De Sutter, G., Cappelle, B., De Clercq, O., Loock, R., & Plevoets, K.. (2017). Towards a corpus-based, statistical approach to translation quality: Measuring and visualizing linguistic deviance in student translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *16*, 25–39.

Towards a corpus-based, statistical approach to translation quality: Measuring and visualizing linguistic deviance in student translations

Gert De Sutter
Ghent University, Belgium
gert.desutter@ugent.be

Bert Cappelle
Université Lille III, France
bert.cappelle@univ-lille3.fr

Orphée De Clercq
Ghent University, Belgium
orphee.declercq@ugent.be

Rudy Loock
Université Lille III, France
rudy.loock@univ-lille3.fr

Koen Plevoets
University of Leuven, Belgium
koen.plevoets@kuleuven.be

In this article we present a corpus-based statistical approach to measuring translation quality, more particularly translation acceptability, by comparing the features of translated and original texts. We discuss initial findings that aim to support and objectify formative quality assessment. To that end, we extract a multitude of linguistic and textual features from both student and professional translation corpora that consist of many different translations by several translators in two different genres (fiction, news) and in two translation directions (English to French and French to Dutch). The numerical information gathered from these corpora is exploratively analysed with Principal Component Analysis, which enables us to identify stable, language-independent linguistic and textual indicators of student translations compared to translations produced by professionals. The differences between these types of translation are subsequently tested by means of ANOVA. The results clearly indicate that the proposed methodology is indeed capable of distinguishing between student and professional translations. It is claimed that this deviant behaviour indicates an overall lower translation quality in student translations: student translations tend to score lower at the acceptability level, that is, they deviate significantly from target-language norms and conventions. In addition, the proposed methodology is capable of assessing the acceptability of an individual student's translation – a smaller linguistic distance between a given student translation and the norm set by the professional translations correlates with higher quality. The methodology is also able to provide objective and concrete feedback about the divergent linguistic dimensions in their text.

1. Introduction

Empirical Translation Studies has undergone major descriptive and theoretical advances in the past few years which have clearly been brought about by what one could call a "methodological shift" from monodimensional comparable corpus analyses (in which the frequency of a given linguistic feature in a corpus of translated texts is compared to its frequency in a corpus of non-translated, original texts; e.g., Olohan & Baker, 2000) to multidimensional empirical analyses (e.g., Evert & Neumann, 2017). This shift includes stricter data control, analysis of both comparable and parallel

data, the use of more advanced statistical techniques and the integration of different methodological designs in order to arrive at so-called "converging evidence" (see, for example, De Sutter, Delaere, & Lefer, 2017). This has led to a better, more accurate, more fine-grained understanding of translation products and processes and to a better theoretical underpinning of Empirical Translation Studies in general. An excellent case in point of this recent shift is the research carried out by Kruger (2015, 2016) in which she combines different methodologies (eye-tracking, keystroke logging, corpus data) and uses statistical techniques "with the aim of developing a comprehensive view of the interrelation between cognitive and social aspects of translation as bilingual language processing in a complex multilingual society" (Kruger, 2015).

So far, however, these advances in Empirical Translation Studies have had a relatively small impact on the more applied branches in Translation Studies. Certainly, the use of corpora in translator training is widespread, and translator aids are being updated on the basis of carefully designed analyses, but it is still fair to say that the full potential of the methodological and analytical resources which are now increasingly being used in Empirical Translation Studies still have to find their way into Applied Translation Studies (see, however, Daems, Vandepitte, Hartsuiker, & Macken, in press).

This article aims to help bridge this gap between theory and practice, thereby supporting Chesterman's (1999) plea that "even if we aim to rid translation research of non-empirical bathwater, we do not have to throw out the prescriptive baby as well" (p. 19). In particular, we investigate the extent to which multifactorial corpus analysis can help the translation teacher to measure translation quality reliably and objectively and to provide clear, specific and understandable feedback to the translation student.

Although the issue is omnipresent and many researchers have tried to provide criteria and methods for assessing translation quality (for an overview, see Daems, 2016, pp. 22–26), it seems that it remains difficult, if not impossible, to define objectively what a good translation is. Theoretical, pedagogical and professional approaches offer diverging solutions (e.g., Secară, 2005; Toudic, Hernandez Morin, Moreau, Barbin, & Phuez, 2014), but what they all have in common is the search for objective criteria. Nevertheless, many studies have already shown that evaluation quality depends heavily on multiple, often irrelevant factors such as the evaluator's personal ideas about translation competence, time pressure, the number of translations to be corrected and their relative order. As a consequence, translation evaluations of one single text may vary considerably between evaluators or even within one evaluator (cf., for example, Anckaert, Eyckmans, & Segers, 2008; Williams, 2009).

We will claim that translation evaluation can benefit from a corpus-based statistical approach in several ways: it is objective and systematic, and it is capable of identifying relevant (hidden) deviant patterns, while at the same time diminishing the influence of irrelevant evaluator- or context-related factors. Such a corpus-based approach, however, will not be capable of evaluating translations fully automatically; rather, it is seen as one tool in the translation teacher's toolbox that should not be used without further interpretation by a highly skilled and experienced teacher, one that should be accompanied by more qualitative evaluation of translation quality, such as that suggested by Bowker (1999, 2000).

This article is organized as follows. After giving a concise overview of previous work on corpus-based translation evaluation in section 2, we present the methodology underlying our statistical approach in section 3. Sections 4 and 5 are devoted to the results of two case studies in which this methodology is used: English-to-French translation of fictional texts and French-to-Dutch translation of news texts. The final section summarizes the main research findings and discusses the implications and future avenues for research on corpus-based evaluation of translation quality.

2. Previous work on corpus-based translation evaluation

Using corpora to assess and improve the quality of translations is nothing new. In general, two different approaches can be discerned: a qualitative, particularistic, user-driven approach, on the one hand, and a quantitative, generalist, feature-driven approach, on the other. Scholars who espouse the

former, qualitative, approach are, for instance, Bowker (2000, 2001) and Hassani (2011). Both share the idea that electronic corpora containing authentic texts can be used as a "benchmark against which the goodness or otherwise of translations could be measured" (Hassani, 2011, p. 352). Moreover, they do not aim at quantifying the overall quality of a given (student) translation; instead, they provide translation teachers and students with a means to assess the appropriateness of specific words and constructions in a given translation, namely, by means of concordance lists in reference corpora. In particular, Bowker (2000, 2001) shows that the use of a specific *evaluation corpus* made up of different corpus types can provide evaluators with the tools to explain what is suitable and what is not in a translated text, with, on the one hand, electronic corpora containing fit-for-purpose source language (SL) and target language (TL) data and, on the other, a so-called 'inappropriate corpus' containing data that differ in terms of style, text type, technicality or publication date. She finds that evaluators are not only able to identify more errors in student translations (compared to non-corpus-based translation evaluation), they are also more confident about the feedback they give, and the feedback is more appreciated and considered more reliable by the students. Hassani (2011) conducted a case study on professional translators working in a news agency, showing that the use of TL monolingual corpora, in this instance English, could improve both translation quality and evaluation, in particular as far as collocations and semantic prosodies are concerned. The approach Hassani and Bowker adopt is tantamount to using electronic corpora as CAT tools to improve translation quality, an idea regularly found in the literature, in particular in relation to translator education to improve the understanding of the source text (ST), of terminological and phraseological choices and of the naturalness of the target text (TT) (see, for example, Bowker, 1998, 1999, 2003; Bowker & Pearson, 2002; Pearson, 2003; Frankenberg-Garcia, 2015; Kübler, 2001, 2003, 2008, 2011a, 2011b; Loock 2016a, 2016b; Ruiz Yepes, 2011; Sánchez-Gijón, 2009; Varantola, 2003; Zanettin, 2012; Zanettin, Bernardini, & Stewart, 2003).

In addition to this approach, the quantitative, generalist, feature-driven approach aims at quantifying overall translation quality based on one or more linguistic features that are deemed relevant for the task. This approach often incorporates the methodology suggested in Baker (1993), in which electronic corpora are used for corpus-based translation studies to measure differences between original language and translated language to uncover differences in the frequencies of specific linguistic features (see, for example, Baker & Olohan, 2000; Cappelle, 2012; Cappelle & Loock, 2017; Xiao, 2010). Besides interpreting these differences in a purely descriptive, theory-oriented manner, some authors have suggested that these deviances can also be considered indicative of translation quality. This idea is found, for instance, in Rabadán, Labrador and Ramon (2009), who discuss the position of adjectives in both original Spanish and Spanish translated from English: "The smaller the disparity between native and translated usage in the use of particular grammatical structures associated with specific meanings, the higher the translation rates for quality" (Rabadán, Labrador, & Ramon, 2009, p. 323). Likewise, Loock, Mariaule and Oster (2014) discuss the frequency of derived adverbs in both original and translated English and French, and relate the differences to the quality of translations: "An under- or over-representation of a given linguistic feature might correspond to a violation of usage constraints …; a good translation should be tending towards linguistic homogenization between original and translated language" (Loock, Mariaule, & Oster, 2014, p. 3). In other words, what these researchers advocate is a correlation between the absence of significant linguistic differences between original and translated language (in other words, linguistic homogenization) and translation quality. This assumed association is not widely shared, though, as one could raise the objection that differences between language varieties (in this study: professional translations and professional non-translations or original language) are often related to the specific sociopsychological characteristics of the translation situation (e.g., perceived norms; Delaere, De Sutter, & Plevoets, 2012) or to the specific cognitive constraints of mediating between different language systems (Vandevoorde, 2016) rather than the association being indicative of defective language use. Many register-variation studies have provided convincing evidence for the association between contextual features and linguistic features, because many linguistic differences can be observed in different contextual varieties. Whatever the exact explanation for the differences between original and translated language may be, it is not very plausible to suggest that the language used by professional translators is of lower quality than that of other professional

writers, especially since the above-mentioned differences are attested for different translators in different genres.

The association between linguistic differences and translation quality is, however, plausible if one compares translation students, on the one hand, and professional translators and other professional writers, on the other. This idea can be found in Loock (2016a, pp. 217–227), where it is suggested that some deviant linguistic characteristics of translations written by advanced students might be symptomatic of the overall quality of their translations. Although the results based on the analysis of a learner corpus of translated fiction texts from English to French are mixed and require further investigation, Loock notices a tendency in students' TTs to overuse highly frequent SL linguistic phenomena, such as the use of derived adverbs in particular, which could serve as a good indicator of the quality of a translated text as a whole.

Such an approach to the use of electronic corpora to evaluate translation quality goes against the traditional descriptive approach typically found in corpus-based Translation Studies, as advocated in Baker's (1993) seminal paper or by Johansson (2007): "It should be stressed that overuse and underuse are meant as *descriptive* terms and do not necessarily imply that there is anything wrong with translated texts where we find evidence of overuse or underuse" (p. 33, italics added).

However, we believe here, in line with Chesterman (2004), for instance, that "descriptivists have perhaps over-reacted against traditional prescriptivism in their desire to place Translation Studies on a more scientific basis" (p. 36), and that the analysis of electronic corpora containing translated texts by professional translators and translation trainees could provide us with valuable information by which to measure and improve translation quality and, more particularly, translation acceptability (in terms of linguistic homogenization) by uncovering differences between the translation behaviour in the two groups.

3. Methodology

The corpus-based statistical approach to translation quality that we propose here is primarily oriented towards measuring TL acceptability (how acceptable is a translation, given the TL norms and conventions?) and not towards measuring adequacy (how adequately does a translation transfer the meaning of the ST?). Our approach comprises five steps: (1) compiling a student translation corpus (i.e., a learner corpus), a professional translation corpus and a reference corpus containing original, non-translated texts written by professional writers; (2) preprocessing the three corpora (including tokenization, lemmatization, part-of-speech tagging and syllabification); (3) extracting a series of linguistic and textual features from the three corpora; (4) performing multivariate statistical analysis, and (5) interpreting the output of this analysis.

The assumption underlying our approach is that the linguistic behaviour of the professional translators and the other professional writers (in the reference corpus) is considered the standard. For this reason, it is important to collect texts written by different professional translators and writers. Setting the standard on the basis of one professional writer only would obviously cause a problem of representativeness; we need to ascertain that the linguistic behaviour of professionals is more or less alike, which accounts for our choice to include texts written by many professional authors and translators. What we expect, then, is that the linguistic behaviour of these professionals will be relatively homogeneous: a standard implies a high degree of homogeneity.[1] The linguistic behaviour of students can then be assessed by means of linguistic distances – by computing the differences from the group of professional writers and translators.

In order to be able to measure these linguistic differences, it is important not to rely on a small set of features, so as to minimize the potentially coincidental effect of a given linguistic feature in a given corpus. Moreover, since our basic interest lies in measuring standard linguistic behaviour in any given language area, and measuring translation quality (acceptability), we opted to include only those general textual and linguistic features that are not related to specific norms or conventions in a given language or language variety. In other words, we chose primarily to extract language-independent features that are not related to a given genre. For the first case study, on English–French

translation, we selected 25 language-independent features and five language-dependent features. For the second case study, on French–Dutch translation, we selected the same 25 language-independent features as in case study 1. The complete list of language-independent features is presented in Table 1. As can be seen, the list contains basic frequency information on different part-of-speech categories (lexico-grammatical features), measures of lexical creativity and originality (e.g., Type-Token Ratio, Lexical Density, hapax legomena), a general word-frequency measure (Zipf) and, finally, the degree of syntagmatic patterning/formulaicity (i.e., the total number of the 100 most frequent 3- and 4-grams). Obviously, at this juncture, when a quantitative corpus-based approach to translation quality is only starting to emerge, the number and nature of the features to be selected is somewhat arbitrary: there is no independently validated set of features to start from.

The analyses presented in the next section will surely give rise to an evaluation of the features in Table 1, and the odds are high that some of the features will turn out to be insignificant (both statistically and conceptually) and that new features will be introduced (see, e.g., the list of features in Evert & Neumann, 2017).

Table 1: List of language-independent features

Type frequency	Freq. of common nouns	Freq. of interjections
Token frequency	Freq. of proper nouns	Freq. of foreign words
Type-Token Ratio	Freq. of adjectives	3-grams (word)
Lexical Density	Freq. of adverbs	3-grams (POS)
Hapax Legomena	Freq. of verbs	4-grams (word)
Dis Legomena	Freq. of pronominals	4-grams (POS)
Tris Legomena	Freq. of conjunctions	Word frequency (Zipf)
Average sentence length	Freq. of prepositions	
Average word length	Freq. of determiners	

To extract these features, it was crucial to preprocess all three corpora linguistically. This preprocessing consisted of three steps: tokenization, part-of-speech tagging and syllabification. For the first two steps we relied on the LeTs preprocessing toolkit (Van de Kauter et al., 2013), which can process a variety of languages. For the third step, deriving syllables, we relied on two different techniques. For English and Dutch we used a classification-based syllabifier (Van Oosten, Tanghe, & Hoste, 2010) and for French the hybrid syllabification method as described in François and Miltsakaki (2012). This latter method works as follows: for words included in Lexique (New, Pallier, Brysbaert, & Ferrand, 2004) we used the gold syllabification included in the dictionary. For all other words, API phonetic representations were generated with espeak (http://espeak.sourceforge.net/), after which the Lexique3 syllabification tool was applied (Pallier, 1999).

Along with the language-independent features, we decided also to include a restricted set of language-dependent features in the first case study. These features were chosen because they were thought to provoke so-called "translationese-prone" errors, that is, errors that are likely to occur in translations as a result of a formally similar linguistic feature in the ST with a significantly higher or lower frequency of use. The language-dependent features selected for case study 1 are: prenominal vs postnominal adjective, the frequency of *dire* "to say", *et "*and" and the de-adjectival suffix: *ment*, "-ly".

All the language-independent and language-dependent features were extracted from the three corpora using custom-made Python scripts (and manually checked where necessary). This resulted in a datamatrix in which every row contains the numerical information of the 25 + 5 features with respect to a given text (either a student or a professional translation or a professional original text).

Every text is therefore represented as a feature vector consisting of the scores of 30 linguistic features as well as the status of the text (student translation, professional translation, professional non-translation), resulting in a 31-dimensional vector.

After extracting the quantitative information from the three corpora, we used principal component analysis (PCA) to inspect the correlation structure of our datamatrix in a lower-dimensional structure. For ease of presentation, we will present two-dimensional plots only in the remainder of this article. These visual representations will elucidate the extent to which professional writers and translators in fact set a homogeneous standard and, if so, which student translations approximate to this standard better. This explorative analysis was subsequently corroborated by an ANOVA of each linguistic feature, where the difference between professional writers and students was statistically tested.

4. Case study 1: English-to-French translation of fictional texts

For the first case study, we compiled three corpora: (1) a corpus of French texts translated from English by advanced students, (2) a corpus of French texts translated from English by professional translators, and (3) a reference corpus of texts originally produced in French. The three corpora are comparable in that they all consist exclusively of fictional texts produced after 1980. For the translations, this is also the case with the STs.[2]

The corpus of student translations contains 39 different text fragments, each translated from English into French by a different student enrolled between 2010 and 2015 in the first year of "MéLexTra", a profession-oriented master's programme with entry selection at the English Department of the University of Lille.[3] As part of their coursework, the students were required to translate into French (the students' mother tongue (L1)) a short story or a chapter from a novel originally written in English. Among the STs are works by Margaret Atwood, Nick Hornby and Irvine Welsh (the last of whom being the only author from whose work not one but two text fragments were chosen for translation, assigned to two different students). The average length of the student translations is 12,735 words (standard deviation: 3,067 words), the shortest and longest text containing 8,149 words and 23,806 words respectively. The 39 texts in this corpus of student translations make up close to half a million words.

The corpus of professional translations (Loock, Lefebvre-Scodeller, & Mariaule, 2012) contains 42 complete novels originally written in English and translated into French. These texts have an average length of 120,585 words (standard deviation: 48,311 words), the shortest being a 31,532-word book by Roald Dahl and the longest a 282,766-word book by Tom Clancy. For each text in the corpus, the ST is by a different author, therefore minimizing the risk that a single set of individual author peculiarities shines through in the translations. For five of the 42 books, we have no precise information on the identity of the translator, but among the 37 texts for which we do, only two are translated by the same translator. In other words, this corpus certainly does not display any bias in translation habits but rather aims to represent as wide a variety of translation styles as one might find among professional translators.

Finally, the smaller reference corpus of original, non-translated French is made up of three entire post-1980 novels written in French: Frédéric Beigbeder's (2003) *Windows on the World* (69,732 words), Marc Lévy's (2000) *Et si c'était vrai …* (64,243 words), and Bernard Werber's (1991) *Les Fourmis* (94,873 words). As with the texts in the corpus of professional translations, the choice of texts in this corpus was largely determined by the ease with which they could be found on the web. Obviously, the number of different texts in the reference corpus will have to be expanded in future studies. Table 2 summarizes the information about the structure of these three corpora.

Table 2: Data used for case study 1

	French translated from English by advanced students	French translated from English by professional translators	Original French
Genre	Fiction	Fiction	Fiction
Period	Post-1980	Post-1980	Post-1980
Number of texts	39	42	3
Number of tokens	471,660	5,280,232	228,848

After computing all 30 linguistic features from these corpora, PCA was used to analyse the data. The results of this PCA are visually represented in Figure 1. As mentioned before, we have limited ourselves to presenting only the first two dimensions (principal components), which encompass almost half of the variation in the original dataset (more specifically, 44.5%).

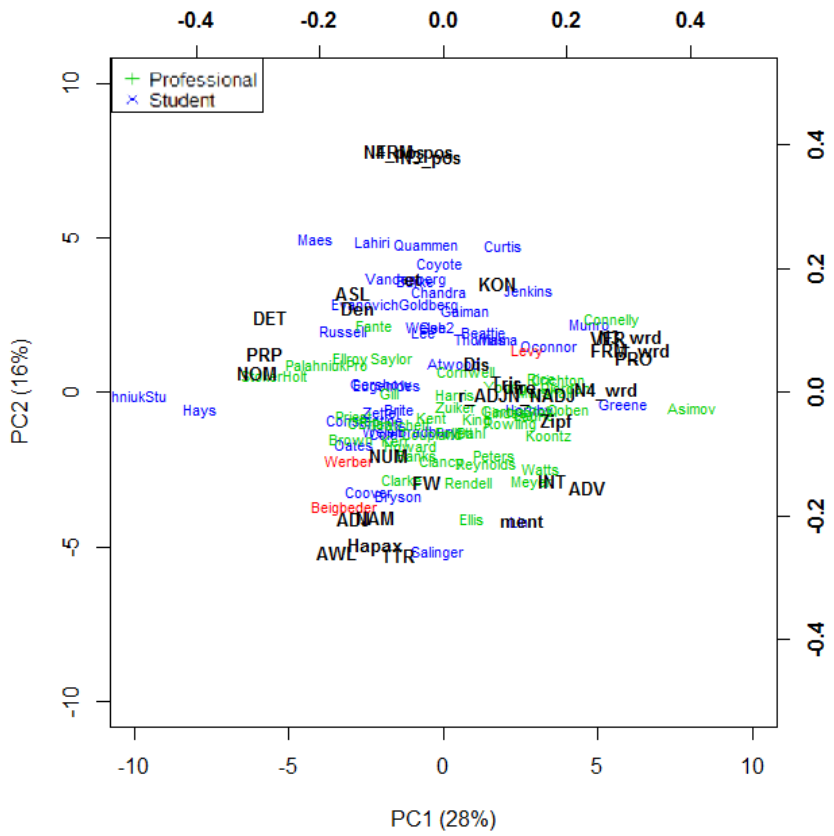

Figure 1: Biplot of the PCA for the English-to-French translations (blue: student translations, green: professional translations, red: original, non-translated professional texts)

Before interpreting this plot, one should note that the abbreviations printed in black represent the 30 selected linguistic features, the names in blue are the student translations, those in green are the professional translations and those in red are the professional originals (non-translations). The numerical values on the *x*- and *y*-axis do not have a straightforward interpretation. What *is* meaningful, however, is the relative position of the different texts vis-à-vis each other and vis-à-vis the linguistic features in the plot: the closer two texts are, the more similar their linguistic characteristics (and vice versa); when a text is close to a given linguistic feature this means that the feature is clearly present in this text.

It should be observed that the picture in Figure 1 is far from clear-cut, as the professional translations together with the professional non-translations do not form a homogeneous cluster clearly separated from the student translations. Nevertheless, one *can* observe that the student translations (marked in blue) and the professional translations (marked in green) form separate clusters, although these clusters are not clearly distinct, with some student translations clearly being part of the professional translators' cluster and vice versa. The three non-translated French texts are situated more or less on the border between the clusters containing student and professional translations. What one can learn from this is that the methodological approach introduced in the previous section is – to some extent – capable of discriminating between texts translated by experienced, professional translators, on the one hand, and inexperienced student translators, on the other. This suggests that student translations do not sufficiently conform to the TL norms and conventions (hence, an acceptability problem, see below), albeit not to such an extent that it is immediately applicable to translation curricula. This can either mean that the methodological procedure is not yet optimal – we need other or more linguistic features to identify the professional standard – or that there is no such thing as a clearly delineated linguistic standard to which all professional translators and writers adhere. Needless to say, more research is needed to find out how feasible this corpus-based statistical approach is, hence our decision to replicate this study on a dataset using another translation direction (French to Dutch) and another genre (news translation). The results of this study are presented in the next section.

Figure 1 also gives us an initial idea of how our methodological approach can be used to assess individual student translations and to provide tailor-made feedback about the linguistic dimensions that clearly deviate from "normal" professional (translators' and non-translators') behaviour. The student translations that are located on the periphery of the plot (Curtis, Maes, Lahiri, Hays, etc.) exhibit linguistic features which place them at a distance from the standard. The student translations at the top periphery of the plot, for instance, apparently show over-use of formulaic patterns, as these are positioned close to 3- and 4-grams (and far away from the professional translations). These translations therefore do not seem to be acceptable according to TL norms, and need further inspection.

Finally, our methodology can also be used to investigate the aggregate behaviour of the students and, more particularly, to identify the linguistic features that are significant indicators of deviant student translation behaviour. Although this information is suggested by Figure 1 (the location of the black-coloured features relative to a given cluster reveals the degree of association), the difference between students and professionals can be tested for each individual linguistic feature by means of ANOVA. Figure 2 depicts all those features that indicated a significant difference between student and professionals.

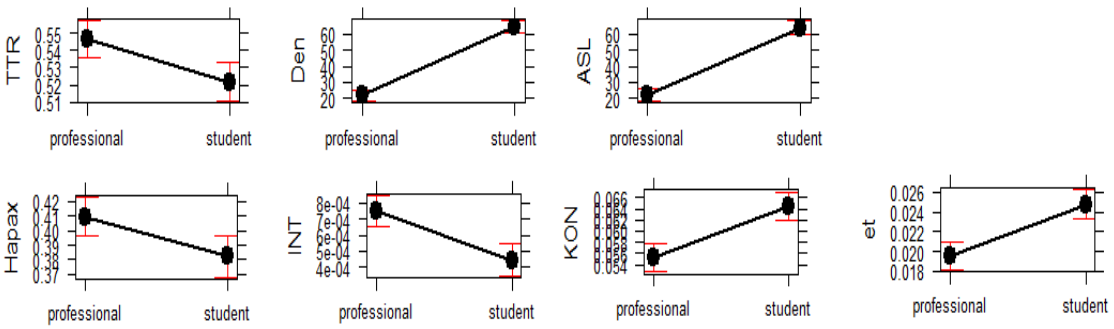

Figure 2: Plots of linguistic indicators with a significant difference between students and professionals (English to French)

It can be observed that only seven features (out of 30) exhibit a significant difference between students and professionals. More particularly, it is clear that the student translators score significantly lower for Type-Token Ratio (TTR) and hapax legomena (i.e., two features commonly associated with lexical creativity) as well as for number of interjections (INT). The student translators score higher for lexical density (Den), average sentence length (ASL), number of conjunctions (KON) and the use of the (semantically empty) connector *et*.

In sum, whereas professional translators outperform students in lexical creativity, students deviate from professionals' behaviour in density and clause length: clauses are longer in student translations and, despite the fact that they contain a higher proportion of lexical to grammatical words, they are more often paratactically connected by means of the vague connector *et* or by any other explicit connecting device.

In the next section, we present the results of a new case study based on the same methodology. This study will elucidate the feasibility and stability of our approach.

5. Case study 2: French-to-Dutch translation of news texts

As in case study 1, we used three corpora: a student translation corpus, a professional translation corpus and a professional non-translation corpus. All the texts are comparable in genre (news) and they were all produced between 2000 and 2015 by native-speakers of Dutch. Both professional corpora were extracted from the publicly available Dutch Parallel Corpus (Macken, De Clercq, & Paulussen, 2011); the student translation corpus was compiled for this study and the translation direction is French to Dutch. Each of the corpora consists of 136 texts of approximately 400 words each, each text being translated or written by a different translator/author. The student translations were written in 2015 by native-speakers of Dutch who enrolled in the second year of the Applied Language Studies programme at Ghent University.[4] As part of their coursework, the students were required to translate into Dutch short news articles (on different topics, such as Ebola, the Mexican army and a hunger strike in Iran) from the French magazine *le Courrier International*.

After extracting all 25 language-independent linguistic features listed in Table 1, PCA was used to analyse the data. The results are represented visually in Figure 3. As in case study 1, the first two principal components are able to encompass almost half of the variation in the original dataset (more particularly, 45.2%).

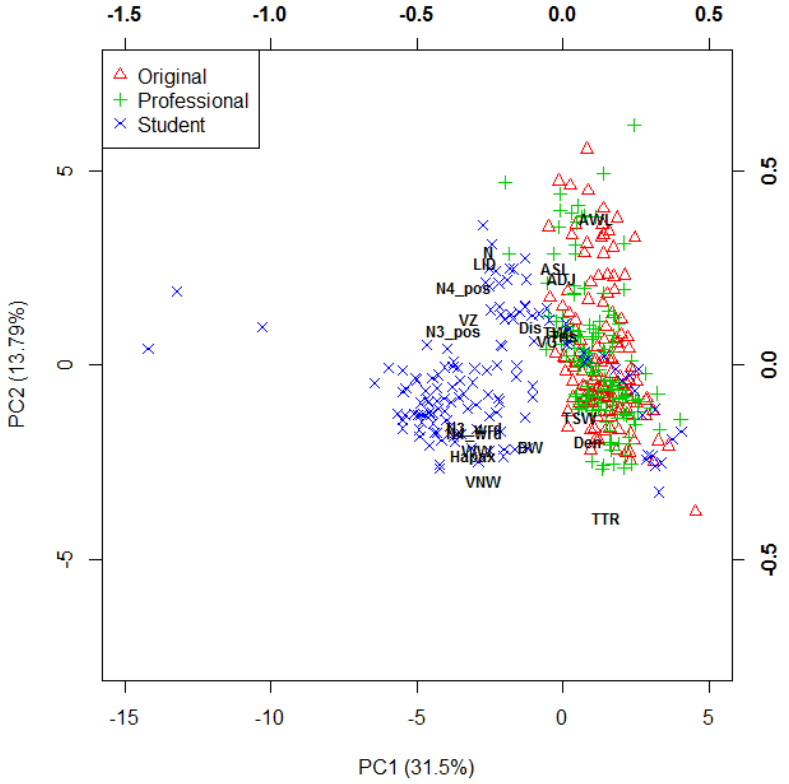

Figure 3: Biplot of the PCA for the French-to-Dutch translations (blue: student translations, green: professional translations, red: original, non-translated professional texts)

Compared to the results presented in case study 1, a clearer picture emerges: most student translations are separated from the translations and originals (non-translations) written by professionals and the texts produced by both types of professional coincide completely. Obviously, the students' and the professionals' clusters are again not perfectly distinct, as some student translations are clearly part of the professionals' cluster; however, the opposite does not occur here. This might signify that the selected linguistic features are better able to identify the underlying linguistic standards to which professionals adhere. Without further research, it is impossible to guess why this is so for case study 2 and to a lesser extent for case study 1; it might have to do with the languages involved or with the genre, to mention only two possible explanations. We will return to this point in the final section.

As we are better able to identify the professionals' standard, it is also more straightforward to evaluate the translation quality of individual student translations: student translations located in (or close to) the professionals' cluster can be considered to conform (almost) completely to the professionals' textual patterns.

Which linguistic features are significant indicators of deviant (aggregate) student translation behaviour in this case study? To answer this question we again perform ANOVA on each individual linguistic feature to test the differences between students and professionals. Figure 4 lists all those features that show a significant difference between students and professionals.

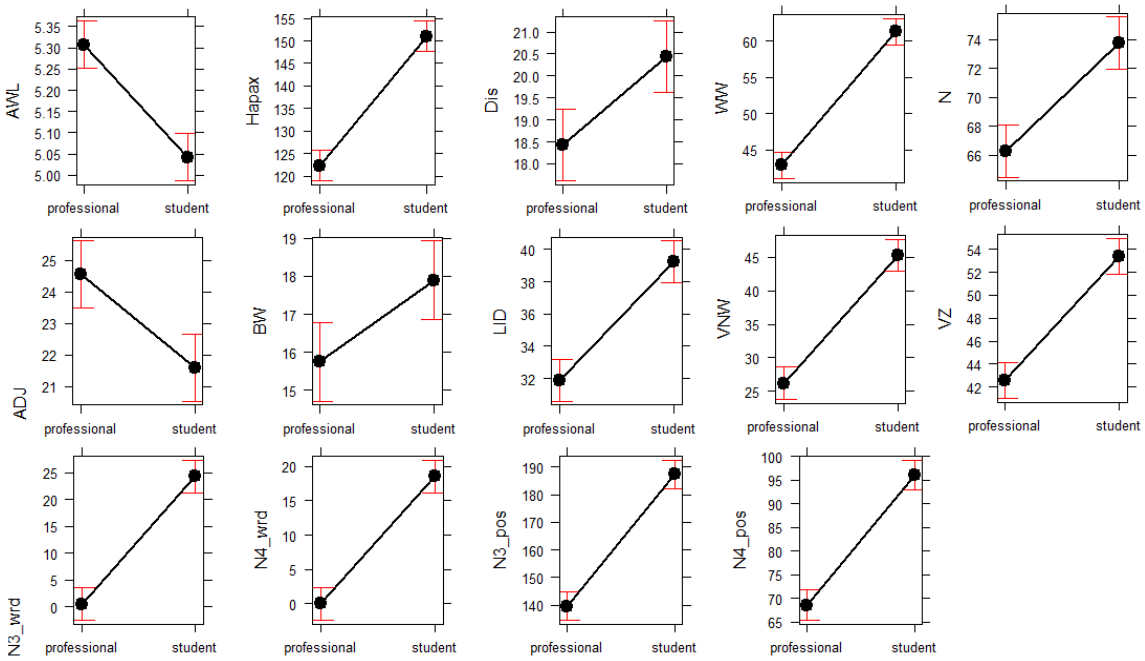

Figure 4: Plot of linguistic features with a significant difference between students and professionals (French to Dutch)

First, it can be seen that many more features, namely, 14 (out of 25), exhibit significant differences between student and professional translations than in case study 1: here, student translations have lower scores for average word length (AWL) and the use of adjectives (ADJ); all the other features yield higher scores for the student translators, including hapax legomena, dis legomena, the frequency of verbs (WW), nouns (N), adverbs (BW), determiners (LID), pronouns (VNW), prepositions (VZ) and tri- and four-grams.

These results do not align well with the results in the previous case study, rendering them difficult to interpret. For instance, in case study 1 it was found that student translations exhibit a lower number of hapax legomena than those of professional translators, whereas we find the opposite effect in case study 2. A feature such as Type-Token Ratio is significant in the first case study but not in the second. Why we obtain these contradictory findings, and what this tells us about the methodological approach, will be answered provisionally in the next section.

6. Discussion and conclusion

In the previous sections, we argued for a corpus-based, statistical approach to translation evaluation, more particularly translation acceptability and feedback based on the distribution of a large set of "basic" linguistic features. These features are thought to escape the conscious attention of most writers: Type-Token Ratio, lexical density, number of adjectives, number of hapax legomena, etc. More specifically, by investigating the distribution of these features in a corpus of student translations compared to their distribution in a corpus of professional writers and translators (by means of PCA and ANOVA), we have shown that such an approach is feasible. By adopting the methodology in two different case studies in which different languages, different genres and different students were involved, we have shown that our methodology is capable of distinguishing between translation students' texts, on the one hand, and professional texts (both by translators and non-translators), on the other. Based on the assumption that the relatively homogeneous linguistic behaviour can be considered a standard, the methodology used in this article can be employed to

identify hidden patterns of deviance in student translations (both individual or aggregate) and therefore to assess the acceptability of student translations.

Nevertheless, it has also become clear that many questions remain unanswered. Even though the methodology is able to identify deviant student translations in two different case studies, it is not always a straightforward matter to find out why certain linguistic features are indicators of deviant student translation behaviour in a given setting and why the effects of the indicators fluctuate across studies. In other words, it is still not completely clear how the selected linguistic features relate exactly to translation quality, and it is therefore unclear how this should be incorporated into feedback to students. In summary, more research needs to be done on new student populations in order to replicate the discriminatory ability of the proposed methodology and to find out whether the differences we have encountered in this study are due to genre differences, translation direction, a student's degree (BA or MA) or any other factor. Moreover, other "basic" linguistic features should be included in the analysis, such as the lexico-grammatical indicators mentioned in Evert and Neumann (2017). Furthermore, the method should be tested in a real translation class situation and post-hoc interviews with teachers and students should be conducted to gauge the perceived value of this approach. Finally, a comparative analysis of our translation quality-assessment procedure and other, non-corpus-based procedures needs to be performed to see how they relate to each other and whether they can be combined in a translation quality toolbox.

Even if one were to find the development of such a toolbox not worth pursuing, we are still convinced that it should at least be feasible to develop a tool that gives students and translation teachers automatic feedback on a student translation. What we have in mind is an application, possibly integrated in an electronic learning environment, which would allow students to upload their translations, after which the teacher and/or the student would then immediately be informed about the text's properties that might otherwise easily escape the teacher's or the student's attention. This application would produce a table with detailed information about how closely the translation approaches a "normal" text with respect to such features as lexical density, sentence length, the frequency of the connector *et*, and so on.

Will translation teachers in the future be replaced by computers? Obviously not. A scoresheet with numerical data about textual features can never be a substitute for a teacher's more carefully considered appreciation of a student's translation assignment. We suggest here that automatically generated feedback may be useful to the teacher, though, as an *instrument* to be used when discussing a student's translation. The way the teacher uses this feedback would not be very different from the way a doctor interprets a patient's blood values: these remain to be interpreted. In addition, some "divergent" values might turn out to be less problematic if they result from particular features of an ST. For instance, if the SL contains many short, staccato-like sentences, then the student should of course attempt to be faithful to that ST style even if doing so results in a translation that stands out as special in the TL. Conversely, there is the possibility that a translation might not exhibit, for a particular parameter, any noticeable difference from an originally produced text in the TL while, in actual fact, the translation grossly fails to represent a stylistic SL property that ideally ought to have been reflected in the translation. A more sophisticated feedback tool should therefore also consider the linguistic properties of the uploaded ST so that any alarming discrepancies between, say, the average sentence length of the ST and that of the TT could also be included in the feedback.

We hope that this article will encourage other researchers to investigate the intricate and challenging issue of measuring translation quality using a corpus-based quantitative methodology, and in so doing help to bridge the gap between Empirical and Applied Translation Studies.

References

Anckaert, P., Eyckmans, J., & Segers, W. (2008). Pour une évaluation normative de la compétence de traduction. *International Journal of Applied Linguistics, 155*, 53–76.

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

Bowker, L. (1998). Using specialized monolingual native-language corpora as a translation resource: A pilot study. *Meta: Journal des traducteurs / Meta: Translators' Journal, 43*(4), 631–651.

Bowker, L. (1999). Exploring the potential of corpora for raising language awareness in student translators. *Language Awareness, 8*(3–4), 160–173.

Bowker, L. (2000). A corpus-based approach to evaluating student translations. *The Translator, 6*(2), 183–210.

Bowker, L. (2001). Towards a methodology for a corpus-based approach to translation evaluation. *Meta: Journal des traducteurs / Meta: Translators' Journal, 46*(2), 345–364.

Bowker, L., & Pearson, J. (2002). Working with specialized language: A practical guide to using corpora. London: Routledge.

Cappelle, B. (2012). English is less rich in manner-of-motion verbs when translated from French. *Across Languages and Cultures, 13*(2), 173–195.

Cappelle, B., & Loock, R. (2017). Typological differences shining through: The case of phrasal verbs in translated English. In G. De Sutter, I. Delaere, & M.-A. Lefer (Eds.), *Empirical Translation Studies: New theoretical and methodological traditions* (pp. 235–263). Berlin: Mouton de Gruyter.

Chesterman, A. (1999). The empirical status of prescriptivism. *Folia Translatologica*, *6*, 9–19.

Chesterman, A. (2004). Beyond the particular. In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (pp. 33–49). Amsterdam: John Benjamins.

Daems, J. (2016). A translation robot for each translator?: A comparative study of manual translation and post-editing of machine translations: process, quality and translator attitude (Unpublished doctoral dissertation). Ghent University, Ghent.

Daems, J., Vandepitte, S., Hartsuiker, R., & Macken, L. (in press). Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators. *Meta: Journal des traducteurs/Meta: Translators' Journal*.

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

De Sutter, G., Delaere, I., & Lefer, M.-A. (Eds.). (2017). *Empirical Translation Studies: New theoretical and methodological traditions*. Berlin: Mouton de Gruyter.

Evert, S., & Neumann, S. (2017). The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. In G. De Sutter, I. Delaere, & M.-A. Lefer (Eds.), *Empirical Translation Studies: New theoretical and methodological traditions* (pp. 47–80). Berlin: Mouton de Gruyter.

François, T., & Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and improving text readability for target reader populations (PITR2012)* (pp. 49–57). Montréal: The Association for Computational Linguistics.

Frankenberg-Garcia, A. (2015). Training translators to use corpora hands-on: Challenges and reactions by a group of 13 students at a UK university. *Corpora, 210*(3), 351–380.

Hassani, G. (2011). A corpus-based evaluation approach to translation improvement. *Meta: Journal des traducteurs / Meta: Translators' Journal, 56*(2), 351–373.

Johansson, S. (2007). *Seeing through multilingual corpora*. Amsterdam: John Benjamins.

Kruger, H. (2015, June). Translation and the intersection of social and cognitive aspects of bilingualism. Paper presented at *Theory, Practice and Innovation: Social, Cognitive and Linguistic Perspectives in the Study of Bilingualism*. University of New South Wales.

Kruger, H. **(**2016). What's happening when nothing's happening?: Combining eyetracking and keylogging to explore cognitive processing during pauses in translation production. *Across Languages and Cultures*, *17*(1), 25–52.

Kübler, N. (2001). Corpora in terminology and translation teaching: Methodological approach. In S. De Cock, G. Gilquin, S. Granger, & S. Petch-Tyson (Eds.), *Proceedings of the ICAME 01 Conference* (pp. 53–55). Louvain-la-Neuve : Centre for English Corpus Linguistics.

Kübler, N. (2003). Corpora and LSP translation. In S. Bernardini, D. Stewart, & F. Zanettin (Eds.), *Corpora in translator education* (pp. 25–42). Manchester: St Jerome.

Kübler, N. (2008). A comparable learner translator corpus: Creation and use. In P. Zweigenbaum (Ed.), *Proceedings of the Comparable Corpora Workshop of the LREC Conferenc*e (pp. 73–78). Marrakesh, Morocco.

Kübler, N. (Ed.). (2011a). Language corpora, teaching, and resources: From theory to practice. Bern: Peter Lang.

Kübler, N. (2011b). Working with corpora for translation teaching in a French-speaking setting. In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston (Eds.), *New trends in corpora and language learning* (pp. 62–80). London: Continuum.

Loock, R. (2016a). *La traductologie de corpus*. Lille: Presses Universitaires du Septentrion.

Loock, R. (2016b). L'utilisation des corpus électroniques chez le traducteur professionnel: Quand? Comment? Pour quoi faire? *ILCEA, 27*. Retrieved from https://ilcea.revues.org/3835.

Loock, R., Lefebvre-Scodeller, C., & Mariaule, M. (2012). *Corpus CorTEx de français littéraire traduit depuis l'anglais*. Retrieved from https://sites.google.com/site/cortexresearchproject/home/corpus-corpora

Loock, R., Mariaule, M., & Oster, C. (2014). Traductologie de corpus et qualité: Étude de cas. *Proceedings of the Tralogy II Conference*. Retrieved from http://lodel.irevues.inist.fr/tralogy/index.php?id=243.

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

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers, 36*(3), 516.

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

Pallier, C. (1999). Syllabation des representations phonetiques de brulex et de lexique. Technical Report, update 2004.

Pearson, J. (2003). Using parallel texts in the translation training environment. In S. Bernardini, D. Stewart, & F. Zanettin (Eds.), *Corpora in translator education* (pp. 15–24). Manchester: St Jerome.

Rabadán, R., Labrador, B., & Ramon, N. (2009). Corpus-based contrastive analysis and translation universals: A tool for translation quality assessment English-Spanish. *Babel, 55*(4), 303–328.

Ruiz Yepes, G. (2011). Parallel corpora in translator education. *Redit,* 7, 65–80.

Sánchez-Gijón, P. (2009). DIY corpora in the specialised translation course. In A. Beeby, P. Rodríguez-Inés, & P. Sánchez-Gijón (Eds), *Corpus use and translating: Corpus use for learning to translate and learning corpus use to translate* (pp. 109–128). Amsterdam: John Benjamins.

Secară, A. (2005). Translation evaluation: A state of the art Survey. *Proceedings of the eCoLoRe-MeLLANGE Workshop*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.126.3654.

Toudic, D., Hernandez Morin, K., Moreau, F., Barbin, F., & Phuez, G. (2014). Du contexte didactique aux pratiques professionnelles: Proposition d'une grille multicritères pour l'évaluation de la qualité en traduction spécialisée.' *ILCEA*, *19*. Retrieved from http://ilcea.revues.org/2517.

Van de Kauter, M., Cooreman, G., Lefever, E., Desmet, B., Macken, L., & Hoste, V. (2013). LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal, 3*, 103–120.

Vandevoorde, L. (2016). On semantic differences: A multivariate corpus-based study of the semantic field of inchoativity in translated and non-translated Dutch (Unpublished doctoral dissertation). Ghent University, Ghent.

Van Oosten, P., Tanghe, D., & Hoste, V. (2010). Towards an improved methodology for automated readability prediction. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)* (pp. 775–782). European Language Resources Association (ELRA).

Varantola, K. (2003). Translators and disposable corpora. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 55–70). Manchester: St. Jerome.

Williams, M. (2009). Translation quality assessment. *Mutatis Mutandis*, *2*(1), 3–23.

Xiao, R. (2010). How different is translated Chinese from native Chinese?: A corpus-based study of translation universals. *International Journal of Corpus Linguistic*s, *15*(1), 5–35.

Zanettin, F. (2012). *Translation-driven corpora*. Manchester: St Jerome.

Zanettin, F., Bernardini, S., & Stewart, D. (Eds.) (2003). *Corpora in translator education*. Manchester: St Jerome.

1 Nevertheless, we might expect some differences between professional translators and non-translators as well, albeit to a lesser extent. Evert and Neumann (2017), for instance, provide evidence for significant differences in lexicogrammatical patterning between professional translations and non-translations (originals).

2 The reason we also restricted the range of publication years of the source texts is that we are keenly aware of potential problems in comparing contemporary translations whose source texts display a rather wide range of stylistic properties because they were written in different periods. For example, a modern French translation of Jane Austen's novels will still be felt to be different in many respects from translations of contemporary English novels.

3 We are grateful to Fabrice Antoine and Corinne Oster (Université Lille) for granting us permission to gather the translated texts for the student translation corpus.

4 We gratefully acknowledge the support of Désirée Schyns (Ghent University) in compiling the student translation corpus.