Translation universals: do they exist?
A corpus-based and NLP approach to convergence

Gloria Corpas*, Ruslan Mitkov**, Naveed Afzal** and Lisette Garcia Moya***

* University of Malaga, Email gcorpas@uma.es
** University of Wolverhampton, Email {r.mitkov, n.afzal}@wlv.ac.uk
*** Centre for Pattern Recognition and Data Mining, Santiago de Cuba, Email lisette.garciamoya@gmail.com

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 transfer, 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 of 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.

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 translation universal of convergence. According to this universal, translated texts tend to be more similar than non-translated texts. The objective of this study is to establish whether this universal is valid with Spanish as target text. To this end, we analyse corpora of translated texts into Spanish and comparable corpora of Spanish non-translated texts. Then we compute similarity between every pair of corpora of translated texts and every pair of corpora of original texts for both languages. The similarity is measured in terms of both style and syntax.

2. Corpora used

According to the convergence universal, translated texts tend to be more similar than non-translated texts. The objective of this study is to verify whether this universal is valid with Spanish as the target language. To this end, we compare pairs of corpora of translated texts as well as pairs of comparable corpora of original, non-translated Spanish texts in terms of style and syntax with a view to establishing whether translated texts are found to be more similar than non-translated texts. We specifically compiled the following corpora for this experiment:

- Corpus of Medical Spanish Translations by Professionals (MSTP)
- Corpus of Medical Spanish Translations by Students (MSTS)
- Corpus of Technical Spanish Translations (TST)
- Corpus of Original Medical Spanish Comparable to Translations by Professionals (MSTPC)
- Corpus of Original Medical Spanish Comparable to Translations by Students (MSTSC)
- Corpus of Original Technical Spanish Comparable to Technical Translations (TSTC)
As stated above, MSTP is comparable to MSTPC, MSTS is comparable to MSTSC and TST is comparable to TSTC. Comparability was a crucial consideration for this study as otherwise any style or syntax comparison would have been compromised.

We compiled the corpora 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. MTSP 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 TL 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: SL segments of TMs different from the ones used to compile the MTSP, a small corpus of diabetes and an ad-hoc virtual corpus compiled to match MTSP 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 MTSP, but with a higher proportion of research papers, product descriptions and PILs. The MSTSC is comparable to the MTSP as they share similar design criteria.

Finally, the TST comprises TL segments of TMs of technical and technological domains (telephony, network services, telecommunications, etc.) and the CRATER Spanish subcorpus. It comprises fragments from users manuals, guides and operating instructions, companies press releases and, to a lesser extent, rules and regulations, standards, projects and monographies. The TSTC has been compiled ad-hoc from evaluated electronic sources. After analysing the TST in terms of text types, domains and topics, we have derived a catalogue of index words and search equations. As a result, we have ended up compiling a corpus which is partially comparable to the TST, 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:

- MSTP: 1,058,122
- MSTS: 780,006
- TST: 1,736,027
- MSTPC: 1,402,172
- MSTSC: 1,164,435
- TSTC: 1,986,651.

3. Methodology

We compared all 3 pairs of translated texts (MSTP-MSTS; MSTS-TST; MSTP-TST) and all 3 pairs of comparable non translated texts (MSTPC-MSTSC; MSTSC-TSTC; MSTPC-TSTC). If the convergence universals holds, we would expect to find higher similarity for pairs of translated texts.

---

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.
Previous studies on universals, unfortunately, have not accounted for what exactly classes as evidence in terms of different features for their validity. Therefore we first have to ask the question when a text or a corpus is more or less similar to another text or corpus. It is important to know what the features or parameters of similarity are so that formal empirical studies can be conducted to compare texts in terms of similarity and more specifically to verify whether translated texts ‘converge’ in that in general are more similar than non-translated texts. In the absence of any such guidelines, the first step to take in this study is to identify features which could be used for measuring similarity of translated or non-translated texts.

We propose to assess to what extent translated or non-translated texts ‘converge’ on the basis of (i) style (stylistic features) and (ii) syntax (syntactic features). This experiment covers the following style characteristics\(^2\): lexical richness (type/token) ratio, lexical density, sentence length, use of simple as opposed to complex sentences, use of aspect, discourse markers as well as conjunctions. Unlike any previous corpus-based work on universals (simplification), we perform stemming of each corpus so that the results related to lexical richness are not compromised in that two morphological variants of a word (e.g. experiment, experiments) are not regarded as two different words. The analysis of general syntactic patterns is unique in that no such previous experiments have been carried out. We perform part-of-speech tagging/shallow parsing\(^3\) for each corpus and compare the sequences of parts of tags which account for the linear syntactic structures. More specifically, vectors of n-grams are compared using cosine and recurrence metrics modelled as permutation tests (Nerbonne and Wiersma, 2006).

### 3.1 Style comparison

**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.

**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 three is more a syntactic concept and (b) we believe the parse three 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

---

\(^2\) 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.

\(^3\) Part-of-speech tagging/shallow parsing is performed using Connexor’s Machinese.
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 sentence is 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.

In order to compute similarity between each pair of translated and non-translated texts, two statistical tests (Chi-Square test and T-test) are employed. Chi-square takes all features used and produces a global score of similarity between the corpora analysed. T-test does not provide a global score but instead compares separate features and establishes any statistically significant difference or not.

### 3.2 Syntax comparison

In this experiment 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 n-gram from one of the corpora with any n-gram from the other. 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 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. Results

This section reports the results of the experiments/comparisons described above and seeks to offer insights whether convergence holds as a universal.

#### 4.1 Style comparison

In order to compare the style of translated texts as well as the style of non-translated texts, we first compute the style features lexical density, lexical richness, the average sentence length, proportion of simple/complex sentences and discourse markers (Table 1).

<table>
<thead>
<tr>
<th>Features</th>
<th>MSTP</th>
<th>MSTS</th>
<th>TST</th>
<th>MSTPC</th>
<th>MSTSC</th>
<th>TSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical Density</td>
<td>0.027954</td>
<td>0.052715</td>
<td>0.020679</td>
<td>0.042505</td>
<td>0.041159</td>
<td>0.025529</td>
</tr>
<tr>
<td>Lexical Richness</td>
<td>0.016929</td>
<td>0.037709</td>
<td>0.013281</td>
<td>0.029992</td>
<td>0.028905</td>
<td>0.015591</td>
</tr>
<tr>
<td>Simple Sentences (%)</td>
<td>0.441768121</td>
<td>0.507205751</td>
<td>0.476949103</td>
<td>0.638889238</td>
<td>0.52120611</td>
<td>0.592110096</td>
</tr>
<tr>
<td>Discourse Markers (Ratio)</td>
<td>0.001268941</td>
<td>0.001852604</td>
<td>0.000763805</td>
<td>0.002022331</td>
<td>0.002099085</td>
<td>0.001649655</td>
</tr>
</tbody>
</table>

Table 1: Stylistic features

Next, we compute similarity between each pair of translated and non-translated texts using the results obtained for the above features (lexical density, lexical richness, sentence length, simple sentences proportion, discourse markers) in two statistical tests: Chi-Square test and T-test. The Chi-Square values obtained for each pair of corpus of translated and non-translated texts are as displayed in Table 2.

<table>
<thead>
<tr>
<th>Translated Corpora</th>
<th>Chi-Square Values</th>
<th>Non-translated Corpora</th>
<th>Chi-Square Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1MSTP → 2MSTS</td>
<td>0.010622566</td>
<td>1MSTPC → 2MSTSC</td>
<td>0.059779549</td>
</tr>
<tr>
<td>1MSTP → 3TST</td>
<td>0.00266151</td>
<td>1MSTPC → 3TSC</td>
<td>0.006140764</td>
</tr>
<tr>
<td>2MSTS → 3TST</td>
<td>0.023731912</td>
<td>2MSTSC → 3TSC</td>
<td>0.07122404</td>
</tr>
<tr>
<td>Total</td>
<td>0.037015988</td>
<td>Total</td>
<td>0.137144352</td>
</tr>
<tr>
<td>Average</td>
<td>0.012338663</td>
<td>Average</td>
<td>0.045714784</td>
</tr>
</tbody>
</table>

Table 2: Ch-Square values

Finally, we conducted T-tests for statistical significance. In order to conduct T-test, each corpus was divided into small subsets of equal size. For each subset the figures for the above stylistic features are computed and compared with the figures of the corresponding subsets of the corpus being compared.

<table>
<thead>
<tr>
<th>Features</th>
<th>Translated Corpora (T-test Values)</th>
<th>Non-translated Corpora (T-test Values)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSTP → MSTS</td>
<td>MSTP → TST</td>
</tr>
<tr>
<td>Lexical Density</td>
<td>0.002545387</td>
<td>0.000123172</td>
</tr>
<tr>
<td>Lexical Richness</td>
<td>0.00066040</td>
<td>0.000006.9792</td>
</tr>
<tr>
<td>Sentence Length</td>
<td>0.011826639</td>
<td>0.522122939</td>
</tr>
<tr>
<td>Simple Sentences</td>
<td>0.057465277</td>
<td>0.673936375</td>
</tr>
<tr>
<td>Discourse Markers</td>
<td>0.001048007</td>
<td>0.005746253</td>
</tr>
</tbody>
</table>

Table 3: T-Test values

4.2 Syntax comparison

We assess syntax similarity (in our case dissimilarity) between each pair of translated and non-translated texts by comparing sequences 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 Rsq 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 (MSTP+MSTS+TST) and the corpus of non-translated texts (MSTPC+MSTSC+TSTC) involves a
total of 18,468 different POS.\textsuperscript{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.

<table>
<thead>
<tr>
<th>Corpora</th>
<th>I-C</th>
<th>R</th>
<th>Rsq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translated texts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSTP - MSTS</td>
<td>0.206015066283</td>
<td>252526.914323</td>
<td>638848591.082</td>
</tr>
<tr>
<td>MSTP - TST</td>
<td>0.337626383799</td>
<td>388466.504863</td>
<td>3146471863.13</td>
</tr>
<tr>
<td>MSTS - TST</td>
<td>0.176310545152</td>
<td>432725.578482</td>
<td>2643068563.82</td>
</tr>
<tr>
<td>Non-Translated texts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSTPC - MSTSC</td>
<td>0.0176469276126</td>
<td>98448.0858054</td>
<td>82218137.9687</td>
</tr>
<tr>
<td>MSTPC - TSC</td>
<td>0.150912596476</td>
<td>364322.217714</td>
<td>851312764.364</td>
</tr>
<tr>
<td>MSTSC - TSC</td>
<td>0.167167511143</td>
<td>372940.61477</td>
<td>1008322991.78</td>
</tr>
</tbody>
</table>

Table 4. Results measuring vector differences

5. Discussion and concluding remarks

The average Chi-Square values\textsuperscript{5} of translated texts are smaller than average Chi-Square values of non-translated texts (Table 2) which implies that the translated texts included in our experiment are more similar than non-translated texts with regard to the stylistic features studies. On the basis of the corpora used and the features employed, it appears that the convergence universal holds on this occasion.

The T-Test values (Table 3) of non-translated texts show that there is no significant difference between any of the above mentioned list of features in MSTPC$\rightarrow$MSTSC pair and the MSTPC$\rightarrow$TSC pair results show that there is a significant difference between lexical density and lexical richness, while in the MSTSC$\rightarrow$TSC pair there is a significant difference among 3 features (lexical richness, sentence length and discourse markers). In case of translated texts the T-Test values of the pair MSTP$\rightarrow$MSTS significantly differ in terms of lexical density and discourse markers and of the pair MSTS$\rightarrow$TST significantly differ lexical density and lexical richness. There is no significant difference between MSTP$\rightarrow$TST.

From the T-test results it is clear that whereas the Chi-square test suggests general greater similarity between translated texts, we can make several interesting observations.

i. There are non-translated texts which are not statistically different in terms of the chosen stylistic features whereas the corresponding comparable corpora of translated texts different statistically with regard to two stylistic features (see the pairs MSTPC$\rightarrow$MSTSC and MSTP$\rightarrow$MSTS respectively)

ii. There are non-translated texts which are statistically different in terms of only one stylistic feature whereas the corresponding comparable corpora of translated texts different statistically with regard to two stylistic features (see the pairs MSTPC$\rightarrow$TSC and MSTP$\rightarrow$TST respectively)

\textsuperscript{4} We compare a total of 8,484 trigrams between MSTP and MSTS, 9,954 trigrams between MSTP and TST and 10,019 between MSTSC 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.

\textsuperscript{5} The smaller Chi-Square value indicates the bigger similarity between the two corpora.

iii. Translated texts could often differ significantly with regard to certain style features (MSTP -> MSTS; MSTS -> TST) of which especially surprising is the lexical density. Whereas difference in the lexical density between student and professional translators could be somehow acceptable, statistical difference in lexical density between professional translators is unexpected.

Therefore, on the basis of our data and with regard to the style features adopted, whereas convergence appears to be broadly holding, we argue that no definite conclusion can be made that convergence is a clear-cut universal due to the above T-test results. In the case of an absolute, clear-cut universal, one would not have expected results such as the ones stated in (i) and (ii) above.

From Table 4 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 differ more syntactically than non-translated texts on our experimental data.

In general, the results do not provide sufficient support to the convergence ‘universal’.

References


