THE DESIGN AND ANALYSIS OF A COMPARABLE CORPUS OF ENGLISH NEWSPAPER ARTICLES

Sara Laviosa-Braithwaite

U.M.I.S.T. Manchester

Definition of corpus

In my paper *corpus* is intended as a collection of complete texts held in machine-readable form for automatic processing by text-retrieval software programs.

Baker's notion of "comparable corpora"

Very generally, comparable corpora are two collections of texts in one and the same language (Baker 1995). One collection contains texts originally produced in a given language, the other includes texts translated into that same language from one or more source languages. Baker proposes that the two collections "should cover a similar domain, variety of language and time span, and be of comparable length". Moreover, "the translation corpus should be representative in terms of the range of original authors and of translators", (Baker ibid).

To the best of my knowledge, the concept of comparable corpora has been applied only in two pieces of research. One is Gellerstam's computer-based study (1986) of translationese, which he defines as "the systematic influence on target language from source language". His corpus consists of two collections: 27 English novels translated from English into Swedish and a number of original Swedish novels with the same total running words. The time span is 1976-1977 and the books belong to the same genre. With regard to the literary level, the author tells us that "they are all good books, not pulp literature". The second investigation is Puurtinen's (1995) manual analysis of the occurrence of finite and non-finite constructions in two non-computerized collections of fairy tales and fantasy stories produced between 1940 and 1993; one includes 40 translations from English into Finnish, the other contains 40 Finnish originals.

In the present paper I apply the concept of comparable corpora to the investigation of a small collection of newspaper articles in order to throw some light on the nature of translation as a variety of linguistic behaviour, more specifically, a type of linguistic behaviour which involves mediating between a source and a target text. My objective is to create a comparable corpus of articles from the *Guardian* newspaper upon which I intend to realise the specific aim of establishing the incidence and impact of the translational features of Simplification and Normalisation, of which more below.

The *Guardian* English Comparable Corpus (G-ECC)

This is composed of two small corpora: the Guardian Translational English Corpus (G-TEC) and the Guardian Non-Translational English Corpus (G-NON-TEC). Both corpora were obtained by downloading selected articles from the *Guardian on CD-ROM* after being granted permission from the *Guardian* Syndication Department.

Composition of G-TEC

G-TEC comprises 37 complete articles published from 19th May to 28th July 1994 in the weekly supplement *Guardian Europe* of the British daily newspaper *Guardian*. The total number of running words is 25,879 and the average text length is 699 words. The articles are English translations of originals written in the following Source Languages: Czech, Danish, Dutch, Finnish, French, German, Greek, Italian, Norwegian, Russian, Spanish and Swedish. Full details regarding the sex and employment status of all the translators are not available. On the basis of general information gathered from both the Editor of *Guardian Europe* and the Translation Agency UPS, I can affirm that the range of translators represented is fairly wide.

The file structure of G-TEC

G-TEC is a collection of ASCII text files contained in one directory with the same name. The filename of each G-TEC text is made up of 8 letters and 3 digits (001 - 037). The first letter is N (for Newspaper); the second is G (for *Guardian*); the third is U (for UK); the fourth and fifth letters indicate the Source Language, the sixth and seventh (F or M) indicate the sex of the author of the source text and that of the translator respectively, the eighth letter indicates whether English is the translator's mother tongue (Y), foreign language (F) or language of habitual use (H). If the author and/or the translator are more than one, the letter T for Team is used in the sex slot. X is used to indicate unobtainable data.

Spell check and annotation of G-TEC texts

Every article is first of all thoroughly checked for spelling errors using the spell checking facility of a word processing package. It is then annotated as follows.

Text mark-up

This is carried out with angle bracket tags inserted in the text. Any section between angled brackets is excluded from the automatic text processing. The following extract is an example of a tagged article. The title and the synopsis, being non-translational, have been marked accordingly and so has the title of the newspaper of the source article. Pictures, tables and diagrams are also marked up.

<omit desc=title RENT ASUNDER reason=non-tc><omit desc=synopsis reason=non-tc Tomas Pudil reports that Czech prime minister Vaclav Klaus's soft spot is a weak point for his policy><omit desc=stnewspaper Lidove Noviny>FOUR and a half years after the Czech Velvet Revolution of 1989, housing - perhaps the most touchy issue in

all modern societies - remains unchanged. It is therefore hardly surprising that both property owners and tenants are at each other's throats. Meanwhile, the government, unable to deliver a long term housing policy, tries to patch up the situation by controlling.....

Annotation of the extra-textual features

This is done by creating a Database file called G-TEC.WDB. Each G-TEC article has a corresponding database record containing the following information:

filename; sex, age, employment status of the translator(s) (whether part-time or full-time, free-lance or employed by a translating agency), nationality at birth and current nationality of the translator(s); word count, type/token ratio, mean sentence length, lexical density, special features (pictures, non-translational sections etc); date of publication, whether or not the translation has been commissioned to a translation agency; Source Language, name and sex of the author(s); name of the newspaper publishing the source article and name of the country.

The Database file enables me to describe the components of the corpus and create ad hoc subcorpora where selected parameters can be controlled for intra-corpus comparative analyses. Another facility provided by the database is the automatic calculation of the Standard Deviation for any of the numerical variables recorded, for example: the lexical density, the average sentence length and the type/token ratio.

Composition, file structure and annotation of G-NON-TEC

G-NON-TEC is made up of 51 complete English articles published from 19th May to 28th July 1994 in the *Guardian* newspaper. The total word count is 25,832 words and the average text length is 550 words. The authors are varied.

The main criterion for selecting these articles was their political content. The texts were therefore extracted from the *Guardian*'s Home News and the Foreign News sections. I used the title for identifying the topic of the article.

G-NON-TEC is a collection of ASCII text files contained in one directory with the same name. The filename of each article is the date of publication. Articles selected from the same newspaper edition are identified by a different letter at the end of the numerical section.

The articles are checked for spelling errors. The annotation is minimal and carried out by simply inserting angle brackets to isolate extra-textual features such as pictures and diagrams, the name of the author and the date of the article.

WordSmith Tools for the automatic analysis of G-ECC

This is a forthcoming Windows-based suite of programs, which is still experimental. It offers the following lexical analysis tools that can be applied to the individual text files as well as to the entire corpus:

CONCORD.EXE - a concordancer that also does collocation, dispersion plots and word cluster analysis

WORDLIST.EXE - makes word-lists and compares them

KEYWORDS.EXE - finds key words using word-lists from WordList
identifies "key key-words"
makes key-word databases
identifies "associates"

SPLITTER.EXE - splits long text files into numerous texts

Using the word-lists given by WORDLIST.EXE, the program calculates the following statistics:

Total number of running words
Type/Token ratio, calculated as average on every 100 words of text
Average word length
Average sentence length
Lexical density

The study reported in the present paper makes use only of this last facility offered by WordSmith Tools.

Research hypotheses

In a previous paper (Laviosa-Braithwaite 1995) I put forward a series of hypotheses consistent with what has been identified intuitively as the universal features of translated texts, that is, features that typically occur in translated text rather than original texts and which are independent of the influence of the specific language pairs involved in the translation. In line with Baker (1993 and 1995), I suggested testing these hypotheses through comparative analyses of a representative and coherent corpus of English texts vis-à-vis a collection of texts translated into English from a variety of source languages.

The present study aims to test the validity of some aspects of two of these features: Simplification and Normalisation.

Simplification

The first hypothesis assumes that the lower the variety of vocabulary in a text, the simpler it is. As a measure of lexical variation I use the **type/token ratio**. This is the ratio of the number of content words to the number of running words in the same text. I therefore hypothesize that the G-TEC corpus has a lower type/token ratio than the G-NON-TEC corpus.

The second hypothesis assumes that the lower the information load in a text the simpler it is. As a measure of information load I use **lexical density**. This is the ratio of the number of content words to the number of running words in a text. I therefore hypothesize that the G-TEC corpus has a lower lexical density than the G-NON-TEC corpus.

The third hypothesis assumes that the lower the average sentence length in a text, the simpler it is. I therefore hypothesize that the average sentence length of G-TEC is lower than that of G-NON-TEC.

Normalisation

Standard Deviation (SD) is a statistical measure of the variability or dispersion of scores around the average value. It measures the extent to which a group lacks homogeneity so that the

higher its value the more heterogeneous the group is. It can be assumed that the lower the SD of the scores relating to the global linguistic features of lexical density, type/token ratio and average sentence length in a corpus, the higher its level of textual conventionality is. In turn, conventionality can be regarded as an aspect of Normalisation. I therefore hypothesize that the Standard Deviation for lexical density, type/token ratio and average sentence length is lower in G-TEC than in G-NON-TEC.

Findings

The results (see Table 1.) show a negligible difference between the two corpora with regard to the type/token ratio. The first hypothesis concerning the feature of Simplicity was therefore not cornfirmed.

Both the scores for lexical density and average sentence length were lower in G-TEC than G-NON-TEC. The second and third hypotheses testing the feature of Simplicity were confirmed.

Standard Deviation for lexical density, type/token ratio and average sentence length was found to be lower in the translational corpus. The hypothesis concerning the feature of Normalisation was confirmed.

Table 1.

	G-NON-TEC	G-TEC
AVERAGE LEXICAL DENSITY	**64.92%**	**58.59%**
SD OF LEXICAL DENSITY	*5.34*	*2.73*
AVERAGE TYPE/TOKEN	**73.95%**	**73.60%**
SD OF TYPE/TOKEN RATIO	*8.06*	*2.37*
AVERAGE SENTENCE LENGTH	**30.56**	**23.44**
SD OF SENTENCE LENGTH	*6.23*	*4.13*

Subsequent investigations and findings

The results obtained prompted me to look further into the relationship between the Source Languages of the translated texts and the corresponding lexical density values. With the sorting facility provided by the database, I selected the Source Language field for the first sorting and the Lexical Density for the second one. I then calculated the average lexical density corres ponding to those languages from which at least two articles had been translated. (See Table 2.) I found that the average lexical density values for texts translated from French, Italian and Russian were very close to the average lexical density for the entire translational corpus. The texts translated from German and Swedish had a lower than average value, while the translations from Czech and Spanish scored higher than average.

Table 2.

SL	LEXICAL DENSITY	
Czech	61.97	
Czech	63	AVG = 62.49
Danish	57.6	
Dutch	65.14	
Finnish	59.63	
French	54.25	
French	57.11	
French	57.54	
French	57.67	
French	57.73	
French	57.8	
French	61.41	
French	61.44	AVG = 58.12
German	54.78	
German	55.17	
German	57.1	
German	57.47	
German	62.23	AVG = 57.35
Greek	58.53	
Italian	50.9	
Italian	56.52	
Italian	56.54	
Italian	56.62	
Italian	58.61	
Italian	59.41	
Italian	60.5	
Italian	61.47	AVG = 58.70
Norwegian	59.26	
Russian	57.97	
Russian	58.99	
Russian	59.1	
Russian	59.44	AVG = 58.88
Spanish	57.84	
Spanish	60.76	
Spanish	62.86	AVG = 60.49
Swedish	56.4	
Swedish	57.09	AVG = 56.75

Evaluation

The two sets of results analysed so far are to be treated with great caution for the following reasons:

- the corpus size is, at this stage of the research, still very small
- no tests to assess the significance of the differences found have been carried out
- the source languages are unequally represented in the translational corpus

It is however useful to discuss possible future outcomes of further research together with their implications. If similar results were obtained with a larger and more representative comparable corpus and if the differences found in the values for lexical density, average sentence length and standard deviation were significant, I would feel justified in pursuing further my initial hypotheses and testing them on a larger corpus of the same genre and/or a corpus consisting of different types of text.

With regard to the second set of results concerning the relationship between source languages and lexical density, two possibilities can be envisaged, providing the corpus is large and there is a fair number of translated texts for each source language. If the differences in lexical density among the various source languages were significant, we may suggest that a relatively low level of lexical density is a feature of translated texts and that this feature is also affected by the source language. If, on the other hand, the differences were not significant, we may safely conclude that a relatively low lexical density is inherent in translation per se, independently of the specific language pairs involved in the mediating process.

Suggestions for further research

The nature of this type of research is essentially descriptive and inductive. Data are regularly collected, described and added to the initial corpus, new hypotheses are put forward and tested, laws of linguistic behaviour are continually formulated and revised in the light of fresh evidence. The present study is to be regarded only as the very beginning of what I hope will become a long term commitment to creating ever larger and varied comparable corpus. It is impossible at this stage to list all the possible studies that can be carried out with this new type of corpus: such a list would be endless.

I will therefore only mention the type of investigation that I am currently carrying out using G-ECC. First of all, I am extending the corpus to include a total of 103 translational articles, equivalent to approximately 70,000 words. I will then extract a sufficient number of non-translational articles with the same total number of words. Immediately after this, I will replicate on this larger corpus the analyses reported in the present paper. Finally, the procedure used to investigate the relationship between source languages and lexical density will be applied also to the type/token ratio, average sentence length and standard deviation.

REFERENCES

BAKER, M. (1993) — "Corpus Linguistics and Translation Studies", in Baker, M., Francis, G. and Tognini-Bonelli, E. (eds), *Text and Technology: In Honour of John Sinclair*. Amsterdam and Philadelphia, John Benjamins.

—— (1995) — "Corpora in Translation Studies An Overview and some Suggestions for Future Research", in *Target*, vol. 7:2, pp. 223-243.

GELLERSTAM, M. (1986) — "Translationese in Swedish novels translated from English", in *Translation Studies in Scandinavia Proceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II Lund 14-15 June 1985, Lund Studies in English 75*, Wollin, L. and Lindquist, Hans (eds) Lund: CWK Gleerup.

Guardian on CD-ROM, Syndication Department, *The Guardian*, 119, Farringdon Road, London EC1R 3ER, UK.

LAVIOSA-BRAITHWAITE, S. (1995) — "Comparable Corpora: Towards a corpus linguistic methodology for the empirical study of translation" in *Proceedings of the 1995 Maastricht-Lodz Duo Colloquium on Translation and Meaning Part 3, 1995.*

PUURTINEN, T. (1995) — *Linguistic Acceptability in Translated Children's Literature,* Joensuu: University of Joensuu Publications in the Humanities No. 15.

WordSmith Tools (forthcoming, Oxford University Press), copyright Dr Mike Scott, Department of English Language and Literature, The University of Liverpool, Modern Languages Building, Liverpool. S69 3BX, UK.