Meta
Journal des traducteurs
Translators' Journal

Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose

Sara Laviosa

Volume 43, Number 4, décembre 1998

L'approche basée sur le corpus
The Corpus-based Approach

URI: https://id.erudit.org/iderudit/003425ar
DOI: https://doi.org/10.7202/003425ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Laviosa, S. (1998). Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta*, *43*(4), 557–570. https://doi.org/10.7202/003425ar

Article abstract

This paper investigates the linguistic nature of English translated texts. The author' corpus consists of a sub-section of the English Comparable Corpus (ECC). It comprises two collections of narrative prose in English: one is made up of translations from a variety of source languages, the other includes original English texts produced during a similar time span. The study reveals four patterns of lexical use in translated versus original texts.

CORE PATTERNS OF LEXICAL USE IN A COMPARABLE CORPUS OF ENGLISH NARRATIVE PROSE

SARA LAVIOSA
UMIST, Manchester, United Kingdom

Résumé

Cet article étudie la nature linguistique des textes anglais traduits. L'auteur part d'une sous-section du English Comparable Corpus (ECC). La sous-section contient deux sélections de proses narratives en anglais : un premier jeu constitué de traductions à partir de diverses langues sources et un second contenant des textes originaux en anglais rédigés à la même époque. L'étude révèle quatre différents modèles d'utilisation lexicale dans les textes traduits par rapport aux originaux.

Abstract

This paper investigates the linguistic nature of English translated texts. The author' corpus consists of a sub-section of the English Comparable Corpus (ECC). It comprises two collections of narrative prose in English: one is made up of translations from a variety of source languages, the other includes original English texts produced during a similar time span. The study reveals four patterns of lexical use in translated versus original texts.

1. INTRODUCTION

The idea of creating a "comparable" corpus of English (Baker 1995: 234) has been recently realised in the design of a monolingual, multi-source-language English Comparable Corpus (ECC).[1] At the time of writing, this corpus represents two text genres:[2] newspapers and narrative prose, and has an overall size of 2 million words.

A previous investigation of a ECC subsection consisting of newspaper articles has shed new light on the linguistic nature of translational English (Laviosa-Braithwaite 1996, 1997; Laviosa 1998). The study shows that, in the British newspapers *The Guardian* and *The European*, the translated articles use a relatively lower proportion of lexical versus grammatical words[3] independently of the source language, as well as a higher proportion of frequent versus less frequent words. Moreover, the 108 most frequent words (or list head) are repeated more often, the nucleus of the words most frequently used is less varied, and the average sentence length is lower.

The analysis of newspaper articles has also shown other interesting differences between translated and original texts. In both newspapers, the translations use the present tense of the auxiliary verbs "to be" and "to have" more frequently. Translated articles are also found to be more homogeneous in respect of lexical density, as shown by their relatively lower variance: a phenomenon I have named "convergence" in an attempt to convey the meaning of clustering of a corpus of translations around the average value of a linguistic feature (Laviosa 1998).

The aim of the present study is to investigate the extent to which the global linguistic patterns discovered in translated newspaper articles are also typical of translated narrative prose, with a view to identifying the distinguishing features of

translational English. The first part of the paper gives details of the composition and general features of the comparable corpus of English narrative texts. This is followed by the statement of hypotheses, the exposition of the results, and discussion. In the final section, I will evaluate my analyses and make suggestions for further research.

2. THE COMPARABLE CORPUS OF NARRATIVE PROSE

2.1. The Translational Component

2.1.1. The publications

The collection of translated narrative comprises 14 published works in total; two are biographies, the rest is fiction. All the works included are complete texts, with the exception of *Forbidden Territory*, which is a sample of 38,714 words. The majority of publishers are British.

Biography

Wittgenstein's Nephew: A friendship by Thomas Bernhard, translated by Ewald Osers from German.
Forbidden Territory by Juan Goytisolo, translated by Peter Bush from Spanish.

Fiction

Memoirs of Leticia Valle by Rosa Chacel, translated by Carol Maier from Spanish.
The Stone of Laughter by Hoda Barakat, translated by Sophie Bennett from Arabic.
The Stone Raft by José Saramago, translated by Giovanni Pontiero from Portuguese.
The Gospel according to Jesus Christ by José Saramago, translated by Giovanni Pontiero from Portuguese.
Discovering the World by Clarice Lispector, translated by Giovanni Pontiero from Portuguese.
Turbulence by Chico Buarque, translated by Peter Bush from Brazilian Portuguese.
Passion by I. U. Tarchetti, translated by Lawrence Venuti from Italian.
Fantastic Tales, short stories mainly by I. U. Tarchetti, translated by Lawrence Venuti mainly from Italian.
The Siren, short stories by Dino Buzzati, translated by Lawrence Venuti from Italian.
Restless Nights, short stories by Dino Buzzati, translated by Lawrence Venuti from Italian.
Lucio's Confession by Mario De Sá Carneiro, translated by Margaret Jull Costa from Portuguese.
The Dedalus Book of Surrealism, short stories by different authors, translated mainly from French and edited by Michael Richardson.

2.1.2. Source languages, translators' gender, and direction of translation

As the tables below show, the Romance languages are by far the best represented both in terms of word count and number of texts[4] (table 1). The proportion of male translators is considerably higher than that of female translators (table 2). The vast majority of translations have been carried out into the mother tongue (table 3).

2.1.3. The process of translation

The majority of the translations have been carried out by highly reputed professional literary translators. All publications have been commissioned by publishers or editors, with the exception of two of Lawrence Venuti's translations (*Fantastic Tales* and *Passion*) which he initiated. Copyright is held either by the translator, the publisher or the editor of a collection. Most of the translations have been edited to varying degrees by either the publisher, the series editor or a freelancer hired by the publisher. Peter Bush's and Giovanni Pontiero's works, in particular, have been produced in close cooperation with the author.

Translational narrative prose			
	No of texts	Total word-count	% of Subcorpus
Germanic			
German	3	35239	3.52
Total	3	35239	3.52
Greek			
Greek	1	1945	0.20
Total	1	1945	0.20
Romance			
French	44	69258	6.93
Italian	6	208382	20.84
Spanish	7	130861	13.09
Portuguese	5	445028	44.50
Brazilian Port.	1	35068	3.51
Total	63	888597	88.87
Semitic			
Arabic	1	72239	7.22
Total	1	72239	7.22
Slavic			
Czech	1	1925	0.19
Total	1	1925	0.19

Table 1
Source languages

Translational narrative prose			
	No of texts	Total word-count	% of Subcorpus
Female	11	163254	16.33
Male	53	829763	82.98
Team	5	6928	0.69

Table 2
Translators' gender

Translational narrative prose			
	No of texts	Total word-count	% of Subcorpus
Into mother tongue	60	955324	95.54
Out of mother tongue	3	5500	0.55
Into language of habitual use	1	33492	3.35
Mixed	4	5233	0.52
Unknown	1	396	0.04

Table 3
Direction of translation

2.2. The Non-translational Component

2.2.1. The publications

The collection of original narrative texts comprises 18 text samples selected from the British National Corpus. As detailed below, three are biographies, the rest is fictions.

Biography

Michael Ramsey: A Life by Owen Chadwick.
C. S. Lewis: A Biography by Andrew Norman Wilson.
Leonard Cohen: Prophet of the Heart by Loranne S. Dorman and Clive L. Rawlins.

Fiction

The Clothes in the Wardrobe by Alice T. Ellis.
The Magic Toyshop by Angela Carter.
The Fifth Child by Doris Lessing.
Condition Black by Gerald Seymour.
The Crow Road by Iain Banks.
Complicity by Iain Banks.
Flaubert's Parrot by Julian Barnes.
A History of the World by Julian Barnes.
Amongst Women by John McGahern.
Time's Arrow by Martin Amis.
The Maid of Buttermere by Melvyn Bragg.
Crystal Rooms by Melvyn Bragg.
Passing on by Penelope Lively.
Bird Song by Sebastian Faulks.
Callanish by William Horwood.

2.3. Comparability between Translated and Original Narrative Works

The translational and the non-translational components of the corpus are comparable with regard to the relative proportion of biography and fiction, time span, distribution of female and male authors, distribution of single and team authorship, and overall size of each component (see table 4 below). Moreover, the target audience of both collections can be characterised as literate, intellectual adults of both sexes. Highly experimental works of fiction and poetry have been excluded from both components of the corpus because these tend to have restricted audiences and are arguably less representative of general original and translational language.

3. STATEMENT OF HYPOTHESES

The lexical and stylistic patterns previously found in translated versus original newspaper articles (Laviosa-Braithwaite 1997; Laviosa 1998) form the basis of two initial sets of hypotheses:

> The translational component of the comparable corpus of narrative texts has a lower lexical density and mean sentence length than the non-translational component.
>
> The translational component of the comparable corpus of narrative texts contains a higher proportion of high frequency words and its list head covers a greater percentage of text with fewer lemmas than the non-translational component.

The comparable corpus of narrative prose		
	Translational	Non-Translational
BIOGRAPHY	7.22%	17.57%
FICTION	92.78%	82.43%
TIME SPAN	1983-1994	1985-1993
AUTHORSHIP:		
- Individual Male	85.18%	72.20%
- Individual Female	14.15%	22.05%
- Team	0.67%	5.75%
TOTAL WORD-COUNT	9999,945	729,349

Table 4
Dimensions of comparability

4. RESULTS CONCERNING THE INITIAL HYPOTHESES

4.1. Lexical Density and Mean Sentence Length[5]

The lexical density is highly significantly lower[6] in translated narrative, while the mean sentence length, contrary to my predictions, is significantly higher[7] (see figure 1 and table 5). The analysis of the individual scores reveals that the texts with the highest mean sentence length are *Wittgenstein's Nephew: A friendship* by Thomas Bernhard, translated by Ewald Osers from German (with an average of 35.41 words per sentence), and *The Gospel according to Jesus Christ* and *The Stone Raft*, both by José Saramago, translated by Giovanni Pontiero from the Portuguese (44.12 and 52.93 words per sentence respectively). Even excluding these texts, the mean sentence length for the translational narrative is still significantly higher than the comparable original group (18.62 vs 15.62).[8]

The first set of hypotheses is therefore confirmed only with regard to lexical density. This is partly consistent with the corresponding results concerning the newspaper subcorpus, which reveal significantly lower values for both lexical density and mean sentence length in translated texts (Laviosa 1998).

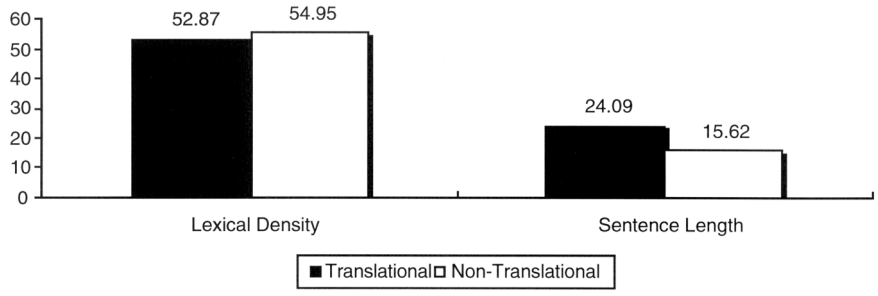

Figure 1
Comparable Corpus of Narrative Lexical Density & Mean Sentence Length

	Translational	Non-translational
LEXICAL DENSITY	52.87439153	54.9536606
Variance	2.552626875	1.8632584
PROPORTION OF HIGH FREQUENCY WORDS	59.736429	58.51277778
Variance	6.8007801	5.0424312
MEAN SENTENCE LENGTH	24.08714286	15.62555556
Variance	137.6326347	3.551449191

Table 5
The comparable corpus of narrative prose: lexical density, proportion of high frequency words, mean sentence length, and variance

4.2. List Heads and Proportion of High Frequency Words

The list head of the translational component represents 56.2% of the entire collection and contains 82 lemmas, while the corresponding percentage for the non-translational collection is 51.6% with 87 lemmas[9] (see figure 2, and appendices 1, 2, 3, and 4). These results are consistent with those obtained for the newspaper subcorpus and suggest that the nucleus of words most frequently used in translational narrative is less varied and accounts for a larger part of the entire component, when compared with the original narrative.

On average, the proportion of high frequency words used is significantly higher in the translated narrative than in the comparable original works[10] (see figure 3). This result is consistent with the corresponding finding concerning the newspaper collections.

The second set of hypotheses is therefore confirmed.

5. ADDITIONAL RESULTS

5.1. Variance[11]

The variance for lexical density and proportion of high frequency words is very low in both samples and only marginally higher in translational narrative. However, it

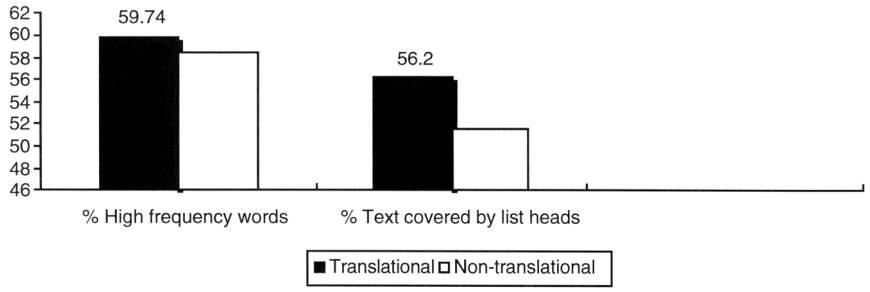

Figure 2
Comparable Corpus of Narrative High frequency words and list heads

is significantly higher in translated text in respect of mean sentence length[12] (see table 5 and figure 3). The variance drops considerably to the value of 21.45286116 when the three translated texts with the highest mean sentence length are removed from the sample, but it is still higher than the corresponding value for the non-translational narrative, though not significantly so.[13]

The results concerning variance do not confirm the greater homogeneity found in translated newspaper articles in respect to lexical density.

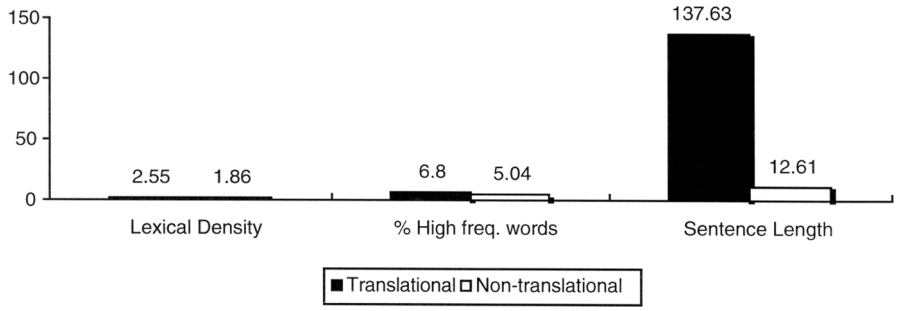

Figure 3
Comparable Corpus of Narrative Differences in Variance

5.2. The Auxiliary Verbs "to be" and "to have"

Only the forms "is," "are," and "has" are more frequent in translational narrative. The form "had" is used more often in non-translational texts. There are negligible differences in the use of the forms "was," "were," and "have" (see table 6). Unlike the corresponding analysis performed on the newspaper collections, there does not seem to be a clear pattern showing a preference in translational texts for the present tense of the auxiliary verbs "to have" and "to be" versus the past tense.

	Translational Position	Non-translational Position
IS	12th	25th
ARE	48th	61st
WAS	9th	8th
WERE	35th	31st
HAS	63rd	-
HAVE	31st	33rd
HAD	21st	13th

Table 6
Position of the auxiliary verbs in the list heads: Narrative Prose

6. DISCUSSION

The evidence provided suggests that translational texts of narrative prose exhibit specific patterns of lexical use arising from the proportion of content words versus grammatical words used; the percentage of frequent versus less frequent vocabulary; the proportion of text represented by the list head, and its lexical variety.

Contrary to my prediction, the mean sentence length is significantly higher in translational narrative texts compared to the originals. This pattern holds even when the three texts with the very highest scores are taken out of the sample on the grounds that they may be highly idiosyncratic and unrepresentative.[14] The different conventions regarding the punctuation of abbreviations and acronyms (see note 4) of translated and original works cannot explain this result because open punctuation is on the whole more common in the non-translational texts. This factor would therefore contribute to reducing, rather than increasing, the mean sentence length of the translational component. Moreover, since the number of texts examined in this study is small, it is not possible to try and assess the influence of the source language using the methodology tested on the newspaper collections (Laviosa-Braithwaite 1997). Further evidence on a much larger and varied corpus is necessary before any plausible explanation can be put forward for this finding.

The relatively low variance in both translational and non-translational texts in respect to lexical density and proportion of high frequency words suggests that both samples are fairly homogeneous. The differences in variance are marginal and suggest that the translational narrative is only slightly more idiosyncratic than the comparable original texts. On the other hand, owing to the exceedingly high scores of three of the works included in the translational sample, translated texts appear to be considerably more idiosyncratic than original ones in respect of mean sentence length. I cautiously hypothesize, pending further evidence from a more varied and larger sample, that the average sentence length may be particularly sensitive, in the narrative subject domain, to the influence of different source languages, as well as the author's particular style.

In any event, the analysis of variance in narrative prose does not confirm the greater homogeneity found in newspaper articles in respect to lexical density. It could be that this phenomenon may be detectable only with a large number of texts. The narrative corpus is, in fact, much larger in size than the newspaper corpus[15] but is made up of fewer texts.[16] This feature of translated text may pertain only to subject fields other than narrative.

In summary, we can say that four main global patterns appear to characterise newspaper and narrative translational texts in English:

i) Translated texts have a relatively lower percentage of content words versus grammatical words (i.e. their lexical density is lower);

ii) The proportion of high frequency words versus low frequency words is relatively higher in translated texts;

iii) The list head of a corpus of translated texts accounts for a larger area of the corpus (i.e. the most frequent words are repeated more often);

iv) The list head of translated texts contains fewer lemmas.

I propose to call these regular features of translated text "core patterns of lexical use" in an attempt to convey the fact that because they occur in two different subject domains, they may prove typical of English translated text in general.

7. EVALUATION AND SUGGESTIONS FOR FURTHER RESEARCH

The investigation carried out in the present study differs from most previous research into the linguistic nature of translational language in several ways. First of all, it has used computerised corpora as data and computerised methods of analysis for processing this data.[17] Secondly, it has been carried out entirely in the target language environment. And finally, it has focused on global patterns of language use, which, by their very nature, cannot be discerned through manual analysis.

At presents, corpus design is limited by the restricted number of publications, the over-representation of Romance languages, and the problem of achieving an adequate level of comparability between translated and original texts (Laviosa 1997). These concerns need to be addressed in future studies. Nevertheless, I think this new methodology can be fruitfully employed to assess the extent to which the core patterns of lexical use are language — subject domain — or modality — specific. The procedure one could adopt for this purpose would consist of creating and comparing ad hoc subcorpora in which each of the latter features is controlled in turn.

Moreover, the annotation of extra-textual attributes, such as the translators' gender or the direction in which the translation is carried out, can be used to study possible links between these variables and the lexical patterning of translated texts. This evidence can, in turn, form the basis for the elaboration of the type of probabilistic and conditional "laws of translational behaviour" proposed by Toury, which are based on the systematic observation of regularities in translation and translating (Toury 1995: 259-279).

What an ECC-based methodology cannot tells us, however, is why certain patterns occur and how they come about. The corpus design and methods of analysis adopted in this research focus on the character of the final product of translation, rather than the processes underlying it. When studying translation as a product entirely in the target language environment, we can only put forward suggestions regarding the possible causes that may have led to certain patterns. In order to find an explanation for our results, we would need to construct and analyse in parallel another corpus that would include the source texts of the translational component of ECC.

In conclusion, I would like to propose that, providing the present corpus is suitably enlarged and rendered more balanced, the core patterns of lexical use identified in the present investigation can be taken as sources of hypotheses to test on a variety of translational text genres and different types of translation (for example, conference, court interpreting, etc.). This is done in order to establish whether and to what extent these regularities are subject field and/or modality specific and/or language specific, or whether they can indeed be considered universal features of translational English.

Notes

1. The English Comparable Corpus consists of two computerised collections of texts in English: one, which I refer to as Translational English Corpus (TEC), comprises translations from a variety of source languages; the other, which I have called Non-Translational English Corpus (NON-TEC), includes original English texts of a similar type produced during a similar time span.

2. The terms subject domain, subject field, text category and text genre are used interchangeably in this study. They all refer to groups of texts considered similar by the corpus compiler or by general consensus on the basis of their extra-linguistic features. I have deliberately chosen to avoid the words "genre" and "text type" on their own, because these are used by Biber and Finegan to refer to two different notions (Biber and Finegan 1986, 1991). According to these scholars, "genres" are "the text categories readily distinguished by speakers of English (e.g. novels, newspaper articles, public speeches)" (Biber and Finegan 1991: 213). The

notion of "genre" is therefore used to characterise texts on the basis of external criteria (Biber and Finegan 1986: 20). "Text types," conversely, are defined in terms of the linguistic characteristics of the texts themselves. They represent sets of texts "that are similar with respect to their linguistic form, irrespective of genre categories" (Biber and Finegan 1986: 20). Their identification therefore depends on the analysis of the predominant linguistic features of the texts, which in the case of Biber's studies is carried out through Factor Analysis. Nakamura (1989, 1991, 1994) makes the same distinction between "genre" and "text type," and uses a statistical method called "Extended HAYASHI's Quantification Method Type III" to describe text types in large corpora (Nakamura 1994: 141).

3. As a measure of the proportion of lexical versus grammatical words, I have used lexical density as defined by Stubbs (1986: 33, 1996: 172). Lexical density is expressed as a percentage and is calculated by subtracting the number of function words in a text from the number of running words (which gives the number of lexical words) and then dividing the result by the number of running words.

4. The word "texts" refers to the actual ASCII Text Files which make up a collection of texts within the corpus. In some cases (*The Dedalus Book of Surrealism* and *Fantastic Tales*), a book has been divided into a number of text files because it contains translations from varied source languages and authors.

5. The term "sentence" is simply used to refer to "the orthographic unit that is contained between full stops" (Halliday 1985: 193). "Sentence length" is the number of words that are comprised between full stops. The computer program used in the present analysis (*WordSmith Tools*) does not identify the full stops in decimals as sentence endings, but those included in abbreviations and acronyms are processed as markers of sentence boundaries. This does not appear to be a problem for the newspaper corpus, since both *The Guardian* and *The European* use open punctuation. On the other hand, different conventions apply to the publications belonging to the narrative corpus. A random check has revealed that translated texts tend to use full punctuation for the most common abbreviations (e.g. Mr. and Mrs.), while in the original texts open punctuation is more common. These differences may very well affect the results of the comparative analyses of average sentence length and will be taken into account when discussing the results (see section 6).

I am aware that the concept of sentence is not straightforward, nor universal. Although Halliday (1985: xxi) maintains that the sentence constitutes a "significant border post" to which writing systems are sensitive, he also recognises that "the sentence itself is an indeterminate category" (Halliday and Hasan 1976: 232). This point of view is substantiated by Baker"s observation that punctuation, which is highly developed in English and is used to indicate breaks in information flow, varies considerably among languages. In Arabic, for example, full stops are frequently found only at the end of paragraphs, so that sentences as such are extremely long and made up mostly of coordinate clauses (Baker 1992: 193, 215 and personal communication, 1996).

6. $t = -3.83978577$ ($p < 0.005$).

7. I have used the Mann-Whitney non-parametric test of significance. The result of this test is: $U = 68$ ($p < 0.025$).

8. $t = 1.886009763$ ($p < 0.1$).

9. The difference between these percentages is highly significant:

$Z = 59.96236118$ ($p < 0.001$).

10. $t = 1.379166319$ ($p < 0.1$).

11. The variance is a statistical measure of the variability or dispersion of scores around the average value. It indicates the degree to which a group lacks homogeneity, so that the higher is the value, the less homogeneous the group is.

12. The results of the F statistic are as follows:

$F(14,18) = 1.369980071$ ($p > 0.05$) for Lexical Density

$F(14,18) = 1.348710539$ ($p > 0.05$) for Proportion of High Frequency Words

$F(14,18) = 10.912147741$ ($p < 0.005$) for Mean Sentence Length.

13. $F(11,18) = 1.700881364$ ($p > 0.05$).

14. With regard to José Saramago's novels in particular, the late Giovanni Pontiero, Saramago's regular English translator, in his preface to *The History of the Siege of Lisbon* writes: "As in his other novels, Saramago's paragraph-long sentences, minimally interrupted by punctuation, challenge the reader to follow his continuous stream of thought, thus permitting a stronger sense of interaction and a more diverse interpretation of phrases and clauses."

15. The overall size of the translational newspaper collections is 74,791 words, whereas the size of the translational biography and fiction combined in the narrative subcorpus is 999,945 words.

16. 102 translated newspaper articles were selected from *The Guardian*; 64 from *The European*. The number of translated narrative works is 14.

17. The program used for these analyses is *WordSmith Tools*. It can be downloaded from: http://www.oup.co.uk, which is OUP's World Wide Web address. It is designed by Mike Scott of the Department of English Language and Literature, University of Liverpool.

REFERENCES

BAKER, Mona (1992): *In Other Words: A Coursebook on Translation*, London and New York, Routledge.

BIBER, Douglas and Edward FINEGAN (1986): "An Initial Typology of English Text Types", Jan Aarts and Willen Meijs (Eds), *Corpus Linguistics II: New Studies in the Analysis and Exploitation of Computer Corpora*, Amsterdam, Rodopi, pp. 19-46.

BIBER, Douglas and Edward FINEGAN (1991): "On the Exploitation of Computerized Corpora in Variation Studies", Karin Aijmer and Bengt Altenberg (Eds), *English Corpus Linguistics*, London and New York, Longman, pp. 204-220.

HALLIDAY, M. A. K. (1985): *An Introduction to Functional Grammar*, London, Edward Arnold.

HALLIDAY, M. A. K. and R. HASAN (1976): *Cohesion in English*, London and New York, Longman.

LAVIOSA, Sara (1997): "How Comparable can 'Comparable Corpora' Be?", *Target*, 9 (2), pp. 289-319.

LAVIOSA, Sara (1998): "The English Comparable Corpus: a Resource and a Methodology", Lynne Bowker, Michael Cronin, Dorothy Kenny and Jennifer Pearson (Eds), *Unity in Diversity? Current Trends in Translation Studies*, Manchester, St. Jerome Publishing.

LAVIOSA-BRAITHWAITE, Sara (1996): *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*, PhD Thesis, Manchester, UMIST.

LAVIOSA-BRAITHWAITE, Sara (1997): "Investigating Simplification in an English Comparable Corpus of Newspaper Articles", Kinga Klaudy and János Kohn (Eds), *Transferre Necesse Est*, Proceedings of the Second International Conference on Current Trends in Studies of Translation and Interpreting 5-7 September, 1996, Budapest, Hungary, Budapest, Scholastica, pp. 531-540.

NAKAMURA, Junsaku (1989): "A Quantitative Study on the Use of Personal Pronouns in the Brown Corpus", *Jacet Bulletin*, Vol. 20, pp. 51-71.

NAKAMURA, Junsaku (1991): "The Relationships among Genres in the LOB Corpus Based upon the Distribution of Grammatical Tags", *Jacet Bulletin*, Vol. 22, pp. 55-74.

NAKAMURA, Junsaku (1994): "Extended HAYASHI"s Quantification Method Type III and its Applications in Corpus Linguistics", *Journal of Language and Literature*, Vol. 1, March 1994, pp. 141-192.

STUBBS, Michael (1986): "Lexical Density: A Technique and Some Findings", Michael Coulthard (Ed.), *Talking about Text. Discourse Analysis*, Monograph No 13, English Language Research, Birmingham, University of Birmingham, pp. 27-42.

STUBBS, Michael (1996): *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*, Oxford and Cambridge, MA, Blackwell.

TOURY, Gideon (1995): *Descriptive Translation Studies and beyond*, Amsterdam and Philadelphia, John Benjamins.

WordSmith Tools (1996): Copyright, Michael Scott, Oxford, Oxford University Press.

APPENDIX 1

LIST HEAD — TEC NARRATIVE PROSE

	WORD	Freq.	%				
				18	NOT	7026	(0.7%)
				19	YOU	7017	(0.7%)
1	THE	58734	(5.9%)	20	ON	6657	(0.7%)
2	TO	28760	(2.9%)	21	HAD	6627	(0.7%)
3	AND	28305	(2.8%)	22	ME	6378	(0.6%)
4	OF	25893	(2.6%)	23	BUT	6044	(0.6%)
5	A	20587	(2.1%)	24	HER	5872	(0.6%)
6	I	20205	(2.0%)	25	BE	5463	(0.5%)
7	IN	16538	(1.7%)	26	AT	5301	(0.5%)
8	THAT	13076	(1.3%)	27	THEY	5146	(0.5%)
9	WAS	11285	(1.1%)	28	SHE	5120	(0.5%)
10	HE	10443	(1.0%)	29	THIS	5104	(0.5%)
11	IT	10232	(1.0%)	30	FROM	4906	(0.5%)
12	IS	8557	(0.9%)	31	HAVE	4854	(0.5%)
13	MY	8194	(0.8%)	32	ONE	4575	(0.5%)
14	WITH	8159	(0.8%)	33	WHICH	4352	(0.4%)
15	HIS	7685	(0.8%)	34	BY	3943	(0.4%)
16	FOR	7488	(0.7%)	35	WERE	3866	(0.4%)
17	AS	7426	(0.7%)	36	WOULD	3862	(0.4%)

37	SO	3769	(0.4%)	73	NOW	1721	(0.2%)
38	HIM	3763	(0.4%)	74	THOSE	1717	(0.2%)
39	WHO	3749	(0.4%)	75	HOW	1701	(0.2%)
40	IF	3745	(0.4%)	76	YOUR	1685	(0.2%)
41	THERE	3691	(0.4%)	77	ANY	1677	(0.2%)
42	WE	3691	(0.4%)	78	KNOW	1657	(0.2%)
43	NO	3591	(0.4%)	79	ITS	1506	(0.2%)
44	ALL	3530	(0.4%)	80	MAN	1486	(0.1%)
45	WHAT	3442	(0.3%)	81	OUR	1486	(0.1%)
46	OR	3283	(0.3%)	82	WITHOUT	1472	(0.1%)
47	WHEN	3202	(0.3%)	83	LIFE	1422	(0.1%)
48	ARE	3198	(0.3%)	84	WHERE	1412	(0.1%)
49	AN	3044	(0.3%)	85	SEE	1411	(0.1%)
50	THEIR	2695	(0.3%)	86	AFTER	1399	(0.1%)
51	LIKE	2553	(0.3%)	87	US	1390	(0.1%)
52	THEM	2529	(0.3%)	88	NEVER	1374	(0.1%)
53	OUT	2512	(0.3%)	89	THAN	1365	(0.1%)
54	COULD	2451	(0.2%)	90	DAY	1303	(0.1%)
55	UP	2442	(0.2%)	91	LITTLE	1297	(0.1%)
56	WILL	2341	(0.2%)	92	BEFORE	1249	(0.1%)
57	MORE	2315	(0.2%)	93	MUCH	1248	(0.1%)
58	THEN	2279	(0.2%)	94	TWO	1220	(0.1%)
59	ABOUT	2269	(0.2%)	95	BACK	1215	(0.1%)
60	ONLY	2234	(0.2%)	96	AM	1210	(0.1%)
61	BEEN	2192	(0.2%)	97	SAY	1192	(0.1%)
62	TIME	2190	(0.2%)	98	DOWN	1183	(0.1%)
63	HAS	2084	(0.2%)	99	PEOPLE	1160	(0.1%)
64	SOME	1994	(0.2%)	100	STILL	1159	(0.1%)
65	DO	1980	(0.2%)	101	THESE	1159	(0.1%)
66	INTO	1939	(0.2%)	102	OVER	1157	(0.1%)
67	EVEN	1915	(0.2%)	103	VERY	1136	(0.1%)
68	SAID	1898	(0.2%)	104	LOVE	1135	(0.1%)
69	BECAUSE	1875	(0.2%)	105	JUST	1131	(0.1%)
70	CAN	1844	(0.2%)	106	FIRST	1127	(0.1%)
71	DID	1838	(0.2%)	107	GO	1120	(0.1%)
72	OTHER	1777	(0.2%)	108	COME	1094	(0.1%)

TOTAL : 56.2%

APPENDIX 2

LEMMATISED LIST HEAD — TEC NARRATIVE PROSE

1	THE		ITS
2	TO	12	WITH
3	AND	13	FOR
4	OF	14	AS
5	A	15	NOT
	AN	16	YOU
6	I		YOUR
	ME	17	ON
	MY	18	HAD
7	IN		HAVE
8	THAT		HAS
9	WAS	19	BUT
	IS	20	HER
	BE		SHE
	WERE	21	AT
	ARE	22	THEY
	BEEN		THEIR
	AM		THEM
10	HE	23	THIS
	HIS		THOSE
	HIM		THESE
11	IT	24	FROM

25	ONE	52	SAID
26	WHICH		SAY
27	BY	53	BECAUSE
28	WOULD	54	OTHER
	WILL	55	NOW
29	SO	56	HOW
30	WHO	57	ANY
31	IF	58	KNOW
32	THERE	59	MAN
33	WE	60	WITHOUT
	OUR	61	LIFE
	US	62	WHERE
34	NO	63	SEE
35	ALL	64	AFTER
36	WHAT	65	NEVER
37	OR	66	THAN
38	WHEN	67	DAY
39	LIKE	68	LITTLE
40	OUT	69	BEFORE
41	COULD	70	MUCH
	CAN	71	TWO
42	UP	72	BACK
43	MORE	73	DOWN
44	THEN	74	PEOPLE
45	ABOUT	75	STILL
46	ONLY	76	OVER
47	TIME	77	VERY
48	SOME	78	LOVE
49	DO	79	JUST
	DID	80	FIRST
50	INTO	81	GO
51	EVEN	82	COME

APPENDIX 3

LIST HEAD — NON-TEC NARRATIVE PROSE

	WORD	**Freq.**	**0%**				
				25	IS	3713	-0.50%
				26	BE	3558	-0.50%
1	THE	42586	-5.80%	27	HIM	3354	-0.50%
2	AND	22178	-3.00%	28	SAID	3118	-0.40%
3	OF	19512	-2.70%	29	FROM	3073	-0.40%
4	TO	18337	-2.50%	30	THIS	3045	-0.40%
5	A	16763	-2.30%	31	WERE	2832	-0.40%
6	HE	12358	-1.70%	32	ALL	2817	-0.40%
7	IN	11965	-1.60%	33	HAVE	2721	-0.40%
8	WAS	11703	-1.60%	34	THERE	2677	-0.40%
9	IT	8658	-1.20%	35	WOULD	2657	-0.40%
10	I	8359	-1.10%	36	BY	2579	-0.40%
11	HIS	7913	-1.10%	37	OR	2564	-0.40%
12	THAT	7836	-1.10%	38	WHICH	2416	-0.30%
13	HAD	6947	-1.00%	39	MY	2322	-0.30%
14	SHE	5898	-0.80%	40	WHAT	2312	-0.30%
15	FOR	5437	-0.70%	41	OUT	2309	-0.30%
16	WITH	5356	-0.70%	42	ONE	2257	-0.30%
17	ON	5320	-0.70%	43	UP	2155	-0.30%
18	HER	5102	-0.70%	44	WE	2143	-0.30%
19	AS	4951	-0.70%	45	SO	2130	-0.30%
20	YOU	4801	-0.70%	46	AN	2115	-0.30%
21	AT	4799	-0.70%	47	BEEN	2050	-0.30%
22	BUT	4391	-0.60%	48	LIKE	2035	-0.30%
23	NOT	4332	-0.60%	49	COULD	2033	-0.30%
24	THEY	3981	-0.50%	50	NO	2000	-0.30%

51	THEM	1982	-0.30%	81	WELL	946	-0.10%
52	ME	1913	-0.30%	82	EVEN	929	-0.10%
53	WHEN	1907	-0.30%	83	NEVER	928	-0.10%
54	THEIR	1812	-0.20%	84	TWO	926	-0.10%
55	IF	1803	-0.20%	85	VERY	920	-0.10%
56	INTO	1766	-0.20%	86	AFTER	894	-0.10%
57	THEN	1745	-0.20%	87	DON'T	894	-0.10%
58	ABOUT	1743	-0.20%	88	OTHER	883	-0.10%
59	WHO	1740	-0.20%	89	YOUR	881	-0.10%
60	MORE	1571	-0.20%	90	GO	868	-0.10%
61	ARE	1448	-0.20%	91	HOW	862	-0.10%
63	BACK	1393	-0.20%	92	IT'S	851	-0.10%
64	DID	1383	-0.20%	93	MUCH	851	-0.10%
65	TIME	1364	-0.20%	94	THINK	841	-0.10%
66	NOW	1305	-0.20%	95	CAME	839	-0.10%
67	DOWN	1285	-0.20%	96	FIRST	832	-0.10%
68	SOME	1238	-0.20%	97	GET	824	-0.10%
69	OVER	1168	-0.20%	98	OWN	820	-0.10%
70	JUST	1114	-0.20%	99	LITTLE	819	-0.10%
71	ONLY	1058	-0.10%	100	MADE	814	-0.10%
72	KNOW	1043	-0.10%	101	BEFORE	813	-0.10%
73	MAN	1040	-0.10%	102	OFF	813	-0.10%
74	THOUGHT	1035	-0.10%	103	AGAIN	811	-0.10%
75	THAN	1029	-0.10%	104	STILL	811	-0.10%
76	SEE	1022	-0.10%	105	LIFE	809	-0.10%
77	WHERE	1022	-0.10%	106	WENT	803	-0.10%
78	WAY	1004	-0.10%	107	MOTHER	790	-0.10%
79	TOO	982	-0.10%	108	PEOPLE	790	-0.10%
80	ITS	981	-0.10%				

TOTAL : 51.60%

APPENDIX 4

LEMMATISED LIST HEAD — NON-TEC NARRATIVE PROSE

1	THE	15	WITH
2	AND	16	ON
3	OF	17	AS
4	TO	18	YOU
5	A		YOUR
	AN	19	AT
6	HE	20	BUT
	HIS	21	NOT
	HIM	22	THEY
7	IN		THEM
8	WAS		THEIR
	IS	23	SAID
	BE	24	FROM
	WERE	25	THIS
	BEEN	26	ALL
	ARE	27	THERE
9	IT	28	WOULD
	ITS	29	BY
	IT'S	30	OR
10	I	31	WHICH
	MY	32	WHAT
	ME	33	OUT
11	THAT	34	ONE
12	HAD	35	UP
	HAVE	36	WE
13	SHE	37	SO
	HER	38	LIKE
14	FOR	39	COULD

40	NO
41	WHEN
42	IF
43	INTO
44	THEN
45	ABOUT
46	WHO
47	MORE
48	DO
	DID
	DON'T
49	BACK
50	TIME
51	NOW
52	DOWN
53	SOME
54	OVER
55	JUST
56	ONLY
57	KNOW
58	MAN
59	THOUGHT
	THINK
60	THAN
61	SEE
62	WHERE
63	WAY
64	TOO
65	WELL
66	EVEN
67	NEVER
68	TWO
69	VERY
70	AFTER
71	OTHER
72	GO
	WENT
73	HOW
74	MUCH
75	CAME
76	FIRST
77	GET
78	OWN
79	LITTLE
80	MADE
81	BEFORE
82	OFF
83	AGAIN
84	STILL
85	LIFE
86	MOTHER
87	PEOPLE