Translation universals: do they exist? A corpus-based NLP study of convergence and simplification

Gloria Corpas Pastor
Department of Translation and Interpreting
University of Malaga
Malaga, Spain
gcorpas@uma.es

Ruslan Mitkov, Naveed Afzal
Research Institute in Information and Language Processing
University of Wolverhampton
Wolverhampton, UK
{r.mitkov,n.afzal}@wlv.ac.uk

Viktor Pekar
Oxford University Press

Great Clarendon St.
Oxford, OX2 6DP, UK
viktor.pekar@oup.com

Abstract

Convergence and *simplification* are two of the so-called universals in translation studies. The first one postulates that translated texts tend to be more similar than non-translated texts. The second one postulates that translated texts are simpler, easier-to-understand than non-translated ones. This paper discusses the results of a project which applies NLP techniques over comparable corpora of translated and non-translated texts in Spanish seeking to establish whether these two universals hold Corpas Pastor (2008).

1 Introduction

Studying the characteristics of translated text or, more specifically, what distinctive features typically translated texts exhibit and how they differ from original, non-translated texts written by native speakers has been a topic of long-standing interest in translation studies. Initial research goes back to Toury (1995) who put forward the laws of growing *standardisation* and the law of interference, but it was Baker (1993, 1996) who formulated many of the so-called universals and proposed the use of corpora to study these. The universals attracted considerable attention from translation experts, but their formulation and initial explanation has been based on intuition and introspection with follow-up corpus research limited to comparatively small-size corpora, literary or newswire texts and semi-manual analysis. In addition, previous research has not provided sufficient guidance as to which are the features which account for these universals to be regarded as valid Corpas Pastor (2008).

In this paper we are taking a completely different and innovative approach by employing robust NLP techniques on corpora of translated texts into Spanish and on comparable corpora of non-translated Spanish in order to investigate the validity of two translation universals, those of *simplification* and *convergence*. The simplification universal is manifested in the fact that translated texts are likely to be simpler, easier-to-understand than non-translated texts. According to the convergence universal, translated texts tend to be more similar to each other than non-translated texts. The objective of this study is to establish whether these two universals hold with Spanish as target text. To this end, we analyse corpora of translated texts into Spanish and comparable corpora of Spanish non-translated texts. With the help of language processing tools we analyse the corpora with respect to a variety of lexical, grammatical, and stylistic characteristics.

2 Corpora used

In the study we examined pairs of comparable corpora from two specialisation domains – the medical and the technical; within the medical domain we worked with two kinds of corpora: translations prepared by professional translators and translations prepared by students. Below is a list of the corpora which were specifically compiled for this experiment:

- Corpus of Medical Translations by Professionals (MTP)
- Corpus of Medical Translations by Students (MTS)
- Corpus of Technical Translations (TT)
- Corpus of Original Medical Comparable to Translations by Professionals (MTPC)
- Corpus of Original Medical Comparable to Translations by Students (MTSC)
- Corpus of Original Technical Comparable to Technical Translations (TTC)

MTP is comparable to MTPC, MTS is comparable to MTSC and TT is comparable to TTC. Comparability was a crucial consideration for this study as otherwise any style or syntax comparison would have been compromised. Corpora were compiled in such a way that comparability was ensured. Design criteria comprise diatopic, diachronic, diasystematic and domain constraints. All translated texts have British or American English as the source language and peninsular Spanish as the target language. Both corpora of translated and non-translated texts have roughly the same size. MTP is composed of biomedical translations performed by professional translators (in-house or freelancers working for certified translation companies in Europe). It is a specialised reference corpus as it does not contain whole documents, but fragments composed of the target language segments of translation memories (TMs). Text types range from research papers in journals to clinical essays, textbooks, product description and PILs, users' guides and instructions for surgical equipment. Its comparable corpus of non-translated biomedical Spanish includes a similar selection of text types and topics. It is a mixed corpus, as it contains fragments and whole documents: source language segments of TMs different from the ones used to compile the MTP, a small corpus of diabetes and an ad-hoc virtual corpus compiled to match MTP as regards sub-domains, topics, level of communicative specialisation and text types. The other corpus of biomedical Spanish is a specialised textual corpus that contains whole documents, i.e. translations by last-year undergraduates in Translation and Interpreting during the academic years 2004-2005, 2005-2006 and 2006-2007. It comprises almost the same text types and topics as the MTP, but with a higher proportion of research papers, product descriptions and PILs. The MTSC is comparable to the MTP as they share similar design criteria.

Finally, the TT comprises target language segments of TMs of technical and technological domains (telephony, network services, telecommunications, etc.) and the CRATER Spanish subcorpus. It comprises fragments from user's manuals, guides and operating instructions, companies' press releases and, to a lesser extent, rules and regulations, standards, projects and monographies. The TTC has been compiled ad-hoc from evaluated electronic sources. After analysing the TT in terms of text types, domains and topics, a catalogue of index words and search equations have been derived. As a result, we have ended up compiling a corpus which is partially comparable to the TT, as it contains whole documents (not just fragments). It should be pointed out that locating this kind of technical documents in peninsular Spanish has proved to be more complicated than finding original medical Spanish, as many texts of this kind are covert translations. We have ensured that only non-translated original technological texts are included by filtering and refining all electronic searches.

The size of the above corpora (no. of tokens) is as follows[1]:

- MTP: 1,058,122
- MTS: 780,006
- TT: 1,736,027
- MPC: 1,402,172
- MTSC: 1,164,435
- TTC: 1,986,651.

Therefore, the corpora of translated Spanish and non-translated Spanish are comparable on the following grounds:

(i) The pairs of translated and non-translated corpora include roughly the same range of text types and forms.

(ii) They belong to the domains and subdomains.

(iii) They exhibit the same level of specialisation and formality.

(iv) They are restricted diatopically to Peninsular Spanish.

[1] Whereas the size of these corpora is small by today's standards, we should not that any previous corpus analysis on translation universals (e.g. Laviosa's (2002) work on simplification) has covered even smaller data.

(v) They were produced during the same span of time (2005-2008).

(vi) They are of a similar size (no. of tokens).

3 Corpus features

Previous studies on universals, unfortunately, have not accounted for what exactly classes as evidence in terms of different text features for their validity. To obtain an objective measure that would quantify the degree to which this or that universal holds, it is important to define features or parameters so that formal empirical studies can be conducted to compare texts in terms of simplification or similarity, and more specifically to verify our hypotheses. In the absence of any such guidelines, the first step to take in this study is to identify features of texts.

We propose to assess these characteristics of corpora on the basis of the following features[2]:

(i) lexical features (lexical richness and lexical density);

(ii) stylistic features (sentence length, use of simple as opposed to complex sentences, use of aspect, discourse markers as well as conjunctions, readability of text);

(iii) syntactic features (patterns of PoS tags).

In the following we describe these features in more detail.

3.1 Lexical features

Lexical density: Lexical density is computed as type/token by dividing the number of types by the total number of tokens present in the corpus. Low lexical density involves a great deal of repetition with the same words occurring again and again. On the other hand, high lexical density means that a more diverse form of language is being employed.

Lexical richness: We argue that lexical density is not indicative of the vocabulary variety of an author as it counts morphological variants of the same word as different word types. However, whereas *student* and *students* may technically be separate words and word types, from lexical point of view they represent the same word. To alleviate this inadequacy, we propose a new measure lexical richness, which is computed as the number of lemmas divided by the number of tokens present in the corpus and accounts for the variety of word use by an author. The lemma of every word is automatically returned by the Connexor parser (Tapanainen and Jarvinen, 1997).

3.2 Stylistic features

Sentence length: Sentence length is a feature deemed to be typical of an individual style. We compute sentence length as the number of tokens in corpus divided by the number of sentences in this corpus. In this study, unlike Study 1, we have opted for not including the parse tree depth as a stylistic feature because (a) the parse tree is more a syntactic concept and (b) we believe the parse tree depth and sentence length are not completely independent features.

Simple sentences vs. complex sentences: We argue that whether the use of predominantly simple or complex sentences, or balanced combination of both, is a relevant feature for the style of an author. In order to count the number of simple or complex sentences we developed an algorithm to automatically identify the type of sentence by counting the number of finite verbs (and their corresponding verbal constructions) in a sentence; sentences with more than one finite verb are classified as complex. Constrictions such as (HABER, TENER or SER) + Past Participle and ESTAR + Gerund are counted as well. Verbs are detected by the Connexor parser, so are past participles and gerunds. We have computed the proportion of cases where simple or complex sentences are used.

Discourse markers: According to Biber (1988, 1995, 2003), the use of discourse markers is another characteristic of someone's style. To this end, using a list of discourse markers in Spanish, we have extracted and calculated the proportion of both discourse markers from the number of all words in a corpus.

Readability: We experiment with three popular text readability measures: Automated Readability Index (ARI), Coleman-Liau Index (CLI), and Flesch-Kincaid Grade Level Readibility Test (FK).

The automated readability index (Smith and Senter 1967) was originally created for U.S. Air Force manuals and technical documents. This readability test is designed to measure the understand ability of a text. The formula for this test is:

[2] Some of these features have been adopted from Biber (1993, 1995); other such as the type of sentences, are our own proposals. It is worth noting that the set of stylistic features is language dependent. For example, the use of active or passive voice would have been more interesting for English or German.

$$ARI = 4.71\frac{c}{w} + 0.5\frac{w}{s2} - 21.43$$

where c is the number of characters, w is the number of words, and s is the number of sentences in the text. The formula estimates the minimum grade level required to understand the text.

M.Coleman and T.L. Liau (1975) presented their readability test to measure the understand ability of a text. Similar to ARI it also relies on characters instead of syllables per word. In order to calculate the Coleman-Liau Index the following formula is used:

$$CLI = 5.89\frac{c}{w} - 0.3\frac{s}{w} - 15.8$$

The Flesch-Kincaid test (Flesch 1948) is designed to indicate comprehension difficulty when reading a passage of academic English. This test relies on syllables per word instead of characters. It is calculated using the following formula:

$$FK = 0.39\frac{w}{s} + 11.8\frac{syl}{w} - 15.59$$

where *syl* describes the number of syllables in the text.

3.3 Syntactic features

We perform part-of-speech tagging/shallow parsing[3] for each corpus and compare the sequences of parts of tags which are meant to reflect the syntactic structure of the sentences. To determine the similarity of two corpora in terms of their syntactic features, vectors of n-grams are compared using cosine and recurrence metrics modelled as permutation tests (Nerbonne and Wiersma, 2006).

In our experiments we compare sequences of POS tags between for every pair of corpora. Sequences of POS tags account for the linear syntactic structure of sentences and the idea behind our general methodology consists of comparing any two corpora taking into account n-grams. Previously, n-grams of POS tags have been used to measure syntactic distance and best results have been reported for n=3 (Nerbonne and Wiersma, 2006). The corpora to be compared are represented as frequency vectors of 3-grams and the measures employed for comparison are the cosine as well as the measures R and R_{sq} which were inspired by the recurrence (R) metric (Kessler, 2001).

4 Hypotheses

Taking previous work on corpus-based studies of translated text as the departing point, we formulated the following set of hypotheses. In accordance with the simplification postulate, we expect translated corpora:

(i) to be characterised by less varied and more familiar vocabulary;

(ii) to contain a greater number of simple sentences than complex ones;

(iii) to contain shorter sentences than sentences of original text;

(iv) to contain fewer discourse markers than original text;

(v) to be generally more readable and easy-to-understand according to established measures of readability.

In accordance with the convergence universal, we expect that the lexical, stylistic, and syntactic features described above (see Section 3) will reveal smaller differences within a set of translated corpora than within a set of original ones. Specifically, we expect that a set of translated texts will exhibit smaller differences in (i) lexical richness and lexical density; (ii) sentence length and proportion of simple sentences; (iii) the use of discourse markers; (iv) the kinds of syntactic constructions are used in the text, between them than to the original texts.

5 Simplification universal

To examine the simplification hypothesis, we computed mean values for lexical and stylistic features for each corpus. For this purpose, each corpus was split into segments with each segment containing 6000 sentences and the means were obtained by averaging the values for individual segments in each corpus. These means were then compared using the unpaired two-tailed t-test. Because syntactic characteristics are compared by computing a similarity measure between corpora, in this experiment we included all the features, except the syntactic ones. The results are presented in Table 1: for each of the 3 pairs of corpora, the table shows the mean for each corpus and the significance level (α) determined using t-test (statistically

[3] Part-of-speech tagging /shallow parsing is performed using Connexor's Machinese (Tapanainen and Jarvinen, 1997).

significant differences are shown in bold).

Features	MTP-MTPC			MTS-MTSC			TT-TTC		
	MTP	MTPC	α	MTS	MTSC	α	TT	TTC	α
Lexical Density	**.027**	**.042**	**0.005**	.052	.041	0.4	**.02**	**.025**	**0.001**
Lexical Richness	**.016**	**.029**	**0.005**	.037	.028	0.4	**.013**	**.015**	**0.001**
Average Sentence Length	25.25	20.70	0.2	28.49	26.44	0.1	**27.29**	**18.12**	**0.001**
Simple Sentences (%)	**.441**	**.638**	**0.01**	.507	.521	0.7	**.476**	**.592**	**0.002**
Discourse Markers (Ratio)	**.0012**	**.002**	**0.05**	.0018	.0021	0.2	**.0007**	**.0016**	**0.002**
ARI	16.85	15.08	0.4	19.14	19.01	0.75	**17.85**	**12.85**	**0.001**
CLI	16.27	16.9	0.3	**17.16**	**18.28**	**0.05**	16.28	15.5	0.1
FK	19.53	18.21	0.5	21.32	21.51	0.5	**20.03**	**15.46**	**0.001**

Table 1: A comparison of mean values of the lexical and stylistic features between corresponding comparable corpora.

Features	Translated Corpora			Non-translated Corpora		
	MTP - MTS	MTS - TT	MTP - TT	MTPC - MTSC	MTSC - TTC	MTPC - TTC
Lexical Density	0.002	0.001	0.079	0.14	0.201	0.001
Lexical Richness	0.001	0.001	0.14	0.14	0.015	0.001
Sentence Length	0.011	0.522	0.202	0.145	0.002	0.368
Simple Sentences	0.057	0.673	0.202	0.096	0.462	0.212
Discourse Markers	0.001	0.005	0.351	0.063	0.001	0.072

Table 2: P-values for differences between corpora, computed using t-test.

6 Convergence universal

To experimentally examine the convergence universal, we compared similarities within a set of translated texts (MTP, MTS, TT) and within a set of comparable non- translated texts (MTPC, MTSC, TTC). See also Corpas et al. (2008).

6.1 Comparison of lexical and stylistic features

As in the previous experiment, we examine lexical and stylistic features separately from the syntactic ones, as the latter involve similarity scores rather than means. We operationalise the dissimilarity within each group of corpora as averages of probabilities for the differences between them, which we compute with the help of two tests: the unpaired t-test for each feature individually and the chi-square test for the entire set of features. Thus, p-values from chi-square tests produce a global score of dissimilarity within a set, while p-values from t-tests give an idea of dissimilarity within the set only with respect to particular features. The means for the lexical and stylistic features are computed over the same corpus segments as in Section 5. Table 2 presents the results of these tests. Table 3 presents global measures of similarities between corpora as computed using chi-square test.

Corpora	p-values
Translated Corpora	
MTP - MTS	0.01
MTP - TT	0.002
MTS - TT	0.023
Average	0.012
Non-translated Corpora	
MTPC - MTSC	0.059
MTPC - TTC	0.006
MTSC - TTC	0.071
Average	0.045

Table 3: P-values for differences between corpora, computed using chi-square test

6.2 Syntax comparison

Furthermore, we assess syntax similarity (in our case dissimilarity) between each pair of translated and non-translated texts by comparing sequences of 3-grams of part-of-speech (POS) tags for every pair of corpora. We first run the Connexor parser to identify all POS tags, then collect frequency vectors of 3-grams whose dissimilarity is compared on the basis of the 1-C (C=cosine), R and R_{sq} measures.

More specifically, for every corpus we build a frequency vector featuring all trigrams of POS tags. For example, the comparison of the frequency vectors of the corpus of all translated texts (MTP+MTS+TT) and the corpus of non-translated texts (MTPC+MTSC+TTC) involves a total of 18,468 different POS.[4] Table 4 below represents the results obtained from comparing the pairs of corpora applying the aforementioned dissimilarity measures. The higher values of the measures employed indicate greater dissimilarity (and less similarity) between two corpora under comparison.

Corpora	1-C	R	R_{sq}
Translated texts			
MTP - MTS	0.206	252526	638848591
MTP - TT	0.337	388466	3146471863
MTS - TT	0.176	432725	2643068563
Non-Translated texts			
MTPC - MTSC	0.017	98448	82218137
MTPC - TTC	0.15	364322	851312764
MTSC - TTC	0.167	372940	1008322991

Table 4: Results measuring syntactic differences

7 Discussion and conclusions

With respect to the simplification hypothesis, it appears to be validated on some, but not all parameters. Indeed, we find that translated texts often exhibit significantly lower lexical density and richness, and seem to be more readable that non-translated texts (however, statistical significance for the readability differences for this could be established only for one corpus pair). Unexpectedly, translated texts displayed significantly smaller proportion of simple sentences and their sentences turned out to also be significantly shorter. With respect to discourse markers, we find that in two pairs out of three, non-translated texts use discourse markers significantly more often. Interestingly, simplification traits are more visible on the technical translation corpora and, to a somewhat lesser degree, on corpora of professionally-produced medical translations (where all the features, except sentence length and simple sentences ratio, indicate simpler wordings and formulations), while simplification cannot be found in texts produced by student translators.

As to the convergence universal, we find that the p-values for pairs of translated corpora are in fact generally smaller than those for pairs of non-translated ones, both determined with the help of the t-test and the chi-square test. Smaller p-values indicate greater probability that the pair of corpora under comparison is different. This is true for most individual features: lexical richness, lexical density, sentence length, and simple sentences ratio. With respect to the former two features, the differences between the translated corpora are greater in two pairs, but in one (MTS – TT) the picture is actually the opposite. In terms of discourse markers, however, translated corpora are indeed more similar to each other, with the exception of the MTP – MTS pair.

Considering p-values computed using chi-square tests over all the features, we see that pairs of translated corpora have consistently smaller p-values that non-translated ones, which, again, contradicts the convergence hypothesis.

With regard to syntactic differences between corpora, from our results it is clear that translated texts differ more in terms of syntax for all compared pairs and from the point of view of all measures (1-C, R and Rsq). It is also clear that the difference of syntax is greater between texts of different domains. On the basis of the above results we can conclude that there is no evidence that convergence holds in terms of syntax. In fact, the results from Table 4 even show that translated texts

[4] We compare a total of 8,484 trigrams between MSTP and MSTS, 9,954 trigrams between MSTP and TST and 10,019 between MSTS and TST. We also compare 8,278 trigrams between MSTPC and MSTSC, 13,297 trigrams between MSTPC and TSC and 13,007 between MSTSC and TSC.

differ more syntactically than non-translated texts on our experimental data.

To sum up, the results of our experiments suggest that simplification does affect translated texts, albeit this is not true with regard to the sentence length and the use of simple vs. complex sentences, and texts produced by non-professional translators do not seem to possess such simplification traits. Another important and unexpected finding of the study is that none of the lexical, stylistic and syntactic features we chose to study the convergence hypothesis could reveal any evidence for its validity. In the experiments performed so far, sentence length and complexity do not appear to reveal much about the simplification and convergence universals, so it would be interesting to identify and investigate new features such as usage of idioms and multi-word units.

This research can be further extended to other languages as well as different domains and identification of more features in translated texts which can be computed using NLP and evaluating these features. The implications for Machine Translation (MT) would be that non-translated text tends to be simpler than translated texts (Table 1). Therefore, in order to enhance MT systems, researchers should aim at analysing non-translated vs. translated comparable corpora, in order to identify the characteristic features of non translated texts and try to reproduce them in the MT output. It should also be said that those features will surely change from one domain to another, so it is necessary to have a genre/restricted register approach.

References

Baker, M. 1993. “Corpus Linguistics and Translation Studies – Implications and Applications”. In: M. Baker, M.G. Francis & E. Tognini-Bonelli (eds.). 1993. *Text and Technology: In Honour of John Sinclair.* Amsterdam & Philadelphia: John Benjamins. 233-250.

Baker, M. 1996. “Corpus-based Translation Studies: The Challenges that Lie Ahead”. In: H. Somers (ed.). 1996. *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*. Amsterdam & Philadelphia: John Benjamins. 175-186.

Biber, D. 1988. *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. 1995. *Dimensions of Register Variation: a Cross-Linguistic Comparison.* Cambridge: Cambridge University Press.

Biber, D. 2003. “Variation among University Spoken and Written Registers: A New Multi-dimensional Analysis”. In: P. Leistyna & C. F. Meyer (eds.). 2003. *Corpus Analysis. Language Structure and Language Use.* Amsterdam & New York: Rodopi. 47-70.

Coleman, M. and Liau, T.L. (1975). *A Computer readability formula designed for machine scoring,* Journal of Applied Psychology, Vol. 60, pp. 283-284.

Corpas Pastor, G. 2008. *Investigar con corpus en traducción: los retos de un nuevo paradigma.* Frankfort am Main, Berlin & New Cork: Peter Lang.

Corpas Pastor, G., Mitkov R., Afzal N., Garcia Moya L. (2008). *Translation Universals: Do they exist? A corpus-based and NLP approach to convergence.* In Proceedings of the LREC (2008) Workshop on "Comparable Corpora". LREC-08. Marrakesh, Morocco.

Flesch, R. 1948. *A new readability yardstick,* Journal of Applied Psychology, Vol. 32, pp. 221-233.

Kessler, B. 2001. *The Significance of Word Lists.* Stanford: CSLI Press.

Laviosa, S. 2002. *Corpus-based Translation Studies. Theory, Findings, Applications.* Amsterdam & New York: Rodopi.

Nerbonne J. & Wiersma, X. 2006. “A Measure of Aggregate Syntactic Distance”. In: J. Nerbonne & E. Hinrichs (eds.) 2006. *Linguistic Distances. Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics*. Sidney, Australia. 82-90.

Smith, E. A. and R.J. Senter 1967. *Automated Readability Index* AMRL-TR, 66-22. Wright-Patterson AFB, OH: Aerospace Medical Division.

Tapanainen, P., Jarvinen, T. 1997. A non-projective dependency parser. In: Proceedings of the 5th Conference of Applied Natural Language Processing, Washington D.C., USA. pp. 64–71.

Toury, G. 1995. Descriptive Translation Studies and Beyond. Amsterdam: John Benjamins.