82

A PRELIMINARY ANALYSIS OF COLLOCATIONAL DIFFERENCES IN MONOLINGUAL COMPARABLE CORPORA

Marco Baroni and Silvia Bernardini

Source: D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 Conference, UCREL*, Technical Paper number 16, Special issue, Lancaster: UCREL, Lancaster University, 2003, pp. 82–91.

1 Introduction

The notion of collocation has enjoyed mixed fortunes in the 50 odd years of its existence. Claimed to be obscure (Lyons 1977: 612), counter-productive (Langendoen 1968: 63ff) and generally useless (Lehrer 1974)[1] by its detractors, the idea that part of the meaning of a word is somehow related to its "word accompaniment, the other word material in which [it is] most commonly or most characteristically embedded" has also had its supporters. These have suggested that words have no meaning out of context, and meaning itself is not contained anywhere, but rather dispersed as the "light of mixed wave-lengths into a spectrum" (Firth 1957(1951): 192).

The intuitive appeal of this view is evident if one thinks of the difficulty a compositional approach to meaning has (even allowing for subcategorisation frames) in explaining the patterned quality of language performance, as found in a corpus, and ultimately the speaker's or writer's effortless routine handling of co(n)textual restrictions. The hypothesis that "everything we say may be in some degree idiomatic – that [. . .] there are affinities among words that continue to reflect the attachments the words had when we learned them, within larger groups" (Bolinger 1976: 102) provides a powerful argument in favour of the empirical study of collocations, with implications

for theoretical, descriptive and applied branches of linguistics. In recent years, notwithstanding the vagueness of the notion and consequent methodological problems in investigating it empirically, the study of collocations has indeed defied difficulties and criticism and sparked renewed interest in a number of areas ranging from computational and corpus linguistics to lexicography, language pedagogy, and crucially for our purposes, translation studies.

The hypotheses that "everything we say may be in some degree idiomatic" (Bolinger, above), and that "actual usage plays a very minor role in one's consciousness of language" (Sinclair 1991: 39) raise a number of interesting questions for translation research. Is there any evidence that translators be aware of collocational restrictions in the source and target languages? Do they show sensitivity to phraseological (a)typicality and restrictedness? These are very complex issues, that can hardly be resolved in one fell swoop. For a start, theoretical as well as methodological problems remain as to what collocations are in the first place,[2] and how best they can be retrieved from corpora and compared (see e.g. Krenn 2000a). Secondly, different types of corpora for the study of translation exist, providing different perspectives on the translation process.

In this paper, we limit our investigation to monolingual comparable corpora (MCC) and present a number of attempts at selecting and comparing collocations across original and translated texts. This study is novel in at least two ways: to the best of our knowledge, no previous investigation of the behaviour of translators through MCC has focused on collocational restrictions, and no study of collocational restrictions in translated texts has attempted to select candidate bigrams automatically.

The paper is organised as follows. Section 2 provides a general background on monolingual comparable corpora and collocation extraction. In section 3 we present a brief study conducted on a small corpus of EU reports in translated and original English. Section 4 describes a more substantial study of a corpus of translated and original world affairs articles in Italian. In section 5 we briefly discuss directions for further work.

2 Background

2.1 Monolingual comparable corpora in translation research

Monolingual comparable corpora are collections of original and translated texts in the same language, assembled according to comparability criteria such as "a similar domain, variety of language and time span [. . .]" (Baker 1995: 234).

According to Frawley (1984: 168–169)

> translation [. . .] is essentially a third code which arises out of the bilateral consideration of the matrix and target codes [. . .] since [it] has a dual lineage, it emerges as a code in its own right, setting its own standards and structural presuppositions and entailments [. . .].

Following Baker (e.g. 1995), work on MCC has adopted a corpus-based research methodology to unveil *universal* features of translation seen not as an individual act of interlinguistic transfer, but as a mediated communicative event, with its own "third code".[3] Thus, rather than focusing on differences between single originals and their translations (i.e. *parallel* texts), MCC allow the analyst to compare collections of originals and translations in the *same* language, unrelated to each other but chosen so as to be broadly comparable.

There is no doubt that MCC provide an innovative research environment in which translation norms (Toury 1995) and strategies (Löscher 1991) can be observed against the backdrop of target language use. Yet they raise substantial methodological problems that risk invalidating the results obtained. The main problem one is confronted with relates to the comparability of the corpus components.

According to one of its compilers (Laviosa 1998a), the two components of the *English Comparable Corpus* (ECC) are comparable with regard to the relative proportion of biography and fiction (i.e. genre), time span, distribution of female and male authors, distribution of single and team authorship, and overall size of each component. Furthermore, continues Laviosa, "the target audience of both collections can be characterised as literate, intellectual adults of both sexes".

The compilers have clearly thought out their design criteria. Yet doubts about the comparability of the two corpora remain. The criteria just mentioned would appear to be derived from monolingual corpus building heuristics. Yet your average translated novel, say, has a much more complex history than your average original novel. First, it derives directly from an existing text (its source text). Disregarding the latter may have somewhat worrying consequences: for example, the source text could have been written much earlier than its translation. In this case, it would not seem unlikely for the latter to display features typical of a diachronically different state of the target language (for further discussion and a concrete example see Bernardini and Conrad 2002).

In more general terms, translation involves two cultures-languages, at times distant from each other, and a set of decisions – often involving substantial investments of time and funds – that have to be taken in order for a work to migrate between the two. It has been suggested that these migrations are subject to socio-cultural norms (Toury 1995). For instance, not everything that is published in the Netherlands is translated into English, far from it; instead,

> Dutch fiction is chosen for translation either in the function of assumed target taste or in that of the status the work has acquired at the source pole, often as a combination of the two.
>
> (Vanderauwera 1985: 132)

Ignoring such socio-cultural factors when setting up a monolingual comparable corpus may result in reduced comparability and doubtful interpretations of the data obtained.

Given the complexity of the translation situation – we have only scratched the surface of the problem here, see Bernardini and Zanettin (forthcoming) for a more detailed discussion –, and the fact that corpus comparability is in itself an untrivial concept (Kilgarriff 2001), it would seem wise to give serious thought to the composition of a MCC, at least at these early stages, and to refrain from the assumption of comparability based on crude situational criteria inherited from monolingual research.

In an attempt to limit the proliferation of variables, and consequent difficulty in interpreting results, the corpora on which the experiments we present are based were selected so as to be maximally comparable.

The first corpus (EU) is a collection of official reports submitted by different EU countries to the EU Commission, describing progress made in the implementation of European guidelines in the area of employment policies.[4] Two versions are typically available, the original in the country's language, and a translation, normally into English. Though the originals were not included in the corpus, they were consulted to make sure that the English texts were indeed translations, not independent texts nor source texts for the pair.

From these texts we constructed a small corpus of originals from Ireland and the United Kingdom and translations from Finland, Italy, Portugal and Sweden.[5] The corpus contains 72,966 words in the original section and 145,932 words in the translation section.

This corpus has its advantages and its disadvantages. On the one hand, translation is carried out into international English, thus limiting the effect of preliminary norms deciding what to translate when (Toury 1995, above). The texts are very homogeneous, in terms of topic, date, register etc. Copyright clearance and text preparation are straightforward, and this is no trivial advantage in a exploratory study like this one, in which the possibility that the material be inadequate for the purposes of the research has to be investigated empirically. On the other hand, these are rather boring texts, translated into a highly conventionalised variety of English – sometimes called *EUese*; thus, doubts may be raised about the generalisability of any results to different translation settings. Furthermore, the corpus is rather small.

The second corpus (LIMES) is a slightly less homogeneous but far larger collection of articles published in the Italian Quarterly *Limes – Rivista italiana di geopolitica* [*Italian geopolitics journal*].[6] The complete collection of articles

published between 1993 and 1999 is included. The total size of the corpus is approximately 3 million words, with translations accounting for slightly less than one third of the total. These have been carried out by approximately 40 different translators, about half male and half female. The total number of authors exceeds 700. Each volume is centred around a theme ("War in Europe"; "Divisions within Islam", "What use is NATO?", "The global bomb"), and normally contains both originals and translations, thus ensuring thematic consistency across the original and translated components. Judging from the topic matter and names of authors and translators, it would appear that source languages for these include, among others, Albanian, Arabic, Chinese, English, French, German, Serbo-Croatian, Russian.

These characteristics of homogeneity (originals and translations address the same audience, deal with virtually the same topics, conform to the same editorial policies) would appear to guarantee an acceptable level of comparability between the subcorpora, such as is rare in MCC. The likely variety of source languages should limit specific source language effects. Similarly, the variety of authors, translators and topics covered should limit the effect of idiosyncracies.

2.2 Collocational restrictions and translation

In order to compare the level of patterned-ness in translated versus original language, it is first of all necessary to retrieve candidate patterns. A common approach to this problem has been the structural one: patterns are defined in terms of sequences of parts of speech, which are then searched either manually, or automatically with the help of a tagged corpus. For example, Gitsaki (1996) adopts this approach in her study of collocations in ESL student written production, deriving candidate structures from a dictionary (Benson *et al.* 1986). Krenn and Evert (see e.g. Krenn 2000b, Krenn and Evert 2001) adopt statistical measures to rank potential collocations that match certain syntactic templates, on the basis of a tagged and partially parsed corpus. Heid (1996: 121) describes "discovery procedures for collocations [. . .] based on a detailed description of the targeted collocations". The starting point in this case is a classification of *lexical functions* as defined by Mel'_uk (e.g. 1996).

Within TS, two recent works have attempted to analyse collocations in original *vs* translated language. They both rely on bilingual (parallel) corpora, adopting a different viewpoint from the one adopted here. Yet the method used for retrieving relevant units of analysis is equally relevant. Kenny's study of "sanitisation" in translation adopts various techniques for spotting creativity in originals before checking how it is rendered in translation. One "node" is identified – the German word *Auge* [eye] – as an impressionistically viable choice, i.e. a word that is frequent enough, enters into fixed expressions, and has been found in previous studies to be the

object of creative manipulation by other writers. Concordances, tables of collocates and lists of clusters are then retrieved for this word. This method is appropriate to investigate whether translators tend to normalise creative collocations, but would appear to be hardly adaptable to the aims of the present study. More relevant to our concerns is Danielsson's (2001) attempt at designing an automatic process whereby UMs (units of meaning) for English and Swedish can be identified in a parallel corpus and compared. She takes a sample of about 200 "interesting" words occurring 200 times or more in her corpora, and automatically works through their downward and upper collocates (Sinclair 1991). This method yields units of varying length containing some of the most frequent words in the corpora. A third method, the longest-linear method, is used to retrieve "units of structure" of the type "the X of the new Y" (ibid: 149). What would appear to be left out by the joint application of these methods is the potentially large set of collocations containing less frequent words, whose significance is not necessarily small.

The collocation extraction methods discussed in this section, all relatively knowledge-intensive, have led to cleaner, easier-to-interpret results than the ones we report below. It might be argued, however, that they place strong interpretative grids over the data (e.g. through intuitive classifications, POS tagging and so forth) that are better avoided at these early stages of research.

We prefer the simple knowledge-free method we describe below (4.2) since we do not know, *a priori*, which collocational structures are most typical of the original and translated languages we are studying, and thus we would not want to bias the results in the direction of a specific subset of collocational templates. Indeed, as we will briefly discuss in 4.5, it appears that some interesting differences between translated and original texts concern frequent syntactic structures that would not traditionally be considered collocations.

No less important, we are interested to see whether it is possible to obtain meaningful results with a method that would be applicable to any language or sub-language, independently of the NLP resources available.

3 A pilot study with the EU corpus

We conducted a pilot study with the EU data in which, following a method similar to those proposed in Kilgariff (2001), we verified that the lists of collocations extracted from subcorpora constructed from parts of the same corpus are more strongly correlated than those extracted from different corpora. This result was obtained with all the collocation extraction and scoring methods we describe below. Thus, the pilot experiment indicates that our measures are sensitive to systematic similarities among corpora more than to the random similarities and differences that we expect to exist in any set of textual data.

However, a more thorough investigation of the EU data along the lines of the one we report below for the LIMES data failed to reveal systematic differences between original and translated documents.

While it is tempting to attribute this failure to the "scripted" nature of the EU report genre, that would tend to reduce the differences between originals and translations, we feel that, because of the small size of the corpus, it is premature to draw any conclusions from this failure. We plan to collect a larger corpus of EU reports, in order to obtain more robust results.

4 Analysis of the LIMES corpus

4.1 Corpus pre-processing

The text extracted from the LIMES database was split into a corpus of articles originally written in Italian and a corpus of articles translated from other languages into Italian using an automated procedure. Interviews and roundtables were discarded, since we were not sure about their status.

The output of the automated procedure was checked and corrected by hand. We also performed other minor clean-ups in a semi-automated fashion.

The original text corpus (O) and the translated text corpus (T) were tokenized in an extremely rudimentary way, removing all non-alphabetic symbols except the apostrophe, that, following the conventions of Italian orthography, was tokenized as part of the preceding word.

After tokenization, the O corpus contained 2,132,060 words and the T corpus contained 895,820 words.

O and T were further subdivided into subcorpora as described in sections 4.3 and 4.4.

4.2 Collocation extraction and ranking

In order to find collocations, we first collected for each subcorpus of interest (see discussion in 4.3 and 4.4 below) candidate bigrams that had the following characteristics: 1) they were made of words that occurred at least twice in all the subcorpora to be compared; 2) they occurred at least 3 times in the relevant subcorpus.

The rationale for the first of these conditions is that we are not interested in differences in collocations that are due to differences in the topics covered by the various articles. We expect that words that are relatively frequent in all the subcorpora being compared are words that are not strongly linked to any particular topic.

The second condition guarantees that we have a list of manageable size, and it is unlikely to exclude any "true" collocation, since collocations are, by definition, frequent ngrams.[7]

We used three *association* measures to rank the lists of bigrams: raw frequency, (point-wise) mutual information (Church and Hanks 1990) and (–2*) log-likelihood ratio (Dunning 1993). Unlike raw frequency, the other two measures take the unigram frequency of the words composing the bigram into account, and favor word combinations whose frequency is higher than what we would expect under assumptions of independence.

Mutual information and log-likelihood ratio are calculated using the formulas given in Manning and Schütze (1999, ch. 5). For discussion of these and other measures of collocativity see also Evert (2001).

We chose not to pick one specific measure as the "right" one given the exploratory nature of this study. Moreover, recent work (e.g. Inkpen and Hirst 2002, Baroni *et al.* 2002) suggests that measures such as mutual information and log-likelihood ratio should be used in combination, as they tend to discover different types of related words.

As a consequence of the collocation extraction and evaluation methods we used, the results reported below are based on a rather generous and vague notion of what counts as a collocation: essentially, any pair of adjacent words that has a high frequency, and/or a higher frequency than what we would expect by chance, is treated as a collocation.

Furthermore, by working with lists of ranked bigrams, we are implicitely assuming that collocativity is gradient, rather than binary.

4.3 Comparing the number of collocations in translated and original text

The first question we were interested in answering was the following: Do translators have a greater tendency to use fixed expressions than original authors? In principle, there could be an effect in either direction: On the one hand, translators could have a tendency to use a simplified language characterized, among other things, by the frequent repetition of the same expressions (a possible effect of the tendency towards explicitness, on which see Schmied and Schäffler 1996). On the other, faithfulness to the source language text, coupled with the fact that many fixed expressions are often not translatable from a language to the other, could lead the translators to use fewer collocations than the creators of original texts.

In order to study this issue, we split the T corpus into 5 subcorpora containing 179,164 words each, and we randomly selected 5 chunks of 179,164 words from the larger O corpus.

From each of the 10 subcorpora created in this way, we extracted candidate bigrams and computed frequency, mutual information and log-likelihood as described above.

First of all, we compared the number of candidate bigrams in the T-subcorpora to the number of candidate bigrams in the O-subcorpora. For the T-subcorpora, the average number of bigrams is 8,094.2 (median: 8,149);

Table 1 Proportion of bigrams ≥ cutoff value.

Measure	*Cutoff*	*T-avg*	*O-avg*	*T-med*	*O-med*	*MW test*
fq	2	25.8	24.87	25.84	24.73	significant
fq	3	6.45	5.8	6.42	5.73	significant
fq	4	1.43	1.09	1.39	1.05	significant
mi	5	25.53	25.07	25.55	25.09	significant
mi	8	4.58	4.75	4.56	4.70	not sig
mi	10	1.05	1.15	1.04	1.16	not sig
llr	3	41.39	40.07	41.56	39.92	significant
llr	4	12.42	11.65	12.35	11.36	significant
llr	5	3.11	2.69	3.02	2.83	significant

for the O-subcorpora the average is 8,044.8 (median: 8,128). According to the results of a two-tailed Mann-Whitney U test (see Siegel 1956, ch. 6), the difference between the two sets is not significant at the $\alpha = 0.05$ level.

In the subsequent analyses, rather than considering simply the number of bigrams extracted from each subcorpus, we looked at the association scores that were assigned to these bigrams. In particular, for each measure *m* and for each cutoff point *c* from a set of cutoff points across the distribution of *m*, we computed the percentage of bigrams in each subcorpus that had an *m*-score equal or greater than *c*. We then compared the percentage of bigrams at or above the various cutoff points in the T- *vs* O-subcorpora.

In Table 1, we report the results of this type of analysis performed at 3 cutoff points for each of the measures (for frequency and log-likelihood ratio, the cutoffs are expressed as logarithms of the actual values). We chose this particular set of cutoff values since they seem to represent well the range of patterns encountered.

For each measure and representative cutoff point, the table reports the average percentage of bigrams with scores at least as high as the cutoff point in the T- and O-subcorpora, the medians, and whether a two-tailed Mann Whitney test comparing the T- and O-subcorpus sets with respect to the relevant percentages was significant at the $\alpha = 0.05$ level.

As the table shows, there is a small but clear tendency for the translated texts to contain a larger number of bigrams with stronger association scores (this is also in line with the results on the absolute number of bigrams we presented above).

The only data that do not go in this direction are those for the "middle'" and "high" mutual information cutoffs. It is interesting that these are also the only levels at which the difference between the groups is not statistically significant, i.e. what we have here is not a reversal of the effect, but a lack of significant effects of the translated/original distinction.

An informal comparison of the top bigrams according to mutual information to the top bigrams according to frequency and log likelihood ratio

suggests that the latter two measures (the first more than the second) tend to pick up bigrams where at least one component is a function word, whereas mutual information tends to pick up bigrams that are closer to our intuitive idea of what a collocation is (frequent/lexicalized N+Adj or V+N structures).

Thus, the difference between translated and original texts detected with frequency, log-likelihood ratio and the lowest mutual information cutoff seems to be due to frequent bigrams that would not normally be treated as collocations. We will come back to this topic in 4.5 below.

4.4 Collocation overlap among translated and original texts

The data reported above provide some (weak) evidence that there are systematic differences between translated and original texts in terms of collocational patterns, but they do not tell us whether such differences are due to a general tendency for translators to use more fixed expressions, or whether there are specific fixed expressions that tend to be favored by translators (or by original writers).

In order to test this second possibility, we conducted another set of experiments in which we measured the degree of overlap and correlation among the collocations found in original and translated texts.

This time, we split the T corpus into 10 subcorpora containing 89,582 words each, and we randomly selected 10 chunks of 89,582 words from the larger O corpus.

We then merged 5 randomly selected T-subcorpora into a 447,910 word "reference" T corpus and 5 randomly selected O-subcorpora into a 447,910 word "reference" O corpus.

The idea, then, was to compare the bigrams found in the unmerged T- and O-subcorpora to the bigrams found in reference T and reference O. If there is a tendency to use similar bigrams in texts of the same type, we should find that the bigrams in the T-subcorpora tend to be closer to those in reference T than to those in reference O, and/or that the bigrams in the O-subcorpora tend to be closer to those in reference O than to those in reference T.[8]

First of all, we looked at the number of candidate bigrams (in the sense of 4.2 above) that the T-subcorpora and the O-subcorpora shared with the reference corpora. To control for the effect of the absolute size of the bigram lists we computed the percentage of shared bigrams over the total number of distinct bigrams in the two lists being compared.

The average percentage of bigrams shared by the T-subcorpora with reference T was 21.28 (median: 21.36); the average percentage shared by the T-subcorpora with reference O was 21.32 (median: 21.11). The average percentage of bigrams shared by the O-subcorpora with reference O was 21.16 (median: 21.20); the average percentage of bigrams shared by the O-subcorpora with reference T was 20.23 (median: 20.25).

These data suggest that there is no strong trend in either direction as far as simple ovelap of the candidate bigram lists goes. This was confirmed by the statistical analysis. We ran Wilcoxon two-tailed matched-pairs signed-rank tests (Siegel 1956, ch. 5) comparing the T-subcorpora percentage overlap with reference T *vs* their overlap with reference O, and comparing the O-subcorpora percentage overlap with each of the reference corpora. Neither test gave significant results at $\alpha = .05$.

We then computed, for each of the association measures, the Spearman rank correlation coefficients (Siegel 1956, ch. 9) between each T- or O-subcorpus and the reference corpora. The correlations were computed by considering only those bigrams that occured both in the list extracted from the relevant subcorpus and in the list extracted from the relevant reference corpus.[9]

The results of these analyses are summarized in Table 2. The first column reports the association measure; the second column reports the subcorpus set; the third column reports the average (in parenthesis: median) of the Spearman coefficients of the correlations between each of the relevant subcorpora and reference T; the fourth column reports the same data for the correlations with reference O; the fifth column reports whether a Wilcoxon test for the corresponding data (correlation coefficients of the subcorpora with reference T *vs* reference O) gave significant results at $\alpha = .05$.

First of all, notice that in general the correlation coefficients between corpora are quite high, in the case of mutual information so high that the uniform results could be due to a ceiling effect.

The results with frequency and log-likelihood ratios go in the expected direction (each set of subcorpora is correlated more strongly with the corresponding reference corpus). However, the differences are very small and they are not statistically significant.

Given the small size of the subcorpora (< 100,000 tokens) and their limited number (5 per set), it seems that an obvious next step with respect to the overlap/correlation analysis would be to test whether the weak trends we have detected are confirmed by an analysis based on a larger data set.

Table 2 Correlations between subcorpora and reference corpora.

Measure	*Subcorpora*	*Avg (med) r with T*	*Avg (med) r with O*	*W test*
fq	T	.63 (.63)	.59 (.60)	not sig
fq	O	.61 (.60)	.62 (.62)	not sig
mi	T	.91 (.91)	.91 (.91)	not sig
mi	O	.91 (.91)	.91 (.91)	not sig
llr	T	.74 (.73)	.72 (.72)	not sig
llr	O	.72 (.72)	.73 (.73)	not sig

4.5 Qualitative analysis

In order to collect a smaller data set for a preliminary qualitative analysis, we extracted the collocations that appear to be most typical of translated texts and the collocations that appear to be most typical of original texts using the following method.

We first computed the average log-likelihood ratio for each bigram in the O-subcorpora and in the T-subcorpora described in 4.3. Then, we computed the log ratio of these two values for each bigram. We put the bigrams with a positive value of this measure equal to or greater than 12 in the list of bigrams typical of original text, and the bigrams with a negative value equal to or greater than 12 in the list of bigrams typical of translated text. The ±12 cut-off point was arbitrarily chosen to limit the data to a manageable amount.

Based on a subjective evaluation of meaningfulness and wellformedness, each set was further divided into two (sub-)sets. Set A contains sequences that appear to be meaningful and well-formed, while set B contains less likely collocation candidates, i.e. incomplete sequences resulting in syntactically ill-formed structures (*termine geopolitica; iniziativa centro; veda nota*); fully-predictable sequences (*suo figlio* ["his/her son"]; *noi europei* ["we Europeans"]; *sarà possibile* ["it will be possible"]) and, somewhat more controversially, content words preceded or followed by function words (usually articles or prepositions), such as *sull'isola* ["on the island"]; *delle riserve* ["of the reserves"]; *proveniente dal* ["coming from the"]. Since this is no attempt at proposing a classification of collocations, we have not provided intepretative labels for these groupings. Table 3 shows the number of bigrams assigned to each set, and their proportion out of the total number of bigrams selected for analysis.

Although the initial number or bigrams selected is substantially larger in the case of translations (203 *vs* 166 in originals), when less meaningful and wellformed sequences are removed only 74 bigrams are left (*vs* 83 in the originals).[10] These can be futher analysed in terms of their topic-dependency, and grouped along the cline "technical-general" into the fuzzy categories *strongly topic-dependent* (containing a geographical term), *weakly topic-dependent* and *topic independent* (general language). The top 5 bigrams from each category are reproduced in Table 4:

Table 3 A tentative classification of bigrams.

	Tot	*Set A*	*%*	*Set B*	*%*
original	166	83	50	83	50
translated	203	74	36.45	129	63.54

Table 4 Examples of bigrams grouped according to topic-dependency.

	ORIGINALS	*Meaning approximately*	*TRANSLATIONS*	*Meaning approximately*
strongly	*lega nord*	Northern League (It poi. party)	*fratelli musulmani*	Muslim brothers
topic-dependent	*lingua russa*	Russian language	*marco tedesco*	German mark
	minoranza italiana	Italian minority	*mar rosso*	Red Sea
	alto adriatico	High Adriatic	*chiesa russa*	Russian church
	centro europea	Central-European	*vicino oriente*	Near East
weakly	*guerra giusta*	right war	*governo federale*	Federal government
topic-dependent	*minoranze etniche*	ethnic minorities	*nucleo centrale*	central nucleus
	opinioni pubbliche	public opinions	*sistema monetario*	monetary system
	spazio vitale	vital space	*istituto orientale*	Oriental institute
	prodotti industriali	industrial products	*autorità federali*	Federal authorities
	basti pensare	suffice it to think	*terza fase*	third phase
topic-independent	*breve periodo*	short period	*porre fine*	put a stop
	chi scrive	the writer [lit. s/be who writes]	*stessa cosa*	same thing
	occorre realizzare	it is necessary to set up	*reso noto*	made public
	scorso anno	last year	*far sì*	make possible

Table 5 Distribution of topic independent and strongly topic-dependent bigrams.

	Topic-independent	*%*	*Strongly topic-dependent*	*%*
original	21	(25,3%)	12	(14,4%)
translated	10	(13,5%)	29	(39,1%)

Table 5, based on a manual count, shows that topic-independent typical sequences are twice as common in originals as in translations, whilst the opposite is true of strongly topic-dependent sequences

The initial impression of a more substantial incidence of repeated patterns in translated vs original language, supported by the number of patterns retrieved, is mitigated by observation of actual instances. It does seem that translated language is repetitive, possibly more repetitive than original language. Yet the two differ in *what* they tend to repeat: translations show a tendency to repeat structural patterns and strongly topic-dependent sequences, whereas originals show a higher incidence of topic-independent sequences, i.e. the more usual lexicalised collocations in the language. The latter may be viewed as instances of those "target-specific features" that according to Mauranen (forthcoming), who analyses Finnish data, tend to be underrepresented in translations with respect to comparable originals.

Closer observation of the bigrams excluded from this analysis (B set) reveals further interesting patterns. For example, the sequences *considerato come* [considered.MASC.SING as] and *considerata come* [considered. FEM.SING as], appear in the list of typically translational expressions. A search for all the variants of the adjective/past participle (masculine singular and plural, feminine singular and plural) retrieves 619 occurrences from the original corpus, and 333 occurrences from the translation corpus. These figures are in accordance with the relative sizes of the two corpora, approximately 2:1. However, if we look at the frequency of the collocate *come* in the first position to the right of the keyword, the proportions change dramatically (table 6).

It might be hypothesised that translators show a preference for using optional *come* ["as"] in this structure. This would be in line with Olohan's findings concerning overuse of optional elements in (English) translation (Olohan 2001).

Table 6 Frequencies of *considerato* and *considerato come.*

	considerato/considerata considerati/considerate	*Total tokens*	*%*	+come *(R1)*	*%*
Original	619	2,096,191	2.9	69	9.8
Translated	333	922,946	3.6	61	20.7

Clearly, caution must be exercised in drawing conclusions, especially in an exploratory study like this one. However, the current findings are promising, hinting at some systematic differences in the use of collocations in closely comparable translated and non-translated texts.

5 Conclusions

We believe that this study has shown that monolingual (closely) comparable corpora are promising resources for the study of collocational restrictions in translated *vs* non-translated language. Simple data-exploration methods coupled with qualitative analyses would appear to be adequate in providing at least preliminary insights in this area.

At the same time, we were only able to detect weak trends in the LIMES corpus, and no effects in the EU corpus. We plan to improve on this via two strategies. On the one hand, we hope that by simply enlarging both corpora, we will be able to identify more robust statistical trends. On the other, we can bootstrap from the data-driven insights presented in this exploratory study to devise knowledge-richer collocation extraction methods. For example, the observations in 4.5 suggest that analysis of frequent bigrams including function words might reveal itself to be particularly relevant in telling translated language apart from original language. This is a strategy we would have not considered had we not performed this preliminary data-driven investigation.

Notes

1 "The main criticism against the lexical approach to co-occurrence is that it does not explain anything. The lexical item is found to collocate with a second item and not with a third, but no explanation is given. Collocations and sets are treated as if combinatorial processes of a language were arbitrary" (Lehrer 1974: 176).

2 A recent book on corpus-based lexical semantics gives the following rather general definition of collocation: " 'collocation' is frequent co-occurrence" (Stubbs 2001: 29).

3 Among the universals proposed are *simplification, explicitation/explicitness, normalization, levelling out, disambiguation* and *standardization* (e.g. Baker 1995, 1996; Schmied and Schäffler 1996; Laviosa 1998a, 1998b; Olohan and Baker 2000; Olohan 2001).

4 The reports are freely available on the Web: http://europa.eu.int/comm/employment_social/news/2001/may/naps2001_en.html.

5 The direction of translation was confirmed by bilingual speakers of English and each of the four languages in question.

6 We would like to thank *Limes* for granting permission to use their CD-ROM for this research.

7 We ran some preliminary experiments in which the bigram collection procedure ignored words from an automatically constructed list of likely function words. We are still in the process of analyzing the results obtained in this way, bat see note 10 below for some short remark on them.

8 We use the reference corpus strategy rather than directly comparing all subcorpora to each other since the latter strategy would yield results that are difficult to interpret, as the samples would not be independent from each other (e.g. the degree of overlap between, say, subcorpora TI and T2, that between TI and T3 and that between T2 and T3 would all have counted as instances of T-to-T comparisons).

9 In general, the percentage overlap between the bigrams in a subcorpus and the bigrams in a reference corpus is around 21%, as we have just seen. Including the 79% of bigrams that are not shared by the compared corpora into the correlation analyses would have been problematic both from a statistical point of view, because of the massive tie problem due to the 0-scores, and from an empirical point of view, since, in the best case, the analyses would have essentially been a replica of the overlap analyses we just presented.

10 This agrees with what we observed in 4.3 concerning the results obtained with mutual information *vs* frequency and log-likelihood ratio, i.e. that the former tends to pick up the most plausible collocations whilst failing to detect differences between originals and translations. The same point emerges from a preliminary analysis of the bigrams extracted after removing the function words (see footnote 7 above). Again, in this kind of data the differences between original and translated language nearly disappeared.

References

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

Baker M 1996 Corpus-based translation studies: the challenges that lie ahead. In Somers H L (ed), *Terminology, LSP and translation*, Amsterdam, Benjamins, 175–186.

Baroni M, Matiasek J, Trost H 2002 Using textual association measures and minimum edit distance to discover morphological relations. Paper presented at the *International Workshop on Computational Approaches to Collocations*. Online: http://sslmit.unibo.it/-baroni/.

Benson M, Benson E, Ilson R 1986 (1997) *The BBI dictionary of English word combinations*. Amsterdam, Benjamins.

Bernardini S, Conrad S 2002 "Multidimensional Analysis and translation". Paper presented at the international conference *Corpora and Discourse*, Camerino 27–29 September 2002.

Bernardini S and Zanettin F forthcoming When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals. In *Proceedings of Translation universals, do they exist? Savonlinna, 19–20 October 2001*.

Bolinger D 1976 Meaning and memory. *Forum Linguisticum* 1(1): 1–10.

Church K, Hanks P 1990 Word association norms, mutual information, and lexicography. *Computational Linguistics* 16: 22–29.

Danielsson P 2001 *The automatic identification of meaningful units in language*. Unpublished PhD Thesis. Göteborg University.

Dunning T 1993 Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61–74.

Evert S 2001 On lexical association measures. Online: http://www.collocations.de/EK/am-html/.

Firth J R 1957 (1951) Modes of meaning. *Papers in linguistics*. London, OUP, 190–215.

Firth J R 1964 (1930) *Speech*. London, OUP.

Frawley W 1984 Prolegomenon to a theory of translation. In Frawley W (ed), *Translation: literary, linguistic and philosophical perspectives*. London and Toronto, Associated University Presses, pp159–175.

Gitsaki C 1996 *The development of ESL collocational knowledge*. Unpublished PhD Thesis. University of Queensland.

Heid U 1996 Using lexical functions for the extraction of collocations from dictionaries and corpora. In Wanner L (ed), *Lexical functions in lexicography and Natural Language Processing*. Amsterdam, Benjamins, pp115–146.

Inkpen D Z, Hirst G 2002 Acquiring collocations for lexical choice between near-synonyms. *SIGLEX Workshop on Unsupervised Lexical Acquisition, 40th meeting of the Association for Computational Linguistics*, Philadelphia, June 2002.

Kenny D 2001 *Lexis and creativity in translation*. Manchester, St. Jerome.

Kilgarriff A 2001 Comparing corpora *International Journal of Corpus Linguistics* 6(1): 1–37.

Krenn B 2000a *The usual suspects: data-oriented models for identification and representation of lexical collocations*. Saarbrücken, Saarland University.

Krenn B 2000b Collocation mining: exploiting corpora for collocation, identification and representation. *KONVENS*, pp209–214.

Krenn B, Evert S 2001 Can we do better than frequency? A case study on extracting PP-verb collocations. *Proceedings of the ACL Workshop on Collocations*, Toulouse.

Langendoen T 1968 *The London School of linguistics: a study of the linguistic theory of B. Malinowski and J. R. Firth*. Cambridge (MA), MIT Press.

Laviosa S 1998a "Core patterns of lexical use in a comparable corpus of English lexical prose. *Meta* 43(4): 557–570. online: http://www.erudit.org/revue/meta/1998/v43/n4/.

Laviosa S 1998b Universals of translation In Baker M (ed), *The Routledge encyclopedia of translation studies*. London, Routledge, pp288–291.

Lehrer A 1974 *Semantic fields and lexical structure*. Amsterdam and London, North Holland.

Löscher W 1991 *Translation performance, translation process, and translation strategies*. Tübingen, Gunter Narr.

Lyons J 1977 *Semantics – Volume I*. Cambridge, Cambridge University Press.

Manning C, Schütze H 1999 *Foundations of statistical natural language processing*. Cambridge (MA), MIT Press.

Mauranen A forthcoming "Where's cultural adaptation?" *In TRAlinea*.

Olohan M 2001 Spelling out the optionals in translation: a corpus study. In Rayson P, Wilson A, McEnery T, Hardie A and Khoja S (eds.) *Proceedings of Corpus Linguistics 2001*, Lancaster, pp423–432.

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

Schmied J, Schäffler H 1996 Explicitness as a universal feature of translation. In Ljung M (ed), *Corpus-based studies in English: papers from the seventeenth International Conference on English Language Research on Computerized Corpora* (ICAME 17). Amsterdam, Rodopi, pp21–36.

Siegel S 1956 Nonparametric statistics for the behavioral sciences. New York, McGraw-Hill.
Sinclair J 1991 *Corpus, concordance, collocation*. Oxford, Oxford University Press.
Stubbs M 2001 *Words and phrases. Corpus studies of lexical semantics*. Oxford, Blackwell.
Toury G 1995 *Descriptive translation studies and beyond*. Amstedam, Benjamins.
Vanderauwera R 1985 *Dutch novels translated into English. The transformation of a "minority" literature*. Amsterdam, Rodopi.