Corpus-based methods for Comparative Translation and Interpreting Studies – mapping differences and similarities with traditional and innovative tools

María Calzada Pérez, Universitat Jaume I, Castellón de la Plana, Spain

Abstract

Comparative Translation and Interpreting Studies ($CTIS_1$) encompass all research processes resulting from the comparison of theories, products, and practices associated with the tasks performed by translators and interpreters during the course of their work. A specific set of comparative methods and tools are derived from Corpus-based Translation and Interpreting Studies ($CTIS_2$). In an attempt to perform $CTIS_1$ from a $CTIS_2$ perspective, this paper employs (modern diachronic) corpus-based ideas (such as priming theory in section 2) and tools (section 3) to build a comparative methodology and analysis with the ECPC archive (section 4). The paper centers around the TIS (very influential) notion of difference and TIS (less productive) concept of similarity with the intention of increasing, in Chesterman's words (Chesterman 2004, 33)), "our understanding of the whole picture."

Keywords

Comparative Translation and Interpreting Studies ($CTIS_1$); Corpus-based Translation and Interpreting Studies ($CTIS_2$), Modern Diachronic Corpus Assisted Discourse Studies (MD-CADS); ECPC; methodology.

1. *CTIS$_1$ and CTIS$_2$*

In a nutshell, Comparative Translation and Interpreting Studies (CTIS$_1$), as a distinct branch within Translation and Interpreting Studies (TIS), would encompass all research methodologies resulting from the comparison of theories, products and practices associated with the task(s) carried out by translators and interpreters in the course of their work, independent of time and space. In both the social sciences and the humanities, "[c]omparison"--according to David Collier (1993, 105)--"is a fundamental tool of analysis. It sharpens our power of description, and plays a central role in concept formation by bringing into focus suggestive similarities and contrasts among cases." Indeed, comparison is pivotal for Translation and Interpreting Studies, to such an extent that it could be argued that a vast terrain within TIS is inherently comparative, certainly since the "paradigmatic change" (Xiao and Yue 2009, 237) brought about by Descriptive Translation Studies (DTS). In 1993, this vast comparative terrain welcomed the then incipient Corpus-based Translation and Interpreting Studies (CTIS$_2$), when Mona Baker (1993, 235) officially announced:

> a turning point in the history of the discipline. I would like to argue that this turning point will come as a direct consequence of access to large corpora of both original and translated texts, and the development of specific methods and tools for interrogating such corpora in ways which are appropriate to the needs of translation scholars.

In that seminal and visionary paper, Mona Baker encourages the creation of translation-related corpora and the development of adequate methods to exploit them. Just a decade later, her words had been heard and her advice had been avidly put into practice. By 2004, Mona Baker (2004, 169) was already noting that CTIS$_2$ “have too much rather than too little to go on.”

At present, over two decades after Mona Baker’s paper, CTIS$_2$ scholars have done (and are doing) their bit. There are certainly many translation-related corpora, which have been extensively reviewed by the literature (e.g., Hu 2016; Laviosa 2002; Olohan 2004). Federico Zanettin (2012, 10) succeeds in capturing this overwhelming proliferation in a particularly clear diagram, reproduced in figure 1:

[Insert Figure 1 here]

To the types represented in the diagram, Zanettin (2012, 11) adds further, illuminating information:

> A bilingual, reciprocal corpus may be graphically represented as a square cut across by diagonal lines, in which the different subcorpora stand at the corners. Multilingual, reciprocal parallel corpora may generate complex models which can be described as a star and diamond configurations (Johansson 2003, 139-142). In a star model there are multiple translations of the same texts in different languages. The diamond model includes source texts in more than two languages as well as their translations in all the other languages.

The existence of “too much” (in Baker’s words above) CTIS$_2$ material has indeed been very fruitful in providing “new ways of looking at translation” (Kenny 1998, 53), which are actually nothing more (or less) than different forms of comparing translation-related texts. According to Alan Partington, Alison Duguid, and Charlotte Taylor (2013, 13) corpus-based comparison may be one of roughly four types: simple, serial, multiple, and diachronic. And CTIS$_2$ has practiced them all. Simple comparisons entail the confrontation of two different subcorpora (like when Moropa 2011 studies a set of texts in English vis-à-vis their Xhosa translations). Serial comparisons involve contrasts between a corpus A and a corpus B, and then between the same corpus A and a corpus C, and so on (like when, for instance, Bosseaux 2006 examines Virginia Woolf's *The Waves* and two of its translations into French). Multiple comparisons occur when a corpus A is set against a pool of subcorpora at once. Partington, Duguid, and Taylor (2013, 13) explain that "those studies which employ the BNC [British National Corpus] or the Bank of English [BoE] as a background or reference corpus are of this multiple-comparison type" (for example, when Dorothy Kenny 2001 double-checks her GEPCOLT results against the BNC, she is performing multiple comparisons). Diachronic comparisons involve the exploration of translation-related corpus throughout time and are still, admittedly, rare (Calzada Pérez 2015, and Calzada Pérez 2016, however, do precisely this with her European Comparable and Parallel Corpus Archive, ECPC).

Thus, CTIS$_2$ comparison(s) has been at the base of much interesting research "allow[ing] us to see both similarities and differences in a perspective that increases our understanding of the whole picture, and also of how this picture relates to other pictures" (Chesterman 2004, 33). In particular, CTIS$_2$ has given a boost to the "search for patterns that identify translation qua translation" (Laviosa 2011, 18). In this sense, two kinds of studies have had a particularly strong impact upon the discipline so far, as Mona Baker (2004) summarizes the inspection of so-called translation universals (TU) and the detection of translators' style. Out of the two, the first has been especially prolific and divisive (Mauranen and Kujamäki 2004). Mona Baker (2004) foresees the development of a third line of study within CTIS$_2$, which has not actually taken hold so far, at least not to the same extent as the analysis of translators' style and certainly not to the same level as TUs: diachronic corpus-based translation studies. With Calzada Pérez (2015) and Calzada Pérez (2016), the present paper aims to contribute to filling this gap.

Hence, in an attempt to perform CTIS_1 from a CTIS_2 perspective, this paper uses (modern diachronic) corpus-based ideas (section 2) and tools (section 3) upon which we build our comparative methodology and analysis with the ECPC corpus (section 4). The paper revolves around TIS (very influential) notion of difference and TIS (less productive) concept of similarity, hoping to increase, in Chesterman's words (see above) "our understanding of the whole picture." We start our proposal then with an idea that was nurtured by corpus studies: the theory of lexical priming.

2. *Lexical Priming: A Corpus-based Theory*

For more than a decade, Michael Hoey has advocated a "radical" (Williams 2006, 327) approach to communication, called "lexical priming." Hoey's theory draws on the psycholinguistic notion of priming (for a review, see Neely 1991), which may be described as the effect of context and prior experience on accessing information from memory (Healy and Proctor 2013, 4:447). For Hoey (2005: 8), "[a]s a word is acquired through encounters with it in speech and writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered." First, he uses priming to approach collocation, which, he argues, remains unaccounted for in other linguistic theories. He notices that if lexis was only a small portion of language that fitted into a syntactic structure (as stated explicitly or implicitly in other theories), then there would be no explanation for collocation, and its pervasive existence would be the result of an enormous and puzzling coincidence. However, real life illustrates that we frequently resort to collocation, which to him (Hoey 2013, 155) implies that, on the one hand, "the brain must be storing language in a manner analogous to (though obviously not identical to) the way a concordance represents language" and, on the other,

> when we encounter language we store it much as we receive it, at least some of the time, and that repeated encounters with a word (or syllable or group of words) in a particular textual and social context, and in association with a particular genre or domain, prime us to associate that word (or syllable or group of words) with that context and that genre or domain. (Hoey 2013, 155)

Hoey uses corpus linguistics to explore priming and collocation, and then exports his methodology to the examination of other linguistic features, such as meaning, grammar, cohesion, sociolinguistics, and so on along similar lines. His method considers priming from two perspectives: (a) that of the primed items ("for example […] all the primings associated with the word *consequence*," Hoey 2005, 14) and (b) that of the relationship among primings ("all the primings that contribute to the production of a sentence," Hoey 2005, 14). However, there is at least a third perspective that Hoey seems to have overlooked: that of the prime per se (such as the word *consequence* in our previous example). In effect, this void results in a limitation in the selection of the prime, which becomes a subjective/random issue (researchers choose whatever prime they intuitively suspect might be of interest) or is bound to specific texts and sentences (researchers use concrete texts on a need basis as cues for prime selection). However, if priming is based on co-communicants' exposure to language, and if this exposure is explored with corpus-based techniques, we could do worse than complement subjectivity and textual specificity (both perfectly informative procedures) with a frequency-based protocol, as we will do in this paper.

Priming is not portrayed as a permanent feature associated with words; on the contrary it “drift[s] in the course of an individual’s lifetime” and is “the engine of language change” (Hoey 2005, 9). It is precisely priming’s drifting capacity that brings Hoey to CTIS$_2$ since “translation is a potential source of drifts” (Hoey 2013, 167). His suggestions are “tentative rather than confident” (Hoey 2013, 153) when analyzing two translations into Portuguese of the first sentence of Bill Bryson’s book *Neither Here nor There*. Hoey (2013) dissects the sentence’s primings from perspectives (a) and (b) and seeks aid from two comparable, monolingual corpora of non-translational English and Portuguese. Identifying primes in Bryson’s ST, Hoey examines the English corpus searching for standard primings; then he focuses on the TT equivalents and queries the Portuguese corpus searching for standard primings; finally, he compares the results. His overall conclusion rings a clear bell in Translation Studies: “The translator has the choice of either preserving the primings of the target language or importing the primings of the source language (or, of course, a mixture of both)” (Hoey 2013, 167). To a certain extent, Hoey arrives at the same conclusion that has been proposed earlier in the field, while enriching it with an understanding of priming.

CTIS$_2$ need not remain content with this conclusion, interesting though it is, but may throw greater light to translational phenomena. However, to achieve this, it must access the span of comparison, exercising all types of corpora (not non-translational, monolingual compilations alone) that now exist (see section 1), and all kinds of CTIS$_2$ tools that are available (not just concordances; see section 3). Of course, this cannot be achieved at once. Nevertheless, every piece of distinct, comparative research, such as Hoey’s, may contribute to enriching the discipline. CTIS$_2$ certainly offers the tools to do so. Laviosa’s work on priming and Anglicisms (2010; 2012; 1014) is an excellent example of enrolling translational corpora to pursue the comparison.

This paper proposes a tentative study inspired by the (modern-diachronic) corpus-based theory of priming that may be tested against future research. In our case, we suggest ways of expanding CTIS$_1$ by focusing on diachronic differences and similarities through the application of traditional and innovative CTIS$_2$ tools. More specifically, we intend the paper to be a means of exploring prime selection. In due course, this study might be complemented with Hoey's (a) and (b) perspectives.

*3. **Corpus-based tools of research***

The most popular, traditional tools employed by corpus-based research derive from an interest in difference across space/genres (variation) and along time (change). Among them: statistical data, as an "indicator of markedness" (Baker 2010, 125); wordlists, as the most basic "point[s] of entry" (Baker 2010, 133) of analysis; keywords as "somewhat more sophisticated" (Baker 2010, 134) means of research; and concordances (with associated information, such as collocates and clusters).

The statistical data corpus linguists use most frequently are:

i) corpus, text, and sentence (average) word length;

ii) standardized type/token ratio (STTR): the ratio of the different words (i.e., types) in the corpus to the its total number of words (tokens). STTR is normally calculated in sets of 1,000 words and then an average is established. It may be used to measure the degree of lexical variety in corpora. Consequently, STTR standard deviation (STTR SD) is a good indication of lexis homegenity/heterogeneity within each corpus.

iii) significance (p-value) figures as measured by log-likelihood or chi-square.

A good source of corpus statistics is the UCREL Log likelihood calculator[1].

Wordlists are lists of corpus words in alphabetical or frequency order, upon which further data (such as keywords) may be generated.

[1] http://ucrel.lancs.ac.uk/llwizard.html

Keywords result from the statistical comparison between the terms (wordlists) of two given (sub)corpora. At present, this comparison is performed upon significance (-p value) tests based on log likelihood or chi square. Gabrielatos and Marchi (2012) offer alternative metrics (e.g., %DIFF) to identify keyness. At any rate, keywords may be calculated by confronting the wordlists of two comparable corpora or by opposing a wordlist to the terms of a reference corpus (such as the BNC or the BoE). Keywords are, then, items of unusual frequency in a given corpus, and relate to its most idiosyncratic features. Hence, they seem to be a source of good, informative, data upon which to identify areas of research interest.

A concordance is a line of words extracted from a corpus under analysis revolving around a node and its immediate context (and further linked to the larger context), as in figure 2 below.

[Insert Figure 2]

Associated with concordances are collocates and clusters. There are many definitions and approaches to the notion of collocation. Space constraints prevent us from reviewing them all in detail here. Suffice it to say, that, generally, "a collocation is a co-occurrence pattern that exists between two items that frequently occur in proximity to one another--but not necessarily adjacently or, indeed, in any fixed order" (McEnery and Hardie 2012, 123). Collocations may be grounded on statistics (in which case the measure used to establish the collocation is determinant; Mutual Influence, or MI, is commonly used) or researchers' intuition (in which case it depends on the manual scanning of concordances). Turning to clusters (otherwise known as lexical bundles), they are "sequences of word forms that commonly go together in natural discourse" (Biber et al. 1999, 992). When dealing with clusters, specialists apply "cut-off points for lexical bundles which 'count'" (Kopaczyk 2012, 86). These are threshold levels below which (quantitative) relevance is seen as diluted. In this paper, we adopt the threshold advocated by Biber, Conrad, and Reppen (1998): only clusters with above 40 occurrences per million words are seen as quantitatively reliable.

Recent corpus-based research (especially, but not only, associated with modern-diachronic studies) has developed a particularly strong interest in similarity, due to the fact that, according to Taylor (2013, 83–84), "by focusing on difference we effectively create a 'blind spot'" that can be overcome "by 'looking in both directions'." Out of this interest, innovative tools and concepts have been created (Baker 2011; Gabrielatos and Baker 2008; Partington, Duguid; and Taylor 2013). By innovative, we specifically mean that these present researchers with relatively new protocols to produce sophisticated kinds of comparative work. Among them: statistically grounded detailed consistency; Scott's (1997:233) key keywords and their associates; or Paul Baker's (2011:66) lockwords, c-collocates and s-collocates (see below).

Statistically grounded detailed consistency is easily calculated with WordSmith Tools 6.0 (WST), one of the most popular, complete, and user-friendly concordancer in the market. Measured with the Dice Coefficient, detailed consistency is "based on the joint frequency … and the word-counts of the [various] texts" (WordSmith Tools online manual)[2]. In other words, detailed consistency does not only tell us the percentage of word overlap covered by different corpora. It also lists the specific terms (words or keywords, depending on the methodology) that overlap in the corpora. Scott (2009) shows how to use it to test the comparability and adequacy of different reference corpora. Calzada Pérez, 2016, merges detailed consistency and keywords in search of European Parliament's idiosyncratic features.

[2] http://www.lexically.net/downloads/version5/HTML/?detailed_consistency_relations.htm

Scott's (1997, 233) key keywords are “words that are key in many texts.” In order to calculate them, you need to generate traditional keywords first and then use batch and database procedures in WST. Bachmann (2011, 87) uses key keywords (appearing in at least half of the texts under analysis) in order to detect “a list of concepts that are representative of the [British Parliament] debates as a whole.” Key keywords have associates (i.e., “words that are key in the same texts as a given key keyword”). Calzada Pérez (2015) briefly explores key keywords and associates.

Paul Baker (2011, 66) defines a lockword as “a word which may change in its meaning or context of usage when we compare a set of diachronic corpora together, yet appears to be relatively static in terms of frequency.” When words are used consistently throughout time, they are good candidates to be considered lockwords. Baker (2011) developed and applied the notion when comparing various corpora from the Brown family, searching for vocabulary change and stability over time. Notice that, when Baker refers to lockwords, he bases this concept upon wordlists. However, it seems logical to argue that the same methodology could be applied to keywords lists, thus obtaining ‘lock keywords’ which, replicating Baker’s definition, could be considered keywords “which may change in its meaning or context of usage when we compare a set of diachronic corpora together, yet appears to be relatively static in terms of frequency.” To the best of our knowledge, lock keywords have not been examined by TIS literature so far. We will do so in this paper.

Finally, c-collocates, or “those [collocates] that would point toward consistent and, arguably, core aspects of the linguistic means” (Gabrielatos and Baker 2008, 11), are coined in order to filter s-collocates (or seasonal collocates). Here are Gabrielatos and Baker (2008, 11–12) explaining their notion and procedure of use:

> To this end, collocations of RASIM were calculated for every annual subcorpus and tabulated alphabetically per year. In order for a collocate to be deemed consistent, we stipulated that it had to be a collocate in at least seven out of the ten annual subcorpora. Only a small percentage of the collocates of RASIM were consistent (5.4 percent, on average), which seems to confirm that the vast majority of collocates were seasonal—that is, related to specific events.

As briefly mentioned above, these (modern diachronic) tools are relatively recent additions to corpus linguistics. The use of expressions such as "relatively static" or "at least seven out of ten annual subcorpora" (above) reveal that their creators (established researchers such as Baker or Gabrielatos) are still experimenting with their methods. And yet, there is no doubt that they provide a different (very informative) gateway to the data, especially when it comes to diachronic comparisons, aiming at similarity and difference. These tools are largely unknown within the CTIS$_2$ community and, we believe, worth exploring within this field, as a first step to potential further studies.

4. *The study*

4.1 *Aims and specific tools chosen for the analysis*

The main aim of this study is to import some of the corpus-assisted tools described in section 3 in order to enlarge CTIS$_2$ as a means of practicing CTIS$_1$. In this case, the focus of the study is on prime selection and may be complemented, in the future, with a close analysis of primed items and priming relations, the way Hoey suggests in section 2 of the present paper. Since it is not possible to use all of these tools in a single paper and since all of them are useful for comparative studies (which is why we summarized them above, to present them to the CTIS$_1$ community), we will choose some of them with a view to drawing some tentative conclusions regarding prime selection and its consequences for translation.

Hence, the present study specifically relies on statistics, wordlists, keywords, detailed consistency, lock keywords, and clusters. Notice that we select a combination of traditional tools (i.e., statistics, wordlists, keywords and clusters) and innovative tools (i.e., detailed consistency and lock keywords). Notice also that, in the latter case, when faced with experimenting decisions, we opt for conservative solutions to strengthen the rigor of our protocols to the maximum (as will be seen below). All of these tools provide useful information that help us specify points of departure for further studies on priming that go beyond the subjective or textual rationale.

4.2 *The ECPC Archive*

The European Comparable and Parallel Corpora of Parliamentary Speeches Archive (ECPC), compiled at the Universitat Jaume I (Spain) (and available to the academic community at http://ecpc.xtrad.uji.es/glossa/html/index.php?corpus=ecpc), is a collection of XML metatextually tagged corpora containing speeches from three European chambers: the European Parliament (EP), the British House of Commons (HC), and the Spanish Congreso de los Diputados (CD). It is one of those (square-shaped), bilingual, bidirectional corpora described by Zanettin (see section 1). ECPC has a Tree-Tagger POS-annotated version and a 'clean' version (with XML but no POS). HC and CD in full contain speeches from 2004 to 2014 while EP in full incorporates day sessions ranging from 15 April 1996 to 25 June 2011 (the date when the EP stopped producing multi-lingual versions of its debates). The present paper makes use of the clean version of subcorpora from the HC and the EP.

The HC subcorpus used here (with a total of 57,502,472 tokens) is made up of day session proceedings from 2005 to 2010, as published in *The Hansard*[3]. The EP subcorpus used for this paper (23,429,149 tokens), named EP_en, is the official English version of EP proceedings also from 2005 to 2010, as published in the *Official Journal of the European Union* [4]. EP_en has original speeches in English (EP_en_ST: 5,645,433 tokens) and translated speeches into English from the totality of the other EU languages (EP_en_TT: 17,783,715 tokens).

Finally, it is worth underscoring that ECPC material consists of parliamentary proceedings, which diverge to a greater or lesser extent from proper speeches. The present study researches into this particular type of language to which MPs and MEPs are exposed.

4.3 *Outline of the methodology*

Using WordSmith Tools 6.0, two sets of comparative analyses are performed below: (a) statistics analyses and (b) lock keyword analyses. These tools and mechanism point at interest areas for the study of priming. Here are the steps followed for each of them.

a) Statistics analyses

Step 1: Statistics for EP_en_ST, EP_en_TT, and HC are generated, confronted and analysed. Notice that statistics are calculated upon whole subcorpora (unlike what happens with steps 2 to 6, where data are generated in individual year subcorpora).

b) Lock keyword analyses

Step 2: Wordlists for each single year between 2005 to 2010 are produced for EP_en_ST, EP_en_TT, and HC. By doing this, we obtain, for example, wordlists

[3] http://www.parliament.uk/business/publications/hansard/commons/

[4] http://www.europarl.europa.eu/plenary/en/debates-video.html

EP_en_ST_05, EP_en_ST_06, EP_en_ST_07, EP_en_ST_08, EP_en_ST_09, EP_en_ST_10. We repeat the procedure with the other two subcorpora.

Step 3: Keyword lists are obtained by confronting previous wordlists for each year between EP_en_ST, EP_en_TT, and HC. Step 3, results in, for example, keyword lists EP_en_ST_vs_TT_05 and EP_en_TT_vs_ST_05. We repeat this procedure with all years and all subcorpora. Thus, we obtain idiosyncratic words for each subcorpora (in each year) vis-à-vis the rest of subcorpora (in the same year). EP_en_ST_vs_TT_05 produces idiosyncratic features of 2005's EP_en_ST. EP_en_TT_vs_ST_05 produces idiosyncratic features of 2005's EP_en_TT.

Step 4: We apply detailed consistency tests to previous keywords lists. As stated above, notice that this procedure is totally innovative since Paul Baker performs his detailed consistency studies upon wordlists. In step 4, for example, we obtain detailed consistency of keywords of EP_en_ST_vs_TT (for years 2005, 2006, 2007, 2008, 2009, 2010). We repeat this procedure with the rest of keyword lists. Thus, we spot those keywords that are most persistent throughout time (i.e., lock keywords). To guarantee steady appearance of keywords over the years, only keywords appearing in at least 4 (out of the 6 six years analysed) are kept for analysis; the rest are discarded. This is a conservative decision within an experimenting protocol.

Step 5: Common lock keywords for EP_en_TT_vs_ST and EP_en_TT_vs_HC are chosen as our point of entry into the data. By doing this, the most stable (idiosyncratic) lock keywords of EP_en_TT are identified (for years 2005-2010). This is a conservative decision within an experimenting protocol.

Step 6: Analyses are performed with these and comparable data. Clusters are employed to aid in the analyses.

4.4 *Stats analysis*

Figure 3 includes basic statistics for subcorpora EP_en_ST, EP_en_TT, and HC during the period between 2005-2010.

[Insert Figure 3 here]

A first set of data concerns token numbers. As we see from the table, and focusing on EP data, there are 3.1 times more translated (EP_en_TT) English language than original (EP_en_ST) English tokens. Some immediate conclusions may be drawn from these data.

First, from 2005 to 2010, (English speaking) MEPs were exposed to an awful lot of translated (vs original) English, which was likely to have some effect upon their linguistic primings. Indeed, we cannot guarantee that MEPs consulted *Official Journal* translations during the period under research (notice they may have followed debates through interpreting only, which is different from proceeding translations). However, it would be logical to argue that more often than not they would have resorted to this written material in order to prepare their work as MEPs.

Second, the number of EP original English tokens (5,645,433) from 2005 to 2010 is significantly greater than the average number of original tokens for each of the 24 EP official languages (976,214,541 tokens). See statistics obtained with the UCREL Log likelihood calculator portrayed by figure 4:

[Insert Figure 4 here]

These results roughly coincide with studies published by *The Guardian* on 21 May, 2014 (for a slightly different period of time, from 2008 to 2012). This shows that MEPs are prone to EP_en_ST primings to a higher extent than to other kinds of original primings.

In the meantime, from 2005 to 2010, HC MPs were clearly exposed to the impressive volume of English tokens exchanged at the British Chamber (57,502,472 tokens), which was much larger than its equivalent at the EP (23,429,149 tokens, regardless of original or translated items). It seems suitable to argue that this enormous volume of linguistic material is also prone to have an effect on British MPs' primings.

Apart from token numbers, figure 3 contains other types of statistics, which serve to partly describe the different sorts of language exposure at the EP and the HC.

Regarding lexis, for example, figures for STTR and STTR SD suggest that, from 2005 to 2010, there was a descending trend in lexis variety and lexis homogeneity from EP translated English (STTR: 42.36; STTR SD: 57.92) through EP original English (STTR: 41.04; STTR SD: 59.06) to HC English (STTR: 38.53; STTR SD: 61.35). This means that, overall, in our subcorpora, the nature of lexical exposure (as a source of lexical priming) stands along a continuum. The more interlinguistic the setting (i.e., EP_en_TT), the more lexically varied and homogenous the exposure. The less multilingual the setting (HC), the less lexically varied and homogeneous the exposure. Notice that figures show that exposure (primings) is more similar between EP original and translated English than between (original) English from the HC and the EP. This points at the existence of euro jargon at the lexical level.

Regarding syntax, and focusing on sentence length and sentence length homogeneity, there is a descending pattern of sentence length from EP translated English (length: 26.73) through EP original English (length: 25.34) to HC (length: 21.43). There is also an ascending pattern of sentence length homogeneity from EP translated English (length SD: 16.06) through EP original English (length SD: 16.05) to HC (length SD: 14.23). This means that, overall, in our subcorpora, the nature of sentence exposure (as a source of priming) also stands along a continuum. The more interlingustic the setting (i.e., EP_en_TT), the greater the exposure to longer, less homogenous sentences. The less multilingual the setting (HC), the greater the exposure to shorter, more homogenous sentences. Notice that exposure (primings) are much more similar between EP original and translated English than between (original) English from the HC and the EP. This points at the existence of euro jargon at the syntactic (sentence length) level.

4.5 *Lock keyword analysis*

By implementing steps 2 to 6 with the tools described above, we aim at identifying distinctive language exposure (as a source of primings or potential primes) in three of the subcorpora under scrutiny (EP_en_ST, EP_en_TT, and HC) throughout time (in years spreading from 2005 to 2010).

Notice that, in this section (and as explained above), we move from wordlists to keywords to lock keywords. Wordlists include all linguistic terms present in the subcorpora. Keywords bring to the fore different kinds of idiosyncratic language exposure (potential primes) between and among the subcorpora. However, it is lock keywords (persisting keywords throughout time), that lead us to an analysis of stable (most characteristic) exposure (potential primes), the kind of exposure that may be of interest to draw some conclusions on priming. There are many studies that may be performed upon lock key pronouns, key adjectives, key nouns, key verbs and other key parts of speech. Due to space constraints, in this paper we focus on those lock keywords that are statistically more prominent in each of the subcorpora against the other two (from 2005 to 2010). In other words, first, we focus on those items in EP_en_TT that appear as key in at least four (out of the six) years that make up the subcorpus, when compared both against EP_en_ST and against HC. We do the same with 2005-2010's key items in EP_en_ST (vs EP_en_TT and HC) and with 2005-2010's key items in HC (vs EP_en_ST and EP_en_TT). This is a conservative decision within an experimenting protocol. Figure 5 shows these lock keywords:

[Insert Figure 5]

All of these results seem worth commenting on and point at differences of linguistic exposure (potential primes) regarding cohesion (‘of’, ‘which’, ‘although’, ‘who’, ‘about’, ‘that’) and ‘appraisal’ (as per Martin and White 2005 and Munday 2012), the latter expressed here through adverbial intensifiers (‘very’, ‘fully’) and modality (‘must’, ‘may’, ‘might’). Due to space constraints, we hereby focus on appraisal and reserve cohesive markers for future work.

According to Martin and White (2005, 35), appraisal is “one of the three major discourse semantic resources construing interpersonal meaning” (the other two being negotiation and involvement). It may be subdivided into three basic domains: attitude, engagement and graduation. Attitude conveys: (a) feelings and emotional reactions (affect), (b) assessment of behaviour (judgement) in the form of social esteem and social sanction and (c) construction of the value of things (appreciation). Engagement refers to the intersubjective positioning or “the ways in which resources such as projection, modality, polarity, concession and various comment adverbials position the speaker/writer with respect to that value position” (Martin and White 2005, 36); engagement may be ‘monoglossic’ and ‘heteroglossic’. Finally, graduation adjusts the intensity or clarity of the evaluation. In the former case (of intensity), ‘force’ is measured along a ‘raise’/‘lower’ continuum; in the latter (of clarity), ‘focus’ is weighed along a ‘sharpen’/‘soften’ gradation. According to Munday (2012, 2) appraisal resources are “critical” for translators and, drawing on Agar (1991, in Munday 2012, 2), they may constitute “rich points”, “defined as ‘locations in discourse where major cultural differences are signalled’.”

Adverbial intensifiers are specifically used for graduation purposes of force. Two of these (‘fully’ and ‘very’) are lock keywords in the EP_en_ST (vs both

EP_en_TT and HC). This means that the use of these intensifiers is a particularly idiosyncratic feature/potential prime of original English at the EP (vs EP translated or HC English). By generating 2-3 word clusters above threshold level (see section 3), we obtain figure 6, reflecting the way in which ‘fully’ is used. Notice it is particularly employed to intensify convictions on the part of the ‘I’ speaker.

[Insert Figure 6]

By generating 3-4 word clusters above threshold level (see section 3), we obtain figure 7, reflecting the way in which ‘very’ is used. Notice it is particularly employed to intensify appreciation in the third (e.g., “It is very important”) or first person singular (e.g., “I very much regret”).

[Insert Figure 7]

Modality markers, such as lock keywords in figure 5 (‘must’ at the EP_en_TT; ‘might’ and ‘may’ at the HC), are appraisal resources of judgement and (more or less ‘monoglossic’/‘heteroglossic’) engagement, of dissimilar (‘raise’/‘lower’) force. According to the grammars, ‘must’ is used in a deontic manner in order to express “strong obligation” (Swan 2010, 327) or even “prohibition” (Swan 2010, 328) when in the negative. It is also used in an epistemic sense, to convey “complete certainty (positive or negative)” (Swan 2010, 327). In this sense, ‘must’ is especially ‘monoglossic’ because the speaker does not allow much margin for receivers to interact with the texts. Our results suggest that, from 2005 to 2010, translation of EP speeches subjected their receivers to significantly high doses of ‘raise’ (i.e., strong), ‘monoglossic’ modality of either a deontic or epistemic kind. Since this use is significantly more prominent in EP translated English than in both EP original and HC English, it might be argued that this is a particularly idiosyncratic feature of translated speech at the EP (vs the other two kinds of Englishes). This heavy exposure to ‘raise’ ‘monoglossic’, modality by EP translated speeches seems to contravene the current behaviour of original English (at least in the UK), where according to Paul Baker (2010, 66) strong modality is in clear decline, something which (drawing on Leech, in Baker, 2010: 66), the scholar attributes to “a number of trends in English, including democratization,” or to “Fairclough’s concepts of personalization (1989) and conversationalization of public discourse (1994)” (Baker 2010: 67), which were often initiated in American English.

In comparison to what happens with EP translated speeches, figure 5 shows that HC English has, among its lock keywords, weaker markers of rather more ‘heteroglossic’ deontic and epistemic modality (i.e., ‘may’ and ‘might’), which allow for greater acknowledgement of the receiver. In fact, as discussed in the grammars, ‘may’ is used to convey “possibility” (Swan 2010, 327) and ‘might’ serves to utter “weak possibility” (Swan 2010, 327) and “weak obligation” (Swan 2010, 328). This ‘lower’, more ‘heteroglossic’ use of modality can be seen as a particularly characteristic feature/potential prime of HC English (vs EP English).

For its part, original EP English does not expose its receivers to any lock key modal-verb at a more statistically significant level than both HC and EP translated English (see figure 5). We cannot, thus, identify EP_en_ST idiosyncratic features/potential primes regarding modality. What we could do, however, is compare original lock key modals with the EP_en_TT and the HC separately, to detect nuances in the use of appraisal from communicative situation to communicative situation.

Hence, if we, first, look at those EP_en_ST lock keywords that result from a comparison against EP_en_TT only, we notice that original speeches use ‘have’ as its most frequent (45,388 tokens) stable key verb from 2005 to 2010. Clusters around ‘have’ above 40 hits per million (see section 3 for Biber, Conrad, and Reppen 1998’s cluster threshold level) provide more information about the use of ‘have’ in EP_en_ST (see figure 8), showing that the most frequent cluster around ‘have’ is ‘we have to’ (2500 hits).

[Insert Figure 8]

In fact, there is a total of 5,524 of ‘have to’ (with a deontic sense) in EP_en_ST (12.17% of total use of ‘have’), suggesting that, in EP_en_ST English, there is a partial replacement of ‘must’ with ‘have to’. Grammars record the main difference between these two modality markers.

> In statements about obligation with *must* the obligation most often comes from the speaker (and in questions, from the hearer). To talk about an obligation that comes from 'outside' (for instance a regulation, or an order from somebody else), we usually prefer *have to* […] Have to can also be used to talk about obligation coming from the speaker or hearer, in the same way as must. This is normal in American English (which uses must less often in this sense), and is becoming very common in British English. (Swan 2010, 336) (Emphasis added).

Hence, whereas translated English at the EP builds judgments upon a particularly ‘raise’, ‘monoglossic’ item, original English softens the stance with the addition of ‘have to’, a rather more ‘heteroglossic’ expression that either takes the outside world into consideration or approaches the HC trend of abandoning ‘must’ (or both). Figure 9 provides all clusters with ‘have to’ (above threshold level) in original English at the EP. Notice that here ‘have to’ is used in connection with the first person plural (e.g., “we have to agree”, “we have to take”) except from when it is used together with verb ‘say’, when it is associated with the first person singular (i.e., “I have to say”)

[Insert figure 9]

If we now compare lock key modals in EP_en_ST and HC, we find ‘must’ as particularly frequent in EP original English. Figure 10 shows clusters above threshold level around ‘must’. Notice that here ‘must’ is either used within impersonalized expressions (e.g., “it must be said”) or in connection with ‘we’ (e.g., “we must take”).

[Insert figure 10]

Interestingly and contrary to what happens with ‘have to’, ‘must’ is now impersonalized when used in connection with ‘say’.

In short, the three corpora studied here may be placed along a ‘monoglossic’/‘heteroglossic’ and ‘raise’/‘lower’ judgement continuum, where HC English opts for the most ‘heteroglossic’ and ‘lower’ kind of modality and translated EP English stands at the other side, with the most ‘monoglossic’, ‘raise’ type. Original English from the European Parliament stands in the middle (like a hinge) with greater doses of monoglossia and force than those from the HC but ‘lower’, more ‘heteroglossic’ modality than translated English, with the addition of ‘have to’.

5. Conclusions

We began this research with the primary aim of enlarging and strengthening CTIS$_1$ via CTIS$_2$. We believe we have accomplished this purpose by proposing and exemplifying the use of traditional and innovative corpus-based tools, some of which have rarely been used before within TIS (Calzada Pérez 2015 and Calzada Pérez 2016). Nevertheless, it should be noted that a portion of CTIS$_2$ strength lies in the fact that it relies heavily on ample, flexible comparisons (see section 1). The use of all of its comparative potential provides new avenues for CTIS$_1$ beyond the adequacy/acceptability debate.

This paper tentatively performs specific forms of serial, diachronic, comparative analyses on the ECPC subcorpora, targeting similarities (along time) and differences (across communicative situations), with the specific goal of identifying new methods of performing prime selection. Through statistics, lock keywords, and clusters these forms of comparison unveil “rich points” (Agar in section 4.5) regarding: (a) the homogeneity and variety of original and translated English in monolingual and multilingual settings; (b) the existence of Euro-jargon; and, most remarkably, (c) the exposure of MPs and MEPs to different types of modality and their potential effects on (or reflections of) dissimilar processes of democratization and Americanization (such as personalization or conversationalization).

Comparison does not end here. Further studies may confront proceedings against speeches, translations against interpretations, or the parliamentary genre against its reflection by the media, to state but three examples. The comparative goals pursued here are not exhausted by prime selection. If nothing else, CTIS$_2$ tools may also contribute to examining priming from Hoey’s (a) and (b) perspectives (see section 2 above). What is apparent though is that the more comparison CTIS$_2$ practices, the more varied and nuanced are the results for CTIS$_2$, in particular, and CTIS$_1$, in general.

Cited references

Bachmann, Ingo. 2011. “Civil Partnership - ‘Gay Marriage in All but Name’: A Corpus-Driven Analysis of Discourses of Same-Sex Relationships in the UK Parliament.” *Corpora* 6 (1): 77–105.

Baker, Mona. 1993. “Corpus Linguistics and Translation Studies — Implications and Applications.” In *Text and Technology*, ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 233–50. Amsterdam/Philadelphia: John Benjamins.

———. 2004. “A Corpus-Based View of Similarity and Difference in Translation.” *International Journal of Corpus Linguistics* 9 (2): 167–93.

Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh Sociolinguistics. Edinburgh: Edinburgh University Press.

———. 2011. “Times May Change, But We Will Always Have Money: Diachronic Variation in Recent British English.” *Journal of English Linguistics* 39 (1): 65–88.

Biber, Douglas, Conrad, Susan, and Reppen, Randi. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Conrad, Susan, and Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. Harlow, England; New York: Pearson Education Limited.

Bosseaux, Charlotte. 2006. "Who's Afraid of Virginia's you: a Corpus-based Study of the French Translations of The Waves." *Meta: Journal des traducteurs* 51 (3): 599–610.

Calzada Pérez, María. 2015. "Peeping into Europe's Liquidity Through Cads and Md-Cads." (Paper presented at *5th International Association for Translators and Intercultural Studies Conference*, Belo Horizonte, Brazil, July 6-10, 2015).

Calzada Pérez, María. 2016. "Corpus-assisted discourse studies at the European Parliament." (Paper presented at Languaging Diversity. *3rd International Conference. Language(s) and Power,* Macerata, Italy, 3-5 March, 2016).

Chesterman, Andrew. 2004. "Beyond the Particular." In *Translation Universals Do They Exist?,* ed. by Anna Mauranen and Pekka Kujamäki, 33–50. Amsterdam/Philadelphia: John Benjamins.

Collier, David. 1993. "The Comparative Method." In *Political Science: The State of the Discipline II,* ed. by Ada W. Finifter, 105–19. Washington, D.C.: American Political Science Association.

Gabrielatos, Costas, and Paul Baker. 2008. "Fleeing, Sneaking, Flooding. A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005." *Journal of English Linguistics* 36 (1): 5–38.

Gabrielatos, Costas, and Anna Marchi. 2012. "Keyness. Appropriate Metrics and Practical Issues." (Paper presented at *CADS International Conference. Corpus-assisted Discourse Studies: More than the sum of Discourse Analysis and computing?,* Università di Bologna, Italy, 13-14 September, 2012). Accessed March 14, 2016. https://repository.edgehill.ac.uk/4196/

Hoey, Michael. 2005. *Lexical Priming: A New Theory of Words and Language*. London; New York: Routledge.

———. 2013. "Lexical Priming and Translation." In *Corpus-Based Translation Studies: Research and Applications,* ed. by Alet Kruger, Kim Wallmach, and Jeremy Munday, 153–68. Continuum Advances in Translation Studies. London; New York: Continuum.

Hu, Kaibao. 2016. *Introducing Corpus-Based Translation Studies*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Kenny, Dorothy. 1998. "Corpora." In *Routledge Encyclopedia of Translation Studies,* ed. by Mona Baker, 50-53. London and New York: Routledge.

———. 2001. *Lexis and Creativity in Translation: A Corpus-Based Study*. Manchester, UK; Northampton MA: St. Jerome Pub.

Kopaczyk, Joanna. 2012. "Applications of the Lexical Bundles Method in Historical Corpus Research." In *Corpus Data across Languages and Disciplines*, ed. by Piotr Cap, 83–95. Lodz Studies in Language 28. Krankfurt Am Main: Peter Lang.

Laviosa, Sara. 2002. *Corpus-Based Translation Studies: Theory, Findings, Applications*. Approaches to Translation Studies 17. Amsterdam: Rodopi.

———. 2010."Corpus-Based Translation Studies 15 Years on: Theory, Findings, Applications". *SYNAPS* 24: 3-12.

———. 2011. “Corpus-Based Translation Studies: Where Does It Come From? Where Is It Going?” In *Corpus-Based Translation Studies: Research and Applications*, ed. by Alet Kruger, Kim Wallmach, and Jeremy Munday, 13–32. Continuum Advances in Translation Studies. London; New York: Continuum.

———. 2012.“Divergent and Convergent Similarity in Corpus Translation Studies”. In *Translations-Wissenschaftliches Kolloquium II*, ed. by Barbara Ahrens, Silvia Hansen-Schirra, Monika Krein-Kühle, Michael Schreiber, Ursula Wienen, 295-305. Bern: Peter Lang.

———. 2014. “Towards the Study of Drifts in the Priming of Anglicisms in Business Communication”. In *Identity in and across Cultures*, ed. by Paola Evangelisti Allori, 185-208. Bern: Peter Lang.

Martin, James R., and Peter R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. New York: Palgrave Macmillan.

Mauranen, Anna, and Pekka Kujamäki, eds. 2004. *Translation Universals. Do they exist?* Amsterdam/Philadelphia: John Benjamins.

McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge; New York: Cambridge University Press.

Moropa, Koliswa. 2011. “"A Link Between Simplification and Explicitation in English-Xhosa Parallel Texts: Do the Morphological Complexities of Xhosa Have and Influence?” In *Corpus-Based Translation Studies: Research and Applications*, ed. by Alet Kruger, Kim Wallmach, and Jeremy Munday, 259–81. London; New York: Continuum.

Munday, Jeremy. 2012. *Evaluation in Translation: Critical Points of Translator Decision-Making*. Milton Park, Abingdon, Oxon; New York: Routledge.

Neely James H. 1991. “Semantic Priming Effects in Visual Word Recognition: A Selective Review of Current Findings and Theories.” In *Basic Process in Reading: Visual Word Recognition*, ed. by Derek Besner and Glyn W. Humphreys, 264–336. Hillsdale: Erlbaum.

Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies*. London; New York: Routledge.

Partington, Alan, Alison Duguid, and Charlotte Taylor, eds. 2013. *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam/Philadelphia: John Benjamins.

Scott, Mike. 1997. “PC Analysis of Key Words - and Key Key Words.” *System* 25 (2): 233–45.

———. 2009. “In Search of a Bad Reference Corpus.” In *What’s in a Word-List?: Investigating Word Frequency and Keyword Extraction*, ed. by Dawn Archer. Ashgate Publishing, Ltd.

Swan, Michael. 2010. *Practical English Usage: [easier, Faster Reference]*. 3. ed., [fully rev.], [Nachdr.]. Oxford: Oxford Univ. Press.

Taylor, Charlotte. 2013. “Searching for Similarity Using Corpus-Assisted Discourse Studies.” *Corpora* 8 (1): 81–113.

Williams, G. 2006. “Michael Hoey. Lexical Priming: A New Theory of Words and Language. London: Routledge. 2005. Xiii + 202 Pages. ISBN 0-415-32863-2.” *International Journal of Lexicography* 19 (3): 327–35.

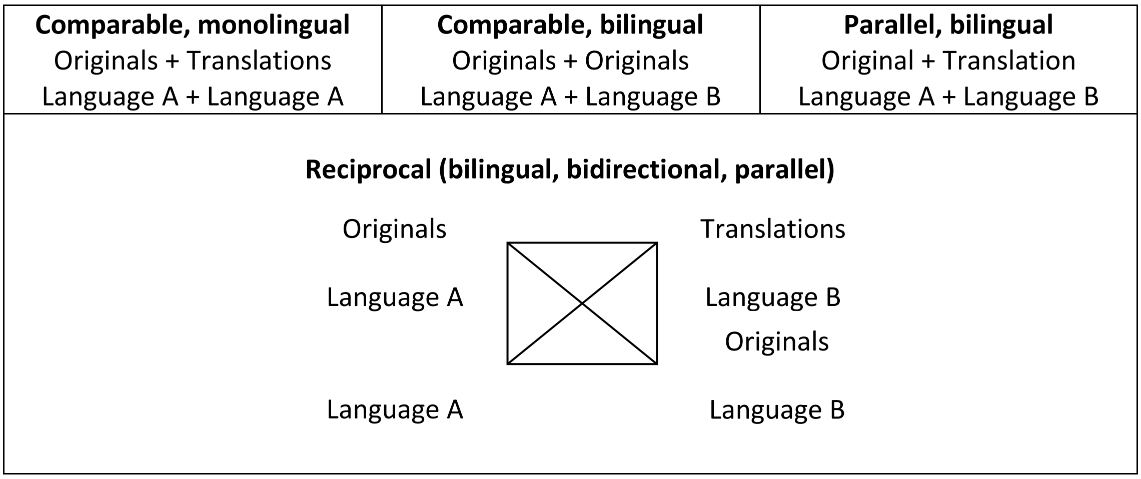
WordSmith Tools. 2015. “Word Smith Tools Manual. Software for Finding Word Patterns.” Accessed March 14, 2016. http://www.lexically.net/downloads/version6/HTML/index.html?getting_started.htm.

Xiao, Richard, and Ming Yue. 2009. “Using Corpora in Translation Studies: The State of the Art.” In *Contemporary Corpus Linguistics*, ed. by Paul Baker, Paperback ed, 237–61. Continuum Studies in Linguistics. London; New York, NY: Continuum.

Zanettin, Federico. 2012. *Translation-Driven Corpora Corpus Resources for Descriptive and Applied Translation Studies*. Translation Practices Explained. Manchester, UK; Kinderhook, NY: St. Jerome Pub.

Figures

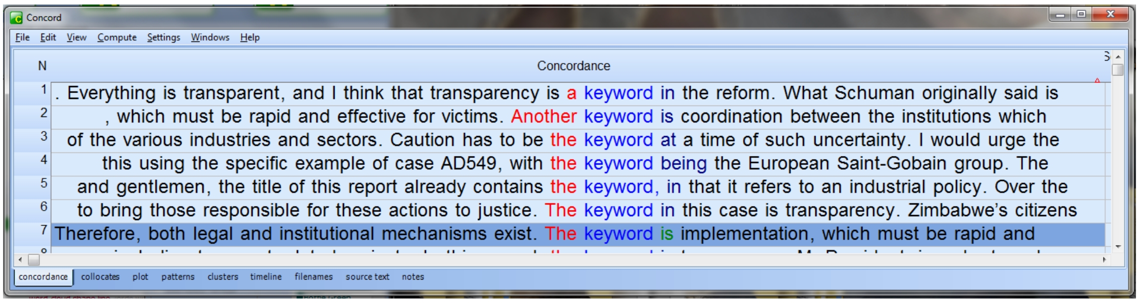
Figure 1: Zanettin’s (2011) typology of translation-related corpora.



Comparable, monolingual Originals + Translations Language A + Language A	**Comparable, bilingual** Originals + Originals Language A + Language B	**Parallel, bilingual** Original + Translation Language A + Language B

Reciprocal (bilingual, bidirectional, parallel)

Figure 2: Concordances.



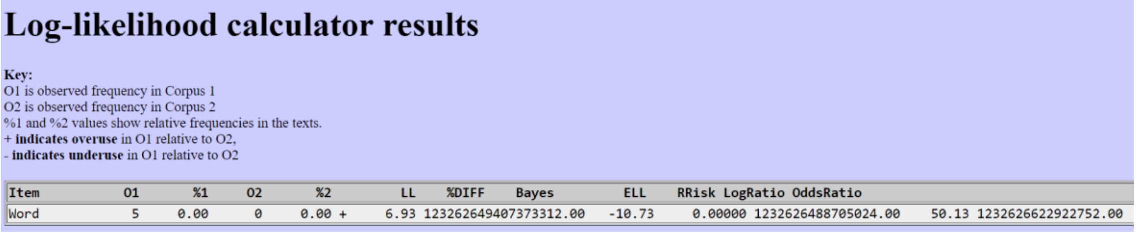
Concord
File Edit View Compute Settings Windows Help
N
Concordance
concordance collocates plot patterns clusters timeline filenames source text notes

1 . Everything is transparent, and I think that transparency is a keyword in the reform. What Schuman originally said is
2 , which must be rapid and effective for victims. Another keyword is coordination between the institutions which
3 of the various industries and sectors. Caution has to be the keyword at a time of such uncertainty. I would urge the
4 this using the specific example of case AD549, with the keyword being the European Saint-Gobain group. The
5 and gentlemen, the title of this report already contains the keyword, in that it refers to an industrial policy. Over the
6 to bring those responsible for these actions to justice. The keyword in this case is transparency. Zimbabwe's citizens
7 Therefore, both legal and institutional mechanisms exist. The keyword is implementation, which must be rapid and

Figure 3: Statistics of HC and EP subcorpora

CORPUS	TOKENS	TYPES	STTR	STTR SD	Sentence	Sentence length	SD
EP_EN_ST	5,645,433	42,656	41.06	59.06	221,171	25.34	16.05
EP_EN_TT	17,783,716	66,373	42.36	57.92	656,846	26.73	16.06
EP_EN_ES	1,162,528	20,197	39.92	60.59	37,092	31.16	20.55
HC	57,502,472	158,272	38.53	61.35	2,615,173	21.43	14.23

Figure 4: Effect size and log likelihood stats for English original vis-à-vis average number of original tokens for the 24 EP official languages.



Log-likelihood calculator results

Key:
O1 is observed frequency in Corpus 1
O2 is observed frequency in Corpus 2
%1 and %2 values show relative frequencies in the texts.
+ **indicates overuse** in O1 relative to O2,
- **indicates underuse** in O1 relative to O2

```
Item         O1       %1      O2        %2          LL      %DIFF     Bayes        ELL     RRisk LogRatio OddsRatio
Word          5     0.00       0     0.00 +       6.93 123262649407373312.00    -10.73     0.00000 1232626488705024.00    50.13 1232626622922752.00
```

Figure 5: Lock keywords.

EP_en_TT lockwords (vs. EP_en_ST and HC)	EP_en_ST lockwords (vs. EP_en_TT and HC)	HC lockwords (vs. EP_en_TT and EP_en_ST).
Must Of which	Very fully	May Might Although Who About That

Figure 6: EP_en_ST clusters around 'fully'.

Cluster	Freq.	Related
I FULLY	501	I FULLY SUPPORT (199),I FULLY AGREE (134),EN I FULLY (60),I FULLY SHARE (54),AND I FULLY (32),PRESIDENT I FULLY (29),I FULLY UNDERSTAND (24),WHICH I FULLY (22),THAT I FULLY (21),I FULLY ENDORSE (16),I FULLY ACCEPT (14),I FULLY RECOGNISE (13)
FULLY SUPPORT	315	I FULLY SUPPORT (199),FULLY SUPPORT THE (171),WE FULLY SUPPORT (44),FULLY SUPPORTS THE (43),COMMISSION FULLY SUPPORTS (31),FULLY SUPPORT THIS (29),GROUP FULLY SUPPORTS (15),FULLY SUPPORTIVE OF (11),WILL FULLY SUPPORT (10), ARE FULLY SUPPORTIVE (10)
THE COMMISSION	244	THE COMMISSION FULLY (91),THE COMMISSION IS (48),THAT THE COMMISSION (30),AND THE COMMISSION (17),THE COMMISSION CAN (15)

Figure 7: EP_en_ST clusters around 'very'.

Cluster	Freq.	Related
IS A VERY	841	IS A VERY IMPORTANT (273),THIS IS A VERY (240),IT IS A VERY (163),IS A VERY GOOD (89),THAT IS A VERY (54),WHICH IS A VERY (43),THERE IS A VERY (42),IS A VERY SERIOUS (32),IS A VERY POSITIVE (19),IS A VERY COMPLEX (17),IS A VERY SENSITIVE (17),IS A VERY INTERESTING (17),IS A VERY STRONG (17),WHAT IS A

		VERY (17),IS A VERY TIMELY (15),IS A VERY DIFFICULT (15),IS A VERY USEFUL (11),IS A VERY WELCOME (11),IS A VERY SIGNIFICANT (10),IS A VERY SMALL (10),IS A VERY CLEAR (10)
IT IS VERY	798	IT IS VERY IMPORTANT (383),THINK IT IS VERY (90),IT IS VERY CLEAR (89),IT IS VERY DIFFICULT (70),THAT IT IS VERY (61),BUT IT IS VERY (39),AND IT IS VERY (33),IT IS VERY MUCH (27),IT IS VERY GOOD (22),BECAUSE IT IS VERY (20),SO IT IS VERY (19),PRESIDENT IT IS VERY (18),IT IS VERY EASY (15)
I AM VERY	728	I AM VERY PLEASED (155),I AM VERY HAPPY (118),I AM VERY GRATEFUL (100),PRESIDENT I AM VERY (88),AND I AM VERY (74),I AM VERY GLAD (61),I AM VERY MUCH (40),THAT I AM VERY (39),I AM VERY PROUD (36),EN I AM VERY (30),I AM VERY CONCERNED (22),I AM VERY SORRY (17),SO I AM VERY (16),I AM VERY SATISFIED (14),I AM VERY DISAPPOINTED (14),I AM VERY CONFIDENT (12),I AM VERY CONSCIOUS (11),WHICH I AM VERY (10),I AM VERY THANKFUL (10),THEREFORE I AM VERY (10)
IS VERY IMPORTANT	687	IT IS VERY IMPORTANT (383),IS VERY IMPORTANT THAT (181),IS VERY IMPORTANT TO (136),IS VERY IMPORTANT FOR (75),THIS IS VERY IMPORTANT (53),IS VERY IMPORTANT AND (49),ON THIS VERY IMPORTANT (48),THAT IS VERY IMPORTANT (46),WHICH IS VERY IMPORTANT (32),THIS VERY IMPORTANT ISSUE (25),IS VERY IMPORTANT IN (25),IN THIS VERY IMPORTANT (23),IS VERY IMPORTANT BUT (17),TO THIS VERY IMPORTANT (15),IS VERY IMPORTANT BECAUSE (14),FOR THIS VERY IMPORTANT (12),IS VERY IMPORTANT AS (11),THIS VERY IMPORTANT REPORT (10)
A VERY IMPORTANT	591	IS A VERY IMPORTANT (273),A VERY IMPORTANT ISSUE (51),A VERY IMPORTANT ROLE (35),A VERY IMPORTANT AND (30),BE A VERY IMPORTANT (29),A VERY IMPORTANT PART (25),A VERY IMPORTANT STEP (23),A VERY IMPORTANT ONE (23),A VERY IMPORTANT POINT (21),HAVE A VERY IMPORTANT (18),WAS A VERY IMPORTANT (16),A VERY IMPORTANT DEBATE (16),A VERY IMPORTANT CONTRIBUTION (16),HAS A VERY IMPORTANT (13),A VERY IMPORTANT PIECE (13),A VERY IMPORTANT REPORT (11),A VERY IMPORTANT ASPECT (11),ALSO A VERY IMPORTANT (11),A VERY IMPORTANT SUBJECT (10),PLAY A VERY IMPORTANT (10),A VERY IMPORTANT QUESTION (10)
IT IS VERY IMPORTANT	383	IT IS VERY (798),IS VERY IMPORTANT (687)
I VERY MUCH	355	I VERY MUCH WELCOME (152),I VERY MUCH HOPE (58),AND I VERY MUCH (39),PRESIDENT I VERY MUCH (37),I VERY MUCH APPRECIATE (25),I VERY MUCH REGRET (20),I VERY MUCH SUPPORT (19),I VERY MUCH AGREE (17),I VERY MUCH LOOK (15),SO I VERY MUCH (14),THAT I VERY MUCH (11),
A VERY GOOD	302	IS A VERY GOOD (89),A VERY GOOD REPORT (29),BE A VERY GOOD (24),A VERY GOOD JOB (23),FOR A VERY GOOD (20),A VERY GOOD IDEA (18),A VERY GOOD

		EXAMPLE (16),HAVE A VERY GOOD (15),A VERY GOOD DEBATE (15),HAD A VERY GOOD (12),DONE A VERY GOOD (11),A VERY GOOD START (11),BEEN A VERY GOOD (11),WAS A VERY GOOD (10)
IS A VERY IMPORTANT	273	IS A VERY (841),A VERY IMPORTANT (591)
THIS IS A	252	THIS IS A VERY (240),THAT THIS IS A (41),PRESIDENT THIS IS A (17),THINK THIS IS A (17),BECAUSE THIS IS A (13),AND THIS IS A (10)
THIS IS A VERY	240	IS A VERY (841),THIS IS A (252)

Figure 8: 3-4 word clusters around ‘have’ in EP_en_ST.

Cluster	Freq.
WE HAVE TO	**2500**
THAT WE HAVE	1815
WE HAVE A	1012
HAVE TO BE	898
TO HAVE A	795
WILL HAVE TO	622
DO NOT HAVE	616
AND WE HAVE	614
WE HAVE BEEN	598
I HAVE TO	560
WE WILL HAVE	531
MEMBER STATES HAVE	513
WE HAVE THE	504
THAT HAVE BEEN	454
WE HAVE HAD	421
WE HAVE SEEN	406
WHAT WE HAVE	391
I HAVE BEEN	364
HAVE TO SAY	362
WE DO NOT	324
THAT I HAVE	312
AS WE HAVE	301
THOSE WHO HAVE	298
THERE HAVE BEEN	298
AS I HAVE	295
THAT WE HAVE TO	291
I HAVE TO SAY	287
WE DO NOT HAVE	285
TO HAVE THE	279
BUT WE HAVE	274
AND I HAVE	274
WE ALSO HAVE	271
WILL HAVE A	269
WHICH WE HAVE	264
WE HAVE NOT	263
WE HAVE HEARD	263
THAT THEY HAVE	257
PRESIDENT I HAVE	255
NOT HAVE THE	251
WE HAVE ALREADY	247
WE HAVE ALSO	237
WE NOW HAVE	235

ALSO HAVE TO	228
THAT YOU HAVE	227
HAVE NOT BEEN	226
WOULD HAVE BEEN	225
WE HAVE TO BE	221

Figure 9: EP_en_ST clusters around 'have to'.

Cluster	Freq.	Related
THAT WE HAVE TO	283	THAT WE HAVE TO BE (16),THINK THAT WE HAVE TO (14),AND THAT WE HAVE TO (12),IS THAT WE HAVE TO (11),MEAN THAT WE HAVE TO (11),AGREE THAT WE HAVE TO (10),THAT WE HAVE TO TAKE (10)
I HAVE TO SAY	283	I HAVE TO SAY THAT (148),BUT I HAVE TO SAY (39),I HAVE TO SAY TO (24),AND I HAVE TO SAY (19),I HAVE TO SAY I (19),PRESIDENT I HAVE TO SAY (16)

Figure 10: Clusters around 'must' in EP_en_ST.

Cluster	Freq.	Related
IT MUST BE	282	IT MUST BE SAID (38),AND IT MUST BE (26),BUT IT MUST BE (23),IT MUST BE A (19),THAT IT MUST BE (12)
AND WE MUST	266	AND WE MUST NOT (21),AND WE MUST ENSURE (20),AND WE MUST BE (16),AND WE MUST TAKE (11),AND WE MUST MAKE (10),AND WE MUST DO (10)
THAT WE MUST	263	THAT WE MUST NOT (21),IS THAT WE MUST (18),BELIEVE THAT WE MUST (17),THAT WE MUST HAVE (13),AND THAT WE MUST (13),AGREE THAT WE MUST (12)
WE MUST NOT	252	WE MUST NOT FORGET (40),BUT WE MUST NOT (23),AND WE MUST NOT (21),THAT WE MUST NOT (21),HOWEVER WE MUST NOT (16),WE MUST NOT LOSE (15),WE MUST NOT ALLOW (15),WE MUST NOT LET (14),WE MUST NOT ONLY (11)