Quaderns de Filologia. Estudis lingüístics. Vol. XVIII (2013) 49-59

WORD-FORMATION IN ORIGINAL AND TRANSLATED ENGLISH: SOURCE LANGUAGE INFLUENCE ON THE USE OF *UN-* AND *-LESS*[*]

Marie-Aude Lefer, Institut libre Marie Haps/Université catholique de Louvain
Bruno Cartoni, Université de Genève

1. Aims and background of the study

This article, which is situated at the intersection of corpus-based contrastive / translation studies and morphology, aims to compare the word-formation features of original and translated language and examine the possible impact of the source language (SL) variable on the word-formation characteristics of translated texts. This will be done by studying the English *-less* and *un-* negative affixes in original texts and texts translated from French, Italian, Dutch and German taken from the Europarl parallel corpus of parliamentary debates (see Koehn, 2005 for a description of the Europarl corpus and Cartoni & Meyer, 2012 for the description of the extraction of directional parallel corpora from Europarl, i.e. parallel corpora where source and target languages are clearly identified).

It is now widely recognized that electronic corpora have dramatically revolutionized the fields of contrastive linguistics and translation studies (see e.g. Granger *et alii*, 2003; Johansson, 2007; Xiao, 2010). The two disciplines make use of similar data to pursue their respective research objectives, which can facilitate their rapprochement (Granger, 2003). One of the most obvious signs of this cross-fertilization is the renewed interest that translation studies have taken in the role played by the SL variable in corpora of translated texts (e.g. Hansen-Schirra *et alii*, 2012).

Back in the early 1990s, Baker (1993, 1995) convincingly called for the exploration of universal features of translation, i.e.

[*] We wish to thank Yves Bestgen (Université catholique de Louvain) for his advice on statistical tests, Paul Rayson (Lancaster University) for POS-tagging the corpora used in this study with CLAWS 8 and the anonymous reviewers for their helpful suggestions. We also gratefully acknowledge the financial support of the European Science Foundation *NetWordS* Research Networking Programme (NetWordS-09-RNP-089).

> features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems (1993: 243).

Baker (ibid., 240) also stressed that descriptive translation studies as a discipline cannot be reduced to "comparative analyses of source and target texts", thereby advocating the use of corpora of translated texts:

> we need to effect a shift in the focus of theoretical research in the discipline, a shift away from comparing either STs [source texts] with TTs [target texts] or language A with language B to comparing text production per se with translation (Baker, 1995: 233).

After some twenty years of intensive and wide-ranging corpus-based research in the field, it now appears that variables other than the translation process itself (e.g. register/genre variation, language pair, translator status) can also provide viable explanations for the observed differences between non-translated, original texts and translated texts (e.g. De Sutter *et alii*, 2012). The research reported on here is part of this new, cross-fertilizing trend in corpus-based contrastive and translation studies that seeks to confront translation-related and language-pair specific features of translated texts.

To date, corpus-based studies in the field of word-formation have looked at translated language with the aim of:

(1) *Assessing the role of the SL in the overuse of some derivational affixes in the target language (TL)*.

Wang & Qin's (2010: 175) study, for example, shows that the Chinese *xing* noun-forming suffix (which denotes properties) is overused in translated fiction compared to original fiction. The authors argue that this tendency is due to the high frequency of *ity*, *-ness* and *-dom* in English source texts compared to *-xing* in non-translated Chinese;

(2) *Examining translation-related, SL- and TL-independent trends such as the normalization of creative lexis in translated language* (cf. Kenny, 2001; Olohan, 2004).

In a translation corpus,

> normalization may be said to occur when translators opt for conventional target language solutions to problems posed by creative or unusual source text features (Kenny, 2001: 66).

Olohan (2004: 108-117), for instance, reports that there are more non-lexicalized, creative uses of English *-ish* (i.e. *-ish* forms not listed in reference dictionaries) in non-translated fiction than in translated fiction (e.g. *mummyish*, *Londonish*);

(3) *Uncovering language-pair specific properties which lead to a marked decrease in the use of certain word-forming devices in target texts compared to their source texts*.

Lefer (2012), which deals with the English *un-* and *in-* negative prefixes and their French translation equivalents, has identified three language-pair specific properties that account for the fact that a SL derivational affix is not translated into a TL affix:

(i) cross-linguistic differences in morphological productivity: many En. *un*-words cannot be translated into Fr. *in*-words because En. *un-* is much more productive than Fr. *in-* (e.g. En. *unconnected* ® Fr. *qui n'avaient aucun lien* 'that had no links', **inconnecté*);
(ii) diverging polysemy: for example, En. *inconsiderate* cannot be translated into Fr. *inconsidéré* when it qualifies people and a simplex adjective such as égoïste 'selfish' has to be used instead; and
(iii) partial phraseological equivalence (e.g. to render En. *unmatched opportunity*, translators often opt for Fr. *occasion unique*, a collocation made up of a noun and a simplex adjective, which is much more frequent than *occasion inégalée*).

These three trends are respectively termed

(1) 'SL-induced morphological increase',
(2) 'translation-inherent morphological decrease' and
(3) 'language-pair specific morphological decrease' in Lefer (2012).

The present article contributes to this growing body of corpus-based research by examining whether the SL variable plays a role in the use of *-less* and *un-* in English texts translated from two Romance languages, French and Italian, and two Germanic languages, German and Dutch. In doing so, we aim to determine whether (and to what extent) the trends identified in translated English are SL-dependent or translation-related.

Our initial hypothesis is that original English and translated English differ in their use of (at least some) derivational features. In addition, we hypothesize

that English translated from Romance languages and English translated from Germanic languages are also markedly different as regards morphological patterns. To test these hypotheses, we focus on the use of two negative affixation patterns:

(1) *un*V*ed*$_{A}$ (i.e. *un*-adjectives formed on the basis of a past participle; e.g. *unachieved*, *uncontested*, *unsuspected*), an English pattern which has been shown to be much more productive than its French or Italian counterparts, i.e. negative word-formation patterns involving a past participle and a negative prefix (e.g. Fr. *inadapté* and It. *inadatto* 'unadapted') (cf. Cartoni & Lefer, 2011). Examples of *un*V*ed*$_{A}$ that lack a morphologically similar counterpart in French and Italian include *unarmed*, *unplanned*, *unpublished*, *unregistered*;

(2) N*less*$_{A}$ (i.e. denominal adjectives ending in *-less*; e.g. *endless*, *meaningless*, *powerless*), as there is no corresponding negative suffix in French or Italian (vs. Du. *-loos* in *ademloos* 'breathless' and Ge. *-los* in *geschmacklos* 'tasteless'). In English texts translated from French and Italian, *-less* can be considered as a 'Unique Item' (Tirkkonen-Condit, 2004), i.e. an element that lacks obvious equivalents in the source languages and can therefore be expected to be underused in translation.

2. DATA AND STATISTICAL METHOD

The study relies on a monolingual comparable corpus of original and translated language made up of five Europarl subcorpora (Koehn, 2005; Cartoni & Meyer, 2012): (1) original English (OE) and (2) translated English (TE), with four SL components: French (FE), Italian (IE), Dutch (DE) and German (GE). The original English corpus contains 1,410,121 running words (which corresponds to 289 speakers). Table 1 provides information on the four translated English components. Note that the number of translators represented in the translated subcorpora is not available in Europarl. The subcorpora used in the study are characterized by a high degree of comparability[1], which greatly reduces the risks related to potential confounding factors, such as genre/register (cf. Baroni & Bernardini, 2006: 262).

[1] This high degree of comparability can be situated at two levels: (1) language-external variables, such as genre (European parliamentary debates), dates (1996-1999), topics (finance, business, environment, health, etc.) and – for the translated components – translator status (professional translators translating into their native language) and translation direction (direct translation from the SL into English, with no pivot language); (2) language-internal features, such as type/token ratio (between 0.024 and 0.027, as calculated over the first 500,000 tokens of each Europarl component) and lexical density (between 0.51 and 0.53), which "is calculated by dividing the number of lexical items by the total number of words in a text or corpus" (Baker, 1995: 237).

TABLE 1: Europarl translated English subcorpora used in the study

	Source language: number of words	Number of speakers	Target language: nb of words
German → English	1,369,535	490	1,527,054
Italian → English	551,145	287	574,992
French → English	1,179,530	244	1,184,278
Dutch → English	836,456	226	837,782

The results presented in Section 3 are based on a comparison-of-means statistical test, the t-test (see Paquot & Bestgen, 2009 on its use in corpus linguistics). We relied exclusively on type counts (as opposed to token counts) as they reflect the range of different *un-* and *-less* derivatives used in each subcorpus rather than the frequency with which certain *un-* or *-less* derivatives are used (which is more likely to be topic-related, e.g. the recurrent use of *unemployed* when discussing unemployment in Europe). Concretely, we computed a ratio for each text of the five Europarl components[2] by dividing the total number of *-less* or *un-* types (with the following breakdown: *un-* nouns, *un-* adjectives, *un*V*ed* adjectives, *un-* adverbs and *un-* verbs) by the total number of types in each text.[3] The ratios were used to calculate means for each of the five Europarl components analyzed, which were then compared, and the t-test was applied to assess the statistical significance of the observed trends. In total, we examined 2,054 and 7,641 *un-* occurrences and 149 and 632 *-less* occurrences in OE and TE respectively.

3. RESULTS AND DISCUSSION

3.1. *Original English vs. translated English: the source language variable*

In this section, we proceed in two steps:

[2] Each text corresponds to a day of parliamentary debates. There are ca. 250 texts per component (OE, FE, IE, DE, GE).

[3] Preliminary analyses of part-of-speech (POS) distribution across the five Europarl components used in this study revealed that POS are not equally distributed in original English and English translated from French, Italian, Dutch and German (cf. Borin & Prütz, 2001). It was therefore decided to use the total number of types per text as a denominator to calculate ratios for the t-test rather than the total number of adjectives / nouns / verbs / adverbs per text.

(1) we first compare original English (OE) and translated English (TE), without taking into account the SL variable (as is often done in translation studies based on monolingual corpora of translated texts) and
(2) we then compare original English (OE) and English translated from German (GE), Dutch (DE), French (FE) and Italian (IE), so as to assess the potential impact of the SL variable on the general trends identified in (1).

Overall, the data show that there are significantly more *-less* types in translated than in original English (e.g. *measureless*, *visionless*, *valueless*; OE: M=.00038; TE: M=.00053)[4]. In other words, as regards *-less*, translated English is lexically richer, i.e. relies on a wider range of different *-less* types, than original English. Contrary to our initial expectation, there are significantly more *-less* types in English translated from French (as well as Dutch and German) than in original English (OE: M=.00038; FE: M=.00061; DE/GE: M=.00054)[5], despite the lack of an equivalent negative suffix in French. By contrast, there is no marked difference in the use of *-less* types in original English and English translated from Italian (OE: M=.00038; IE: M=.00044)[6]. This is represented in Figure 1.

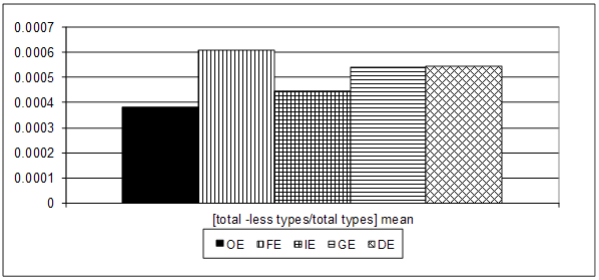

FIGURE 1: *-less* in original English (OE) and English translated from French (FE), Italian (IE), German (GE) and Dutch (DE)

[4] There is a significant difference in the scores for OE (M=.00038, SD=.00060) and TE (M=.00053, SD=.00082); t(1211)=2.735, p=0.006.

[5] There is a significant difference in the scores for OE (M=.00038, SD=.00060) and
(1) FE (M=.00061, SD=.00080); t(489)=3.534, p=0.020;
(2) GE (M=.00054, SD=.00072); t(492)=2.612, p=0.009;
(3) DE (M=.00054, SD=.00094); t(486)=2.247, p=0.025, respectively.

[6] There is no significant difference in the scores for OE (M=.00038, SD=.00060) and IE (M=.00044, SD=.00081); t(473)=0.964, p=0.336.

To sum up, we observe that there are more *-less* types in translated English than in original English, with the exception of English translated from Italian. However, in view of the relatively small size of the IE corpus compared to the other SL subcorpora used in this study, this SL-dependent trend will need to be checked against more data.

Partially similar trends emerge for the *un-* prefix. Contrary to *less*, there are no differences between original and translated English in the use of *un-* noun, *un-* adjective (including the *un*V*ed* pattern) and *un-* verb types when the four SL subcorpora are considered as a whole. The only significant difference we found is that there are more *un-* adverb types in translated English than in original English (e.g. *unceasingly*, *undemocratically*, *undiplomatically*; OE: M=.0007; TE: M=.0011)[7], with no SL influence.

When we compare original English and English translated from French, Italian, German and Dutch, we find two major SL-related contrasts, and they exclusively concern *un-* adjectives (see Figure 2):

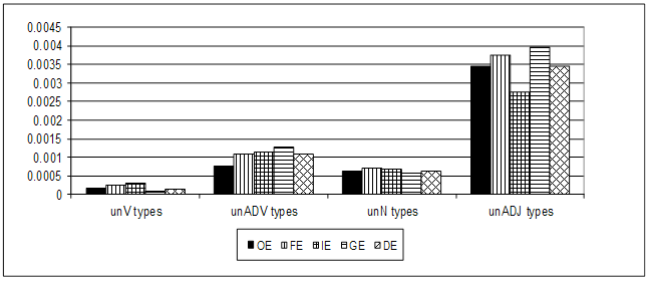

FIGURE 2: *un-* in original English (OE) and English translated from French (FE), Italian (IE), German (GE) and Dutch (DE)

[7] There is no significant difference in the scores for *un-* noun types in OE (M=.00063, SD=.00077) and TE (M=.00065, SD=.00092); t(1211)=0.283, p=0.777; *un-* adjective types in OE (M=.00345, SD=.00220) and TE (M=.00349, SD=.00247); t(1211)=0.224, p=0.823; *un*V*ed* types in OE (M=.00098, SD=.00110) and TE (M=.00100, SD=.00119); t(1211)=0.198, p=0.843; *un-* verb types in OE (M=.00015, SD=.00045) and TE (M=.00020, SD=.00151); t(1211)=0.460, p=0.646. There is a significant difference in the scores for *un-* adverb types in OE (M=.00077, SD=.00086) and TE (M=.00114, SD=.00115); t(1211)=4.670, p<0.000.

(1) there are fewer *un-* adjective types in English translated from Italian than in original English[8] (OE: M=.00345; IE: M=.00275) and
(2) there are more *un-* adjective types, including more *un*V*ed* types (see Figure 3), in English translated from German than in original English (e.g. *unenthusiastic*, *undebated*; *un-* adjectives: OE: M=.00345; GE: M=.00396; *un*V*ed*$_A$: OE: M=.00098; GE: M=.00129)[9]. These SL-dependent patterns in the use of *un-* adjectives are obscured when all four SLs are considered together, i.e. when OE is compared to TE (cf. above). However, when FE, IE, DE and GE are considered individually and compared to OE, we observe an underuse of the *un-* adjective prefixation pattern in IE and an overuse in GE, while no significant differences are observed between original English and English translated from French or Dutch. It is also important to note that the use of *un-* noun and verb types (e.g. *uncertainty*, *unconsciousness*, *unblock*, *unlock*) is similar in original and translated English, whatever the SL component considered. In other words, SL influence is not at play here.

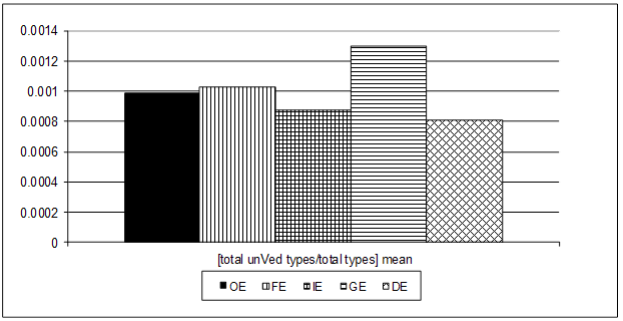

FIGURE 3: *un*V*ed* adjectives in original English (OE) and English translated from French (FE), Italian (IE), German (GE) and Dutch (DE)

[8] This reflects a more general tendency, i.e. that there are fewer adjective types in IE (M=.1614, SD=.0291) than in OE (M=.1696, SD=.0236); t(473)=3.345, p=0.001.

[9] There is a significant difference in the scores for *un-* adjective types in OE (M=.00345, SD=.00220) and IE (M=.00275, SD=.00257); t(473)=3.170, p=0.002 and in OE (M=.00345, SD=.00220) and GE (M=.00396, SD=.00242); t(492)=2.433, p=0.015. There is also a significant difference in the scores for *un*V*ed* types in OE (M=.00098, SD=.00110) and GE (M=.00129, SD=.00118); t(492)=3.018, p=0.003.

3.2. *A qualitative look at* un*V*ed$_A$ *in English texts translated from German*

In this section, we briefly zoom in on a SL-related phenomenon outlined in Section 3.1: the wider variety of *un*V*ed* adjective types in English texts translated from German than in original texts. The analysis of German-English aligned segments[10] reveals that in 52% of the cases (202 out of 388 occurrences), the *un*V*ed*$_A$ pattern is triggered by its German equivalent pattern (i.e. the German *un-* prefix attached to a past participle base). Examples include Ge. *unbelastet* → En. *untaxed*, Ge. *ungekühlt* → En. *unrefrigerated*, Ge. *ungelöst* → En. *unresolved/unsolved*, Ge. *ungewollt* → En. *unwanted*. The negative suffixes *-los* (e.g. Ge. *beispiellos* → En. *unprecedented*) and *-frei* (e.g. Ge. *straffrei* → En. *unpunished*), for their part, trigger the *un*V*ed*$_A$ occurrences examined here in 4% of the cases. As these cross-linguistic morphological correspondence percentages are not particularly high[11], it is difficult to assert at this stage that no other variable than SL influence is at play here. In any case, we observe that the overuse of the *un*V*ed*$_A$ pattern in English translated from German compared to original English does not systematically stem from morphological correspondences between German *un-*, *-los* and *-frei* and English *un-*.

4. Concluding remarks

Our study has shown that the use of *-less* and *un-* types in European parliamentary debates is partially sensitive to SL influence, which demonstrates the importance of considering the SL variable in monolingual corpora of translated texts. Surprisingly, we also found that *-less* types are overused in translated English compared to original English (the only exception being English translated from Italian). This latter trend, provided it is generalizable to other SLs, could be interpreted as a case of 'translation-inherent morphological increase', i.e. an increase in the number of morphologically complex words (i.e. derivatives, compounds, conversions, blends, etc.) in translated texts in L_y compared to non-translated texts in L_y, which is inherent in the translation

[10] We examined the German items that led to the use of the *un*V*ed*$_A$ pattern in translated English. When a given *un*V*ed*$_A$ type was used more than once in the same Europarl text, only the first occurrence was taken into account, while the others were discarded from our final dataset. In addition, it should be noted that the occurrences of the high-frequency *unemployed* – which is by far the most frequent *un*V*ed* adjective in the Europarl corpus – were all disregarded, as they systematically correspond to German *arbeitslos*.

[11] There are 160 cases where there is no negative affixation trigger in the German source text, e.g. *Gesundheit* hat etwas zu tun mit *life science, mit Biotechnologie*. → *Health* is not UNRELATED to *the life sciences and biotechnology*.

process and which, consequently, cannot be attributed to SL_x interference or other confounding factors (cf. Lefer, 2012: 152-153).

Naturally, the corpus approach proposed in this paper needs to be adopted in other genres (see Kruger & van Rooy, 2012), in other languages and for other word-formation elements so as to assess the generalizability of the findings. We also need to uncover the exact reasons behind the trends identified here. This can be done by looking more closely at the SL items that trigger the use of *less* and *un-* in translated English.

Even though we have only scratched the tip of the iceberg, our study shows how monolingual comparable corpora of original and translated language as well as parallel corpora can be used to investigate SL-related and translation-related word-formation phenomena in translation, an area which, to date, remains under-researched in corpus-based contrastive and translation studies.

BIBLIOGRAPHY

Baker, M. (1993): "Corpus linguistics and translation studies. Implications and applications". In: Baker, M. *et alii* (eds.), *Text and Technology. In Honour of John Sinclair*. Amsterdam & Philadelphia, PI: John Benjamins; 233-250.

Baker, M. (1995): "Corpora in translation studies: an overview and some suggestions for future research". *Target* 7, 2: 223-243.

Baroni, M. & Bernardini, S. (2006): "A new approach to the study of translationese: machine-learning the difference between original and translated text". *Literary and Linguistic Computing* 21, 3: 259-274.

Borin, L. & Prütz, K. (2001): "Through a glass darkly: part-of-speech distribution in original and translated text". In: Daelemans, W. *et alii* (eds.), *Language and Computers, Computational Linguistics in the Netherlands 2000*. Amsterdam: Rodopi; 30-44.

Cartoni, B. & Lefer, M.-A. (2011): "Negation and lexical morphology across languages: insights from a trilingual translation corpus". *Poznan Studies in Contemporary Linguistics* 47, 4: 795-843.

Cartoni, B. & Meyer, T. (2012): "Extracting directional and comparable corpora from a multilingual corpus for translation studies". *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), [Istanbul, Turkey, May 2012]. ELRA (European Language Resources Association); 1-27. <www.lrec-conf.org/proceedings/lrec2012/pdf/ 188_Paper.pdf>.

de Sutter, G.; Delaere, I. & Plevoets, K. (2012): "Lexical lectometry in corpus-based translation studies. Combining profile-based correspondence analysis and logistic regression modeling". In: Oakes, M. & Meng, J. (eds.),

Quantitative Methods in Corpus-Based Translation Studies. Amsterdam & Philadelphia, PI: John Benjamins; 325-346.

Granger, S. (2003): "The corpus approach: a common way forward for contrastive linguistics and translation studies?". In: Granger, S. *et alii* (eds.), 17-29.

Granger, S. *et alii* (eds.) (2003): *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam: Rodopi.

Hansen-Schirra, S. *et alii* (2012): *Cross-Linguistic Corpora for the Study of Translation: Insights from the Language Pair English-German*. Berlin: de Gruyter.

Johansson, S. (2007): *Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies*. Amsterdam & Philadelphia, PI: John Benjamins.

Kenny, D. (2001): *Lexis and Creativity in Translation. A Corpus-Based Study*. Manchester: St. Jerome.

Koehn, P. (2005): "Europarl: a parallel corpus for statistical machine translation". *The Tenth Machine Translation Summit. Proceedings of Conference* [Phuket, Thailand, September 12-16, 2005]. <www.mt-archive.info/MTS-2005-Koehn.pdf>; 8 pag.

Kruger, H. & van Rooy, B. (2012): "Register and the features of translated language". *Across Languages and Cultures* 13, 1: 33-65.

Lefer, M.-A. (2012): "Word-formation in translated language: The impact of language-pair specific features and genre variation". *Across Languages and Cultures* 13, 2: 145-172.

Olohan, M. (2004): *Introducing Corpora in Translation Studies*. London: Routledge.

Paquot, M. & Bestgen, Y. (2009): "Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction". In: Jucker, A.H. *et alii* (eds.), *Corpora: Pragmatics and Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29)*: Amsterdam: Rodopi; 247-269.

Tirkkonen-Condit, S. (2004): "Unique items: over- or under-represented in translated language?". In: Mauranen, A. & Kujamäki, P. (eds.), *Translation Universals: do They Exist?* Amsterdam & Philadelphia, PI: John Benjamins; 177-184.

Wang, K. & Qin, H. (2010): "A parallel corpus-based study of translational Chinese". In: Xiao, R. (ed.), *Using Corpora in Contrastive and Translation Studies*. Newcastle upon Tyne: Cambridge Scholars; 164-181.