LEAVE IT OUT! USING A COMPARABLE CORPUS TO INVESTIGATE ASPECTS OF EXPLICITATION IN TRANSLATION

Maeve Olohan
Centre for Translation and Intercultural Studies, UMIST

1. Introduction

Corpus-based translation studies is a relatively new area of research within translation studies, motivated by an interest in the study of translated texts as instances of language use in their own right. This is in contrast to the not uncommon perception of translations as 'deviant' language use, a view which has generally led to the exclusion of translated texts from most reference corpora (Baker 1999). While translations have been seen as useful in parallel bilingual or multilingual corpora, this has usually been for contrastive linguistic analysis which has studied the relationship between source and target language systems or usage. Parallel corpora are naturally also of interest to the translation scholar as they facilitate investigation of the relationship between a translation and its source text. Recent work using corpora in translation studies has, however, been more concerned with building corpora of translations so that the patterns of use of language in translations may be studied. The first corpus of this nature was the Translational English Corpus at UMIST which, since its inception, has provided a model for a number of similar projects for other languages. TEC consists

exclusively of translations, in English, from a variety of source languages and of a range of genres.[1]

One of the fundamental concepts in corpus-based translation studies has been the notion of comparable corpus, defined by Baker (1995: 234) as "two separate collections of texts in the same language: one corpus consists of original texts in the language in question and the other consists of translations in that language from a given source language or languages…both corpora should cover a similar domain, variety of language and time span, and be of comparable length". Baker's initial groundbreaking work posited a number of features of translation which could be investigated using comparable corpora (Baker 1996), for example, that translations tend to be more explicit on a number of levels than original texts, and that they simplify and normalise or standardise in a number of ways. Much of the corpus-based work carried out to date has focused on syntactic or lexical features of translated and original texts which may provide evidence of these processes of explicitation, simplification or normalisation. It should be stressed that, while translators may at times consciously strive to produce translations which are more explicit or simplified or normalised in some way, the use of comparable corpora also allows us to investigate aspects of translators' use of language which are not the result of deliberate, controlled processes and of which translators may not be aware.

2. Data from comparable corpora

The data analysed in this paper are extracted, on the one hand, from the fiction and biography components of the Translational English Corpus (TEC), currently housed at the Centre for Translation and Intercultural Studies in Manchester, and on the other hand, from a comparable corpus made up of selected texts from the imaginative writing section of the BNC. Table 1 provides details of both corpora. The Translational English Corpus is being added to

all the time, which means that successive studies present data from TEC at different stages in its growth. Thus, the study in Olohan and Baker (2000) on reporting *that* with verbs SAY and TELL, makes use of data from TEC when it was smaller than it is at the time of writing, with the BNC subcorpus used for comparison then also correspondingly smaller than the one referred to in this paper.

	BNC	**TEC**
Tokens	6,382,557	6,238,635
Types	74,346	75,780
Content	40,000-word extracts of original writing in English from the imaginative section of the BNC	Full texts of works of fiction and biography translated into English from a range of languages

Table 1: Features of BNC and TEC subcorpora used in this study

3. Explicitation

The analyses reported on here arose from an interest in studying processes of explicitation in translation, where explicitation refers to the spelling out in the target text of information which is only implicit in a source text. This has long been considered a feature of translation and has been investigated by a number of scholars (e.g. Vanderauwera 1985, Blum-Kulka 1986) who have identified different means or techniques by which translators make information explicit, e.g. using supplementary explanatory phrases, resolving source text ambiguities, making greater use of repetitions and other cohesive devices. A focus of the research reported on in this paper are subconscious processes of explicitation and their realisation in linguistic forms in translated texts. Since the starting point is the linguistic form, we have concentrated on, firstly, optional syntactic features, hypothesising that, if explicitation is genuinely an inherent feature of translation, translated text might manifest a higher

frequency of the use of optional syntactic elements than written works in the same language, i.e. translations may render grammatical relations more explicit more often – and perhaps in linguistic environments where there is no obvious justification for doing so – than authors in English. The second part of the analysis focuses on a set of pronouns and their occurrences with common verbs. This analysis is based on the hypothesis that a significantly lighter use of pronouns in the translated texts than in the BNC texts may be related to more nominal repetition in TEC, which could be investigated further and which may point to a higher level of explicitation in translation than in the non-translated texts.

4. Optional syntactic features in English

Linguists present the optional syntactic features of English in different ways; we opted to base this study on Dixon's (1991: 68-71) omission conventions for English, presented in summary form as follows:

A. Omission of subject NP
B. Omission of complementiser *that*
C. Omission of relative pronoun *wh-/that*
D. Omission of *to be* from complement clause
E. Omission of predicate
F. Omission of modal *should* from a THAT complement
G. Omission of preposition before complementisers *that*, *for* and *to*
H. Omission of complementiser *to*
I. Omission of *after/while* in *(after) having* and *(while) *ing*
J. Omission of *in order*

These features span a range of linguistic phenomena, from frequently occurring relative pronouns to much less common constructions (e.g. *to be* in complement clause), and they do not focus exclusively on optionality of omission. As will be obvious from the discussion below, they also vary considerably in terms of

their identification and quantifiability in a corpus which is neither tagged nor parsed. My analysis will therefore focus on features B, C, H, I and J. In some instances, omission is difficult to measure, but occurrence, i.e. inclusion, can be traced and compared across corpora to give an indication of similarities or differences in usage of the longer surface form in both corpora.

4.1. Omission of complementiser *that*

Dixon states that "the initial *that* may often be omitted from a complement clause when it immediately follows the main clause predicate (or predicate-plus-object-NP where the predicate head is *promise* or *threaten*" (1991: 70). An extensive analysis of the use of *that/zero*-connective with reporting verbs SAY and TELL, with reference to TEC and BNC, is presented in Olohan and Baker (2000). The results are summarised in Tables 1 and 2 in which both the absolute values (i.e. occurrences) and the percentages for each form are presented:

Form	*say* (TEC)	*say* (BNC)	*said* (TEC)	*said* (BNC)	*says* (TEC)	*says* (BNC)	*saying* (TEC)	*saying* (BNC)
that	316 **55.5%**	323 **26.5%**	267 **46.5%**	183 **19.2%**	116 **40.4%**	64 **12.8%**	76 **67.3%**	142 **43.0%**
zero	253 **44.5%**	895 **73.5%**	307 **53.5%**	771 **80.8%**	171 **59.6%**	435 **87.2%**	37 **32.7%**	188 **57.0%**

Table 2: SAY + *that/zero* in BNC and TEC

Form	***tell*** (TEC)	*tell* (BNC)	*told* (TEC)	*told* (BNC)	*tells* (TEC)	*tells* (BNC)	*telling* (TEC)	*telling* (BNC)
that	247 **62.8%**	300 **38.2%**	353 **60%**	584 **43.6%**	55 **68.7%**	28 **37.5%**	64 **73.6%**	85 **42.3%**
zero	146 **37.2%**	486 **61.8%**	233 **40%**	755 **56.4%**	25 **31.3%**	52 **62.5%**	23 **26.4%**	115 **57.7%**

Table 3: TELL + *that/zero* in BNC and TEC

It is immediately clear that the *that*-connective is far more frequent in TEC than in BNC. With the exception of *said* and *says*, *that* occurs more often than *zero* for all forms of SAY and TELL in TEC. By contrast, the *zero*-connective is more frequent for all forms of both verbs in the BNC corpus. These differences have been proven to be statistically significant. Furthermore, the results of the SAY and TELL study were consistent with findings by Burnett (1999) who reviewed use of the verbs SUGGEST, ADMIT, CLAIM, THINK, BELIEVE, HOPE and KNOW in TEC and BNC . While that study did not include all forms of these verbs, the data available show that the *that*-connective is far more common than the *zero*-connective in translated than in original English for forms of all seven of the verbs investigated. Although Olohan and Baker (2000) highlight the relative vagueness with which omission and inclusion of *that* are accounted for in the linguistics literature, and the lack of guidance on this in reference works for users of English, there are clear patterns of usage in contemporary English writing, as evidenced in the BNC corpus, and there is an equally clear contrast between these patterns and those perceived in the English of the literary translation contained in TEC.

A brief analysis of one of the verbs suggested by Dixon, namely PROMISE, serves as further illustration and corroboration of this pattern. Table 4 shows that, although the number of instances of promise + *that/zero* were almost identical in the two corpora (135 in BNC and 131 in TEC), the relationship between *that* and *zero* in TEC (*that* = 67.9%, *zero* = 32.1%) is almost directly inverse to that in BNC (*that* = 34.1%, *zero* = 67.9%).

	PROMISE (TEC)	PROMISE (BNC)
that	89 **67.9%**	46 **34.1%**
zero	42 **32.1%**	89 **65.9%**

Table 4: PROMISE + *that/zero* in BNC and TEC

4.2. Omission of relative pronoun *wh-/that*

The occurrences of these frequently occurring relative pronouns are difficult to measure in an untagged corpus. Thus far, total counts of occurrence of *which* have been taken, with 11,201 in BNC and 23,607 in TEC. A first step in discarding irrelevant instances was to identify sentence-initial and sentence-final/clause-final *which*. Their removal leaves 10,457 concordance lines in BNC and 22,483 in TEC, indicating considerably higher usage of *which* in TEC. Further detailed analysis of these instances will be required to identify the occurrences in relative clauses where the co-referential NP is not in subject function in the relative clause, i.e. where omission could have taken place.

A study of *who is* and *who's* reveals that the overall use of *who* is considerably higher in TEC occurrences) than in BNC:

Form	BNC	TEC
who's	419	337
who is	339	824
Total	**758**	**1,161**

Table 5: Occurrences of *who's* and *who is* in BNC and TEC

Through closer analysis of the concordances for *who's* and *who is*, it was possible to ascertain that, while a similar number of *who's* and *who is* occurrences are clearly questions, TEC has a very significant number of *who* being used as a relative pronoun rather than as an interrogative (Table 6). 44% of BNC occurrences of *who's* or *who is* are interrogative, compared with only 15% of total TEC occurrences. The non-interrogative occurrences have not yet been analysed further to identify those instances where *who* could have been omitted, i.e. occurrences in relative clauses where the co-referential NP is not the subject of the relative clause.

Form	BNC	TEC
who's (interrogative)	187	78
who is (interrogative)	150	102
Total (interrogative)	**337**	**180**
who's (relative)	232	259
who is (relative)	189	722
Total (relative)	**421**	**981**

Table 6: Interrogative and relative pronoun occurrences of *who's* and *who is* in BNC and TEC

Similarly, in the use of *who've*, *who have*, *who'd*, *who did*, *who had* and *who would* (Tables 7 and 8), we can see that, here too, TEC has a significantly higher overall occurrence of the *who* form. Closer investigation of the co-text, which would be required to differentiate interrogative from relative usage, and to determine the optional vs. non-optional nature of the relative pronoun in each case, has not yet been carried out for these forms. If, as was the case with *who's* and *who is*, future analysis reveals heavier use of the relative pronoun in TEC than in BNC, this would provide further evidence of a greater propensity to form relative clauses in TEC, which may, in turn, be seen as indicative of a greater degree of explicitness of clausal relations in translation.

Form	BNC	TEC
who've	27	35
who have	135	405
Total	**162**	**440**

Table 7: Occurrences of *who've* and *who have* in BNC and TEC

Form	**BNC**	**TEC**
who'd	246	236
who did	127	194
who would	386	437
who had	2,003	2,477
Total	**2,762**	**3,344**

Table 8: Occurrences of *who've* and *who have* in BNC and TEC4.3. Omission of complementiser *to*

According to Dixon (1991), the complementiser *to* is optional following HELP or KNOW. The form *help* was analysed, first discarding all uses of *help* as noun, as reflexive verb, verb + ING complement and verb + preposition, and then looking at occurrences of *help* (*) (*) to in detail (Table 9).

Form	BNC		TEC	
	Total occurrences	Relevant occurrences	Total occurrences	Relevant occurrences
Occurrences of *help*	2,374	300	1,792	365
help + to	62	26	72	38
*help + * + to*	67	50	98	80
help + + * + to*	19	3	35	19
Total help (+*) (+*) + *to*		**79**		**137**
help(+*)(+*) + zero		**229**		**228**

Table 9: *help* (+*) (+*) + *to* in BNC and TEC

These data tell us that although the word form *help* is more frequent in TEC, its verbal use in both corpora is quite similar with *help* (+*) (+*) + *to/zero* occurring slightly more often in TEC than in BNC, of which the complementiser *to* is used in 37.5% of TEC instances, compared with 26% of the BNC occurrences.

4.4. Omission in *(while) *ing* and *(after) having + participle*

As in the case of other features discussed here, we can more readily measure occurrence of these features rather than omission. Concordances of *while *ing* were pruned, discarding constructions such as *all the while *ing*, *after/in/for a while *ing*, *worth your while *ing*. The *while *ing* construction is much more frequent in TEC overall and for the feature investigated here (Table 10).

Form	BNC	TEC
Total *while *ing* concordances	150	360
Relevant concordances	138	330

Table 10: *while *ing* in BNC and TEC

A count of *after *ing *ed* (which obviously does not take irregularly formed past participles into account) also shows a tendency for TEC to use this construction more frequently than BNC (Table 11).

Form	BNC	TEC
*after *ing *ed*	11	65

Table 11: *after *ing *ed* in BNC and TEC

4.5. Omission of *in order*

Dixon (1991) states that *in order* is usually omitted before *to* and may occasionally be omitted before *for* or *that*. While the investigation of every instance of the items *to, that* and *for* to see whether an *in order* has been omitted is not practical, we can easily measure usage of *in order to*, *in order for* and *in order that* and compare results from the two corpora. This investigation yields the following (Table 12):

Form	BNC	TEC
in order to	250	1,225
in order for	1	14
in order that	12	18
Total	**263**	**1,257**

Table 12: *in order to/for/that* in BNC and TEC

This does not conclusively prove that *in order* has been omitted more often in BNC but certainly indicates that the longer forms of the conjunctions appear with markedly higher frequency in TEC.

5. Some patterns of personal pronoun usage

Tables 13-18 present a selection of data on occurrences of personal pronouns from the BNC subcorpus and TEC. These comprise frequencies of personal pronouns occurring with verb forms *will*, *have*, *am*, *is*, *has* and *are*, both within verb contractions and within non-contracted forms. These data show that, when used in conjunction with these particular verb forms, personal pronouns *I*, *you*, *he*, *she*, *we* and *they* are more common in the BNC subcorpus than in TEC. The differences are very striking in the case of *I*, *you*, *she* and *we*, but less marked for *he* and *they*. (The significance of the patterns of contraction of these and other forms in the two corpora is discussed in some detail in Olohan (forthcoming).)

Form	BNC	TEC
I'll	4,267	1,807
I will	771	659
I've	4,135	2,070
I have	2,591	3,717
I'm	9,279	4,254
I am	2,366	3,671
Total	**23,409**	**16,178**

Table 13: Occurrences of *I* with *will*, *have* and *am*, contracted and non-contracted

Form	BNC	TEC
you'll	1,666	874
you will	812	804
you've	1,096	918
you have	1,651	1,368
you're	4,550	2,239
you are	2,075	2,639
Total	**11,850**	**8,842**

Table 14: Occurrences of *you* with *will, have* and *are*, contracted and non-contracted

Form	BNC	TEC
he'll	674	463
he will	288	538
he's	3,936	1,926
he has	661	1,342
he is	1,082	1,846
Total	**6,641**	**6,115**

Table 15: Occurrences of *he* with *will, has* and *is*, contracted and non-contracted

Form	BNC	TEC
she'll	404	190
she will	174	252
she's	2,550	1,142
she has	356	667
she is	652	1,095
Total	**4,136**	**3,346**

Table 16: Occurrences of she with will, has and is, contracted and non-contracted

Form	BNC	TEC
we'll	1,247	786
we will	262	212
we've	997	618
we have	938	1,045
we're	1,503	1,034
we are	783	1,148
Total	**5,730**	**4,843**

Table 17: Occurrences of *we* with *will*, *have* and *are*, contracted and non-contracted

Form	BNC	TEC
they'll	548	286
they will	279	398
they've	509	344
they have	493	814
they're	1,686	927
they are	989	1,680
Total	**4,504**	**4,449**

Table 18: Occurrences of *they* with *will*, *have* and *are*, contracted and non-contracted

6. Conclusions

The SAY and TELL (*that*) study (Olohan and Baker 2000) mentioned above was a first step in concretely investigating subconscious processes of explicitation in translation using a comparable corpus. Explicitation in translation had been discussed previously in the translation studies literature (e.g. Blum-Kulka 1986), but attention has often been focused on conscious processes of explicitation.

Following on from the SAY and TELL (*that*) study, other optional syntactic elements were considered to be of interest for further investigation, based on the hypothesis that explicitation will usually involve the use of a longer surface form in preference to a shorter one, leaving less room for ambiguity. This paper presents preliminary quantitative analysis of a number of other optional syntactic structures in TEC and BNC, and the results point towards a general tendency for syntactic explicitation in TEC. Furthermore, this tendency not to omit optional syntactic elements may be considered subliminal or subconscious, rather than a result of deliberate decision-making of which the translator is aware.

Olohan and Baker (2000) pointed out that the optional *that* data discussed in that paper revealed potentially different patterns in other features, such as use of modifiers, pronominal forms, modal constructions etc. in TEC compared with the BNC. Thus, although we can investigate a specific syntactic or lexical structure in terms of overall occurrence and of its usage within the narrow context of a concordance line, it is important to consider the wider issue of co-occurrence and interdependency of features. A study of contracted forms (Olohan forthcoming) investigates the relationship between *that*-deletion and contractions, a relationship suggested by Biber's register analysis (Biber 1988 and Biber et al. 1998). According to the co-occurrence patterns which Biber proposes as underlying the five major dimensions of English, *that*-deletion and contractions are in the top three features at the positive end of one scale: Dimension 1, representing 'Involved vs. Informational Production'. These features, and others grouped with them in Dimension 1, are thus likely to co-occur in texts of shared function and are associated with "involved, non-informational focus, related to a primarily interactive or affective purpose and on-line production circumstances" (Biber et al. 1998: 149). They may constitute a reduced surface form which results in a "more generalized, less explicit content" (ibid.).

Relating this to the data showing higher prevalence of optional syntactic forms in TEC, we could thus posit that the BNC writing is

more involved, more generalised, less explicit, less edited than the literary texts of TEC. The surface form of the translations is not reduced to the same extent as that of the BNC texts. Second and first person pronouns are also features of 'involved production' in Biber's Dimension 1, and although the quantitative data on pronominal use given above represent only a very small subset of total pronominal usage, it is interesting to note that the greatest difference between the BNC texts and the TEC texts occurs in the use of *you* and *I-*, the BNC occurrences constituting a 26% and 30% increase on the TEC occurrences respectively.

It is thus clear from this preliminary analysis that there is scope to analyse each of the optional syntactic features outlined here in greater detail, using a qualitative co-text analysis in combination with these quantitative data. The latter data, although in some cases consisting of rather crude measures, illustrate the usefulness of investigating optional syntactic elements in a comparable corpus study as a starting point from which to approach explicitation in translation. Secondly, the findings on pronoun usage, combined with previous data for *that*-deletion and contractions, indicate that it may be possible to distinguish further the TEC texts from those of the BNC subcorpus by investigating occurrences and co-occurrences of other features of 'involved production' in both corpora, and then relating them to the profiles developed by Biber (1988 and 1995) for different genres of written English.

Note

1. For further details and remote access to TEC, see the Research section of the CTIS website: www.umist.ac.uk/ctis.

References

Baker, Mona (1995) "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research", *Target* 7(2): 223-243.

Baker, Mona (1996) "Corpus-based Translation Studies: The Challenges that Lie ahead", in Harold Somers (ed.) *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*. Amsterdam and Philadelphia, John Benjamins, pp 175-186.

Baker, Mona (1999) "The Role of Corpora in Investigating the Linguistic Behaviour of Translators", *International Journal of Corpus Linguistics* 4(2): 281-298.

Biber, Douglas (1988) *Variation across Speech and Writing*. Cambridge: CUP.

Biber, Douglas (1995) *Dimensions of Register Variation: A Cross-Linguistic Study*. Cambridge: CUP.

Biber, Douglas, Susan Conrad and Randi Reppen (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: CUP.

Blum-Kulka, Shoshana (1986) "Shifts of Cohesion and Coherence in Translation", in Juliane House and Shoshana Blum-Kulka (eds.) *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*. Tübingen: Gunter Narr, pp 17-35.

Burnett, Scott (1999) *A Corpus-based Study of Translational English*. Manchester: unpublished MSc dissertation, UMIST.

Dixon, R.M.W. (1991) *A New Approach to English Grammar, on Semantic Principles*. Oxford: Clarendon Press.

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

Laviosa, Sara (1998) "The English Comparable Corpus: A Resource and a Methodology", in Lynne Bowker, Michael Cronin, Dorothy Kenny and Jennifer Pearson (eds.) *Unity in Diversity: Current Trends in Translation Studies*. Manchester: St. Jerome, pp 101-112..

Olohan, Maeve and Mona Baker (2000) "Reporting *that* in Translated English: Evidence for Subconscious Processes of Explicitation?", *Across Languages and Cultures*, 1(2): 141-158.

Vanderauwera, Ria (1985) *Dutch Novels Translated into English: The Transformation of a 'Minority' Literature*. Amsterdam: Rodopi.