\alpha = 0.01$ is considered to be highly statistically significant.

The results are reported in plots where the standard error of the mean is shown by error bars. Where statistical significance is reported, a post-hoc Tukey test, one of the standard comparison tests following the ANOVA (see Cantos Gómez 2013: 55), has been conducted to examine which corpora differ from each other for the given variable. To just compare two corpora, I have used the Mann-Whitney test, which, unlike the often used *t*-test, does not assume normal distribution of the data (Kilgarriff 2001: 104).

4 Analysis

4.1 Explicitation

4.1.1 Frequency of *dass* complementisers

The causes for the omission of *dass* ('that') in written German, generally referred to as "declarative complementiser drop" (Reis 1995: 33), have not been conclusively explored to my knowledge. There is widespread agreement that the verbs allowing the omission of *dass* are the same as the verbs known as "bridge verbs" (Grewendorf 1989: 54; Müller 1993: 362–363; Steinbach 2002: 8), though

this has been refuted by Reis (1995). The omission of the complementiser is less straightforward than in English because *dass* is not always optional, depending on the semantics both of the subclause and the particular verb or noun involved (Müller 1993; Gärtner & Steinbach 1994; for an overview of some literature, see Lapshinova-Koltunski 2010: 30). Verbs that require a finite extension using *dass* in German have English counterparts that allow both finite and non-finite extensions (Fischer 1997: 214). English, on the other hand, tends to require non-finiteness more often than finiteness (Fischer 1997: 214). In German, it is only with some verbs that the same content can be expressed both with *dass* and with a coordinate clause.

For this analysis, I selected the most common German verbs and nominalisations that can take a *dass* complement. The selection was based on Jones & Tschirner (2006), who draw on the Leipzig/BYU Corpus of Contemporary German to provide a list of the 4000 most common German words. From the 2500 most frequent German words (occurring with a frequency of at least 30 instances per million words), I have compiled a list of the most common verbs and nominalisations that can be complementised both by a *dass*-clause and a main or infinitive clause according to the E-VALBU valency dictionary for German (Schumacher et al. 2004).[1] The resulting list is shown in Table 2.

I have considered *dass* to be omitted when the verb or nominalisation was followed by either an infinitive clause with *zu* or by a finite main clause because those constructions can be replaced by a *dass* clause. If the verb or nominalisation was followed by a subordinate, verb-final clause, such as a clause introduced by another conjunction like *wie* ('how'), *was* ('what'), *wo* ('where') or *ob* ('if'), the construction was not counted as an omission of *dass* because *dass* cannot replace those conjunctions.

Regarding the analysis of items that were used with a *dass* clause, the ANOVA test reports a highly statistically significant difference among the mean frequencies (see Figure 1), which is confirmed by a Kruskal-Wallis test ($H = 11.8\ (df = 2),\ p = .0027$). A post-hoc Tukey test reveals that there is a significant difference ($p < .05$) between both the unedited and the edited translations, where *dass* is present at a frequency of just under 17.5 instances per 10,000 words, and the non-translated articles, where it occurs at a frequency of around 9 instances per 10,000 words.

Constructions where the items under analysis were used with an alternative to *dass* occur with a frequency of around 7.5 to 9.5 instances per 10,000 words in each subcorpus, and there is no significant difference as reported by the ANOVA (see

[1] Available at http://hypermedia.ids-mannheim.de/evalbu/index.html.

Table 2: Verbs and nouns with *dass*

sagen 'to say'	*wissen* 'to know'	*mitteilen* 'to inform'
merken 'to notice'	*glauben* 'to believe'	*meinen* 'to think'
schreiben 'to write'	*erklären* 'to explain'	*vorstellen* 'to imagine'
lesen 'to read'	*vermuten* 'to suspect'	*bedeuten* 'to mean'
hören 'to hear'	*fordern* 'to demand'	*erwarten* 'to expect'
spüren 'to sense'	*heißen* 'to be called'	*drohen* 'to threaten'
angeben 'to claim'	*behaupten* 'to claim'	*schätzen* 'to estimate'
fürchten 'to fear'	*annehmen* 'to assume'	*vorschlagen* 'to suggest'
finden 'to find'	*vereinbaren* 'to agree'	*befürchten* 'to fear'
sehen 'to see'	*zugeben* 'to admit'	*einräumen* 'to admit'
denken 'to think'	*erzählen* 'to narrate'	*scheinen* 'to seem'
hoffen 'to hope'	*ausgehen von* 'to assume'	*wünschen* 'to wish'
betonen 'to stress'	*versprechen* 'to promise'	*beschließen* 'to decide'
fühlen 'to feel'	*ausrichten* 'to tell'	
Meinung 'opinion'	*Forderung* 'demand'	*Eindruck* 'impression'
Ansicht 'view'	*Überzeugung* 'conviction'	*Auffassung* 'view'
Hoffnung 'hope'	*Vermutung* 'assumption'	*Behauptung* 'claim'
Befürchtung 'worry'		

Figure 1) and confirmed by a Kruskal-Wallis test ($H = 1.74\ (df = 2),\ p = .419$).

These findings seem to support the view that translations are more explicit than the non-translated articles as the frequency of the use of *dass* in translated texts stands out. The editors do not seem to have made any substantial changes to this feature.

4.1.2 Frequency of linking adverbials

Linking adverbials make links between the clauses they connect more explicit (House 2015a). A more frequent use of linking adverbials would thus increase the degree of explicitation in a text. To compile a list of the most frequent linking adverbials in German, I first extracted all the linking adverbials ("konnektintegrierbare Konnektoren", that is, connectors that can be integrated into one of the clauses they connect, see Pasch et al. 2003: 487) according to Pasch et al. (2003: 504–509). To limit the range of adverbials to those that specify links between clauses, I have only chosen those that can occur both between clauses (*Null* po-

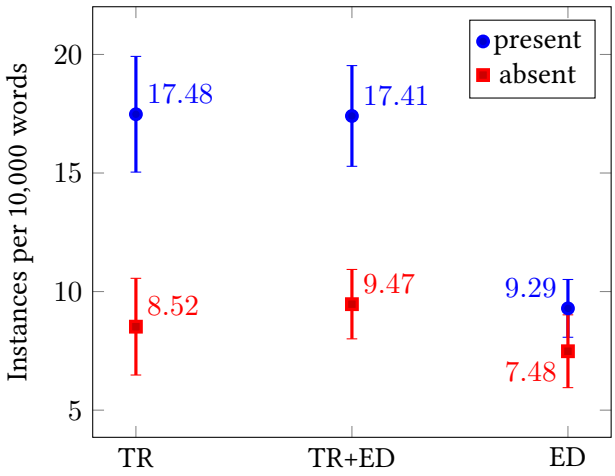

Figure 1: Mean normalised frequencies of *dass* clauses ($F(2, 78) = 5.56$, $p = .0055$) and coordinate clause alternatives to *dass* ($F(2, 78) = 0.34$, $p = .7128$)

sition) and in the final element of the sentence (*Nachfeld* position) according to Pasch et al. (2003: 504–509). I have further eliminated all pronominal adverbs, as these will be analysed separately in Section 4.1.3.

The final list (see Table 3) only includes those linking adverbials whose frequency class in the *Deutscher Wortschatz* reference corpus (Quasthoff, Goldhahn & Heyer 2013) from the Leipzig Corpora Collection is no higher than 16.[2]

The results are shown in Figure 2. Published and manuscript translations show a basically identical frequency of 9.2 linking adverbials per 1,000 words, whereas the non-translated texts only have 8.5 per 1,000 words. While this may support the existing hypothesis that translations are more explicit than non-translations, the difference is not statistically significant according to the ANOVA. Further research on German where a different set of linking adverbials is analysed might lead to a different result, but for the present analysis it must be concluded that the subcorpora do not differ significantly in terms of linking adverbials.

It is beyond the scope of this chapter to consider the frequency of individual linking adverbials, but it would be interesting for further research to investigate whether any linking adverbials are used specifically in translated texts or non-

[2] The corpus, which is available at corpora.uni-leipzig.de, assigns words to frequency classes from 0 to 24, from most to least frequent. See Quasthoff, Goldhahn & Heyer (2013: 2) for details on how the frequency class is calculated.

Table 3: Linking adverbials

allerdings 'indeed'	*also* 'thus'
ander(e)nfalls 'otherwise'	*and(e)rerseits* 'or else'
anders/genau(er)/kurz/nebenbei gesagt 'in other words/(more) precisely/ briefly/by the way'	*ansonsten* 'otherwise'
aus diesem Grund 'for this reason'	*außerdem* 'in addition'
beispielsweise/bspw. 'for instance'	*bloß* 'however'
dagegen 'on the other hand'	*das heißt* 'that is'
dessen ungeachtet 'notwithstanding'	*dennoch* 'still'
einerseits 'on the one hand'	*ergo* 'thus'
erstens, zweitens... 'first, second'	*folglich* 'therefore'
freilich 'of course'	*gleichwohl* 'nevertheless'
hingegen 'on the other hand'	*im Gegensatz zu/dazu* 'contrarily'
im Übrigen 'what's more'	*immerhin* 'at least'
indes(sen) 'meanwhile'	*infolgedessen* 'consequently'
insbesondere 'especially'	*insofern* 'for that matter'
insoweit 'as far as'	*jedenfalls* 'in any case'
jedoch 'however'	*mit anderen Worten* 'in other words'
mithin 'thus'	*nämlich* 'namely'
nichtsdestotrotz 'notwithstanding'	*obendrein* 'on top of that'
ohnehin 'in any case'	*schließlich* 'after all'
sodann 'consequently'	*stattdessen* 'in spite of that'
überdies 'what's more'	*übrigens* 'by the way'
unterdessen 'meanwhile'	*vielmehr* 'rather'
vor allem 'above all'	*währenddessen* 'meanwhile'
weiterhin 'in addition'	*wiederum* 'on the other hand'
wohlgemerkt 'let me add'	*zudem* 'plus'
zum Beispiel/z. B. 'for example'	*zum einen* 'on the one hand'
zumal 'given that'	*zumindest* 'at least'
zunächst 'initially'	*zusammenfassend* 'to sum up'
zwar/und zwar 'it's true that/namely'	*...erweise* '...ly'

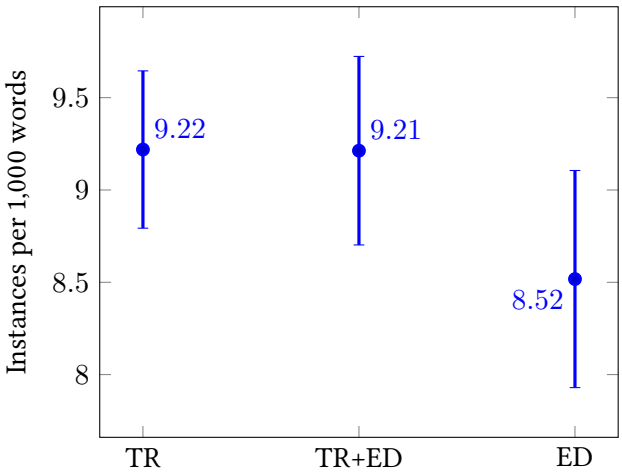

Figure 2: Mean normalised frequency of linking adverbials ($F(2, 78) = 0.62,\ p = .5406$)

translated texts.

4.1.3 Frequency of pronominal adverbs

Bisiada (2014: 14–15) has found that pronominal adverbs such as *darum* ('therefore'), *daraus* ('from that') or *darüber hinaus* ('on top of that') are regularly introduced when sentences are split, both by translators and editors. The introduction of pronominal adverbs to the text clarifies cohesive relations (Kunz & Lapshinova-Koltunski 2015) and is thus an explicitating addition to the text.

Pronominal adverbs have been extracted by a search for the tag PAV, which stands for *pronominal adverb* in the Stuttgart-Tübingen tagset. The absolute occurrences were then converted to normalised frequencies.

Figure 3 shows that in the translated texts, pronominal adverbs occur at a rate of 9.4 instances per 1,000 words, while in the non-translated texts, they only occur at a rate of 8 instances per 1,000 words, which would give further support to the hypothesis that translated texts are more explicit.

However, the statistics do not quite allow this conclusion. The ANOVA test argues for a statistically significant difference (see Figure 3), and the post-hoc Tukey test places the difference between the non-translations and both translated subcorpora ($p < .05$). According to the Kruskal-Wallis test, however, the difference between the normalised frequencies in the three corpora is not statist-

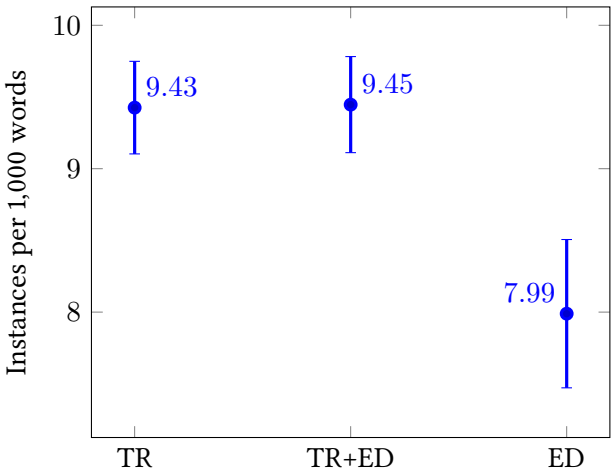

Figure 3: Mean normalised frequency of pronominal adverbs ($F(2, 78) = 4.33,\ p = .0165$)

ically significant ($H = 4.07\ (df = 2),\ p = .1307$). As stated in Section 3.2, the Kruskal-Wallis test takes precedence for data that is not entirely normally distributed. Thus, while translated texts seem to contain more pronominal adverbs than non-translated texts, that difference is not statistically significant.

4.1.4 Conjunction vs preposition ratio

Steiner (2001: 26) suggests measuring the ratio of conjunction vs preposition as a way of testing the grammatical metaphoricity of a text. The greater the ratio, that is, the more conjunctions a text has in relation to prepositions, the less metaphorical and the more explicit it is (Steiner 2001: 26).

For the present analysis, the tagged corpora were searched for the Stuttgart-Tübingen tags indicating conjunctions (KOUI, KOUS, KON, KOKOM) and prepositions (APPR, APPRART). Figure 4 shows that the ratio, while highest in published translations, is rather similar in all three corpora, between 0.63 and 0.68. It is perhaps interesting to note that editors seem to have made the text more explicit by increasing the ratio. Overall, however, the ANOVA reports no statistically significant difference between the conjunction vs preposition ratios of the three subcorpora.

To sum up this section, there seem to be more similarities between the two translated subcorpora than between the two edited subcorpora. However, the the

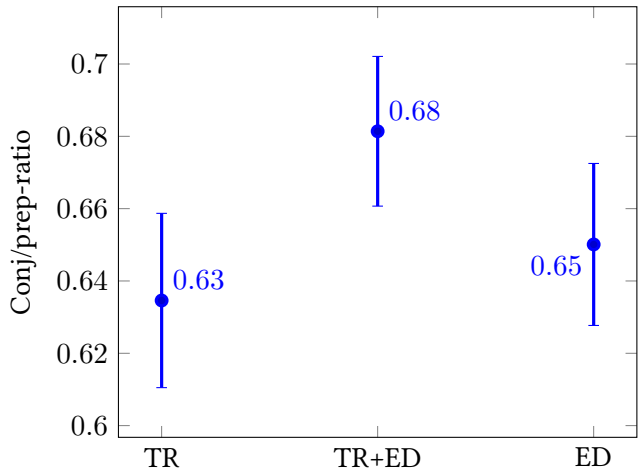

Figure 4: Mean ratio of conjunction vs preposition ($F(2, 78) = 1.12,\ p = .3283$)

operationalisations under analysis show no statistically significant differences between the three subcorpora, except regarding the use of *dass* clauses.

4.2 Normalisation/conservatism

4.2.1 Passive alternatives

The use of passive alternatives is considered a typical feature of German. Passive alternatives have been used increasingly often in professional and scientific discourse to replace the passive while keeping the language economical (see Gang 1997). They also occur more often in German non-translated texts than in English ones (Teich 2003: 181). Therefore, a higher amount of passive alternatives would indicate a higher degree of normalisation.

Three different passive constructions have been chosen for analysis: modal passives (combinations of *lassen* ('to let') and a reflexive verb; see König & Gast 2012: 162), clauses containing the impersonal pronoun *man* (Durrell 2003: 237; Teich 2003: 94) and modal infinitives, where *sein* is used with an infinitive phrase (Durrell 2003: 238; Teich 2003: 93; König & Gast 2012: 161).

To obtain the frequencies of modal passives, I have searched for instances of *lassen* and then manually reduced this list to instances where reflexive verbs were used as passive alternatives. Instances of *man* were simply counted. As for the modal infinitives, the subcorpora were searched for the STTS tags PTKZU and VVIZU to obtain instances of the pre- and intrainfinitival *zu*. The resulting list

of infinitive phrases was reduced to those where *sein* is used.

The ANOVA test finds no significant overall difference between the three subcorpora (see Figure 5). The data is not normally distributed, but the backup Kruskal-Wallis test confirms the observation ($H = 0.45$ $(df = 2)$, $p = .7985$).

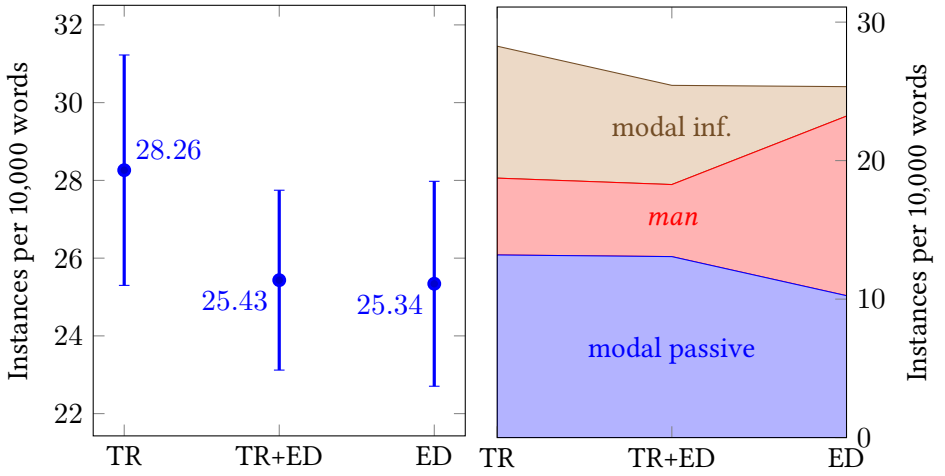

Figure 5: Left: Mean normalised frequency of passive alternatives ($F(2, 78) = 0.39$, $p = .6784$); Right: Mean normalised frequencies of modal infinitives, *man* and the modal passive.

A closer look at the passive alternatives, however, reveals some differences between the translated and the non-translated texts. Both unedited and edited translations use *man* statistically significantly less often than the non-translated texts ($H = 9.01$ $(df = 2)$, $p = .0111$). In the translations, *man* occurs at a frequency of around 5 instances per 10,000 words, while in the non-translations, it occurs at around 13 instances per 10,000 words. A post-hoc Tukey test confirms that this is statistically significant at the $p < .01$ level.

At the same time, both unedited and edited translations use the modal infinitive more often than non-translated texts. Modal infinitives occur at a frequency of 2 instances per 10,000 words in non-translated texts, and at a frequency of 9.5 and 7 instances per 10,000 words in the manuscript and published translations, respectively. The difference between the unedited and edited translations is statistically insignificant (Mann-Whitney $U = 303$, $p > .05$), but seems to show that editors have approximated the frequency of modal infinitives to that of non-translated texts.

4.2.2 Degree of unconventional language use

Translators have been claimed to be more conservative in their language use than authors of non-translated texts (Bernardini & Ferraresi 2011: 242). Kruger (2012) conducts her analysis by searching for hapax legomena (words that occur only once in a text) and then filtering out "lexicalised" words by using the spell checker and online dictionary in Microsoft Word.

That seems like a somewhat unconvincing method to decide which words count as lexicalised. Some words may be used regularly, but may not occur in a dictionary and thus would not count as lexicalised. German has extensive means of compounding, making it even easier to coin new words. A further problem with using hapax legomena as a tool for analysing idiosyncracy of the lexis is that even the most unconventional or innovative words will not appear in the analysis if they are used a second time somewhere in the text.

Nevertheless, the analysis presented here also takes the initial step of isolating hapax legomena using AntConc. From the resulting lists, words that feature in the Hunspell dictionary[3], abbreviations, web addresses, proper names and untranslated job titles have been filtered out. I have only considered English words as loan words if they were found in the text "as is", that is without quotation marks or explanations. Like the lexicalisation issue, the question of whether or not something is a loan word is difficult to answer (Heller 2002).

Instead of pursuing the notion of lexicalisation any further, I have instead analysed the remaining words based on their frequency in the *Deutscher Wortschatz* reference corpus (Quasthoff, Goldhahn & Heyer 2013) from the Leipzig Corpora Collection (see Section 4.1.2). For the present purposes, I have reduced the list to lemmas in the frequency classes 18 or above, which means they are outside the 200,000 most frequent words in German.

Even with those parameters, the methodology remains somewhat problematic. Technical terms that are not in the dictionary might be infrequent in the reference corpus and thus be considered idiosyncratic language use. However, overall, the method does what it should by measuring the different frequencies with which unconventional words are used in the texts.

Keeping the mentioned drawbacks in mind, the analysis shows quite clearly that non-translated articles make more use of unconventional or less established words than the translated texts (Figure 6). The difference is most pronounced in the case of lexical items that are not attested in the Leipzig corpus, which occur at a frequency of less than 5 instances per 10,000 words in the translated texts, but

[3] Available at: http://hunspell.sourceforge.net/

at a frequency of 18.5 instances per 10,000 words in the non-translated articles. The rather large error bars for the non-translated texts indicate that the actual values depend largely on the individual style of the author.

A further interesting aspect is that unattested lexical items and those at frequency classes 21–24 occur less frequently in the edited translations than in the manuscript ones. That seems to indicate that editors attempt to make the text more conservative by removing unconventional words.

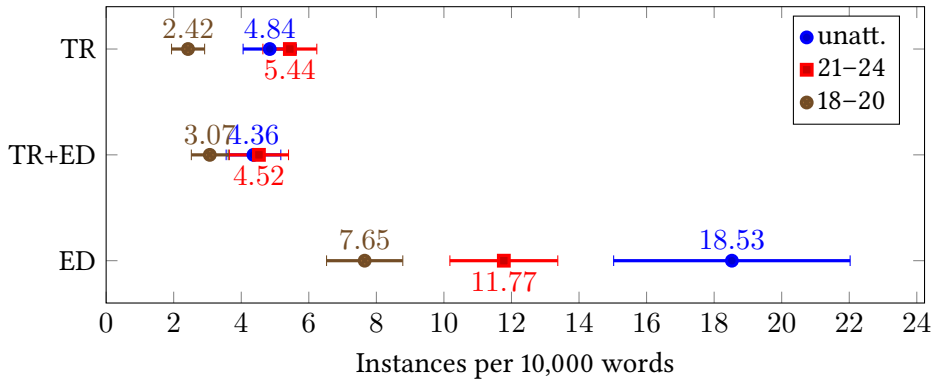

Figure 6: Mean normalised frequency of hapax legomena, unattested in the Leipzig corpus ($F(2, 78) = 14.34,\ p < .0001$), at frequency classes 21–24 ($F(2, 78) = 11.82,\ p < .0001$) and 18–20 ($F(2, 78) = 13.45,\ p < .0001$)

The difference according to the ANOVA is highly significant for all three levels of frequency in the Leipzig corpus. It is confirmed by the Kruskal-Wallis test ($H = 17.01$ ($df = 2$), $p < .0001$ for unattested words, $H = 15.92$ ($df = 2$), $p < .0001$ for the frequency class 21–24 and $H = 16.22$ ($df = 2$), $p < .0001$ for the frequency class 18–20). A post-hoc Tukey test shows that while the difference between translations and non-translations is significant at the ($p < .01$) level in all three cases, there is no statistically significant difference between edited and unedited translations.

Nevertheless, with regard to the unattested words and those in the frequency classes of 21–24, there are fewer instances per 10,000 words in the edited translations compared to the unedited translations. Although that difference is not statistically significant according to the post-hoc Tukey test, it may still indicate that editors think more conservatively when it comes to editing translations, whereas they leave more room for creativity to authors of non-translated texts.

That is why it is only in non-translated texts that we find coinages and innovative compounds such as *Gedankenwerker* ('thought worker'), *glatterklären* ('smooth something out by explanation'), *Abwarter* ('someone who hangs back and waits'), *Lächelanordnungen* ('orders to smile') and *lebenssprühend* ('sparking with life').

4.2.3 Frequency of lexical bundles

Kruger argues that the usage of lexical bundles, "'prefabricated', conventionalised language unit[s]" is "indicative of more normalised or conservative language use" (Kruger 2012: 365). Adopting her method to study lexical bundles, I have created a list of the most common trigrams in each corpus. Trigrams that occurred with a frequency of less than 0.01% in each subcorpus were removed. Proper nouns such as *Harvard Business School* and subject-specific trigrams such as *Triple Bottom Line* were also removed.

Unlike Kruger, I have not removed individual subject-specific words. Given the fact that all texts form part of the same genre, I see no reason to exclude trigrams that contain subject-specific words, as they may be part of the particular jargon or conventionalised discourse of that language community. As a result, the list of the 28 trigrams that are investigated in this section (see Table 4) contains some less general trigrams than the one used by Kruger (2012: 365).

Table 4: Trigrams selected for investigation

in der Regel 'normally'	*nicht nur* ART 'not just the'
bei der Entwicklung 'while developing'	*bei der Arbeit* 'at work'
für das Unternehmen 'for the company'	*in den letzten* 'in the last'
auf diese Weise 'in this way'	*aus diesem Grund* 'for this reason'
eine Reihe von 'a range of'	*in den vergangenen* 'in the past'
die Zahl der 'the number of'	*dass die Mitarbeiter* 'that the staff'
davon überzeugt, dass 'convinced that'	*in der Praxis* 'practically'
zum Beispiel ART 'for example the'	*in der Lage* 'able to'
handelt es sich 'is about'	*für den Kunden* 'for the customer'
die Mitarbeiter, die 'employees who'	*in Bezug auf* 'in relation to'
Auswirkungen auf ART 'effects on'	*dass sich* ART 'that REFL'
für den Erfolg 'for success'	*in diesem Fall* 'in this case'
mit ihren Mitarbeitern 'with its staff'	*Art und Weise* 'way'
im Laufe der 'over the course of'	*sich heraus, dass* 'turns out that'

The ANOVA reveals that there is a highly significant difference among the three subcorpora (see Figure 7), which is confirmed by a Kruskal-Wallis test ($H = 23.02$ $(df = 2)$, $p < .0001$). It is evident from the figure that this difference is found between the non-translations and the two subcorpora of manuscript and published translations. In the latter, the investigated trigrams occur with rather similar frequencies of 3.5 and 3.6 instances per 1,000 words, while in the non-translated articles, they only occur at a frequency of 2.1 instances per 1,000 words. The post-hoc Tukey test confirms that this is a significant difference at the $p < .01$ level.

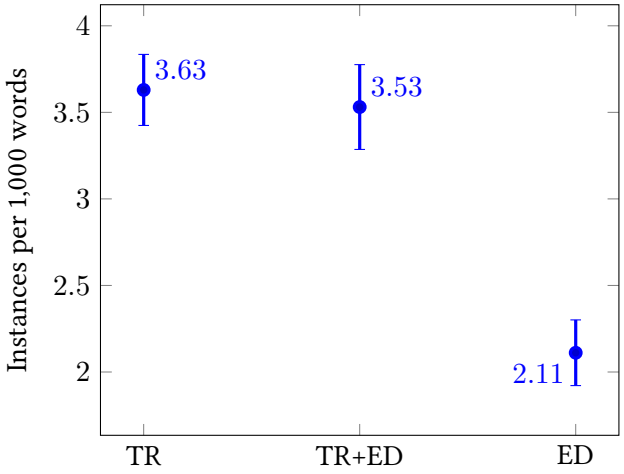

Figure 7: Mean normalised frequency of trigrams ($F(2, 78) = 15.66$, $p < .0001$)

Based on the higher occurrence of common trigrams in the translated texts, it would seem that translators are more conservative in their language use, and that editors have not intervened in this respect. The analysis of selected trigrams applied here is limited to analysing the frequency of specific tokens, while "obscuring differences in terms of the number of bundle types in the three subcorpora" (Kruger 2012: 384).

Thus, in order to strengthen the analysis of conservative or normalised language use in the present corpus, I have conducted a general collocational analysis. As different measures of collocational association tend to produce different types of associations, the strength of an analysis is increased by studying several measures of association (Baroni & Bernardini 2003: 373). The present analysis is therefore based on the log-likelihood and mutual information (for more information on these measures, see Manning & Schütze 1999: ch. 5.3–5.4).

For this analysis, I have used Ted Pedersen's Ngram Statistics Package (Banerjee & Pedersen 2003).[4] Based on the method used by Baroni & Bernardini (2003), the percentages of trigrams at or above certain cut-off points were compared for each subcorpus. High association scores mean a higher degree of collocational expression and thus, according to the present hypothesis, a more normalised language use.

For the log-likelihood ratio, three cut-off points were chosen. Log-likelihood ratios can be looked up directly in the table of the χ^2 distribution (Manning & Schütze 1999: 174). Thus, the cut-off points chosen here are the critical values 18.47, 23.51 and 28.47 given in the table for four degrees of freedom,[5] which correspond to the confidence levels $\alpha = 0.001$, $\alpha = 0.0001$ and $\alpha = 0.00001$.

For the mutual information score, Baroni & Bernardini (2003) use pointwise mutual information. In my case, the results produced by a pointwise mutual information analysis did not seem to be a good representation of actual trigrams in the corpus (see Manning & Schütze 1999: 178–183 for a criticism of pointwise mutual information as a measure of association), so I chose to calculate the (true) mutual information score instead. As the scores are all quite low, there is only one cut-off point at a mutual information score of 0.01.

The results are shown in Figure 8. Surprisingly, the mean percentage of trigrams with a log-likelihood ratio at or above the specified cut-off points is higher in the non-translated texts than in the translated texts, though the difference seems to disappear if the cut-off point is set to a higher value and thus a stricter confidence level. The non-translated texts also have a higher mutual information score than the translations.

The statistical tests confirm this observation. For the lowest cut-off point in the log-likelihood ratio, a score of 18.47, the ANOVA reports a highly statistically significant difference, confirmed by the Kruskal-Wallis test ($H = 22.48$ $(df = 2)$, $p < .0001$). The post-hoc Tukey test confirms that the percentage value of the non-translations is significantly higher than those of both manuscript and published translations at the $p < .01$ level.

For the next cut-off point, 23.51, the ANOVA shows a highly statistically significant difference, confirmed by the Kruskal-Wallis test ($H = 8.53$ $(df = 2)$, $p = .0141$). The post-hoc Tukey test reveals that the value of the non-translations is still significantly higher than that of manuscript translations at the $p < .01$ level, but not higher than that of edited translations. Also, there is no statistically sig-

[4] Available at http://ngram.sourceforge.net

[5] There are four degree of freedom because there are four independent values: one per word in the trigram and the total number of trigrams.

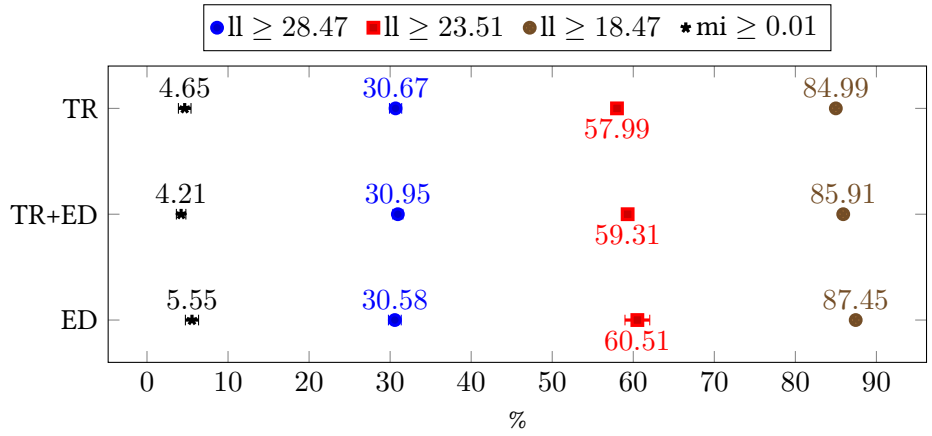

Figure 8: Mean percentages of trigrams at or above the log-likelihood ratios 28.47 ($F(2, 78) = 0.07,\ p = .9325$), 23.51 ($F(2, 78) = 5.04,\ p = .0087$), 18.47 ($F(2, 78) = 13.23,\ p < .0001$), and the mutual information score 0.01 ($F(2, 78) = 0.87,\ p = .423$)

nificant difference between the two translated subcorpora.

Regarding the highest cut-off point in the log-likelihood ratio, a score of 28.47, there is no significant difference between the corpora. In the case of the mutual information score, the ANOVA yields no significant difference either.

The results from the collocational analysis support the tendency observed so far in this section, that translating creates a greater similarity between texts than editing. However, the results do not seem to confirm the hypothesis that translations use more collocational and thus normalised language, which was supported by the finding that translations use a set of very frequent trigrams more often than non-translations. A technical explanation may be that the values yielded by the lower cut-off points are simply not very meaningful; after all, at the highest cut-off point, the difference disappears.

Another possible explanation might be that the use of translation memories favours a set of fixed, recurring phrases which are then used with a high frequency in the translations. That would explain why the set of trigrams chosen above occurs more often in translated than in non-translated texts. The latter, however, make more use of collocational language in general, which provides evidence against the hypothesis that translators use more normalised language. An explanation for this might be that writers have greater lexical freedom, and thus adopt specific collocations while translators are bound to the source text

and thus less free in their language use.

To sum up this section, the analysis of unconventional language use and of lexical bundles argues that there are greater differences between the two translated texts on the one hand and the non-translated articles on the other. In other words, the editing process does not significantly change the features of the language of translation, which make the text differ from a non-translated text.

4.3 Simplification

4.3.1 Lexical diversity

For the analysis of lexical diversity, Kruger (2012) uses the standardised type-token ratio. I use the moving-average type-token ratio (MATTR) instead, which is a more robust measure of lexical diversity than the STTR because it is not affected by text length and takes into account changes within the text (Covington & McFall 2010: 96). I adopt a 500 word window as suggested for stylometric analyses by the authors (Covington & McFall 2010: 97). The MATTR was calculated using the R package koRpus by Meik Michalke.[6]

The ANOVA yields a highly statistically significant difference between the corpora (see Figure 9). The distribution of the TR+ED subcorpus is skewed, but the Kruskal Wallis test confirms a highly statistically significant difference among the corpora ($H = 18.88$ $(df = 2)$, $p < .0001$). A post-hoc Tukey test reveals that the mean MATTR of the manuscript translations is significantly lower at the ($p < .01$) level than the mean MATTRs of both edited texts. The edited texts among themselves do not show a significant difference in mean MATTR.

The findings argue that the manuscript translations are lexically less diverse than both non-translations and their published versions, which shows that the editors have intervened significantly to increase lexical diversity. The assumption that translations have a less varied vocabulary and are therefore simpler is supported. This analysis exemplifies the value of comparing manuscript and edited translations, as in a traditional corpus design, the fact that the actual translations have a much lower lexical diversity value would not have surfaced.

4.3.2 Word and sentence length

Word and sentence length were also calculated with the R package koRpus. Word length operationalises simplicity because more specific or formal words are usually longer while more frequent words are shorter (Kruger 2012: 366; Biber 1991),

[6] http://reaktanz.de/?c=hacking&s=koRpus

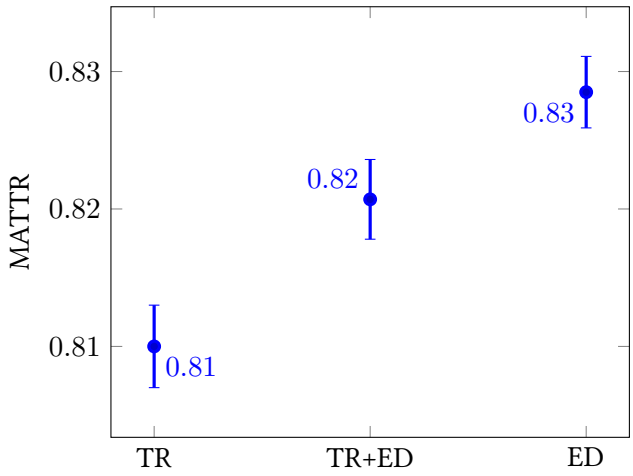

Figure 9: Moving-average type-token ratio ($F(2, 78) = 10.76,\ p = .0018$)

which seems especially true in the case of German (Bengt, Eeg-Olofsson & van de Weijer 2004: 46).

Sentence length is usually considered to be an indicator of simplification, as sentences in translated texts tend to be shorter (Laviosa 2002). As longer sentences are deemed harder to understand (though this may be problematic generalisation; see the discussion in Bisiada 2013: 165–169) it is assumed that translators split sentences to improve readability (Vintar & Hansen-Schirra 2005: 212; Bisiada 2014: 21). Simplification as a translation universal may therefore be operationalised by measuring sentence length.

In terms of the mean word length, the ANOVA reports no statistically significant difference between the three subcorpora (see the graph on the left in Figure 10).

For the mean sentence lengths in the subcorpora, contrary to what is usually assumed, it seems that the sentences in the manuscript translations are longer than those in the edited translations, and even more so than those in the non-translations (see the right graph in Figure 10).

The difference is highly statistically significant according to the ANOVA, and confirmed by the Kruskal-Wallis test ($H = 20.64\ (df = 2),\ p < .0001$). A post-hoc Tukey test shows that the manuscript translations differ from both edited texts. Sentences in manuscript translations are highly significantly ($p < .01$) longer than in the non-translations and significantly ($p < .05$) longer than in the published translations. The two edited texts do not exhibit a statistically signific-

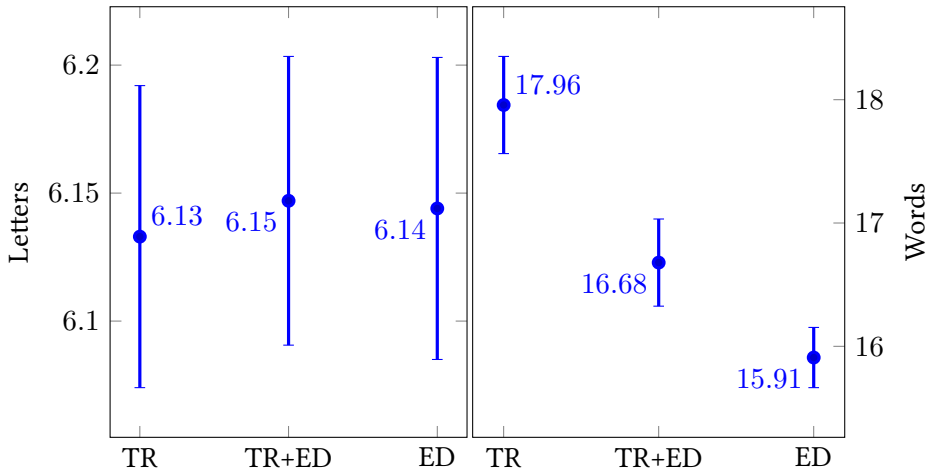

Figure 10: Left: Mean word length ($F(2, 78) = 0.01,\ p = .9901$); Right: Mean sentence length ($F(2, 78) = 9.44,\ p = .0002$)

ant difference to each other.

It appears that the editors have brought the translated texts closer to the average sentence length that is exhibited by the non-translated texts. An analysis of the manuscript non-translations would be useful here to see whether editors have shortened the sentences in those texts as well. The strong editorial influence with regard to sentence length further underlines the need to differentiate manuscripts from edited versions when making statements about the features of translated language.

This section has produced results that contrast with those from the two sections on explicitation and normalisation/conservatism in that there seem to be greater similarities between the edited translations and the non-translated articles. This suggests that, with regard to simplification, the editing process has rendered the language in the translated texts more similar to that encountered in the non-translated texts.

5 Summary and discussion

The analysis in this chapter has produced a large amount of data and claims which I hope will be checked and confirmed or rejected by other scholars, so that we discover more about the effect of editing on translation. Table 5 provides an overview of the mean values that have resulted from the analysis in this chapter.

For completeness' sake, standard deviations are also supplied in brackets.

Table 5: Overview of values with standard deviation in brackets

Variable	TR	TR+ED	ED	p
dass present	17.48 (12.68)	17.41 (11.03)	9.30 (6.33)	<.01
dass absent	8.52 (10.59)	9.47 (7.60)	7.48 (7.98)	>.05
PAV	9.43 (1.68)	9.45 (1.74)	7.99 (2.69)	>.05
Linking	9.22 (2.21)	9.21 (2.65)	8.52 (3.06)	>.05
Conj vs Prep	0.63 (0.13)	0.68 (0.11)	0.65 (0.12)	>.05
Passive alt.	28.26 (15.40)	25.43 (12.02)	25.34 (13.69)	>.05
Unconv. (unatt.)	4.84 (4.10)	4.36 (4.21)	18.53 (18.19)	<.01
Unconv. (21–24)	5.44 (4.15)	4.52 (4.55)	11.77 (8.30)	<.01
Unconv. (18–20)	2.42 (2.55)	3.07 (2.84)	7.65 (5.87)	<.01
Trigrams	3.62 (1.07)	3.53 (1.27	2.11 (0.99)	<.01
Trigr. (ll, cut-off low)	84.99 (2.07)	85.91 (1.51)	87.45 (1.68)	<.01
Trigr. (ll, cut-off mid)	57.99 (3.37)	59.31 (2.55)	60.51 (2.75)	<.01
Trigr. (ll, cut-off high)	30.67 (3.92)	30.95 (3.23)	30.58 (4.11)	>.05
Trigr. (mi)	4.65 (4.04)	4.21 (3.05)	5.55 (4.22)	>.05
MATTR	0.81 (0.02)	0.82 (0.02)	0.83 (0.01)	<.01
WL	6.13 (0.31)	6.15 (0.29)	6.14 (0.31)	>.05
SL	17.96 (2.05)	16.68 (1.84)	15.91 (1.27)	<.01

Cells in colour represent values that are different from the values of the other corpora. If the colour is green, it means that the value behaves as expected under the universal in question; if the colour is red, it means that the value runs counter to expectations and does not support the usual hypothesis attributed to that universal. The lighter colour means that the difference is statistically significant and a darker colour means that the difference was shown to be highly statistically significant. Values in the colourless cells are not significantly different to each other.

In the case of explicitation, most variables analysed here show no difference to each other, so that the features across the three subcorpora are mostly the same, except there are fewer explicitations using *dass* clauses in the non-translated articles. The differences are thus restricted to cases of less economical surface realisation, a "borderline case" of explicitation that may more usefully be considered as "expansion" (Krüger 2015: 239). Alternatively, the more frequent presence of

dass in translated texts may be a sign of conservative language use if we accept the claim that a German finite *dass*-clause is preferred over a non-finite construction (Fischer 1997: 215; 2013: 337), though this claim has not yet been backed up by evidence.

As the use of normalised or conservative language is concerned, the operationalisations analysed here suggest that there is a difference between translated and non-translated language. The latter makes more use of unconventional language and differs in the use of collocations. As regards the latter, it seems that translators use a set of recurring trigrams more frequently than writers of non-translated articles, but overall, the latter seem to use more collocational language.

In terms of the universal of simplification, differences have been observed between manuscript translations on the one hand and edited translations as well as non-translations on the other. This seems to show that editors' influence has been strongest in this respect. An explanation for this might be that simplification is operationalised by mainly quantitative features, which makes it easier for editors to change the text in order to approximate its language to the non-translated articles. More research on other genres and text types should test editors' influence on sentence length. This may explore the question of whether the finding that (published) translations have shorter sentences on average can actually be related to translated language, or whether it should instead be attributed to the influence of editors who try to improve the readability of a text.

As for the hypothesis of universals of mediated discourse, the study produces little evidence in favour of "mediation universals". Verifying that theory in the present study would have required the data to show more similarities between the TR+ED and the ED subcorpora, and for there to be more differences between the TR and the TR+ED subcorpora.

Instead, and similarly to what was reported by Kruger (2012), the editing stage seems to have little effect on the features measured here. That does not mean that changes to the text are negligible, but rather that editors do not intervene in such a way to make the articles more like the non-translated articles. With regard to simplification, however, my findings differ from those reported in Kruger (2012), as editors have made significant changes in this respect.

Based on the present findings, it could be argued that editing is largely a simplifying activity, with editors trying to apply quantitative strategies to make the text more comprehensible (on this issue, see Müller-Feldmeth et al. 2015). The editorial style of course depends to a great extent on genre. Texts edited for commercial publications need to be more reader-friendly than, say, reports or parliament communications.

From a practical perspective, the findings show that editors' intervention is restricted to three features: they make sentences shorter (by splitting them, as reported for German in Bisiada 2014)), they increase lexical diversity and they increase collocational language use to some extent. They also seem to reduce the frequency of alternative passive constructions, though the difference in this case is not statistically significant. I discuss this issue in more detail in Bisiada (In prep.: ch. 4).

With regard to the omission of *dass* and the unconventional words, editors seem to have made the text more unlike the non-translated texts. While again the differences are not statistically significant, this may mean that when editing translations, editors are actually more conservative and restrictive in terms of the non-standard expressions they let pass than when they are editing non-translated articles. In this respect, translations may improve or at least be more consistent with non-translated articles if editors gave translators some more freedom and allowed more unconventional language use.

To empirically strengthen the discipline of translation studies, more transparent and replicable research is needed. I have tried to provide such a study in this chapter, and hope to have offered a range of avenues for further research. As was shown in this chapter, the study of editing can greatly enhance our view of the translation process by differentiating features that really are attributable to translation from those that are introduced by other agents who have influence on the text.

Acknowledgments

This paper has benefitted from a discussion on Academia.edu, where I have made the manuscript available to invite criticism and suggestions from the scholarly community, with the aim of trialling a kind of "community peer review". I would like to thank everyone who participated in this session, especially my colleagues Ralph Krüger and Ekaterina Lapshinova-Koltunski for their critical reading, suggestions and detailed feedback on this paper.

I thank Michael Heinrichs at the translation company Rheinschrift for his efforts with the publishing company to allow me to obtain the manuscript translations for research purposes.

I am equally indebted to the editors at the *Harvard Business Manager*, especially Britta Domke, for their interest in my research and for giving me valuable insights into their workflow.

For all statistical calculations, I have used Richard Lowry's comprehensive

yet accessible website VassarStats (http://www.vassarstats.net) and would like to thank him for making this excellent tool freely available.

References

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: John Benjamins.

Banerjee, Satanjeev & Ted Pedersen. 2003. The design, implementation, and use of the Ngram Statistics Package. *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*. 370–381.

Baroni, Marco & Silvia Bernardini. 2003. A preliminary analysis of collocational differences in monolingual comparable corpora. *Proceedings of the Corpus Linguistics 2003 Conference*. 82–91.

Becher, Viktor. 2010. Abandoning the notion of "Translation-Inherent" explicitation: Against a dogma of translation studies. *Across Languages and Cultures* 11(1). 1–28.

Bengt, Sigurd, Mats Eeg-Olofsson & Joost van de Weijer. 2004. Word length, sentence length and frequency – Zipf revisited. *Studia Linguistica* 58(1). 37–52.

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

Biber, Douglas. 1991. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bisiada, Mario. 2013. *From hypotaxis to parataxis: An investigation of English–German syntactic convergence in translation*. University of Manchester PhD thesis.

Bisiada, Mario. 2014. "Lösen Sie Schachtelsätze möglichst auf": The impact of editorial guidelines on sentence splitting in German business article translations. *Applied Linguistics* Advance online access. 1–24.

Bisiada, Mario. In prep. *The editor's invisibility*. Berlin: Language Science Press.

Burnard, Lou. 2009. *What is the BNC?* http://www.natcorp.ox.ac.uk/corpus/index.xml (30 October, 2015).

Butler, Christopher. 1985. *Statistics in linguistics*. Oxford: Blackwell.

Cantos Gómez, Pascual. 2013. *Statistical methods in language and linguistic research*. Sheffield: Equinox.

Chesterman, Andrew. 2004. Hypotheses about translation universals. In Gyde Hansen, 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.

Covington, Michael A. & Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics* 17(2). 94–100.

Drugan, Joanna. 2013. *Quality in professional translation: Assessment and improvement*. London: Bloomsbury.

Durrell, Martin. 2003. *Using German: A guide to contemporary usage*. 2nd edn. Cambridge: Cambridge University Press.

Fischer, Klaus. 1997. *German-English verb valency: A contrastive analysis*. Tübingen: Gunter Narr.

Fischer, Klaus. 2013. *Satzstrukturen im Deutschen und im Englischen: Typologie und Textrealisierung*. Berlin: Akademie-Verlag.

Gang, Gook-Jin. 1997. *Passivsynonyme als Elemente der wissenschaftlichen Fachsprache im Deutschen*. Frankfurt/M.: Peter Lang.

Gärtner, Hans-Martin & Markus Steinbach. 1994. Economy, verb second, and the SVO-SOV distinction. *Working Papers in Scandinavian Syntax* 53. 1–59.

Gaspari, Federico & 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 upon Tyne: Cambridge Scholars.

Grewendorf, Günther. 1989. *Ergativity in german*. Dordrecht: Foris.

Heller, Klaus. 2002. Was ist ein Fremdwort?: Sprachwissenschaftliche Aspekte seiner Definition. In Rudolf Hoberg (ed.), *Deutsch – Englisch – Europäisch: Impulse für eine neue Sprachpolitik*, 184–198. Mannheim: Dudenverlag.

House, Juliane. 2008. Beyond intervention: Universals in translation? *trans-kom* 1(1). 6–19.

House, Juliane. 2015a. Global English, discourse and translation: Linking constructions in English and German popular science texts. *Target* 27(3). 370–386.

House, Juliane. 2015b. *Translation quality assessment: Past and present*. London: Routledge.

Jones, Randall L. & Erwin Tschirner. 2006. *A frequency dictionary of German*. Abingdon: Routledge.

Kilgarriff, Adam. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6(1). 97–133.
König, Ekkehard & Volker Gast. 2012. *Understanding English-German contrasts*. 3rd edn. Berlin: Erich Schmidt Verlag.
Kruger, Haidee. 2012. A corpus-based study of the mediation effect in translated and edited language. *Target* 24(2). 355–388.
Krüger, Ralph. 2015. *The interface between scientific and technical translation studies and cognitive linguistics*. Berlin: Frank & Timme.
Kunz, Kerstin & Ekaterina Lapshinova-Koltunski. 2015. Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies* 14(1). 258–288.
Lanstyák, István & Pál Heltai. 2012. Universals in language contact and translation. *Across Languages and Cultures* 13(1). 99–121.
Lapshinova-Koltunski, Ekaterina. 2010. *German clause-embedding predicates: An extraction and classification approach*. Universität Stuttgart PhD thesis.
Laviosa, Sara. 2002. *Corpus-based translation studies: Theory, findings, applications*. Amsterdam: Rodopi.
Lefevere, André. 1992. *Translation, rewriting, and the manipulation of literary fame*. London: Routledge.
Lowry, Richard. 2012. *Concepts and applications of inferential statistics*. http://vassarstats.net/textbook/ (26 August, 2015).
Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.
Mauranen, Anna & Pekka Kujamäki (eds.). 2004. *Translation universals: Do they exist?* Amsterdam: John Benjamins.
Mossop, Brian. 2014. *Revising and editing for translators*. 3rd edn. Abingdon: Routledge.
Müller, Gereon. 1993. *On deriving movement type assymetries*. Universität Tübingen PhD thesis.
Müller-Feldmeth, Daniel, Uli Held, Peter Auer, Sandra Hansen-Morath, Silvia Hansen-Schirra, Karin Maksymski, Sascha Wolfer & Lars Konieczny. 2015. Investigating comprehensibility of German popular science writing. In Karin Maksymski, Silke Gutermuth & Silvia Hansen-Schirra (eds.), *Translation and comprehensibility*, 227–261. Berlin: Frank & Timme.
Muñoz Martín, Ricardo. 2010. On paradigms and cognitive translatology. In Gregory M. Shreve & Erik Angelone (eds.), *Translation and cognition*, 169–187. Amsterdam: John Benjamins.

Norris, Sigrid & Rodney H. Jones. 2005. *Discourse in action: Introducing mediated discourse analysis*. Abingdon: Routledge.

Pasch, Renate, Ursula Braußе, Eva Breindl & Ulrich Hermann Waßner. 2003. *Handbuch der deutschen Konnektoren: Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln)*. Berlin: de Gruyter.

Pym, Anthony. 2005. Explaining explicitation. In Krisztina Karoly & Ágata Fóris (eds.), *New trends in translation studies: In honour of kinga klaudy*, 29–34. Budapest: Akadémia Kiadó.

Pym, Anthony. 2008. On toury's laws of how translators translate. In Anthony Pym, Miriam Shlesinger & Daniel Simeoni (eds.), *Descriptive translation studies and beyond: Investigations in honor of gideon toury*, 311–328. Amsterdam: John Benjamins.

Quasthoff, Uwe, Dirk Goldhahn & Gerhard Heyer. 2013. Deutscher Wortschatz 2012. *Technical Report Series on Corpus Building* 1.

Rabadán, Rosa & Marlén Izquierdo. 2013. A corpus-based analysis of English affixal negation translated into Spanish. In Karin Aijmeer & Bengt Altenberg (eds.), *Advances in corpus-based contrastive linguistics: Studies in honour of Stig Johansson*, 57–82. Amsterdam: John Benjamins.

Reis, Marga. 1995. Wer glaubst du hat recht?: On so-called extractions from verb-second clauses and verb-first parenthetical constructions in German. *Sprache & Pragmatik* 36. 27–83.

Robertson, Colin. 2010. Legal-linguistic revision of EU legislative texts. In Maurizio Gotti & Christopher Williams (eds.), *Legal discourse across languages and cultures*, 51–73. Frankfurt/M.: Peter Lang.

Saldanha, Gabriela & Sharon O'Brien. 2013. *Research methodologies in translation studies*. Abingdon: Routledge.

Schäffner, Christina. 2012. Rethinking transediting. *Meta* 57(4). 866–883.

Schiller, Anne, Simone Teufel, Christine Stöckert & Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS*. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf (29 October, 2015).

Schmid, Helmut. 1995. Improvements in part-of-speech tagging with an application to German. *Proceedings of the ACL SIGDAT-Workshop*. 1–9.

Schumacher, Helmut, Jacqueline Kubczak, Renate Schmidt & Vera de Ruiter. 2004. *VALBU – Valenzwörterbuch deutscher Verben*. Tübingen: Gunter Narr.

Scollon, Ron. 2001. *Mediated discourse: The nexus of practice*. London: Routledge.

Steinbach, Markus. 2002. *Middle voice: A comparative study in the syntax-semantics interface of German*. Amsterdam: John Benjamins.

Steiner, Erich. 2001. Translations English–German: Investigating the relative importance of systemic contrasts and of the text-type "Translation". *SPRIKreports* 7. 1–48.

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

Ulrych, Margherita. 2009. Translation and editing as mediated discourse: Focus on the recipient. In Rodica Dimitriu & Miriam Shlesinger (eds.), *Translators and their readers: In homage to Eugene A. Nida*, 219–234. Brussels: Les Editions du Hazard.

Ulrych, Margherita & Simona Anselmi. 2008. Towards a corpus-based distinction between language-specific and universal features of mediated discourse. In Aurelia Martelli & Virginia Pulcini (eds.), *Investigating English with corpora: Studies in honour of Maria Teresa Prat*, 257–273. Monza: Polimetrica.

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. Frankfurt/M.: Peter Lang.

Utka, Andrius. 2004. Phases of translation corpus: Compilation and analysis. *International Journal of Corpus Linguistics* 9(2). 195–224.

Vintar, Špela & Silvia Hansen-Schirra. 2005. Cognates: Free rides, false friends or stylistic devices? A corpus-based comparative study. In Geoff Barnbrook, Pernilla Danielsson & Michaela Mahlberg (eds.), *Meaningful texts: The extraction of semantic information from monolingual and multilingual corpora*, 208–221. London: Continuum.

Xiao, Richard & Xianyao Hu. 2015. *Corpus-based studies of translational Chinese in English–Chinese translation.* Shanghai: Jiao Tong University Press.

Zanettin, Federico. 2014. Corpora in translation. In Juliane House (ed.), *Translation: A multidisciplinary approach*, 178–199. Basingstoke: Palgrave Macmillan.

Change your backtitle in localmetadata.tex

Change your blurb in localmetadata.tex

DRAFT

of 24th November 2015, 11:28