Xianyao Hu*, Richard Xiao and Andrew Hardie

How do English translations differ from non-translated English writings? A multi-feature statistical model for linguistic variation analysis

DOI 10.1515/cllt-2014-0047

Abstract: This paper discusses the debatable hypotheses of "Translation Universals", i. e. the recurring common features of translated texts in relation to original utterances. We propose that, if translational language does have some distinctive linguistic features in contrast to non-translated writings in the same language, those differences should be statistically significant, consistently distributed and systematically co-occurring across registers and genres. Based on the balanced Corpus of Translational English (COTE) and its non-translated English counterpart, the Freiburg-LOB corpus of British English (FLOB), and by deploying a multi-feature statistical analysis on 96 lexical, syntactic and textual features, we try to pinpoint those distinctive features in translated English texts. We also propose that the stylo-statistical model developed in this study will be effective not only in analysing the translational variation of English but also be capable of clustering those variational features into a "translational" dimension which will facilitate a crosslinguistic comparison of translational languages (e. g. translational Chinese) to test the Translation Universals hypotheses.

Keywords: Translation Universals, translational English, linguistic variation, multi-feature analysis

1 Introduction

This paper aims to explore the debatable hypotheses of "Translation Universals" (TUs), as first proposed by Baker (1993: 243) and defined as "the universal features of translation, that is, features which typically occur in translated text rather than original utterances and which are not the result of interference from

***Corresponding author: Xianyao Hu,** College of International Studies, Southwest University, Chongqing, China, E-mail: huxyao@hotmail.com
Richard Xiao: E-mail: r.xiao@lancaster.ac.uk, **Andrew Hardie:** E-mail: a.hardie@lancaster.ac.uk, Department of Linguistics and English Language, Lancaster University, Lancaster, UK

specific linguistic systems". Although some critics dislike the idea that an activity such as translation, which is traditionally considered to be only a secondary form of linguistic behaviour, could have universals (e. g. Tymoczko 1998; Malmkjaer 2005; House 2008; Pym 2008), it is clear that – if confirmed – the TU hypotheses would have great epistemic value and explanatory power.

Discrepancies between translated and non-translated texts in the same Target Language have been explained superficially or pejoratively in earlier research, either in terms of "translationese", caused by the translator's incompetence (e. g. Newmark 1991), or in terms of inappropriate "interference" from the Source Language (e. g. Toury 1979: 226, 1995: 274). However, neither of these two explanations adequately recognizes the commonalities of translation as a universal human linguistic practice through history and across societies. If we focus on the nature of translation as an activity or practice, however, rather than as simply product or content, then it would seem uncontroversial that any consistent features of the practice may result in consistent – that is, universal – features of the linguistic product.

Thus, the concept of Translation Universals is in line with views of translational language as an "inter-language" (Toury 1979), the "third language" (Duff 1981: 12), the "third code" (Frawley 1996: 168) or a "hybrid language" (Schäffner and Adab 2001). Work on TUs thus constitutes an attempt to look into what translating really is, that is, to view translating as a universally similar and cognitively distinct activity across languages and cultures. Under the TU framework, translational language is worthy of serious investigation in its own right, because systematic description of the distinctive linguistic features of translational language is essential to understanding the fundamental mechanism and principles of the process of translation.

As a relatively young theory and new paradigm in translation studies, the TU hypotheses have been scrutinized, questioned, and strongly criticized or rejected by some (e. g. House 2008; Becher 2010), despite growing interest and research in this area (e. g. Kenny 2001; Laviosa 2002; Granger et al. 2003; Mauranen and Kujamäki 2004; Hansen 2003; Mauranen 2004; Olohan 2004; Anderman and Rogers 2008; Chesterman 2010; Xiao 2010; Grabowski 2012) over the past two decades.

Among those strong criticisms against the TU hypotheses, House's (2008: 11) blunt rejection of TU is representative, as she argues that "the quest of translation universals is in essence futile." This assessment is based on her observations that (1) universals of language also apply to translation; (2) translation is inherently language-pair specific; (3) the directionality of translation affects universality; (4) language is genre-specific; and (5) the diachronic development of texts can also affect universality (House 2008: 11–12). The first of these points

relates to the term "universals", which can be considered premature or over-generalized in this context. As noted in Chesterman (2010), the problem is in fact terminological because universals "must be understood in a weaker sense in translation research"; they are "general tendencies or patterns or indeed simply generalizations qualified and conditioned as necessary" (Chesterman 2010: 44) rather than absolute restrictions as in the classical (Chomskyan) understanding of what a "universal" is. House's other four points all concern the specificities of translation which in her view fall beyond the explanations of translation universals, for us, however, it is these specificities that make generalizations regarding translational language different from those in other linguistic disciplines. To put it another way, what House regards as problems for the notion of translation universals may be better seen as the determining factors or conditions of general (if not universal) tendencies of translation.

There is also doubt regarding the vague or too general definitions of specific TU features, such as Simplification, Explicitation and Normalization (see Xiao and Dai 2014 for a full review). For example, the most widely studied and debated translation universal, Explicitation, i. e. the idea that translations tend to be more explicit than their originals (Blum-Kulka 1986), has been defined and interpreted so differently that makes it impossible to compare results (e. g. Becher 2010; cf. Chesterman 2010). In addition, evidence proffered in support of a certain universal often co-exists with counter-evidence, so that contradictory conclusions have been made regarding features at different levels, in different genres, and in different languages. For example, Laviosa (1998) finds proof of the Simplification hypothesis in an investigation of the lexical patterns in translated English, whereas by contrast Mauranen (2004) notes that there are more "unconventional collocations" of words in translated English, which is evidence against the Simplification hypothesis. Finally, the research into TU hypotheses has so far been largely Eurocentric, in that "existing evidence in support of the proposed TU hypotheses has mostly come from Translational English and related European languages" (Xiao and Dai 2014: 3). Criticisms and doubts such as these have been addressed in a number of different ways in corpus-based descriptive translation studies (e. g. Mauranen et al. 2004; Chesterman 2010), the debate, however, is an on-going one. In this paper, our focus is on the empirical study of translational English, we would therefore adopt the methodological assumption that the TU hypotheses are valid and attempt to find linguistic consistencies that support these hypotheses. Nevertheless, we are open to the possibility of failure in this endeavour – which would then point to the TU hypotheses lacking validity.

Apart from the theoretical issues, the issue of methodology also poses a dilemma for the TU research. A central question before we embark on

crosslinguistic research into translation universals should be whether or not there are general tendencies or typical features in translated texts across genres or registers in the same language. In other words, are the distinctive features that we find between translated and non-translated texts statistically significant, systematically and consistently distributed across registers? This must necessarily be addressed first because, if no feature were found to be distinctive, stable and systematic in a given single translational language, it would not be reasonable to expect to find crosslinguistic universal features. Of course, it is not inconceivable that a feature (or trend) might be found to be a consistent marker of translation across languages, but only in one register. Methodologically, however, it is hard to distinguish such features or trends from general register variation, so we will not address this possibility further here. To date, a major methodological defect in TU research seems to be the lack of a unified analytic model, which has caused much confusion in discussion of linguistic features at various levels, in different genres, or in different languages. Researchers interested in translation universals not only tend to define those universal features differently – they also tend to address different linguistic features (at the lexical, phrasal, syntactical, textual and/or stylistic levels) measured in different terms and studied in different genres. Clearly, TU research (like, in fact, much other linguistic research) is now entering an era of multivariate data as defined by Rencher (2002: 1) "a collection of data where several measurements (or variables) have been made on a number of objects or individuals (or units)". Multivariate data naturally calls for more robust statistical (multivariate) methods for more meaningful and efficient analysis. In fact, several recent studies use multivariate techniques to analyse translated texts (e. g. Jenset and McGillivray 2012; Delaere and Sutter 2013; Diwersy et al. 2014).

Another reason for the lack of a unified analytic model is the absence of a representative balanced corpus of translational English, with a comparable non-translated counterpart, which could facilitate a systematic account of the universal and typical features of the translational variety across various register and genres in that language. For example, the well-known Translational English Corpus (TEC) is not balanced, since it consists of only four text types, namely fiction, biography, news and inflight magazines (Baker 2014). With an unbalanced corpus structure and its exclusion of many major English genres, the TEC cannot be fully representative of the translational variety of English. Moreover, there is no corpus of non-translated English which is sufficiently comparable to the TEC in terms of sampling period, size and composition for contrastive analysis (although subparts may be comparable, e. g. the fiction part of the TEC with the fiction part of the British National Corpus) (see Olohan 2002). While a few exceptions of balanced translational corpora exist (e. g. Čulo et al.

forthcoming; Macken et al. 2011), English's status as a world language, and in consequence the most common single target language for translations, makes plain the need for such a balanced corpus of translational English.

In the present paper, we provide some solutions to this methodological problem by addressing the following research questions, which we think are central to a comprehensive account of translational English:

1. Are there any statistically distinctive linguistic features of translational texts in relation to their non-translational English counterparts?
2. Do these features distribute consistently across text categories, registers or genres?
3. Do these features co-occur systematically to realize the shared function of a "translational" dimension?

Based on a balanced corpus of translational English and a comparable non-translational corpus, we will use a new stylo-statistical model, which is built on Biber's (1988) multidimensional approach to identify the cross-register distinctive features of translational English. Section 2 briefly introduces the two comparable corpora used in the research, namely the Corpus of Translational English (COTE) and the existing non-translational counterpart, the Freiburg–LOB Corpus of British English (FLOB). Section 3 provides a description of the model and describes the procedure of corpus analysis, by discussing the selection, retrieval and quantitative standardization of a total of 96 linguistic features at different levels and in different categories. Furthermore, the stylo-statistical analytic framework will be outlined briefly. In Section 4, we present the findings of our analyses: the statistically significant features of translated texts in comparison to non-translated texts, the distribution of these features across registers, as well as the convergence of these features through factor analysis. On the basis of this, we will discuss the links between these typical translational features and the TU hypotheses. Finally, Section 5 concludes by summarizing the major findings and discussing their implications, particularly the possibility of applying the stylo-statistical model established in this study to the analysis of translational varieties of other languages (especially unrelated languages such as Chinese) to test the hypothesized crosslinguistic universals.

2 The corpus of translational English

With the aim of identifying the distinctive linguistic features of translational English and investigating variations in the use of these features across registers,

we created a one-million-word balanced and representative corpus of translated texts, i. e. the Corpus of Translational English (COTE), which is designed as a translational counterpart for the Freiburg–LOB (FLOB) corpus of British English (Hundt et al. 1998). These two comparable corpora enable the present study to investigate not only the distinctive features of translational English in comparison with non-translational English but also variation in these features across various registers.

COTE is intended to match FLOB and other "Brown Family" corpora (including the 1960s Brown and LOB corpora as well as the corresponding 1990s corpora Frown and FLOB)[1] as closely as possible in size and composition, but represents written translational English published in the 1990s. The particular time period of the 1990s was chosen in order to make the results regarding translational English comparable with earlier studies based on translational Chinese from the same period. For example, Xiao and Dai (2014) present a comprehensive investigation of translational Chinese based on two comparable Chinese corpora, the (non-translational) Lancaster Corpus of Mandarin Chinese (LCMC) and the ZJU Corpus of Translational Chinese (ZCTC). Both these corpora follow the FLOB corpus design and are designed to represent non-translated and translated Chinese texts published in the 1990s.

COTE follows the same corpus design of the "Brown Family"; that is, it contains 500 text samples of around 2,000 words distributed across 15 genres. (The total token count is somewhat higher than one million because the above figures include punctuation marks as separate tokens; however, we did not count punctuation marks as words when measuring the 2,000 word sample length, following the usual practice for the Brown Family.) Table 1 lists the genres with the numbers of texts, words and paragraphs in COTE. It should be noted that the News component of COTE is a combination of three Brown Family genres (i. e. reportage, editorial and review) due to difficulties encountered in collecting and classifying translated texts in these genres.[2]

In this research, we compare the translational and non-translational English varieties as represented by the 1,000 text samples in the COTE and FLOB corpora, totalling approximately 2.3 million tokens. For the sake of simplicity of statistical analysis, the 15 genres are grouped into four major registers: (1)

1 Where LOB/FLOB differs from Brown/Frown in composition, COTE follows LOB/FLOB.

2 Translated texts in the News genres differ significantly in length and subject; some are very short and have to be combined into texts of about 2,000 words; some are too diversified to be properly classified into one single genre. To ensure the comparability with FLOB, we combined the short texts into the same number of texts (i. e. 88 texts) of similar length, but did not classify them into three genres.

Table 1: COTE corpus design.

Type	Register	Code	Genre	Texts	Tokens	Sentences	Paragraphs
Non-literary	News (17.6 %)	ABC	News (reportage, editorial, review)	88	207,613	8,789	4,458
	General prose (41.2 %)	D	Religious writing	17	43,135	1,587	356
		E	Skills, trades and hobbies	38	95,342	5,010	1,951
		F	Popular lore	44	112,348	4,728	1,081
		G	Essays and biography	77	196,660	8,266	1,422
		H	Official documents	30	91,431	4,863	3,101
	Academic prose (16.0 %)	J	Academic prose	80	197,560	7,618	1,961
Literary	Fiction (25.2 %)	K	General fiction	29	77,707	4,157	949
		L	Mystery and detective stories	24	61,725	3,716	959
		M	Science fiction	6	15,369	899	286
		N	Adventure fiction	29	74,056	3,943	1,198
		P	Romantic fiction	29	74,790	4,094	1,122
		R	Humour	9	28,246	2,124	999
Total				500	1,275,982	59,794	19,723

News (A-C, 17.6 %); (2) General prose (D-H, 41.2 %); (3) Academic prose (J, 16.0 %); and (4) Fiction (K-R, 25.2 %). These are furthermore grouped at a higher level into two broad categories, Non-literary (74.8 %) and Literary (25.2 %) texts. Although our contrastive analyses of translational and non-translational English will be mostly based on the major registers and broad categories, it is still possible to discuss variation of features across the more fine-grained genres where necessary.

While COTE is designed to be comparable to FLOB in size, composition, sampling period, part-of-speech annotation and other technicalities, it is also a translational corpus. For that reason, as much translation-specific metadata as possible is included in the text headers, for instance, the Source Language, translator, date and source of publication. This makes it possible to define and compare different subcorpora according to various metadata categories for a range of research purposes. For instance, one obvious analysis which COTE enables (and which we pursue elsewhere: Hu et al. forthcoming) to investigate the impact of various source languages on translational variation of English: the texts in COTE are translated from over 18 source languages which vary in how they differ from English in terms of both genetic distance and relative socio-political status to English. It could be interesting to many people in translation and language contact studies to see what common and/or different roles those

source languages play in making translational English distinguishable to non-translational English.

3 Methodology

The methodology taken in this paper is a composite multivariate statistical model. It is inspired by Biber's (1988) multidimensional approach to linguistic variation across speech and writing, but incorporates additional statistical measures such as the non-parametric test, multiple comparisons (e. g. mean rank, mean, median), and an enhanced model of factor analysis. Two major methodological steps are involved: (1) the selection, retrieval and standardization of linguistic features for analysis; (2) the implementation of the statistical model.

3.1 Selection, retrieval and standardization of linguistic features

As it is not possible to know which linguistic features are sufficiently strong or statistically significant before analysis, we follow Biber's advice to include as many linguistic features as possible initially. Our analysis is based on a total of 96 linguistic features, including the 67 features used by Biber (1988) as well as other features which have been widely discussed in previous studies of TUs (e. g. Kenny 2001; Laviosa 2002; Granger et al. 2003; Hansen 2003; Mauranen et al. 2004; Mauranen 2004; Olohan 2004; Anderman and Rogers 2008; Xiao 2010). This constitutes, we would argue, a valid extension of Biber's original approach, in which the feature list was in part based on features found in the literature to be important to register distinctions. Applying the same basic concept, we add to Biber's list further features the previous research in the literature finds to be important markers of translation (see above). While it is of course possible that relevant features exist that are not on our list because we were not aware of them, the same criticism could also be applied to Biber's original work, which has proven both adequate and extensible in subsequent register research. The Appendix of this paper lists all the features that are included in our analysis, which can be roughly grouped into three types according to their source:

1. Biber's (1988: 223–245) 67 features in 16 categories (A–P);
2. Textual features discussed in the literature cited above (Category Q), including such features as *Lexical Density* (LD, i. e. the proportion of all content words, namely, nouns, verbs, adjectives and adverbs), *Grammatical*

Explicitness (GEX, the proportion of all function words, including *possessive pronoun* (APPGE), *articles* (AT), *before-clause marker* (BCL), *conjunction* (C), *determiner* (D), *existential there* (EX), *genitive marker* (GEM), *preposition* (I), and *infinitive marker* (INFTO)), *punctuation* (PUNC), *short length words* (STW, ≤ 3 letters), *long words* (LNW, ≥ 7 letters), *Average Sentence Length* (ASL), *Average Paragraph Length* (APL), *Average Sentence Section Length* (ASSL, i. e. the number of tokens divided by the number of all punctuation marks[3]) and *the proportion of the 10 most frequent words* (TOP10);

3. Other features which are not included in the first two groups but which, we hypothesized, might be important for differentiating translation and non-translated writing (Category R), for example, *reformulation marker* (RFFM) and *foreign word* (FW), among others.

Note that the nature of factor analysis is such that, if the features in Group 3 turn out not to be relevant, that does not endanger the analysis, since any such features will simply not be added to a factor which contains relevant features. While the features in the above three groups could also be categorized in terms of their functions (as in Biber's original categories) and levels (lexical, phrasal, syntactic, textual), we are more interested, in this study, in discovering possible links between these features and the TU hypotheses. For example, the Simplification hypothesis, i. e. the "tendency to simplify the language used in translation" (Baker 1996: 181–182) can be translated into the proposition that we will observe lower information load (lexical simplification), reduced abstractness, and less frequent use of subordination (syntactic simplification) in translated text. Similarly the Explicitation hypothesis (Blum-Kulka 1986:19) would lead us to expect translation to be characterized by strengthened grammatical or cohesive explicitness, realized by the overrepresentation of prepositional phrases, determiners and conjunctions, a dispreference for reduced forms or contractions, and extended length of sentences and paragraphs. The Normalization hypothesis, which refers to "the tendency to conform to patterns and practices that are typical of the target language" (Baker 1996: 183), would lead us to predict translated texts to be characterized with overuse of the most frequent words, reduced creativity (for example, less frequent use of hapax legomena (cf. Kenny 2001), i. e. words that occur only once in the corpus), and

3 Average Sentence Section Length is different from Average Sentence Length in that the former is calculated from the distance between intra-sentential punctuation marks (colon, semi-colon, comma, etc.) rather than the distance between sentence-separator punctuation marks, on which the latter is based. ASSL is a feature of possible relevance to translation because we might expect explicitation in translation to lead to greater use of intra-sentential punctuation.

a dispreference for atypical or abnormal usages in translated texts. It must be noted that these connections are highly hypothetical and exploratory at this stage. Hence we do not presuppose that these features actually do distinguish between non-translational and translational text. The distinctiveness, consistency and convergence of these translation-specific features or TU candidates will be discussed on the basis of our contrastive statistical analyses (see Section 4.4).

A range of corpus linguistic tools and techniques have been used to automatically extract feature frequencies, including the online corpus query processor CQPweb (Hardie 2012) and a number of stand-alone software tools. We used Nini's (2013) Multidimensional Analysis Tagger (Version 1.1) to retrieve the 67 features drawn from Biber's (1988) list. The statistics related to textual features were mostly extracted using WordSmith Tools (Scott 2014), and the remaining features were retrieved via regular expression queries in CQPweb. Prior to the statistical analysis, the raw frequencies were standardized for each text to a common basis of frequency per 100 words.

3.2 The multi-feature statistical model of analysis

As noted in Section 1, one of the problems of previous corpus-based studies on TUs (e. g. Mauranen et al. 2004) is the lack of a unified analytical framework to identify most, if not all, of the linguistic features that differentiate translated from non-translated texts. It would be desirable if such a framework could also show the respective importance of particular linguistic variables within this differentiation. Using such a framework, the conflicts between contradictory evidence at various levels as demonstrated in previous TU research (see Section 1) can be resolved. This being the case, Biber's (1988) multidimensional approach is an optimal basis for our method, because it covers an extensive – and, as we illustrate above, extensible – range of linguistic features and develops a statistical model to reduce those features into a few underlying dimensions in terms of co-occurrence and shared functions.

Inspired by Biber's multidimensional approach, our model for statistical analysis of translational English will make use of multiple linguistic features as defined by Biber (1988: 223–245) as well as other features which have been extensively discussed in corpus-based TU research (see Section 3.1). The basic idea of our model is similar to that of Biber's approach, namely, to include as many features as possible initially, at the lowest possible level of grouping, without presupposing the distinctiveness or importance of those features (cf. Biber 1995). But our model differs from Biber's in that (1) we focus on the

statistical distinctiveness of the features between translational and non-translational/native English by carrying out a nonparametric test; (2) we examine the consistency of those features by comparing group means, mean ranks and medians; (3) we carry out both exploratory and confirmatory factor analysis to see whether or not these features can be grouped to represent a "translational" dimension. The procedure for analysis is as follows.

Firstly, by using the frequency data for all the linguistic features discussed in Section 3.1, we examine whether or not the distribution of each feature in the translated English texts is statistically distinctive from that in the corresponding non-translated texts. This is done by carrying out a non-parametric test of statistical significance to detect differences in the distribution of a feature in two independent samples. A non-parametric test was necessary because parametric analyses, such as the ANOVA (analysis of variance), have strict prerequisites regarding the normal distribution and variances of the data which, in this case, may not be fulfilled. A non-parametric test does not have these prerequisites but is still capable of identifying the distributional distinctiveness of a feature in our analysis. The test we used is the Mann-Whitney U test,[4] also known as the Wilcoxon rank-sum test; this is the most frequently used non-parametric test for use with skewed data. It is used to test whether two independent samples – in this case, the translational versus non-translational texts – are from the same or identical distributions; the null hypothesis (H0) in this case is that they are from identical distributions, and thus do not distinguish translational and non-translational language. Only when the Mann-Whitney U test provides evidence against this null hypothesis do we consider a feature to be "distinctive".

Secondly, we examine the consistency of each feature. By "consistency", we mean the consistency in the direction of the differences between the translational and non-translational texts across broad text categories (Non-literary versus Literary texts), across major registers (News, General Prose, Academic Prose and Fiction), and across the 13 genres (with the 3 news genres combined). We regard as "consistent features" those with either all-positive or all-negative differences between the translational and non-translational texts across text categories, registers and genres. The list of distinctive and consistent linguistic features will be used as a checklist for the convergence analysis that follows.

Thirdly, factor analysis will be conducted on the basis of the normalized frequencies of all 96 features in each of the 1,000 translated and non-translated texts, with the aim of identifying any potential co-occurrence patterns. Our goal

4 The statistical software we use for statistical significance testing, factor analysis, and so on is IBM SPSS version 21.

in carrying out a factor analysis is similar to its purpose in Biber's approach, in that we aim to ascertain whether there is evidence of a "translational" dimension that embraces the distinctive and consistent features identified previously. However, unlike Biber, we are not interested, for present purposes, in detailed investigation of any other dimensions that emerge. We would expect dimensions related to genre/register variation to emerge, as in Biber's study, because of the variety of texts in the two corpora, regardless of whether or not a translational dimension can be identified. The factor analysis thus functions as a double-check of whether the distinctive and consistent features co-occur or co-vary to realize their shared function of differentiating translational from non-translational English texts.

4 Findings and discussion

4.1 Statistically distinctive features

Using the normalized frequencies of all 96 linguistic features for the 1,000 texts in COTE and FLOB, we applied the Mann-Whitney U test, as discussed above, to identify features for which it is possible to reject the null hypothesis that the distribution of that feature is the same in translational as in non-translational texts. A feature where this null hypothesis is rejected at the $p < 0.05$ level of statistical significance is therefore regarded as a "statistically distinctive feature". As well as testing the data for the two corpora considered as units (the "overall level" of the analysis), we also applied the U test to examine the distinctiveness of each feature just between the translational and non-translational subcorpora formed by our two broad text categories and the four major registers. Table 2 gives the results of this process of statistical testing.

Due to limits of space, only the 29 features which are distinctive at the overall level and in the two broad categories are shown in Table 2. The results of the Mann-Whitney U test indicate that:

1. In the overall tests comparing the entirety of each corpus, 62 out of 96 features are statistically distinctive between translational and non-translational English texts;
2. 29 features are distinctive at the overall level and in both the Non-literary and Literary categories (see Columns 2–3);
3. Only six features are distinctive at the overall level, in both the broad categories and in all four major registers, namely, *possessive pronoun*

Table 2: The Mann-Whitney U test for distinctive linguistic features between translation and non-translation.

		Overall [62]		NonLit [60]		Lit [58]		News [73]		General [46]		Academic [45]		Fiction [59]	
		p	H_0	*p*	H_0	*p*	H_0	*p*	H_0	*p*	H_0	*p*	H_0	*p*	H_0
1	APPGE	0.002	×	0.005	×	0.001	×	0.009	×	0.010	×	0.000	×	0.001	×
2	EX	0.000	×	0.000	×	0.001	×	0.000	×	0.000	×	0.001	×	0.001	×
3	GEX	0.000	×	0.000	×	0.000	×	0.000	×	0.007	×	0.000	×	0.000	×
4	LD	0.000	×	0.000	×	0.001	×	0.047	×	0.000	×	0.000	×	0.001	×
5	TOP10	0.000	×	0.000	×	0.000	×	0.000	×	0.002	×	0.000	×	0.000	×
6	WZPRES	0.000	×	0.000	×	0.027	×	0.029	×	0.000	×	0.002	×	0.027	×
7	APL	0.000	×	0.010	×	0.000	×	0.000	×	0.038	×	0.109	○	0.000	×
8	CONT	0.000	×	0.000	×	0.000	×	0.000	×	0.000	×	0.319	○	0.000	×
9	DEMO	0.000	×	0.000	×	0.000	×	0.000	×	0.000	×	0.054	○	0.000	×
10	DPAR	0.000	×	0.000	×	0.012	×	0.000	×	0.000	×	0.647	○	0.012	×
11	EMPH	0.000	×	0.000	×	0.032	×	0.000	×	0.000	×	0.188	○	0.032	×
12	PLACE	0.000	×	0.000	×	0.003	×	0.000	×	0.049	×	0.866	○	0.003	×
13	R	0.000	×	0.000	×	0.027	×	0.000	×	0.046	×	0.199	○	0.027	×
14	THATD	0.000	×	0.000	×	0.000	×	0.005	×	0.000	×	0.143	○	0.000	×
15	ASL	0.000	×	0.000	×	0.000	×	0.000	×	0.748	○	0.023	×	0.000	×
16	AT	0.000	×	0.000	×	0.005	×	0.000	×	0.921	○	0.000	×	0.005	×
17	I	0.000	×	0.000	×	0.000	×	0.000	×	0.428	○	0.002	×	0.000	×
18	LNW	0.037	×	0.015	×	0.003	×	0.000	×	0.063	○	0.004	×	0.003	×
19	PIN	0.000	×	0.000	×	0.000	×	0.000	×	0.256	○	0.001	×	0.000	×
20	V	0.000	×	0.000	×	0.001	×	0.000	×	0.930	○	0.011	×	0.001	×
21	NOMZ	0.002	×	0.000	×	0.010	×	0.000	×	0.440	○	0.101	○	0.010	×
22	PROD	0.000	×	0.000	×	0.000	×	0.000	×	0.161	○	0.811	○	0.000	×
23	RP	0.001	×	0.000	×	0.040	×	0.000	×	0.098	○	0.132	○	0.040	×
24	STPR	0.000	×	0.000	×	0.000	×	0.000	×	0.124	○	0.764	○	0.000	×
25	WZPAST	0.001	×	0.012	×	0.004	×	0.000	×	0.992	○	0.339	○	0.004	×
26	ANDC	0.000	×	0.000	×	0.007	×	0.124	○	0.000	×	0.003	×	0.007	×
27	GEM	0.000	×	0.000	×	0.001	×	0.173	○	0.000	×	0.407	○	0.001	×
28	SERE	0.001	×	0.013	×	0.010	×	0.223	○	0.486	○	0.002	×	0.010	×
29	WDT	0.000	×	0.049	×	0.000	×	0.077	○	0.428	○	0.075	○	0.000	×

The significance level is $p < 0.05$. "×" means the null hypothesis H_0 is rejected; "○" means H_0 is retained. See Appendix for feature abbreviations.

(APPGE), *existential there* (EX), *Grammatical Explicitness* (GEX), *Lexical Density* (LD), *total frequency of top 10 most frequent words* (TOP10) and *present participial WHIZ deletion relatives* (WZPRES) (see Appendix 1 for feature abbreviations);

4. Other features vary in distinctiveness within registers: 73 features are distinctive in News, the largest number; while General Prose, Academic Prose and Fiction respectively have 46, 45 and 59 statistically distinctive features between translated and non-translated English texts.

When a certain feature is non-distinctive in a given register – for example, *demonstratives* (DEMO) and *emphatics* (EMPH) do not discriminate between translational and non-translational texts in Academic Prose – this does not mean that there is no difference in that feature at the overall level or in other registers, but only means that the distinctiveness of the feature *in that particular register* is not statistically significant; it may indeed be significant at the overall level or in other registers. Likewise, overall distinctiveness of a feature does not guarantee that it will be significantly distinctive in all specific registers. For instance, *independent clause coordinatio*n (ANDC) is distinctive at the overall level and in three registers, but not in the fourth (the News). By observing and comparing the distinctiveness of these features between translational and non-translational/native texts at the overall level and in the broad text categories and registers, we are able to build up a general picture of the features that differentiate translational English in relation to non-translational English. However, this picture still lacks sufficient details regarding the consistency and systematicity of these differences. We will discuss the consistent and systematic patterns of feature differences in the following sections.

4.2 The consistency of feature differences

Our second research question concerns the consistency of differences between translated and non-translated texts. As noted in Section 3.2, by "consistency of differences", we refer to a consistent tendency for either over- or under-representation of a particular linguistic feature in translated texts in relation to non-translated texts, irrespective of text categories, registers or genres. In other words, we aim to identify potential features which are consistently overused or underused in translation in comparison with non-translation across all or most text categories. In assessing consistency, we do not take into account the significance of a difference, as determined in the previous analysis – merely its direction.

To discover these consistent differences of linguistic features, we need to investigate and compare the occurrence of the same features across different text categories, registers and genres, where each group of texts at any of these levels differs in scope and consists of a varying number of individual sample texts. It is the usual practice to discuss the central tendency and dispersion of a given feature by observing its mean and standard deviation across different groups. In our case, for instance, we could in theory look at the difference between the mean value of a feature in the group of translational texts, and

the mean value of the same feature in the group of non-translational texts. However, the mean and standard deviation may not be reliable unless the data is assumed to be normally distributed, which is rarely the case in language use. Hence, in the present study we look at differences in two other statistical measures: the mean rank and the median. The mean rank is the mean of the ranks (or ordinal levels) given to a variable (i. e. a single feature) in a group of texts in the Mann-Whitney U test. As the rank is based on the original frequency values, a higher mean rank of a group can be interpreted as reflecting (generally) a higher frequency of the underlying feature in a given group of texts, and contrariwise a lower mean rank implies (generally) a lower frequency of the underlying feature. The merit of the mean rank score is that it offers a unified scale for comparing linguistic features measured on drastically different scales. For example, *Average Word Length* (AWL) is measured in terms of the number of letters, while *Average Sentence Length* (ASL) is the average number of words in a sentence, and most other features are given as instances per hundred words. As for the median, like the mean it is a popular measure of describing the central tendency of a dataset, but is defined as the numerical value separating the higher half of a population from the lower half. For this reason, the median is more robust in the presence of outlier values which might exert an excessive influence on the mean value, and thus is particularly suitable for skewed (non-normal) distributions.

Table 3 presents the 37 features (out of 62 statistically distinctive features in the previous section) which are "consistent" in terms of their median distributions across categories, registers and genres. The features are sorted according to their mean rank differences between the two groups (translation and non-translation) in the whole of each corpus. The mean rank difference shows the distance between the two groups: the larger it is, the more striking the difference between translation and non-translation. A negative value means the translation group has a lower frequency of the feature in question, while a positive value means the translation group has a higher frequency. The *Z* scores and significance levels (*p values*) show that these mean rank differences are statistically significant. The five features at the top of the list are *emphatics* (EMPH), *Lexical Density* (LD), *total frequency of the 10 most frequent words* (TOP10), *demonstrative* (DEMO) and *existential there* (EX); this indicates that these features are the strongest, and most significant, distinguishing features between translation and non-translation at the overall level.

The mean differences are also given for reference in Table 3, although we are mainly interested in the median differences and their distributions in various text categories. Clearly, the mean rank, the mean and the median all agree with one another in the polarity of the difference between translation and non-

Table 3: Consistency of differences between translation and non-translation.

#	Features	Mean rank diff.	*z*	*p*	Mean diff.	Median diff.	Consistency		
							Category	Register	Genre
1	EMPH	−183.82	−10.063	0.000	−0.15	−0.15	−	−	2
2	LD	−171.31	−9.378	0.000	−2.05	−2.82	−	1	2
3	TOP10	160.75	−8.800	0.000	1.72	1.97	+	+	3
4	DEMO	153.19	−8.386	0.000	0.18	0.14	+	+	2
5	EX	−151.32	−8.285	0.000	−0.06	−0.08	−	−	−
6	CONT	−125.23	−7.314	0.000	−0.28	−0.04	−	−	1
7	CONC	−129.18	−7.158	0.000	−0.02	−0.04	−	−	3
8	GEX	130.58	−7.148	0.000	1.52	1.37	+	+	1
9	GEM	−119.31	−6.532	0.000	−0.13	−0.18	−	−	−
10	THATD	−116.10	−6.361	0.000	−0.07	−0.02	−	−	3
11	BEMA	−114.95	−6.293	0.000	−0.18	−0.22	−	−	4
12	ANDC	112.35	−6.151	0.000	0.16	0.12	+	+	2
13	TSUB	110.10	−6.079	0.000	0.05	0.07	+	+	3
14	PROD	−107.71	−5.985	0.000	−0.03	0.00	−	1	3
15	STPR	−102.17	−5.654	0.000	−0.02	0.00	−	−	1
16	R	−102.44	−5.608	0.000	−0.61	−0.48	−	−	3
17	PIN	98.08	−5.370	0.000	0.64	0.67	+	1	4
18	WZPRES	−98.02	−5.366	0.000	−0.03	−0.05	−	−	4
19	INFTO	−97.34	−5.329	0.000	−0.16	−0.16	−	−	2
20	ASL	95.25	−5.214	0.000	3.16	2.67	+	+	2
21	PLACE	−94.10	−5.152	0.000	−0.05	−0.06	−	1	4
22	DPAR	−86.94	−5.151	0.000	−0.01	0.00	−	−	3
23	I	91.91	−5.032	0.000	0.68	0.60	+	+	3
24	APL	87.54	−4.792	0.000	23.30	16.61	+	+	3
25	SMP	−80.82	−4.470	0.000	−0.01	−0.04	−	1	3
26	V	−80.95	−4.432	0.000	−0.74	−0.93	−	−	2
27	WDT	67.04	−3.670	0.000	0.07	0.08	+	+	2
28	GER	−63.75	−3.490	0.000	−0.06	−0.09	−	−	4
29	RP	−62.58	−3.426	0.001	−0.05	−0.04	−	1	4
30	SERE	60.64	−3.333	0.001	0.03	0.03	+	+	4
31	WZPAST	60.66	−3.322	0.001	0.05	0.03	+	+	3
32	PIT	−57.66	−3.157	0.002	−0.07	−0.11	−	1	3
33	APPGE	57.58	−3.152	0.002	0.19	0.13	+	1	4
34	PIRE	54.34	−2.990	0.003	0.02	0.03	+	+	2
35	CONJ	53.48	−2.928	0.003	0.04	0.03	+	1	3
36	FW	40.42	−2.432	0.015	−0.01	0.03	+	1	4
37	WP	37.02	−2.027	0.043	0.05	0.04	+	1	3

The *z* value and *p* value (<0.05) resulting from the Mann-Whitney U test are given in the list. Differences between translation and non-translation in mean rank, mean and median are shown. "+" and "−" mark features where translation has a consistently higher or lower frequency than non-translation.

translation, suggesting that the three statistical measures all show the same general tendencies of overrepresentation or underrepresentation for the different features.

The positive and negative signs in the last three columns in Table 3 indicate a consistent pattern of higher or lower median differences between translation and non-translation across the different broad text categories, registers and genres. That is, when the median of a linguistic feature is *higher* in translation than in non-translation across *most* categories, registers and genres, a positive sign ("+") is placed next to the feature, whereas while a negative sign ("–") is placed next to a feature with lower median in translation than in non-translation across categories, registers and genres. Where a number 1 to 4 is given instead, this indicates the number of registers or genres that are exceptions to the overall pattern. For example, *demonstratives* (DEMO) has "+" signs under Category and Register, which means the translation group has a higher median than the non-translation group in both the Literary and Non-literary categories and in all the four registers, that is, translational English has a higher frequency of demonstratives irrespective of the text category or register. But the number under Genre for DEMO is 2, indicating that there are two genres (out of 13) in which the translation group does not have a higher median.

As such, we can directly observe the polarity of the differences between translational and non-translational English, and more importantly, the consistency of these differences across all the subdivisions of the two corpora. Most features, except for *existential there* (EX) and *genitive marker* (GEM), have at least one genre which goes against the consistent pattern of the other genres. However, given the contingency of natural linguistic phenomena, a few exceptions cannot be understood as undermining the central tendency of the majority of genres. In this analysis, we tolerate up to four exceptions out of 13 genres and one exception out of four registers before rejecting the consistent pattern.

4.3 The co-occurrence of features: A "translational" dimension?

Our third research question is whether or not some features co-occur systematically to realize the shared function of marking translational language, i. e., in Biber's terms, a "translational" dimension. Biber's multidimensional approach is based on the assumption "that strong co-occurrence patterns of linguistic features mark underlying functional dimensions" (Biber 1988: 13). The Translation Universals hypotheses essentially outline, in a generalized form, what are theorized to be the distinctive features of translational language as a linguistic

variety. Therefore, if translated texts do share typical linguistic features driven by a functional distinction between translated and non-translated language, those features can be reasonably expected to co-occur or co-vary, forming a translational dimension.

Biber (1988) uses factor analysis as the primary statistical means to identify co-occurrence patterns among linguistic features. An important data reduction method, factor analysis reduces a large number of original variables (in our case the frequencies of individual features) to a small set of underlying components or factors. As Biber (1988: 79) observes, "[e]ach factor represents an area of high shared variance in the data, a grouping of linguistic features that co-occur with a high frequency". Biber (1988) established seven dimensions of variation across speech and writing. We could, in theory, have used Biber's dimensions to compare translational and non-translational language in our data, without repeating the factor analysis. However, this would not have allowed us to find evidence for an actual translational dimension. For this reason, we apply factor analysis to the 96 features in the hope that it will establish as a factor or factors a group or groups of features that represent the distinction between translational and non-translational varieties of English.

The factor analysis in our analytic model involves the following procedures: (1) assessing statistically the suitability of the data for factor analysis; (2) making a decision on extraction and rotation methods; (3) extracting factor loadings for variables in each component or factor; (4) interpreting and labelling the resulting factors; 5) analysis of the functions of the dimensions that the factors represent. We will explain each of these steps in turn.

Not all datasets are suitable for factor analysis. A pair of tests, i. e. KMO Measure of Sampling Adequacy and Bartlett's Test of Sphericity, are standardly applied to determine whether it is appropriate to apply factor analysis to a given dataset. While KMO measures whether the correlations among variables are small, it is generally agreed that factor analysis should not be applied to data with a KMO score smaller than 0.5. Bartlett's Test of Sphericity determines whether or not the correlation matrix is an identity matrix, which would indicate that the factorial model is inappropriate. Ideally, Bartlett's test should yield a large chi-square score with statistical significance (p <0.05) to validate the factorial model (for detailed explanation, see IBM 2012: 154). Table 4 shows

Table 4: KMO and Bartlett's tests.

KMO measure of sampling adequacy		0.764
Bartlett's test of sphericity	Approx. Chi-square	93,454.96
	df	4,465
	Sig.	0.000

the results of KMO and Barlett's tests. Since KMO = 0.764 and Barlett's test yield a *p*-value of less than 0.001, our data is suitable for factor analysis. Among the 96 features, there is one feature, R (all adverbs), that is "nonpositive definite" (NPD), which means that some of the eigenvalues of the correlation matrix are not positive numbers, and therefore cannot be processed for factor extraction (cf. Wothke 1993). This feature is therefore excluded from further factor analysis, which is implemented on the remaining 95 features in the 1,000 texts.

The initial stage of factor analysis produces a large number of factor components representing aspects of co-variation in the data. Table 5 lists the initial eigenvalues of the first seven components that emerged in our factor analysis (the eigenvalue is a measure of the component's strength), together with the respective and cumulative percentages of total variance explained by each component. As Table 5 and the scree plot in Figure 1 show, the first seven components account for a total of 48.6 % of the variance in the dataset, with the first component accounting for 21.7 % and the second component 9.19 %. No component after the first seven explains more than 2 % of the variance. Hence, we extract only the first seven components as factors. The final step of the factor analysis is to rotate the factors; we used the Varimax method of rotation (cf. IBM 2012: 156), which is the most commonly used method.

Table 5: Factor components and variance explained.

	Initial Eigenvalues		
	Total	**% of Variance**	**Cumulative %**
1	20.66	21.75	21.75
2	8.73	9.19	30.94
3	4.85	5.10	36.04
4	3.74	3.94	39.97
5	3.62	3.81	43.78
6	2.39	2.51	46.30
7	2.19	2.31	48.60

The most important output of the factor analysis is the rotated factor matrix, which shows the factorial structure and indicates the weight of each linguistic feature loaded on each of the seven factors. The loading or weight of a feature on a factor reflects the importance of that feature in constructing the factor; it indicates "the strength of the co-occurrence relationship between the feature in question and the factor as a whole" (Biber 1988: 85). The closer the loading is to 1, the stronger the co-occurrence relationship is, and the more representative the feature is on the factor. The polarity of the loadings, on the other hand, indicates

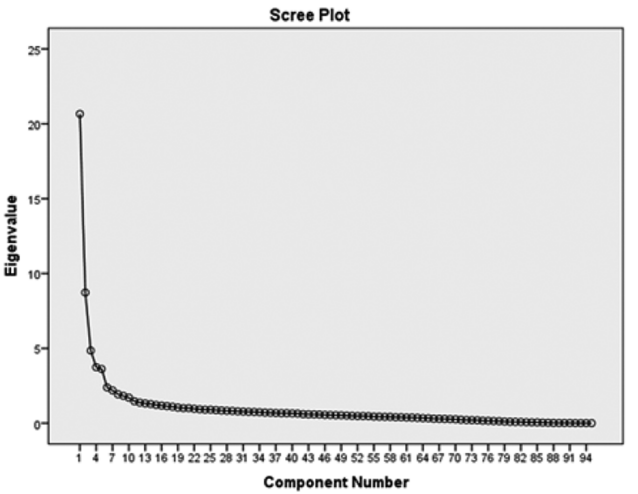

Figure 1: The scree plot of factor analysis.

the polarity of association between feature and factor. Features with positive loadings will tend to be more frequent when the score on a particular factor is high, whereas those with negative weights will tend to be less frequent when the score on a particular factor is high.

Table 6 is a simplified representation of the rotated factor matrix, simplified by excluding loadings with an absolute value less than 0.30, the commonly used cut-off point for statistical significance in factor analysis. Some features are included in more than one factor, because they all have large factor loadings (>0.30) on the respective factors. Biber's method of moving from a factor analysis to dimensions of variation involves linking each feature only to the factor on which it is most highly weighted. We did not do this, since we judged it unwise to prejudge the issue of whether or not a feature might be important to more than one dimension. This is one important respect in which our approach differs from Biber's multidimensional method.

The final step in the analysis is to interpret the seven factors as dimensions. In Biber's (1988) terms, we assume that the features that are loaded on a particular factor tend to co-occur to form a dimension that realizes a shared function. By reviewing the features loaded on a particular factor and their possible shared functions, we are able to interpret the factors using the same approach as Biber. This interpretation of the extracted factors is the key process in multidimensional analysis. As Biber (1988: 92) notes, "an underlying functional dimension is sought to explain the co-occurrence pattern among features

Table 6: The rotated factorial structure and factor loadings of linguistic features.

F	Label (based on interpretation)	*N*	Linguistic features (factor loadings)
1	Involved vs. informational production	47	P (0.915), STW (0.795), V (0.685), TPP3 (0.667), XX0 (0.645), FPP1 (0.641), PRIV (0.632), CONT (0.629), SPP2 (0.616), APPGE (0.612), WP (0.591), WRB (0.584), VBD (0.579), PUNC (0.552), WHCL (0.527), RP (0.52), STPR (0.517), DPAR (0.501), PROD (0.492), WHQU (0.485), INPR (0.465), THATD (0.412), ANDC (0.322), CS (0.311), COND (0.31), PIT (0.309), GER (–0.328), ASL (–0.384), WZPAST (–0.401), BYPA (–0.405), M (–0.412), GEX (–0.44), PASS (–0.473), TOP10 (–0.484), AT (–0.491), PHC (–0.524), ASSL (–0.594), PIN (–0.636), I (–0.653), NOMZ (–0.653), NN (–0.71), ATTRJ (–0.755), J (–0.805), LD (–0.821), LNW (–0.839), N (–0.844), AWL (–0.891)
2	Translation vs. non-translation	26	GEX (0.797), WDT (0.68), TOP10 (0.668), PIRE (0.599), ASL (0.596), I (0.586), PIN (0.584), DEMO (0.54), DT (0.513), APL (0.49), AT (0.471), WHOBJ (0.467), CS (0.424), CONJ (0.386), WHSUB (0.339), SERE (0.331), TOBJ (0.327), WP$ (0.318), ASSL (0.308), PASS (0.306), RP (–0.305), STPR (–0.313), THATD (–0.338), LD (–0.344), PUNC (–0.417), CONT (–0.481)
3	Descriptive concern	27	BEMA (0.679), PRED (0.626), EMPH (0.621), SPAU (0.565), DT (0.532), CS (0.524), DWNT (0.516), PIT (0.485), POMD (0.476), DEMP (0.423), AMP (0.421), THAC (0.391), VPRT (0.377), XX0 (0.359), EX (0.359), STTR (0.358), SYNE (0.348), SMP (0.345), CONC (0.32), CONJ (0.31), WDT (0.308), DEMO (0.303), CAUS (0.301), WZPAST (–0.382), N (–0.415), NN (–0.441), M (–0.454)
4	Persuasive concern	12	MD (0.797), PRMD (0.578), INFTO (0.527), SUAV (0.526), NEMD (0.507), COND (0.481), V (0.457), POMD (0.455), ASSL (0.349), NOMZ (0.332), PASTP (–0.409), STTR (–0.49)
5	Narrative vs. non-narrative concern	15	VBD (0.564), PEAS (0.526), STTR (0.468), APPGE (0.46), TPP3 (0.451), ASSL (0.431), TIME (0.422), RP (0.386), INFTO (0.336), DEMP (–0.333), FW (–0.374), PUNC (–0.399), REFM (–0.458), CONJ (–0.501), VPRT (–0.576)
6	Discourse representation	5	THVC (0.653), PUBV (0.631), TOBJ (0.454), GEM (0.428), PLACE (–0.388)
7	Coordinating discourse relation	3	CC (0.795), PHC (0.628), ANDC (0.515)

identified by the factor. That is, it is claimed that a cluster of features co-occur frequently in texts because they are serving some common function in those texts". How then can the seven factors that emerge from our data best be interpreted?

As shown in Table 6, Factor 1 has 47 features with factor loadings greater than 0.30, among which 27 have positive loadings and 20 have negative loadings. *Pronoun* (P), *short words* (STW), *verb* (V), *third person pronoun* (TPP3), *analytic negation* (XX0) are the top five features with highest positive loadings; by contrast, *Average Word Length* (AWL), *noun* (N), *all other nouns* (NN), *long words* (LNW), *Lexical Density* (LD) take the most strongest opposite end of negative loadings. It is fairly clear that the features with positive loadings on this factor are quite typical of spoken language and, in Biber's terms, are associated with involved concerns. By contrast, the features with negative loadings are more typical of written texts with informational concerns. This factor is very similar to Biber's (1988: 104–108) first dimension, "Involved versus informational production" and we give it the same label.

Factor 2 has 26 features, with 20 of them having positive loadings and 6 of them having negative loadings. *Total frequency of function words* (GEX), *Wh-determiner* (WDT), *Total frequency of the 10 most frequent words* (TOP10), *pied-piping relative clauses* (PRIE) and *Average Sentence Length* (ASL) have the strongest positive loadings, while *stranded prepositions* (STPR), *that deletion* (THATD), *Lexical Density* (LD), *punctuation marks* (PUNC), and *contraction* (CONT) have the strongest negative loadings. Based on the lists of distinctive and consistent translational features which emerged in Sections 4.1 and 4.2, it is very encouraging to see that many of the same features co-occur on this factor as co-varying. We thus hypothesize this factor to represent a dimension of "Translation vs. non-translation". We will verify this hypothesis shortly, and further discuss all the features of Factor 2 in detail in Section 4.4, after a brief discussion of the remaining factors.

Factor 3 brings together 27 features; *be as main verb* (BEMA), *predicative adjectives* (PRED), *emphatics* (EMPH), *split auxiliaries* (SPAU), *determiners* (DT), *subordinating conjunctions* (CS) and *downtoners* (DWNT) have the strongest positive loadings, but only four features, i. e. *past participial WHIZ deletion relatives* (WZPAST), *all nouns* (N), *all other nouns* (NN) and *numbers* (M) have negative loadings. On the one hand, it seems to us that the features with large positive loadings seem to have a common descriptive function, as non-auxiliary *be* and predicative adjectives are often used to describe a static situation while emphatics, downtoners, auxiliaries and other grammatical items are usually used to improve the accuracy of depiction; on the other hand, numbers and various types of nouns are mostly used to introduce new information. Hence this factor is named "Descriptive concern". This is admittedly a drastic oversimplification, but as noted above, the non-translational dimensions are not our primary concern, and we therefore did not need to go beyond this cursory consideration.

Factor 4, which gathers 12 features, seems very similar to Biber's (1988: 111) Dimension 4 "Overt expression of persuasion", in that all kinds of *modals* (MD, PRMD, and NEMD), *infinitives* (INFTO), *suasive verbs* (SUAV), and *conditional subordination* (COND) are brought together to underlie this dimension. We will follow Biber's interpretation of the shared function of these features to name the factor "Persuasion concern".

Factor 5 has 15 features with nine positive and six negative loadings. Features such as *past tense verbs* (VBD), *perfect participial verbs* (PEAS) and *third person pronouns* (TPP3) are in fact the same features that are most strongly loaded on Biber's Dimension 2 "Narrative versus non-narrative concerns" (Biber 1988: 108–109). In addition, the co-occurrence of these most typical narrative devices with greater *variability of word types* (STTR), *longer sentential sections* (ASSL), and more frequent *time adverbials* (TIME) and *particles* (RP) also suggests that this factor distinguishes narrativity (as present in fictional texts especially). It is accordingly labelled after Biber's Dimension 2.

Factor 6 and Factor 7 only have a few features loaded on them, indicating that they carry less weight than the other factors and are thus tentative in nature. We will name Factor 6 "Discourse representation" because of its inclusion of *public verbs* (PUBV), *that verb complements* (THVC), and *that relative clauses on object position* (TOBJ) in this dimension. Public verbs are typically speech act verbs often used in the representation (or quoting) of speech, thought or other public discourse, which tend to co-occur with the other features positively loaded on this factor. Finally, Factor 7, which has only three features with positive loadings – *coordinating conjunctions* (CC), *phrasal coordination* (PHC) and *independent clause coordination* (ANDC) – can be named "Coordinating discourse relation". Subordinating and coordinating are two types of discourse relations that correspond to the hierarchical versus non-hierarchical organization of discourse units (Fabricius-Hansen and Ramm 2008: 16). According to the Segmented Discourse Representation Theory (SDRT), which models discourse coherence and the incremental construction of discourse representations, while subordinating changes the "granularity" of description in discourse (e. g. by elaborating on some element in previous text), coordinating continues the description without changing granularity (Asher 2004; Asher and Lascarides 2003).

It should be noted that all the labels and interpretations of the seven factors are tentative by nature. They are labelled as such only for the purpose of discussion. As noted earlier, it is not our primary aim in this paper to describe the functional dimensions of English comprehensively. Our analysis is focused on Factor 2 and the dimension distinguishing translational and non-translational English that we hypothesize it to represent.

In sum, Factors 1, 4 and 5 in our model are similar to Biber's Dimensions 1, 4 and 2 respectively; Factors 3, 6 and 7 do not map to a single Biberian dimension, but do seem explicable as dimensions by the (admittedly superficial) functional analysis above. Factor 2 is the most interesting from our perspective, as it brings together features that were highlighted in our previous analyses of distinctiveness and consistency of features across the translated and non-translated varieties. We have thus hypothesized that these distinctive and consistent features form a "translational" dimension, reflected by Factor 2. Our last step is to confirm the existence of this dimension, using an analysis of factor scores.

We calculate the factor score for each of the texts. Factor scores are also used by Biber (1988: 93–97) to confirm factor interpretations. If a hypothetical dimension is valid, the factor scores of a set of texts should be explicable in terms of the proposed shared function of these texts. For our purposes, it is only necessary to apply this analysis to Factor 2 to ascertain whether that Factor is optimally discriminatory between non-translational and translational texts. By taking the mean factor scores of the 500 translated texts, and subtracting the mean factor scores of the 500 non-translated texts, we can obtain overall factor score differences between the two groups for each of the seven factors as shown in Figure 2. In the figure, a positive difference indicates that the mean factor score in the translation group is higher than in the non-translation group, and versa negative difference indicate the opposite. The larger the absolute value of the factor score difference, the greater the difference between the two groups. Factor 2 indeed shows the greatest difference in mean factor scores between the two groups of texts, by a substantial margin (0.73 compared to 0.17). Hence it can be safely concluded that Factor 2 captures the greatest difference between translational English and non-translational English. In other words, the dimension represented by Factor 2 does indeed seem to be the "translational" dimension that we have hypothesized.

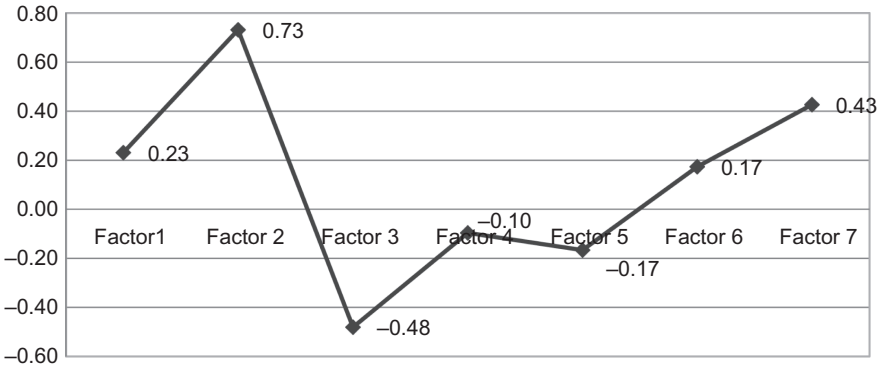

Figure 2: Overall factor score differences between translation and non-translation.

4.4 The typical features of translational English

Let us now return to the question asked in the title of this study: how do English translations differ from non-translated English writings? More specifically, in what way is the linguistic variety of translational language distinctive from the use of language in general? Our position throughout has been that this question cannot be properly answered by comparison of the frequencies of a small group of selected features. Rather, we have presented an approach to addressing the nature of this variety by looking at quantitative differences in *many* features and asking the following questions: (1) Are these differences statistically distinctive? (2) Do they distribute consistently across text categories, registers and genres? (3) Do they systematically co-vary together to realize a shared function? These questions speak to the three prerequisites for the recognition of a linguistic variety: distinctiveness, consistency, and systematicity.

We reviewed all the linguistic features, and the outcomes of the analyses we have so far applied to them, with the aim of pinpointing those typical features of translational English which are statistically distinctive, consistently distributed, and systematically co-occurring. The results of this review are given in Table 7, which lists a total of 38 features. These features are divided into three groups (shown with the dotted lines) based on how many of our conditions they fulfil. Within each group, the features in each group are sorted according to the respective loadings in Factor 2.

1. Group 1 contains 18 features which have an absolute value of factor loadings above 0.30 and a significance level from the U test p <0.05, and which are consistent across text categories, register and genres (with up to 4 exceptional genres).
2. Group 2 contains eight features included in this group which have large factor loadings (>0.30), but six of them are not statistically significant as indicated by their U test results (p >0.05). The two features (CS and WHSUB) that are significant in this group and all other features do not show consistency across categories or genres.
3. Group 3 contains 13 features which are consistently distributed and significantly distinctive, but have relatively smaller factor loadings (absolute value <0.30).

Only the features in Group 1 satisfy all three of our conditions. The features in Group 2 have high factor loadings, but are not consistently distributed (and some failed the U test). Group 3 features have low factor weights (that is, they do not affect the factor scores much). Thus, if we stick to a strict standard, only the first 18 features can be accepted as the typical features of translational English.

Table 7: Features loaded on Factor 2.

		Mean rank diff.	*p*	Consistency			Factor 2 Loadings
				Category	Register	Genre	
1	GEX	130.576	0.000	+	+	1	0.797
2	WDT	67.036	0.000	+	+	2	0.680
3	TOP10	160.748	0.000	+	+	3	0.668
4	PIRE	54.344	0.003	+	+	2	0.599
5	ASL	95.25	0.000	+	+	2	0.596
6	I	91.912	0.000	+	+	3	0.586
7	PIN	98.084	0.000	+	1	4	0.584
8	DEMO	153.192	0.000	+	+	2	0.540
9	APL	87.542	0.000	+	+	3	0.490
10	CONT	−125.23	0.000	−	−	1	−0.481
11	AT	93.43	0.000	+	1	4	0.471
12	PUNC	−83.96	0.000	−	1	4	−0.417
13	CONJ	53.476	0.003	+	1	3	0.386
14	LD	−171.31	0.000	−	1	2	−0.344
15	THATD	−116.098	0.000	−	−	3	−0.338
16	SERE	60.636	0.001	+	+	4	0.331
17	STPR	−102.17	0.000	−	−	1	−0.313
18	RP	−62.582	0.001	−	1	4	−0.305
19	DT	21.35	0.242	*	*	*	0.513
20	WHOBJ	−8.424	0.631	*	*	*	0.467
21	CS	53.104	0.004	+	*	*	0.424
22	WHSUB	−74.274	0.000	*	*	*	0.339
23	TOBJ	24.264	0.183	*	*	3	0.327
24	WP$	23.078	0.150	*	*	4	0.318
25	ASSL	−12.146	0.506	*	*	*	0.308
26	PASS	−9.62	0.598	*	*	*	0.306
27	WZPAST	60.658	0.001	+	+	3	0.282
28	DPAR	−86.936	0.000	−	−	3	−0.280
29	PLACE	−94.104	0.000	−	1	4	−0.231
30	PROD	−107.71	0.000	−	1	3	−0.230
31	V	−80.952	0.000	−	−	2	−0.212
32	TSUB	110.098	0.000	+	+	3	0.210
33	EMPH	−183.818	0.000	−	−	2	−0.209
34	ANDC	112.354	0.000	+	+	2	0.166
35	WP	37.022	0.043	+	1	3	0.161
36	EX	−151.32	0.000	−	−	−	−0.131
37	SMP	−80.816	0.000	−	1	3	0.126
38	GEM	−119.314	0.000	−	−	−	−0.122
39	REFM	31.482	0.059	+	+	2	0.106

"+"/"−" indicates a consistently higher or lower difference between translation and non-translation while "*" indicates no consistency.

However, if we take a broader view, the features in Group 3 and the two significant features in Group 2 could be included, since they meet the (primary) criterion of significance in the U test.

Let us follow the stricter criterion, and consider the 18 features that meet all three conditions. Among these, six features have both negative factor loadings and negative mean rank differences, which means that these features – namely, *particles* (RP), *stranded prepositions* (STPR), *subordinator-that deletion* (THATD), *Lexical Density* (LD), *punctuation* (PUNC) and *contraction* (CONT) – tend to be significantly underused or under-represented in translational English in comparison with non-translational English.

By contrast, the other twelve features all have large positive loadings and positive mean rank differences, indicating that they tend to be overused in translation relative to non-translation. These features include *Grammatical Explicitness* (GEX), *WH-determiner* (WDT), *the total frequency of 10 most frequent words* (TOP10), *pied-piping relative clauses* (PIRE), *Average Sentence Length* (ASL), *all prepositions* (I), *presentational phrases* (PIN), *demonstratives* (DEMO), *Average Paragraph Length* (APL), *articles* (AT), *conjuncts* (CONJ) and *sentence relatives* (SERE).

4.5 The Translation Universals hypotheses revisited

The three criteria employed above constitute a rigorous filter on what features we have admitted as typical of translational English, our statistical model has helped to identify 18 features that are distinctive, consistent and systematic. These features are obviously specific to English (lexico-) grammar and thus do not themselves constitute universals in any sense. However, we would argue that an examination of the nature of these features can provide us with a first step towards building evidential support for the Translation Universals hypotheses. If these hypotheses are correct, then we should be able to identify one or more TU hypotheses as constituting a motivation or explanation for the features that constitute translational English. If we cannot link the features of translational English to the TU hypotheses in this way, then we have failed to support those hypotheses and may question whether continued adherence to this model is justified.

What, then, is the relationship between the typical features of translational English and the TU hypotheses? To answer this question, we take a more detailed look at these features in functional terms. For this purpose, it is possible to consider the communicative functions of the 18 features and put them into more general groups according to these functions. The results of this analysis – a list of what we argue to be the distinctive, consistent and systematic communicative functions underlying the 18 features – are shown in Table 8.

Table 8: Functional characteristics of translational English.

Functional characteristics	The 18 typical linguistic features
1 Reduced information load	– *Lexical Density (proportion of lexical words)*
2 Overrepresentation of the most frequent words	– *Total frequency of the ten most frequent words*
3 Less preference for reduced forms	– *Subordinator-that deletion* – *Contractions*
4 Overrepresentation of function words	– *Grammatical Explicitness (proportion of function words)* – *All preposition tokens* – *All prepositional phrases* – *Demonstratives* – *All article tokens*
5 Extension of sentences and paragraphs	– *Average Sentence Length* – *Average Paragraph Length* – *less preference for punctuation marks*
6 Overrepresentation of relative structures and markers of logical relation	– *WH-determiners* – *Pied-piping relative clauses* – *Sentence relatives* – *Conjuncts*
7 Underrepresentation of some particular items	– *Particles* – *Stranded prepositions*

It is not difficult to find connections between these functional characteristics of translational English and the TU hypotheses as reviewed in Section 3. For example, the hypothesis of Simplification can be seen as motivating the reduced information load and the overreliance on the most frequent words in translational English. The Explicitation hypothesis can be considered an explanation for (1) the reduction of shortened forms, (2) the overrepresentation of function words, and (3) the extension of average sentence and paragraph length. There are, not surprisingly, other features that do not admit of an explanation in terms of the best-known translation universals; for instance, the overrepresentation of relative structures and markers of logical relation (WH-determiner, pied-piping relative clauses, sentence relatives and conjuncts) and the underrepresentation of some grammatical categories (particles and stranded prepositions). How might we begin to explain these? There are four trends which we might consider indicative of a formal style: (1) high use of relative structures; (2) a preference for pied piping over preposition stranding; (3) a preference for single verbs (e. g. exit) over phrasal verbs (e. g. go out) and (4) a preference for conjuncts (e. g.

alternatively). Could it therefore be tentatively hypothesized that translational English is stylistically more formal than non-translational English? This could be explicable in the context of the translation process in terms of incomplete reflection of source-language informality in the target-language output. In fact, some authorities already consider formality to be a potential translation universal feature (e. g. Baker 1996, 2004; Becher 2010). However, we of course cannot make a strong claim for a new TU of Formality without further research into the function of these features at a very detailed level.

The connections between the distinctive, consistent and systematic features of translational English and the theoretical TU hypotheses are encouraging, because it strengthens our expectation that crosslinguistic research will be able to identify a higher level of universals, general tendencies that may be manifest (of course, in different ways structurally) in different language systems. For the sake of discussion, we will refer to this type of universals as crosslinguistic universals in contrast to the monolinguistic universals of the translational variety of any given single language. We would argue that the analytical model proposed and implemented above could be used to investigate the translational varieties of other languages, identifying further sets of monolinguistic universals for those languages. Thereafter, we anticipate that we could find connections between the crosslinguistic TU hypotheses (at the level of general tendencies in communicative function) and the monolinguistic universals (at the level of typical features of the translational variety). Previous studies of translational Chinese (e. g. Xiao and Dai 2014) have demonstrated some evidence for certain features not dissimilar to those we have explored here in translational English; but naturally, much more research in multiple other languages would be required to approach any kind of certainty regarding the connections between mono- and crosslinguistic universals

5 Conclusion

In this paper, we have sought to identify the typical linguistic features of the translational English by developing a multi-feature statistical analytic model on the basis of our balanced Corpus of Translational English and the comparable non-translational corpus, FLOB. A wide range of linguistic features were analysed across registers and genres in order to find the rigorously defined statistically significant, consistently distributed and systematically co-occurring linguistic features. We have seen that, using a strict criterion for the features we admit, there are still 18 linguistic features that seem to be typical of translational English. The connections between these typical features and TU

hypotheses were also discussed. Some of these features and their respective functions can be matched to the Simplification and Explicitation hypotheses; others may suggest a need for one or more other hypotheses to be formulated. It is our hope that the findings of the present study will cast new light on the theoretically controversial TU hypotheses and the methodological debates currently prominent in corpus-based Translation Studies.

Methodologically, this paper has shown that the multivariate statistical model showcased in this study constitutes a useful addition to the toolbox of corpus-based translation studies (see Neumann 2014 for a congruent argument). As briefly reviewed in the introduction, there are other recent multivariate studies carried out on translated texts, which demonstrate as a whole the power of multivariate methods (factor analysis, principal component analysis, correspondence analysis, etc.) in linguistic and translation research. Our study, while it follows the same track, is different from those studies in that ours has taken into account as many linguistic features/variables as possible and we do not limit the analyses to one particular universal, whereas those previous studies focus either on one or a few linguistic features (e. g. Jenset and McGillivray 2012 on affix productivity of translated English) or on a single translation universal (e. g. Delaere and Sutter 2013 on normalization). This multivariate analytical model should be equally applicable to other languages, as is Biber's method from which ours draws much inspiration (see Biber 1995). It can be reasonably argued that a comprehensive and systematic description of the translational variety of the first one, then many particular languages should be the foundation of general hypotheses on crosslinguistic universals. Our exploration in this paper of links between the descriptive evidence (typical linguistic features) and theoretical generalizations (Translation Universals) will facilitate future work to validate similar theoretical proposals across different languages, whether closely or distantly related. Finally, this analytical model may prove equally effective not only in analysing translational variation but also in analysing other forms of linguistic variation, such as L2 production, learner production, regional variants, and so on.

Acknowledgements: This research was undertaken as part of the UK ESRC-funded project "Comparable and Parallel Corpus Approaches to the Third Code: English and Chinese Perspectives" (ES/K010107/1). We are also obliged to the support of the National Social Science Fund of China for the research project (11CYY010) and the Ministry of Education of China under its Program for New Century Excellent Talents in University (grant reference NCET-11-0460).

Funding: National Social Science Fund of China, (Grant/Award Number: "11CYY010") Program for New Century Excellent Talents in University, Ministry of Education of China, (Grant/Award Number: "NCET-11-0460") Economic and Social Research Council, (Grant/Award Number: "ES/K010107/1")

References

Anderman, Gunilla M. & Margaret Rogers. 2008. *Incorporating corpora: The linguist and the translator*. Clevedon: Multilingual Matters.

Asher, Nicholas. 2004. Discourse topic. *Theoretical Linguistics* 30. 163–202.

Asher, Nicholas & Alex Lascarides. 2003. *Logics of conversation*. Cambridge: Cambridge University Press.

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.

Baker, Mona. 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.

Baker, Mona. 2004. A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics* 9(2). 167–193.

Baker, Mona. 2014. Translational English Corpus (TEC). Available at: http://www.llc.manchester.ac.uk/ctis/research/english-corpus/ (accessed 1 September 2014).

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

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Blum-Kulka, Shoshana. 1986. Shifts of cohesion and coherence in translation. In J. House & Shoshana Blum-Kulka (eds.), *Interlingual and intercultural communication. Discourse and cognition in translation and second acquisition studies*, 61–71. Tűbingen: Gunter Narr.

Chesterman, Andrew. 2010. Why study translation universals? *Acta Translatologica Helsingiensia* 1. 38–48.

Čulo, Oliver, Silvia Hansen-Schirra, Stella Neumann & Karin Maksymski. Forthcoming. Querying the CroCo corpus for translation shifts. Beyond corpus construction: Exploitation and maintenance of parallel corpora. In: Silvia Hansen-Schirra, Stella Neumann & Oliver Čulo (eds.), *Beyond corpus construction: Exploitation and maintenance of parallel corpora*. Special Issue of the *International Journal of Corpus Linguistics*.

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

Diwersy, Sascha, Stefan Evert & Stella 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. Linguae et Litterae: Publications of the School of Language and Literature, Freiburg Institute for Advanced Studies. De Gruyter, Berlin.
Duff, Alan. 1981. *The third language: Recurrent problems of translation into English*. Oxford: Pergamon.
Fabricius-Hansen, Cathrine & Wiebke Ramm. 2008. Editors' introduction: Subordination and coordination from different perspectives. In C. Fabricius-Hansen & W. Ramm (eds.), *'Subordination' versus 'coordination' in sentence and text: A cross-linguistic perspective*, 1–30. Amsterdam: John Benjamins.
Frawley, William. 1996. Prolegomenon to a theory of translation. In William Frawley (ed.), *Translation: Literary, linguistic and philosophical perspectives*, 159–175. London: Associated University Press.
Grabowski, Lukasz. 2012. On translation universals in selected contemporary Polish literary translations. *Studies in Polish Linguistics* (7). 165–183.
Granger, Sylviane, Jacques Lerot & Stephanie Petch-Tyson. 2003. *Corpus-based approaches to contrastive linguistics and translation studies*. Amsterdam: Rodopi.
Hardie, Andrew. 2012. CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3). 380–409.
Hansen, Silvia. 2003. *The nature of translated text: An interdisciplinary methodology for the investigation of the specific properties of translations*. Saarbrücken: DFKI/Universität des Saarlandes.
House, Juliane. 2008. Beyond intervention: Universals in translation. *Trans-kom*, 1(1). 6–19.
Hundt, Marianne, Andrea Sand & Rainer Siemund. 1998. *Manual of information to accompany the Freiburg-LOB corpus of British English*. Freiburg: University of Freiburg.
Hu, Xianyao, Richard Xiao & Andrew Hardie. Forthcoming. General tendencies and variations of source language interference in translational English.
IBM Corporation. 2012. *IBM SPSS Statistics Base 21*. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/21.0/en/client/Manuals/IBM_SPSS_Statistics_Base.pdf (accessed 15 September 2014).
Jenset Gard, B. & Barbara McGillivray. 2012. Multivariate analyses of affix productivity in translated English. In Oakes, Micheal P. and Ji, Meng (eds.) *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*, 301–323. Amsterdam: John Benjamins.
Kenny, Dorothy. 2001. *Lexis and creativity in translation. A corpus-based study*. Manchester: St. Jerome Publishing.
Laviosa, Sara. 1998. Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43(4). 557–570.
Laviosa, Sara. 2002. *Corpus-based translation studies. Theory, findings, applications*. Amsterdam: Rodopi.
Macken, L.O., De Clercq H. & Paulussen. 2011. Dutch parallel corpus: a balanced copyright-cleared parallel corpus. *META* 56(2). 374–390. Les Presses de l'université de Montréal.
Malmkjaer, Kirsten. 2005. Norms and nature in Translation Studies. *SYNAPS* 16. 13–19
Mauranen, Anna. 2004. Corpora, universals and interference. In A. Mauranen & P. Kujamäki (eds.), *Translation Universals: Do they exist?* 65–82. Amsterdam: John Benjamins.
Mauranen, Anna & Pekka Kujamäki. 2004. *Translation Universals: Do they exist?* Amsterdam: John Benjamins.

Neumann, Stella. 2014. Beyond translation properties: The contribution of corpus studies to empirical translation theory. Paper presented at the Fourth Symposium of Using Corpora in Contrastive and Translation Studies, Lancaster University, 24–26 July.
Newmark, Peter. 1991. *About translation*. Clevedon: Multilingual Matters.
Nini, Andrea. 2013. Multidimensional Analysis Tagger 1.0 Manual [Online]. https://sites.google.com/site/multidimensionaltagger/about (accessed 18 December 2013).
Oakes, M. P. & JiMeng. 2012. *Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research*. John Benjamins.
Olohan, Meave. 2002. Comparable corpora in translation research: Overview of recent analyses using the translational English corpus. In *LREC Language Resources in Translation Work and Research Workshop Proceedings*, 5–9.
Olohan, Meave. 2004. *Introducing Corpora in Translation Studies*. London & New York Routledge.
Pym, Anthony. 2008. On Toury's laws of how translators translate. In Anthony Pym, Miriam Shlesinger & Daniel Simeoni (eds.), *Beyond descriptive translation studies: Investigations in homage to Gideon Toury*, 311–328. Amsterdam: John Benjamins.
Rencher, Alvin C. 2002. *Methods of multivariate analysis,* 2nd edn. New York: Wiley Interscience.
Schäffner, Christina & Beverly Adab. 2001. The idea of the hybrid text in translation: Contact as conflict. *Across languages and cultures*, 2(2). 167–180.
Scott, Mike. 2014. The WordSmith Tools. http://lexically.net/LexicalAnalysisSoftware/index.html (accessed 15 September 2014).
Toury, Gideon. 1979. Interlanguage and its manifestations in translation. *Meta* 24(2). 223–231.
Toury, Gideon. 1995. *Descriptive translation studies and beyond*. Amsterdam & Philadelphia: John Benjamins.
Tymoczko, Maria. 1998. Computerised corpora and the future of translation studies. *Meta* 43(4). 652–660.
Wothke, W. 1993. Nonpositive definite matrices in structural modelling. In K. A. Bollen & J. S. Long (eds.), *Testing structural equation models*, 256–293. Newbury Park, CA: Sage.
Xiao, Richard. 2010. *Using corpora in contrastive and translation studies*. Newcastle: Cambridge Scholars Publishing.
Xiao, Richard & Guangrong Dai. 2014. Lexical and grammatical properties of translational Chinese: Translation universal hypotheses re-evaluated from the Chinese perspective. *Corpus Linguistics and Linguistic Theory* 10 (1). 11–55

Appendix: Linguistic features analysed

(A) TENSE AND ASPECT MARKERS		
1	VBD	Past tense
2	PEAS	Perfect aspect
3	VPRT	Present tense
(B) PLACE AND TIME ADVERBIALS		
4	PLACE	Place adverbials
5	TIME	Time adverbials
(C) PRONOUNS AND PROVERBS		
6	FPP1	First person pronouns
7	SPP2	Second person pronouns
8	TPP3	Third person pronouns
9	PIT	pronoun it
10	DEMP	Demonstrative pronouns
11	INPR	Indefinite pronouns
12	PROD	Pro-verb do
(D) QUESTIONS		
13	WHQU	WH-questions
(E) NORMINAL FORMS		
14	NOMZ	Nominalizations
15	GER	Gerunds
16	NN	Total other nouns
(F)PASSIVES		
17	PASS	Agentless passives
18	BYPA	By-passives
(G) STATIVE FORMS		
19	BEMA	Be as main verb
20	EX	Existential there
(H) SUBORDINATION		
21	THVC	That verb complements
22	THAC	That adjective complements
23	WHCL	WH-clauses
24	TO	Infinitives
25	PRESP	Present participial clauses
26	PASTP	Past participial clauses
27	WZPAST	Past participial WHIZ deletion relatives
28	WZPRES	Present participial WHIZ deletion relatives
29	TSUB	That relative clauses on subject position
30	TOBJ	That relative clauses on object position
31	WHSUB	WH relative clauses on subject position
32	WHOBJ	WH relative clauses on object position
33	PIRE	Pied-piping relative clauses
34	SERE	Sentence relatives
35	CAUS	Causative adverbial subordinators
36	CONC	Concessive adverbial subordinators
37	COND	Conditional adverbial subordinators
38	OSUB	Other adverbial subordinators
(I) PREPOSITIONAL PHRASES, ADJECTIVES AND ADVERBS		
39	PIN	Total prepositional phrases
40	ATTRJ	Attributive adjectives
41	PRED	Predictive adjectives
42	R	Total adverbs
(J) LEXICAL SPECIFICITY		
43	STTR	Standardized Type/Token Ratio
44	AWL	Average word length
(K) LEXICAL CLASSES		
45	CONJ	Conjuncts
46	DWNT	Downtoners
47	HDG	Hedges
48	AMP	Amplifiers
49	EMPH	Emphatics
50	DPAR	Discourse particles
51	DEMO	Demonstratives
(L) MODALS		
52	POMD	Possibility modals
53	NEMD	Necessity modals
54	PRMD	Predictive modals
(M) SPECIALIZED VERB CLASSES		
55	PUBV	Public verbs
56	PRIV	Private verbs
57	SUAV	Suasive verbs

(*continued*)

(*continued*)

58	SMP	Seem/appear
(N) REDUCED FORMS AND DISPREFERRED STRUCTURES		
59	CONT	Contractions
60	THATD	Subordinator-that deletion
61	STPR	Stranded prepositions
62	SPIN	Split infinitives
63	SPAU	Split auxiliaries
(O) COORDINATION		
64	PHC	Phrasal coordination
65	ANDC	Independent clause coordination
(P) NEGATION		
66	SYNE	Synthetic negation
67	XX0	Analytic negation
(Q) OVERALL TEXTUAL FEATURES		
68	ASL	average sentence length
69	APL	average paragraph length
70	ASSL	average sentence section length
71	STW	Short words (≤letters)
72	LNW	long words (≥letters)
73	TOP10	Highest frequency words
74	LD	lexical density (proportion of lexical words)
75	GEX	proportion of function words
76	PUNC	punctuation
(R) OTHER FEATURES		
77	N	Noun
78	V	Verb
79	J	Adjective
80	M	Number
81	P	Pronoun
82	I	Proposition
83	APPGE	possessive pronoun
84	AT	Articles
85	CC	coordinating conjunction
86	CS	subordinating conjunction
87	DT	Determiner
88	GEM	genitive marker
89	REFM	reformulation marker
90	FW	foreign word
91	MD	Models
92	RP	Particle
93	WDT	Wh-determiner
94	WP	Wh-pronoun
95	WP$	possessive Wh-pronoun
96	WRB	Wh-adverb

Bionotes

Xianyao Hu

Xianyao Hu currently holds a professorship in the College of International Studies at Southwest University in China. He worked as a research associate in the Department of Linguistics and English Language at Lancaster University in 2014. He got his Ph.D. in Translation Studies from East China Normal University in 2006, and had worked as post-doctoral researcher at Beijing Foreign Studies University and Fulbright visiting scholar at the University of California Los Angeles.

Richard Xiao

Richard Xiao is Professor of Linguistics at Zhejiang University in China as well as Reader in Corpus Linguistics and Chinese Linguistics (Honorary) in the Department of Linguistics and

English Language at Lancaster University in the UK. His main research interests cover corpus linguistics, contrastive and translation studies of English and Chinese, and tense and aspect theory. His recent books in these areas include *Aspect in Mandarin Chinese* (John Benjamins, 2004), *Corpus-Based Language Studies* (Routledge, 2006), *A Frequency Dictionary of Mandarin Chinese* (Routledge, 2009), *Using Corpora in Contrastive and Translation Studies* (Cambridge Scholars, 2010), *Corpus-Based Contrastive Studies of English and Chinese* (Routledge, 2010) and *Corpus-Based Studies of Translational Chinese in English-Chinese Translation* (Springer 2015).

Andrew Hardie

Andrew Hardie is Senior Lecturer in the Department of Linguistics and English Language at Lancaster University in the UK. He is Deputy Director of the ESRC Centre for Corpus Approaches to Social Science. His major specialism is corpus linguistics – specifically, the methodology of corpus linguistics, and how it can be applied to different areas of study in linguistics and beyond. He is also interested in the use of corpus-based methods to study languages other than English, especially the languages of Asia, with an especial focus on issues in descriptive and theoretical grammar.