The features of non-literary translated language: a pilot study

Haidee Kruger, Bertus van Rooy

North-West University, Vaal Triangle Campus

1. Introduction

Since Baker's (1993, 1995, 1996) first suggestion in the mid-1990s that comparable corpora of translated and non-translated texts in the same language may be used to investigate the features of translated language that are independent of source-language influence, there have been numerous research studies exploring this proposition. Most of these studies utilise Baker's (1996:176-177) fourfold categorisation of hypothesised features of translated language (simplification, explicitation, normalisation/conservatism, and levelling out) operationalising the investigation of the hypothesised features in various ways. Much of this research has focused on translated English, specifically in the British context, but there have also been numerous studies focusing on other languages. However, to our knowledge, the features of translated English produced in contexts other than the British have rarely been investigated (Williams, 2004 is an exception).

An overview of existing studies of the features of translated language suggests support for some of the initial hypotheses. Studies of **explicitation** on the linguistic as well as content levels (e.g. Mutesayire, 2004; Olohan, 2003; Olohan & Baker, 2000; Williams, 2004) have found evidence for the more frequent use of more explicit surface realisations as well as more explicit presentation of propositional relationships in translated language. Studies such as that of Kenny (2001), with a parallel corpus, and Williams (2004) and Baker (2007) with comparable corpora, have found that translated language tends to be more **conservative or normalised** in terms of lexical features. **Simplification** has been investigated by, amongst others, Laviosa (1998) and Williams (2004), using measures of lexical variety, lexical density, and other measures of complexity (such as mean sentence length). These studies have found some evidence to suggest that simplification is a feature of translated language; however, findings have not been consistent. The last feature, **levelling out**, has not been extensively investigated, and the investigations which have been done have yielded inconclusive findings (Olohan, 2004:100). This feature is defined by Baker (1996:177, 184) as the tendency of translated language to converge or "cluster" around the centre of a continuum, seeking the middle ground between extremes.

An overview of important studies in the field leads to a number of observations begging further investigation. Firstly, the hypothesised features have most frequently been investigated in isolation, or in restricted combination, with limited attention to the co-occurrence of features. Furthermore, many of the existing studies have made use of corpora of literary or other imaginative texts (though there are exceptions), and there have not been many attempts to investigate the relationship of register to the hypothesised features of translated language in a systematic way. The pilot study reported on in this paper is a first attempt to investigate the two questions that arise from the above: (1) What are the occurrence patterns for the different hypothesised features of translated language, investigated together? (2) What is the relationship between register and the features of translated language?

Based on existing research, and adding some additional features, the features set out in Table 1 were selected for investigation.

Table 1. Features selected for investigation

Feature category	**Feature subcategory**	**Feature**
1. Explicitation	1.1 More complete/less economical surface realisation in translation	A. Frequency of use of optional complementiser *that*
		B. Frequency of use of full forms rather than contracted forms
	1.2 More explicit relations between conceptual propositions in text	C. Frequency of linking adverbials
2. Normalisation/conservatism		D. Frequency of coinages and loanwords
		E. Frequency of lexical bundles
3. Simplification		F. Lexical diversity
		G. Mean word length

It was, in the first instance, hypothesised that the occurrence of these linguistic features would demonstrate significant differences in the two corpora utilised in the study, reflecting overall more explicit, more conservative, and simplified language use in the corpus of translated English than in the comparable corpus of original English writing. As a starting point for factoring in the variable of register, it was further hypothesised that the frequency of these features in the translation corpus would show no significant effect for the relationship between corpus and register – in other words, the translation-related features would not be strongly linked to register variation. This has the collateral effect of suggesting a broader hypothesis that in the translation corpus less register variation, or sensitivity to register, will occur, as a result of the interference of translation. This links to the proposed feature of levelling out, which in this study is therefore operationalised by investigating variation in the distribution of the features of translated language across register.

2. Methodology

2.1 Corpus composition

The translation corpus consists of texts translated into English in South Africa, mostly from Afrikaans, but also from some other European languages. The texts were obtained from the language-service offices of two South African universities, the North-West University (Vaal Triangle Campus) and Stellenbosch University, and from a private language-service agency. The texts are all full texts, and vary in length from about 50 words to about 20 000 words. The shorter texts were combined into longer text units, to avoid the analysis problems associated with very short texts. The aim was to work with text units of no less than 1000 words. Care was taken to combine similar text types.

Published texts as well as ephemera are included in the corpus, and various levels of translation expertise are represented. The texts included date from 1998 to 2009. The corpus includes translations by first- and second-language speakers of English. Lastly, and importantly for the purposes of the current investigation, a variety of text types are represented, including popular spiritual books, minutes, memoranda, reports, pharmaceutical information sheets, academic articles, newspaper reports, instructions, formal letters and policy documents. In this paper, the translation corpus will be referred to as the Corpus of Translated English, or CTE.

The comparable corpus of original English writing produced in South Africa used as reference corpus was compiled from available written texts in the International Corpus of

English (ICE) for South Africa (see ICE, 2010 for more information). The written part of the ICE corpus for South Africa (ICE-SA) consists of texts produced mainly by native English speakers, or otherwise educated users of the standard variety. At the time of the study, ICE-SA was not yet complete. A sample of texts that had already been edited was used to compile the reference corpus of original writing for this investigation, subject to the same procedure as the CTE-texts. In other words, where text units were very short, they were combined into bigger text units of at least 1000 words.

All the ICE corpora share the same design (see ICE, 2010), and include text types such as academic articles and books, news editorials, newspaper reports, formal letters, instructional administrative material, popular books and instructional material associated with skills and hobbies. The subset of the ICE corpus used in this paper was selected to be as comparable as possible with the CTE corpus, within the constraints of the availability of texts. The subset of the ICE corpus used as reference corpus is referred to as ICE-SA in this paper.

Table 2. Corpus composition

Register	**Translation corpus** **CTE**	**Reference corpus** **ICE-SA**
Academic	27310	13856
Instructional	108866	36461
Letters	19677	15672
Persuasive	43012	30720
Popular	38984	54577
Reportage	34000	41239
TOTAL	**271849**	**192525**

The texts in both corpora were categorised according to register, using the standard ICE labels. In the written register, the following six registers from ICE-SA were selected since texts with a similar register were included in the CTE corpus: letters, academic writing, popular writing, reportage, instructional writing and persuasive writing.

2.2 Data collection and processing

Data were collected using the Concord and WordList functions in WordSmith Tools, and a combination of automatic and manual sorting.

2.2.1 Frequency of use of the optional complementiser *that*

Verbs taking a *that* complement clause were used as search nodes. All the verbs classified by Biber *et al.* (1999:663-666) as notably common and relatively common verbs were selected for investigation. These verbs are shown in Table 3.

Table 3. Verbs used as search nodes for the investigation of *that* omission

	Mental verbs	**Speech act verbs**	**Other communication verbs**
Notably common (more than 100 instances per million words)	BELIEVE FEEL FIND GUESS KNOW SEE THINK	SAY	SHOW SUGGEST

Relatively common (more than 20 instances per million words)	ASSUME CONCLUDE DECIDE DOUBT EXPECT HEAR HOPE IMAGINE MEAN NOTICE READ REALIS(Z)E RECOGNIS(Z)E REMEMBER SUPPOSE UNDERSTAND WISH	ADMIT AGREE ANNOUNCE ARGUE BET INSIST TELL	ENSURE INDICATE PROVE

Concordances were drawn for all verb forms, in both corpora. Irrelevant entries were manually discarded, and concordance lines were then manually sorted and tagged to identify the instances of VERB + *that* and VERB + *zero*.

2.2.2 Frequency of use of full forms rather than contracted forms

It was decided to investigate only verb contractions (with pronouns) and negative contractions, and only those forms which do occur at least once in contracted form in a corpus. The list of contractions and full forms investigated is presented in Table 4.

Table 4. Contractions and full forms investigated

CTE		**ICE-SA**	
Contractions	**Full forms**	**Contractions**	**Full forms**
aren’t	are not	aren’t	are not
can’t	cannot	can’t	cannot
couldn’t	could not	couldn’t	could not
didn’t	did not	didn’t	did not
doesn’t	does not	don’t	do not
don’t	do not	doesn’t	does not
hadn’t	had not	hadn’t	had not
hasn’t	has not	hasn’t	has not
haven’t	have not	haven’t	have not
he’ll	he will	he’s	he is
he’s	he is	I’m	I am
here’s	here is	I’ve	I have
I’d	I would	I’ll	I shall, I will
I’ll	I shall, I will	I’d	I would
I’m	I am	isn’t	is not
I’ve	I have	it’s	it has, it is
isn’t	is not	let’s	let us
it’s	it is	mustn’t	must not

let’s	let us	she’d	she had
mustn’t	must not	she’s	she has, she is
needn’t	need not	she’ll	she will
she’ll	she will	shouldn’t	should not
she’s	she is	that’s	that is
shouldn’t	should not	there’s	there is
that’s	that is	there’ll	there will
there’s	there is	there’d	there would
they’ll	they will	they’re	they are
wasn’t	was not	they’d	they would
we’d	we would	wasn’t	was not
we’ll	we shall, we will	we’re	we are
we’re	we are	we’d	we had
we’ve	we have	we’ve	we have
weren’t	were not	we’ll	we shall, we will
what’s	what is	what’s	what is
who’re	who are	who’d	who had
who’s	who is	who’s	who is
won’t	will not	won’t	will not
wouldn’t	would not	wouldn’t	would not
you’d	you had, you would	you’re	you are
you’ll	you will	you’ve	you have
you’re	you are	you’ll	you will
you’ve	you have		

2.2.3 Frequency of linking adverbials

For the purposes of this pilot, two kinds of linking adverbials were chosen for investigation (apposition and result/inference, see Biber *et al.*, 1999:875-879), on the basis that they appear to be the most likely kinds of linking adverbials translators might use to explicitise the relationships between the propositions in the text. For each category, typical adverbials were selected, as set out in Table 5.

Table 5. Linking adverbials selected for investigation (from Biber *et al.*, 1999:875-879, and Mutesayire, 2004)

Appositive linking adverbials	**Linking adverbials of result/inference**
in other words that is i.e. that is to say which is to say namely to be exact to be precise to be more exact to be more precise	therefore consequently thus as a result hence as a consequence in consequence

2.2.4 Frequency of coinages and unlexicalised loanwords

Since coinages and loanwords are likely to occur infrequently, it was decided to use hapax legomena as an initial search set. Hapaxes in each corpus were extracted. The spelling checker in Microsoft Word was used as a first measure to remove all lexicalised entries.

Following this, all proper nouns, acronyms, abbreviations, spelling errors, parts of e-mails, etc. were deleted. The remaining entries were checked using the online dictionary Wordweb as well as the Internet. All lexicalised items were removed from the list, and remaining entries were tagged as coinages or loanwords.

2.2.5 Frequency of common lexical bundles

In order to determine a set of search terms, a list of trigrams in each of the two corpora was generated. A selection from the full lists was then made. We deleted all trigrams occuring with a frequency of less than 0.01% in each full corpus and in fewer than 2% of texts in each corpus, as well as all trigrams containing proper nouns and clearly subject-specific words (e.g. “university”).

The reduced lists from both corpora were combined, and trigrams common to both lists were selected for investigation. This yielded the 23 trigrams set out in Table 6.

Table 6. Trigrams selected for investigation

a number of
as well as
at the end
be able to
have to be
in order to
in terms of
it is a
it is not
members of the
one of the
part of the
some of the
terms of the
the end of
the fact that
the right to
the use of
there is a
there is no
to ensure that
will have to
would like to

2.2.6 Lexical diversity

Lexical diversity was measured using type-token ratio, as computed by WordSmith Tools. Since the text lengths were quite varied, standardised type-token ratio was used. This was done by recalculating the type-token ratio for every 1000 words in the text, and then computing the average for the entire text.

2.2.7 Mean word length

Mean word length in the two corpora as computed by WordSmith Tools was used.

2.2.8 Statistical analysis

In the analysis of results, ANOVAs (Analysis of Variance) were used. Throughout the analysis, we sought to determine if there were any significant main effects for either corpus (CTE or ICE-SA) or register. Such main effects for corpus would be straightforward evidence for significant differences between the corpora. Beyond the main effects, we also looked for significant interactions between the two factor variables corpus and register, to determine if the differences between the CTE and ICE-SA were specific to some registers, but not to others. In all cases, two levels of statistical significance were used as guidelines, $p<0.05$ and $p<0.001$.

3. Results and discussion

3.1 Frequency of the optional complementiser *that*

Clauses with the complementiser *that* present (see Figure 1) do no show significant main effects for either individual registers (F(5, 202)=0.52, p=0.76) or for the two corpora (F(1, 202)=0.15, p=0.70), nor is there a significant interaction between the two factors (F(5, 202)=0.56, p=0.73). Overall, the mean frequencies of *that* complement clauses are quite similar, at 2.51 (sd=7.07) for CTE and 2.07 (sd=1.41) for ICE-SA. However, the picture is not complete until we look at the results for clauses in which the *that* is omitted.

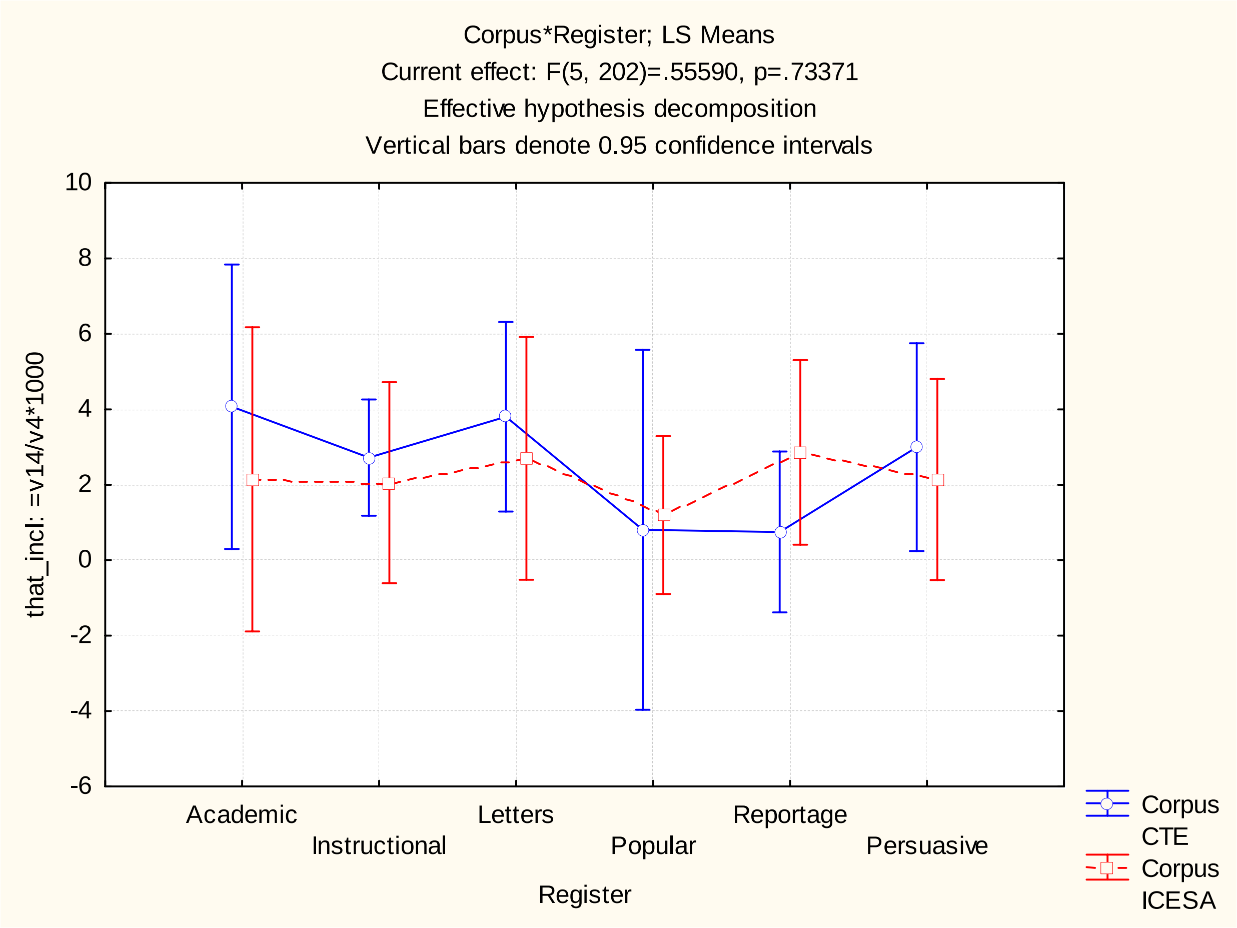

Figure 1. Mean frequencies of *that* complement clauses in the two corpora and six registers

For clauses where *that* is omitted (see Figure 2), the interaction between the factor variables corpus and register yields a significant effect (F(5, 202)=8.96, p<0.001). There are also significant main effects for the two factors on their own, with ICE-SA containing significantly more instances of omitted *that* complement clauses. In CTE, the mean frequency of omitted *that* clauses is a mere 0.29 (sd=0.86), which rises to 1.40 (sd=2.06) in ICE-SA. This is a very significant main effect, with F(1, 202)=22.56 (p<0.001). While registers also show a significant main effect (F(5, 202)=5.63, p<0.001), it is in particular the registers of reportage and persuasive writing, and to a lesser extent business letters and popular writing where the CTE avoids omitting the *that* complementiser consistently, while ICE-SA makes relatively more liberal use of this option. The most spectacular result, for reportage, has to be taken with a little caution, however, since the option of finite complement clauses is not a very frequent one in the corpus. Clauses with the *that* complement present are also much less frequent in the CTE. However, for the other three registers with higher frequencies of omitted *that* clauses, the effect is real, since the frequencies of the unreduced full clause with overt complementiser are quite similar.

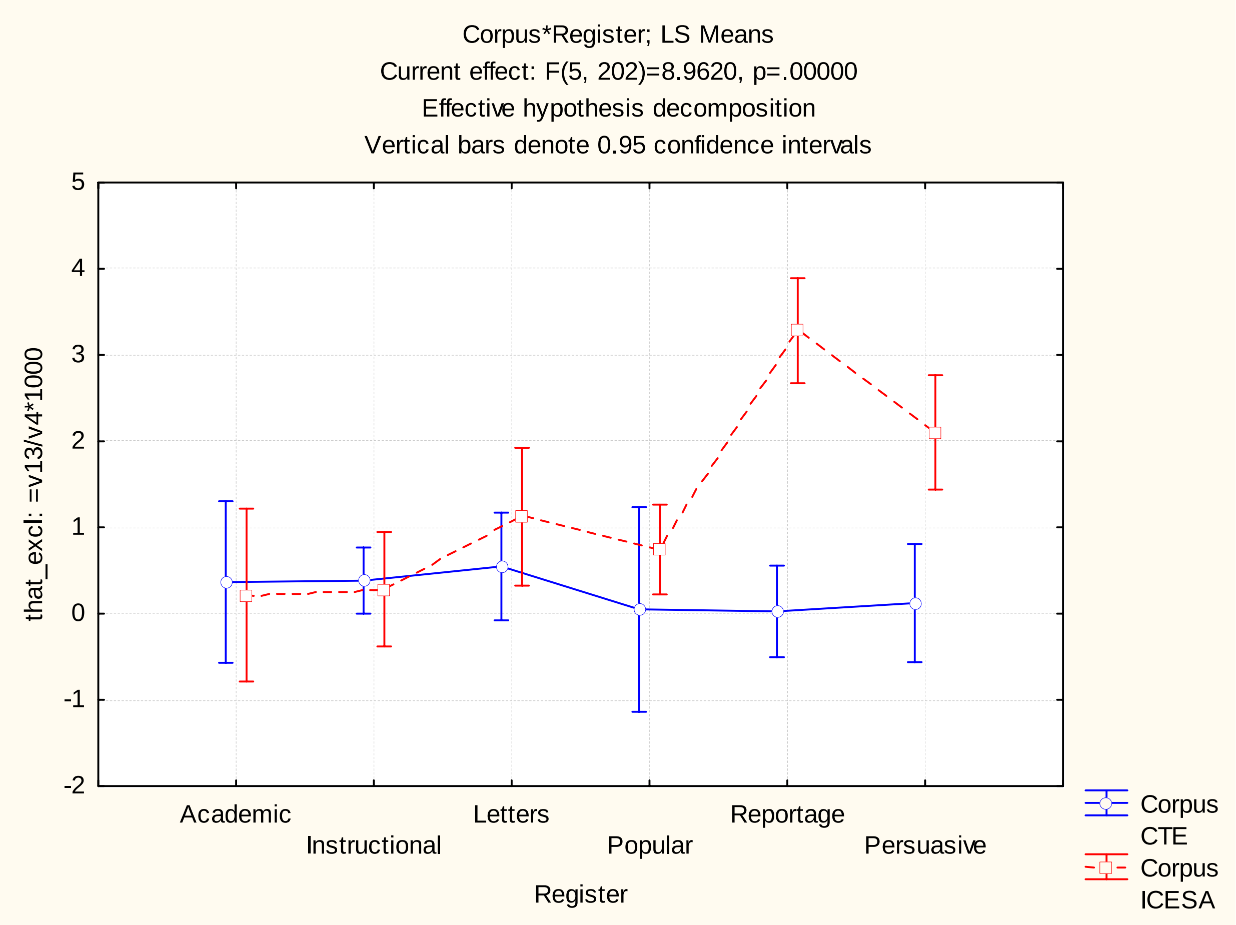

Figure 2. Mean frequencies of finite complement clauses omitting that in the two corpora and six registers

3.2 Frequency of full forms rather than contracted forms

The results for contracted forms (see Figure 3) fail to reach statistical significance (F(5, 202)=1.98, p=.08), although a few noticeable differences can be detected. The much higher frequency of the contracted forms in the CTE register of persuasive writing has to be seen against a similar high frequency of uncontracted alternatives in the same register (see Figure 4). However, the much higher difference of contracted forms in popular writing in ICE-SA is not offset by the higher frequency of uncontracted forms in the CTE, and hence indicates a difference between the two corpora in this register only, but the effect needs to be understood against the wide standard deviation and does not provide a general characterisation of all texts in the popular writing register, only a subgroup. Apart from letters and persuasive writing, the CTE tends to avoid contractions, while the pattern for ICE-SA is a little more varied.

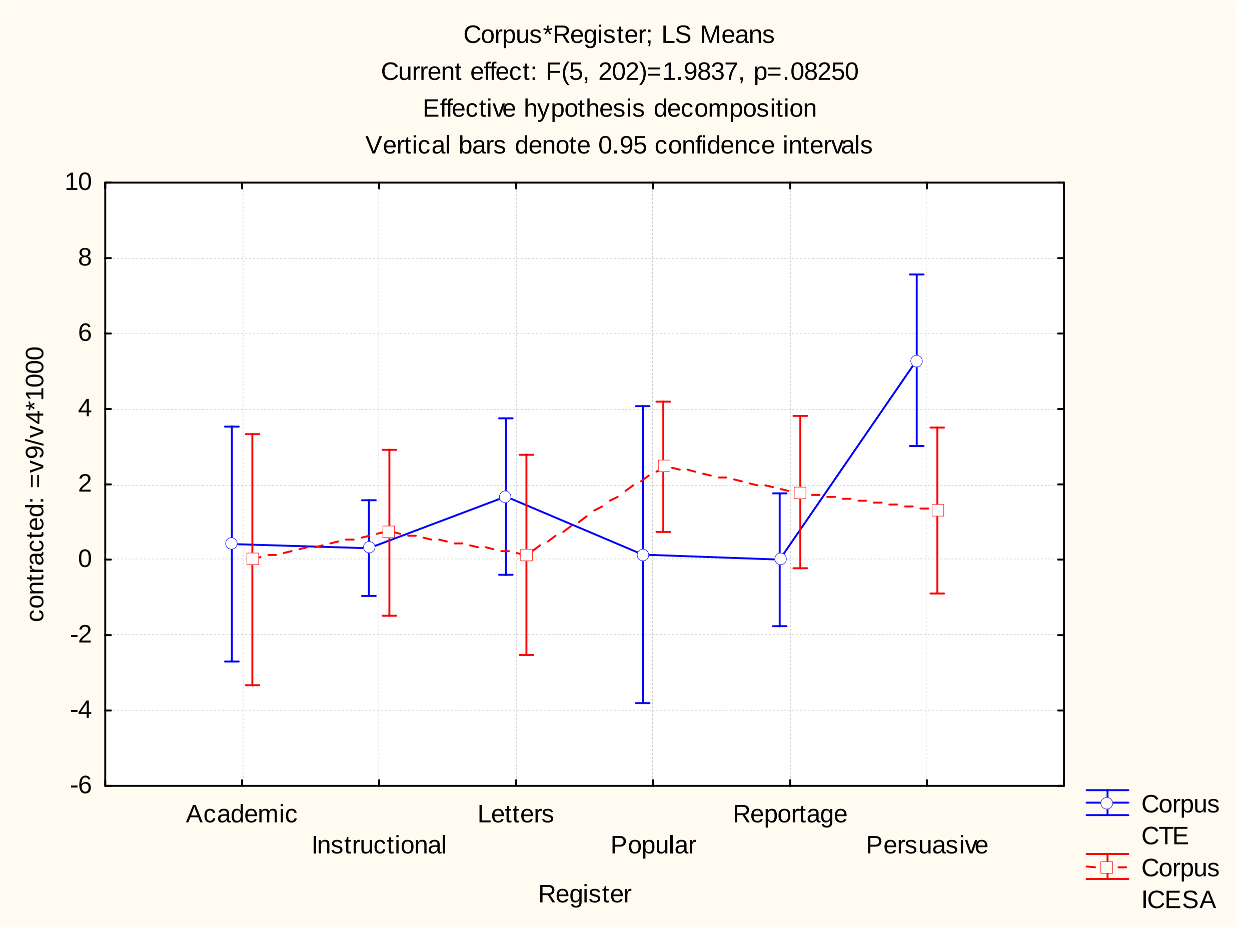

Figure 3. Mean frequencies of contracted forms in the two corpora and six registers

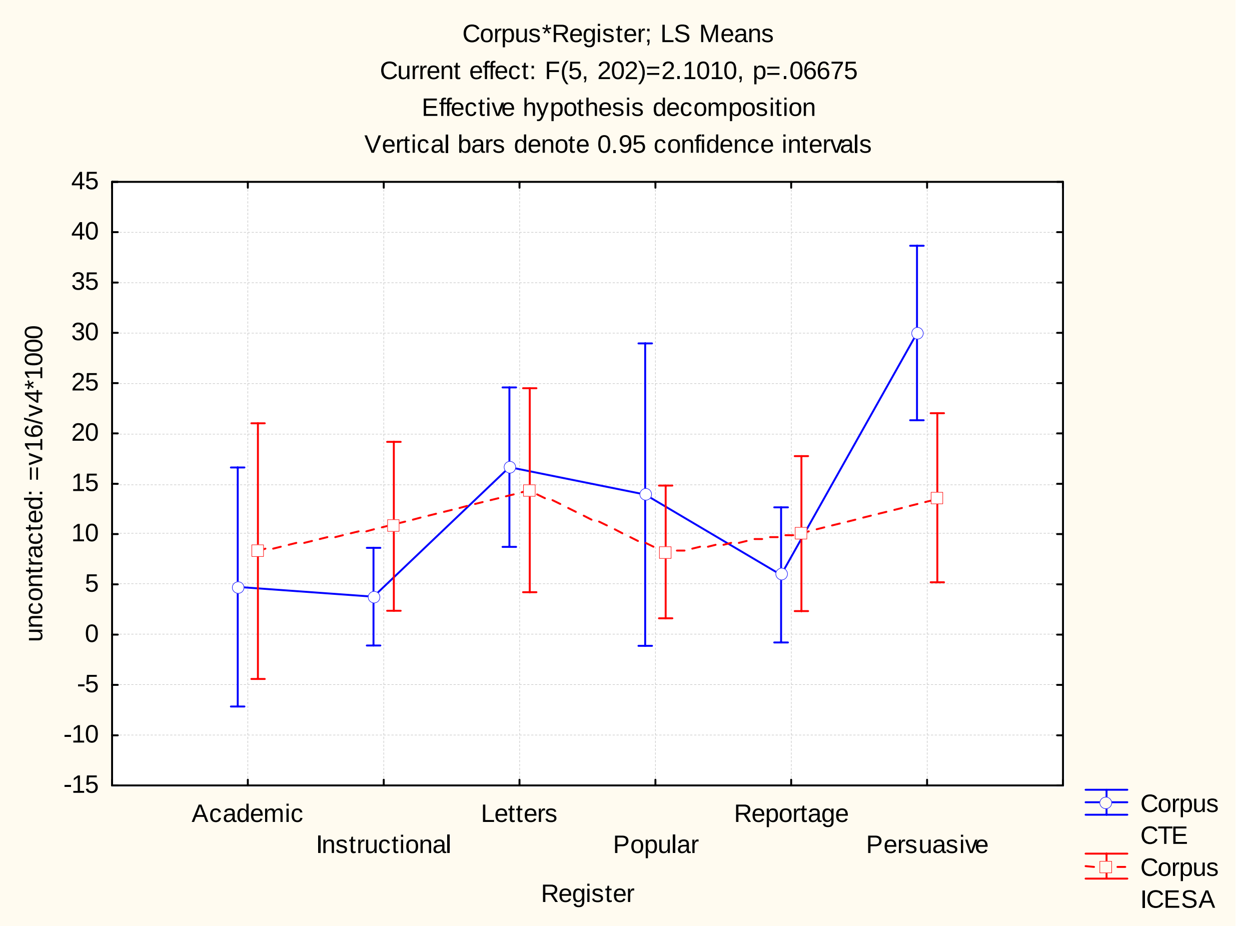

Figure 4. Mean frequencies of uncontracted forms in the two corpora and six registers

Thus, while one may conclude some difference in contracted forms, and would like to claim that the CTE confirms the overall hypothesis of less varied registers, using an ANOVA, the interactions between the factors corpus and register are found not to be significant. The null hypothesis of no difference cannot be rejected.

3.3 Frequency of linking adverbials

The frequency of linking adverbials (see Figure 5) shows a main effect for register (F(5, 202)=10.73, p<0.001), with the academic register using linking adverbials much more frequently than any other register, and instructional writing also making more use of this resource than the remaining registers. Overall, the CTE has slightly more linking adverbials than ICE-SA – an average of 1.18 per thousand words (sd=1.28) against 0.86 (sd=1.13), but the main effect is not statistically significant (F(1, 202)=1.14, p=0.29). The interaction between corpus and register fails to reach statistical significance too (Current effect: F(5, 202)=0.60, p=0.65).

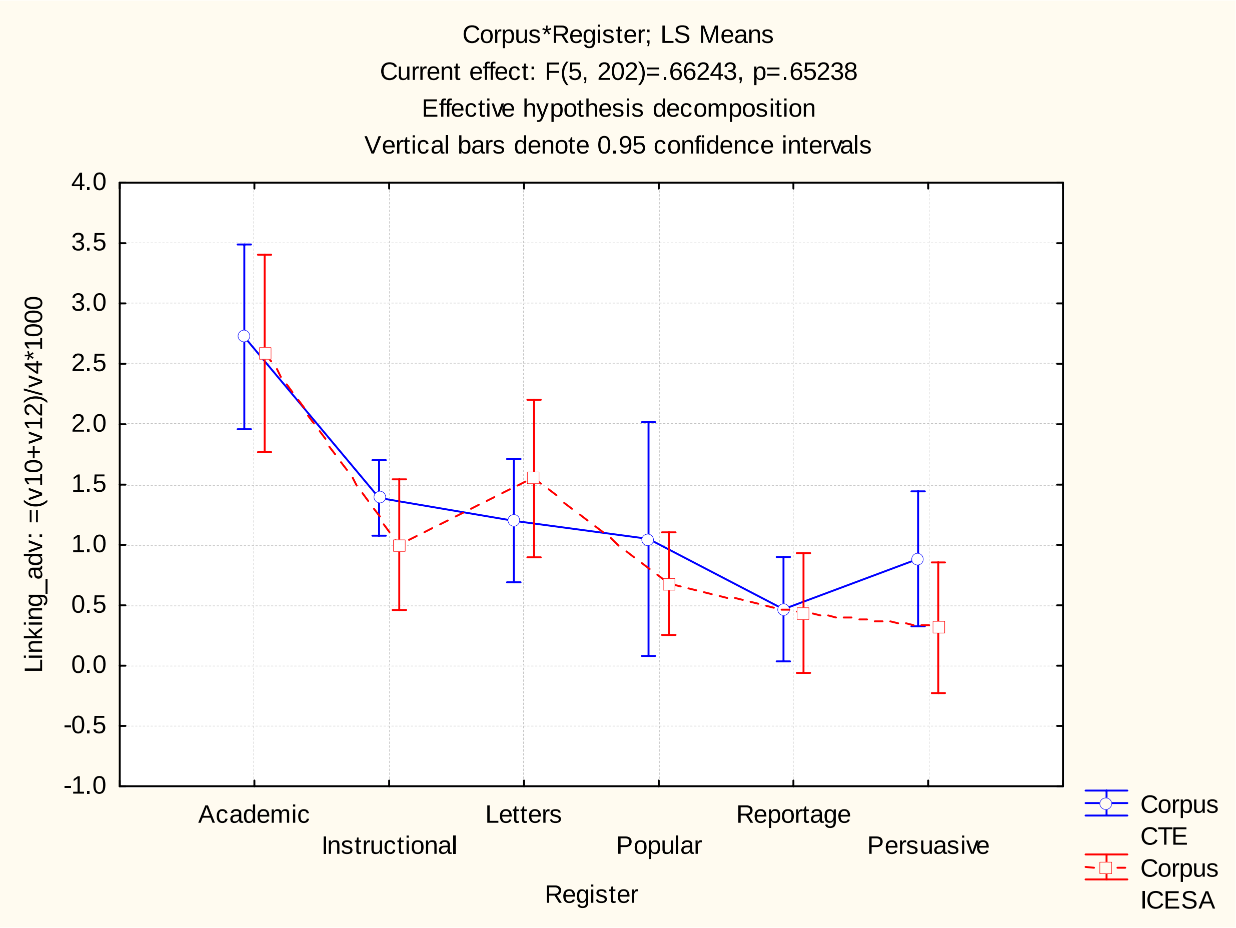

Figure 5. Mean frequencies of linking adverbials in the two corpora and six registers

Figure 5 makes it clear how much stronger the register effect is than the interaction between corpus and register. At best, one may detect a slight tendency among the CTE translators to be more pedantic in the academic and especially the instructional registers, by making the textual relations more explicit. In the other four registers, and even to an extent in the academic register itself, however, these differences are small and tend to cancel each other out. Neither corpus can be said to reveal limited variation; the registers with higher frequencies of linking adverbials are the ones expected to make use of this type of resource.

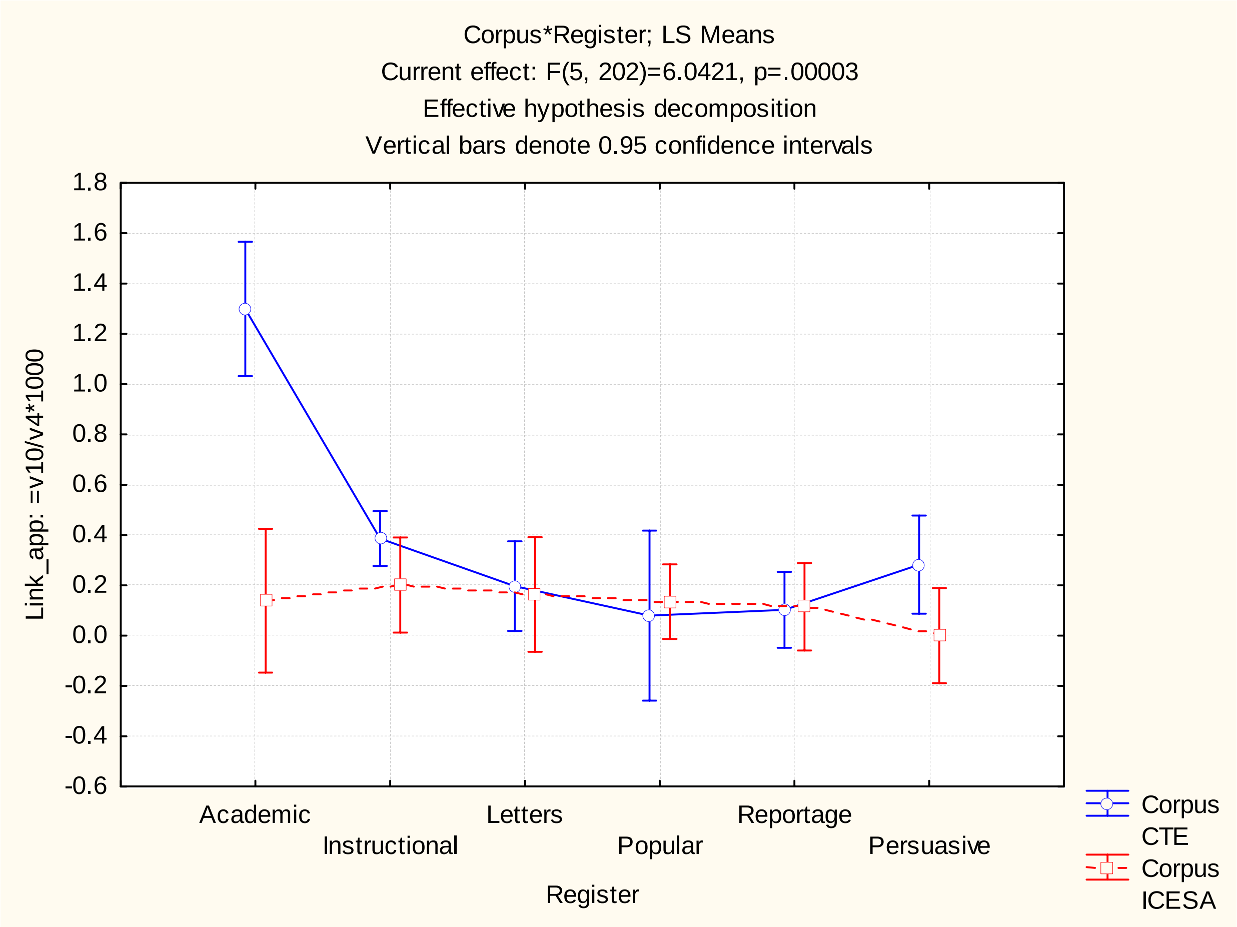

Figure 6. Mean frequencies of appositive linking adverbials in the two corpora and six registers

A closer examination of the two categories of linking adverbials investigated shows that the more frequent of the two categories, result and inference (mean in CTE=0.85, sd=1.01; mean in ICE-SA=0.86, sd=1.13), shows a significant main effect for register, similar to the overall category mean (F(5, 202)=8.99, p<0.001), with no effect for corpus (F(1, 202)=0.25, p=0.62), nor a significant interaction between corpus and register (F(5, 202) 1.67, p=0.14). However, the appositive linking adverbials is used significantly more often in the CTE as an explication device, with mean in CTE=0.33 (sd=0.51) and mean in ICE-SA=0.12 (sd=0.32). The effect is statistically significant (F(1,202)=18.01, p<0.001). The interaction between register and corpus is also statistically significant (F(5,202)=6.04, p<0.001), and is specifically detectable in the more frequent use of this adverbial subtype in academic writing by the CTE, while most other registers in both corpora make very little use of this option, as is shown in Figure 6.

Typical uses of appositive linking adverbials in the academic register in CTE include instances such as the following:

- *Cinema experienced optimum conditions for spreading almost exclusively in those countries which were then the most developed industrial nations – that is France, the USA or Germany.*
- *The group of psychologists agreed that career specificity was an important predictor (i.e. that the relationship between study and work success was stronger in certain professional fields than in others).*

The interaction between register and corpus just fails to reach statistical significance, but is indicative of the underlying hypothesis that translated texts are more explicit in presenting information. The more subtle nuance that the data lead us to add is that this effect is salient only in registers that are intrinsically informational from the start, and not an effect that cuts across all registers. In a sense, however, this subtle nuance falsifies the broader hypothesis that translated texts are less register sensitive and aim for a relatively uniform middle register.

3.4 Frequency of coinages and loanwords

Coinages on the whole are slightly more frequent in ICE-SA (mean=0.32 per 1000 words, sd=0.53) than in the CTE (mean=0.25 per 1000 words, sd=0.52), but the main effect for corpus is not significant (F(1, 202)=0.58, p=0.45). Any possible interaction for register seems to be cancelled out in that some CTE registers (letters, persuasive writing and reportage) make more frequent use of coinages than ICE-SA, but the converse is true for the other three registers (instructional, popular and academic writing). In consequence, the interaction between the two factor variables is not significant either (F(5, 202)=1.90, p=0.10). The one difference that seems more compelling than any of the others is in the register of popular writing. Here, the CTE almost entirely avoids loanwords and coinages, but ICE-SA uses them more than in any other register.

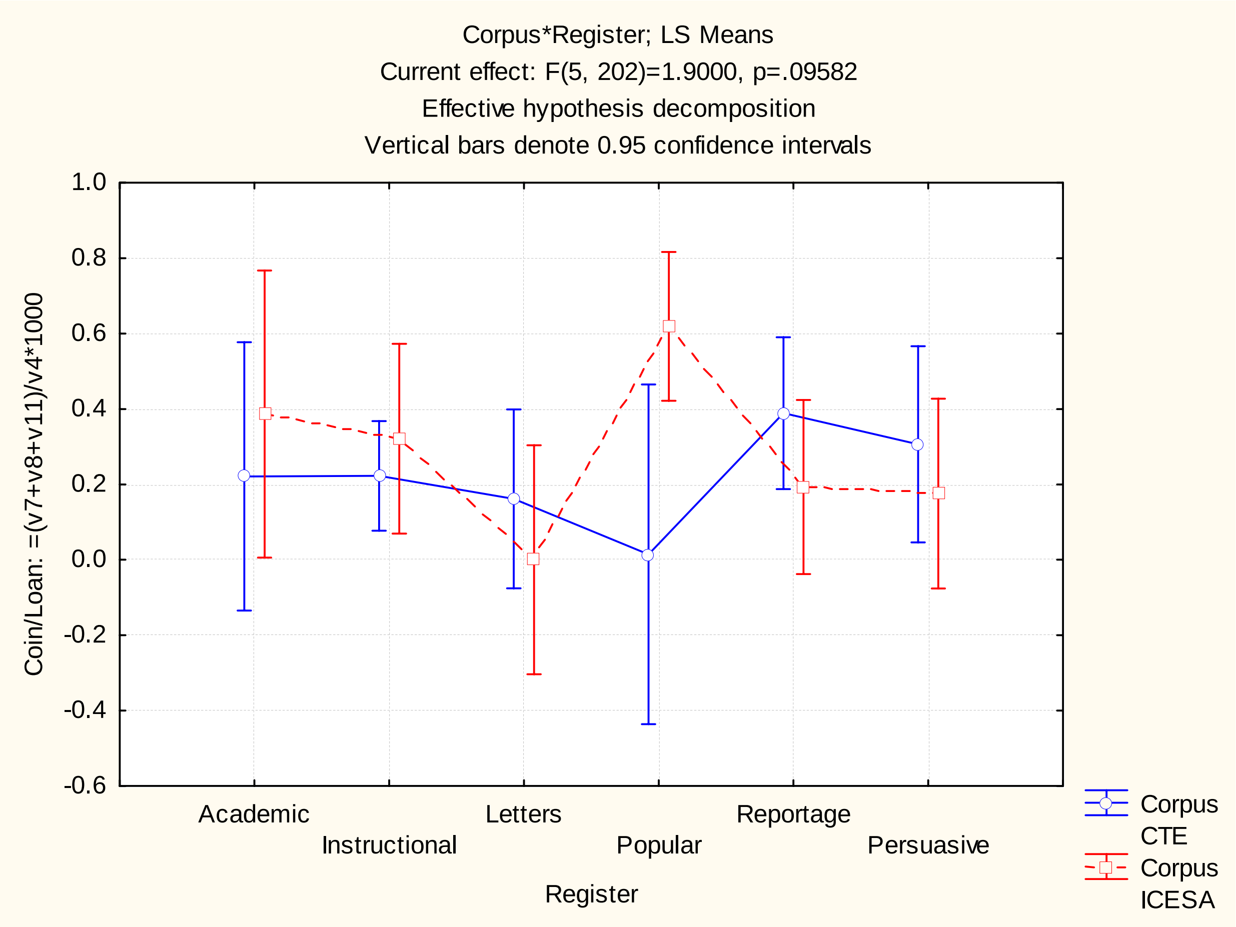

Figure 7. Mean frequencies of coinages and loanwords in the two corpora and six registers

Some of the coinages occurring in ICE-SA include: *outsurfing*, *uncucumber-like*, *amabeating* and *passless*. Loanwords from various South African languages occur, like *vetkoekies*, *verkrampte*, *bundu*, *amakhosi*, *kragdadigheid*, *lapa*, *witblits*, *skottel*, *umlungu* and *Engelse*. Coinages in CTE generally appear less creative, including *recurriculising*, *confocally*, *centralised-decentral*, and *ironhard*. Loanwords appear similarly less diversified in CTE, including mostly words from Afrikaans, such as *boereorkes*, *wors* and *spitbraai*, with only *kwaito* borrowed from an African language.

3.5 Frequency of lexical bundles

Trigrams show a very clear main effect of register, as shown in Figure 8. Letters and instructional writing make liberal use of trigrams, with more than 6 per 1000 words. The main effect for register is statistically significant (F(5, 202)=6.55, p<0.001).

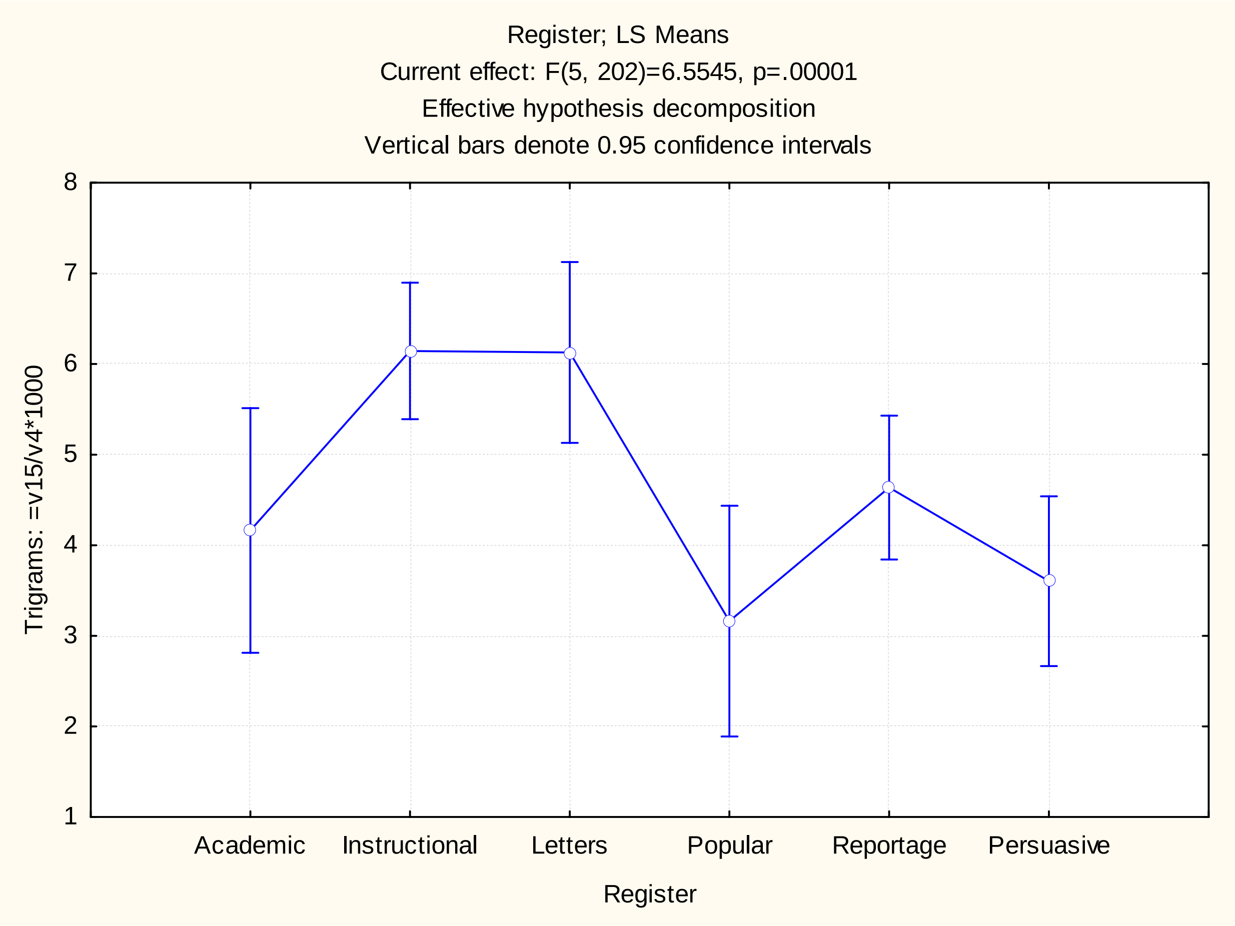

Figure 8. Mean frequencies of all trigrams in the six registers

The two corpora are, however, quite similar in the way they manipulate trigrams in the six registers. Therefore, as shown in Figure 9, there is no significant interaction between register and corpus (F(5, 202)=0.88, p=0.49). If anything, CTE shows a little more variability, in that the registers with high frequencies of trigrams show even higher frequencies for CTE, while in the registers with lower frequencies, CTE shows even less use of trigrams. Thus, no hypothesis about less differentiated registers is supported by the trigram data, since the line for the CTE in Figure 9 is not flat, or even closer to a straight line than the line for ICE-SA.

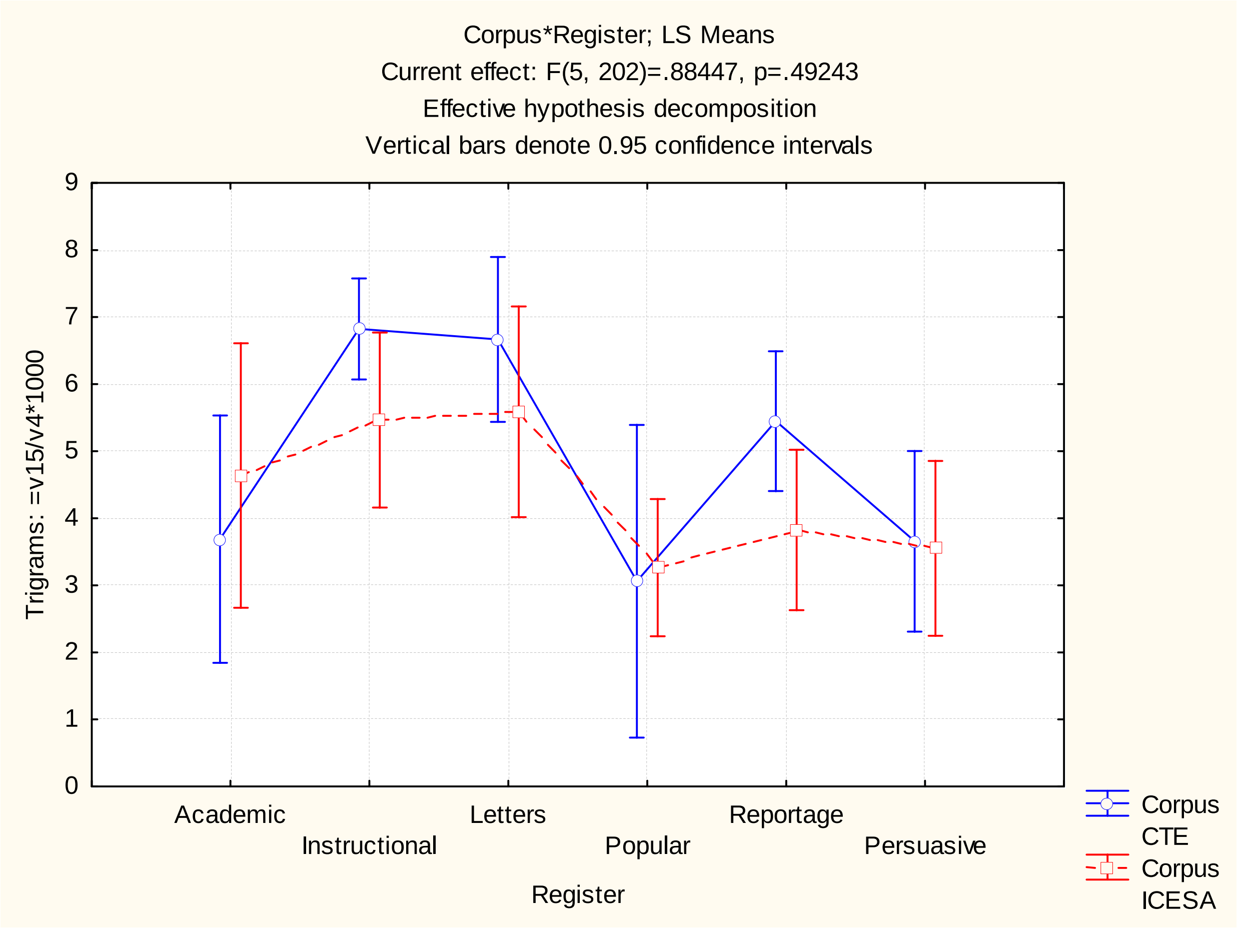

Figure 9. Mean frequencies of all trigrams in the six registers and two corpora

3.6 Lexical diversity

Type-token ratios show a significant main effect for both corpus (F(1, 200)=27.02, p<0.001) and for register (F (5, 200)=19.60, p<0.001). There is therefore a straightforward effect of translated versus non-translated texts in the type-token ratio. The average standardised type-token ratio per text is 38.83 unique lexical items per 100 words (sd=5.31) for the CTE, and much higher at 44.02 for ICE-SA (sd=4.63). This simply means that ICE-SA uses a much more varied vocabulary than CTE. There is an independent effect for register as well, as can be seen in Figure 10, but the interaction between register and corpus does not yield anything of significance (see Figure 11, F(5, 312)=1.28, p=0.27). It thus seems as if these two effects are more or less independent of each other. The register sensitivity in the CTE is similar to ICE-SA, but the overall diversity is simply lower across the board.

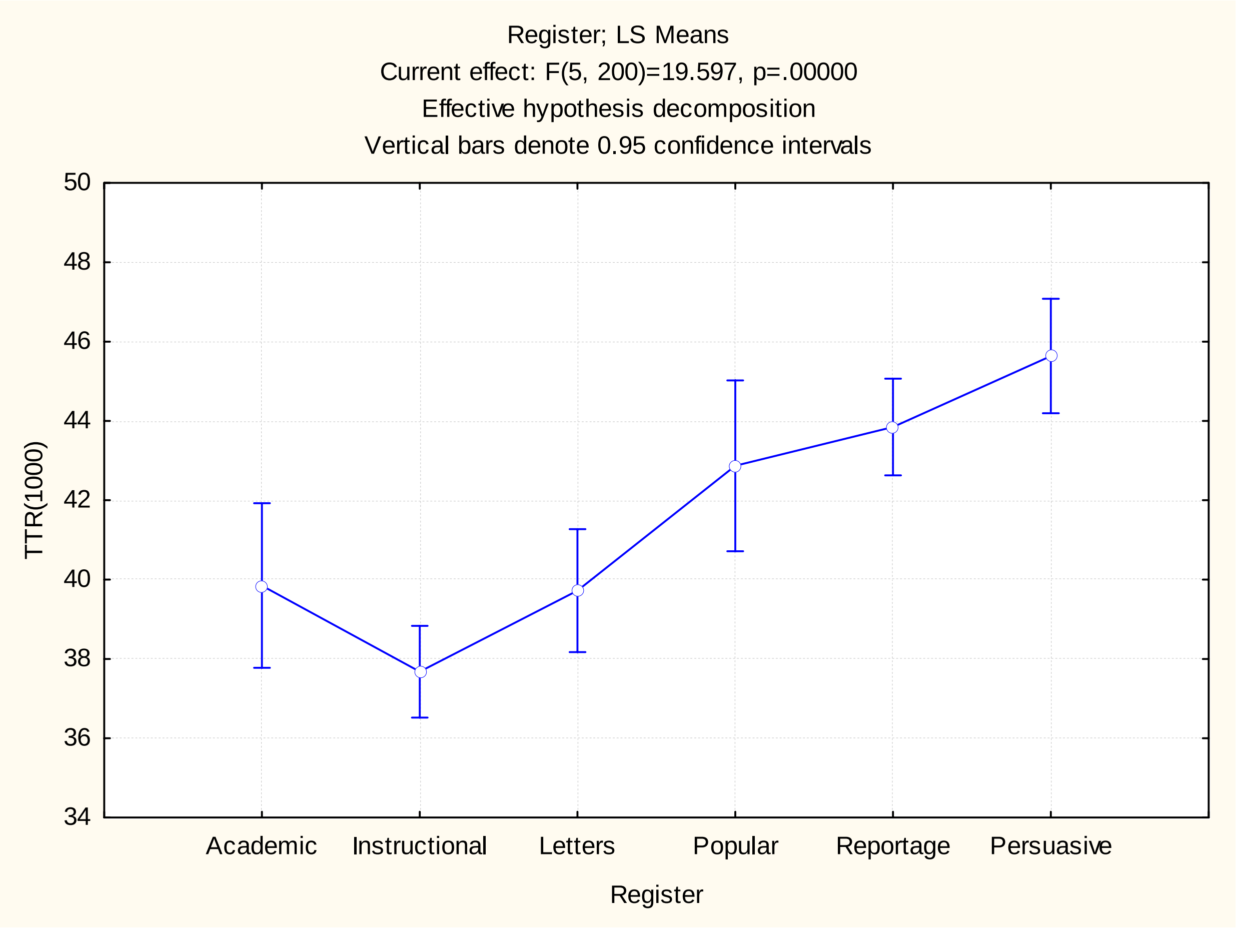

Figure 10. Means of standardised type-token ratios for the six registers

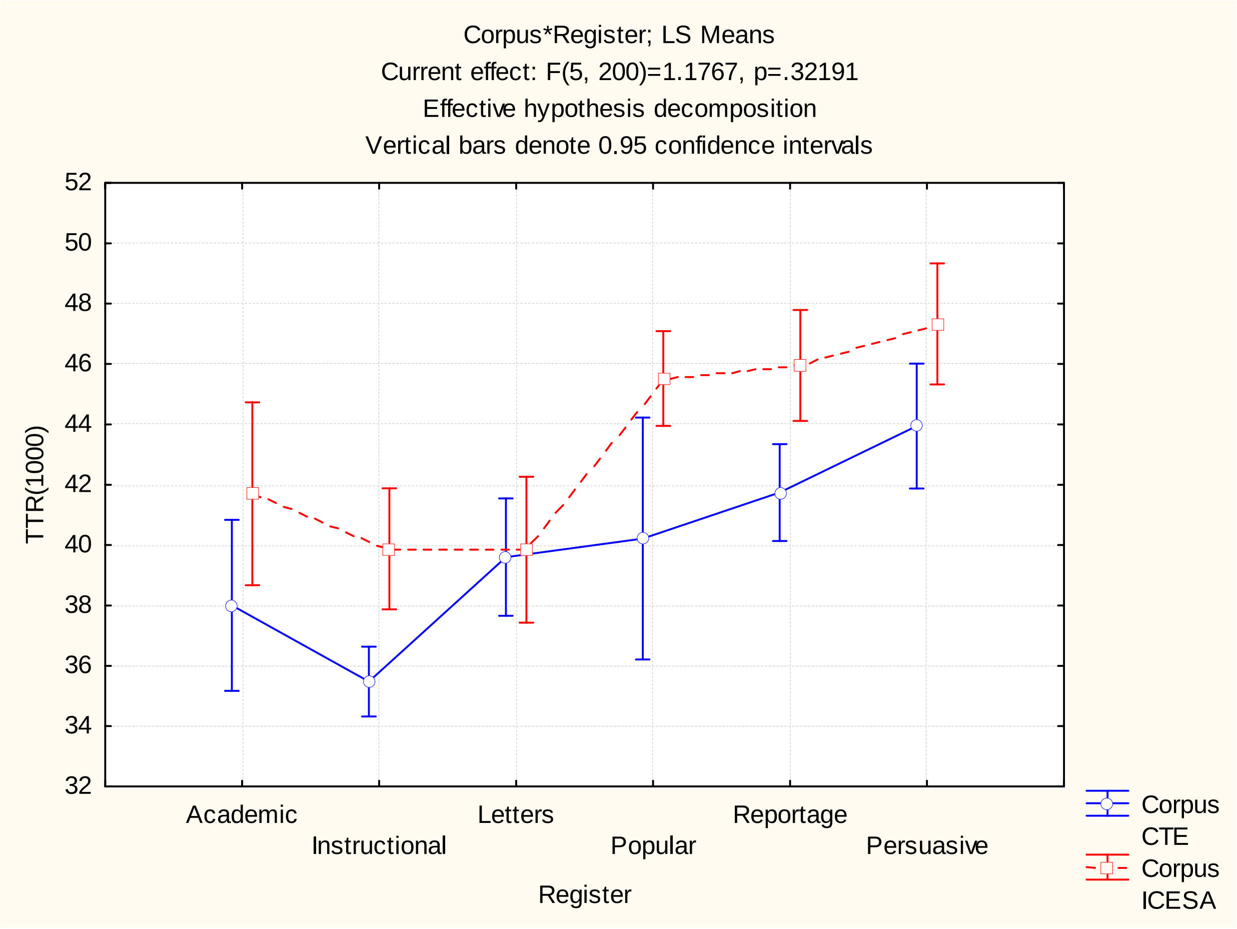

Figure 11. Means of standardised type-token ratios for the six registers and two corpora

3.7 Word length

The average for word length is surprisingly higher in the CTE than in ICE-SA (CTE=4.98 characters per word (sd=0.30); ICE-SA=4.86 (sd=0.30)), but the difference is so slight that it the main effect is not statistically significant (F(1, 202)=0.19, p=0.66). The main effect for register is statistically significant (F(5, 202)=6.15, p<0.001), which can be attributed to the much higher average word length in the academic writing register, and much lower average word length in popular writing (see Figure 12).

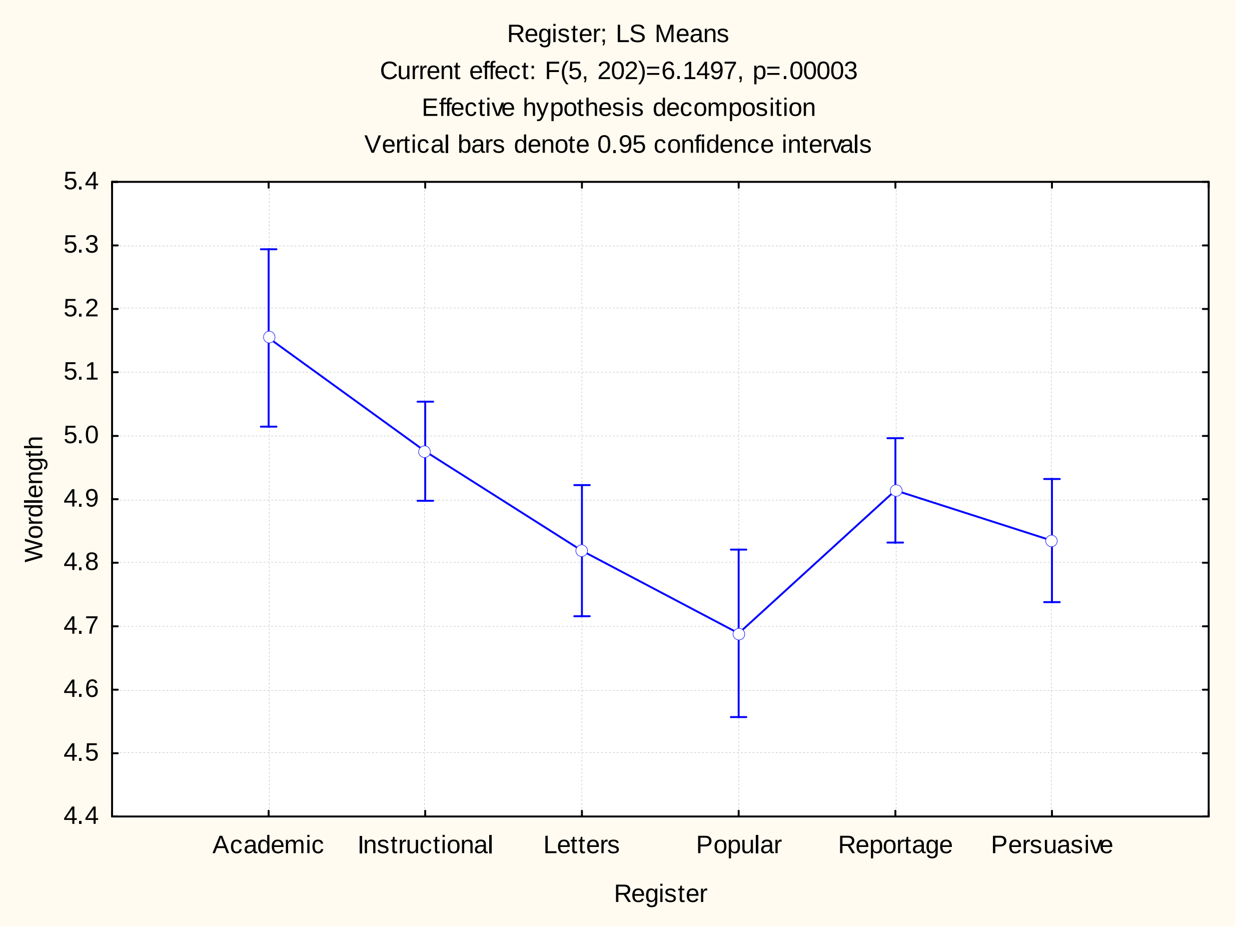

Figure 12. Means of word length for the six registers

The interaction between register and word length is statistically significant as well (F(5, 202)=2.77, p<0.05), as is represented in Figure 13. In particular, the popular writing register in CTE is an outlier, where the average word length (4.53, sd=0.19) is much lower than in any other register in any corpus. Thus, simplification is a prominent feature of this register in terms of word length too. The other registers reveal similar values for academic writing and letters, but a more varied pattern in CTE than in ICE overall; hence no support for a hypothesis that the translated data are less varied across the registers.

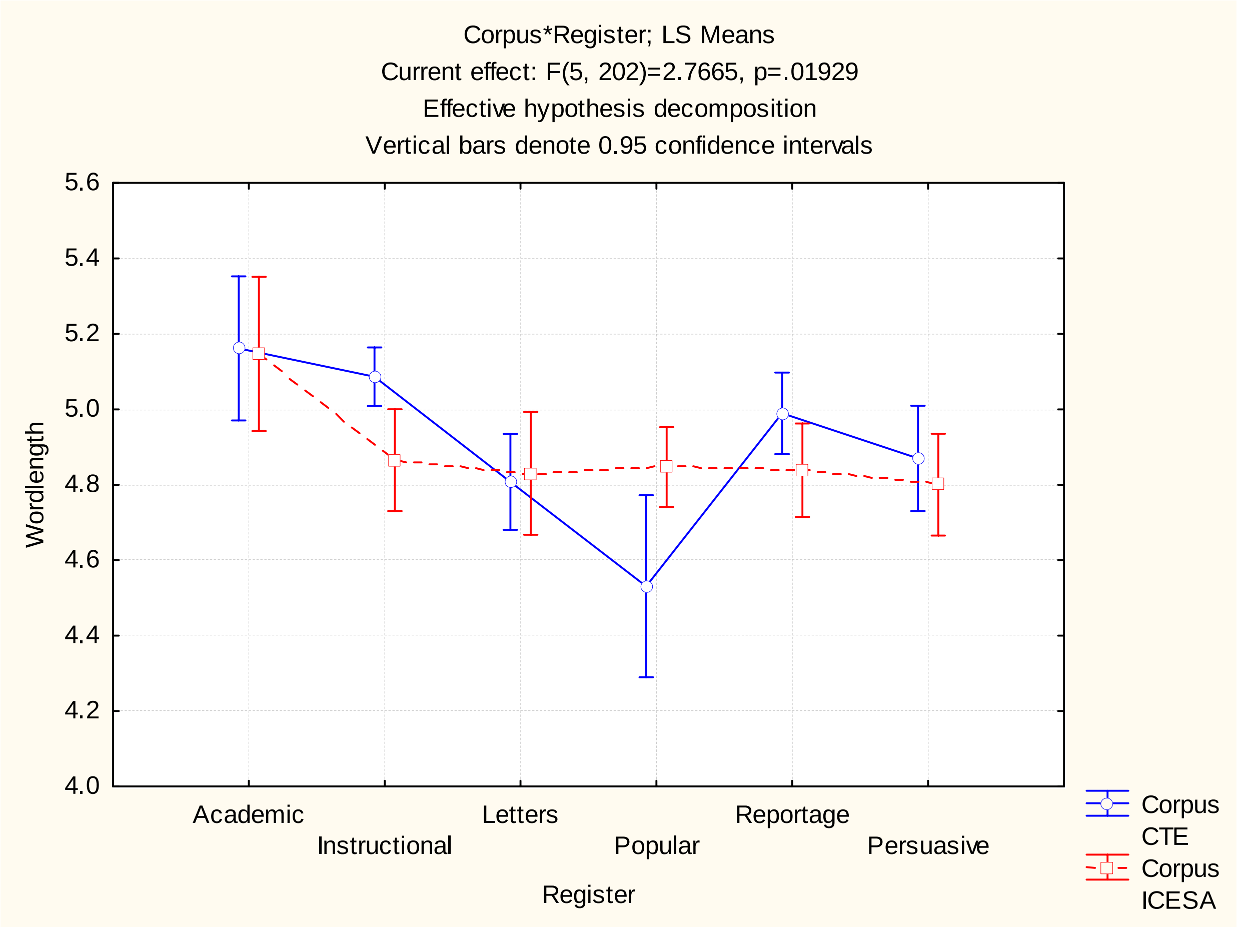

Figure 13. Means of word length for the six registers and two corpora

4. Conclusions, caveats, possibilities

There thus appears to be limited support for the first hypothesis, that the linguistic features selected for investigation would demonstrate significant differences in the two corpora, reflecting more explicit, more conservative, and simplified language use in the corpus of translated English than in the comparable corpus of original English writing.

The findings of this pilot study suggest statistically significant differences between the two corpora for only two of the features investigated: the use of the optional *that* complementiser, and lexical variety. That omission is significantly less frequent in CTE, and standardised TTR is lower in CTE. The case for features unique to translated language is thus statistically supported by findings for only these two features, which, overall, suggests limited support for the first hypothesis in this paper. However, some of the other findings of this paper do suggest that some features associated with, particularly, explicitation and conservatism may warrant further, more careful, investigation.

The second hypothesis, that in CTE less register variation or sensitivity to register will occur as a result of the interference of translation, is not supported by the findings, overall. Register differences occur irrespective of corpus for the following features investigated: TTR, word length, trigrams, linking adverbials and *that* omission – though in some instances, as in the case of that omission, there is less register variation in the CTE. No feature demonstrates a flat distribution across register for CTE and not for ICE-SA, which means that the overall hypothesis of undifferentiated registers in translated language is not supported. However, more subtle effects can be observed. Firstly, popular writing in CTE does not seem to demonstrate the same extent of informality as is the case in this register in ICE-SA, as shown by the frequency of features such as that omission, contraction, the use of coinages and loanwords, and word length. Secondly, translated academic writing appears to be excessively pedantic in its use of appositive linking adverbials, with perhaps some further support from the use of that complement clauses in academic writing (see Figure 1).

The findings of this study should be read against the background of some reservations about the corpora. Both corpora still require some refinement, and, as pointed out earlier, there are some concerns about balance and representativeness that need to be addressed. With some alterations to the corpora, and the findings of the pilot as guideline, the full study will aim to investigate the relationship between the features of translated language and register in more detail and with a greater degree of certitude.

References

Baker, Mona. 1993. Corpus linguistics and translation studies: implications and applications. (*In* Baker, Mona, Francis, Gill & Tognini-Bonelli, Elena., *eds*. Text and technology: in honour of John Sinclair. Amsterdam: John Benjamins. p. 17-45.)

Baker, Mona. 1995. Corpora in translation studies: an overview and some suggestions for future research. *Target*, 7:223-243.

Baker, Mona. 1996. Corpus-based translation studies: the challenges that lie ahead. (*In* Somers, H., *ed.* Terminology, LSP and translation: studies in language engineering, in honour of Juan C. Sager. Amsterdam: John Benjamins. p. 175-186.)

Baker, Mona. 2007. Patterns of idiomaticity in translated vs. non-translated text. *Belgian Journal of Linguistics*, 21(1):11-21.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. Longman grammar of spoken and written English. London: Pearson Education.

ICE (International Corpus of English). 2010. International Corpus of English: home page. http://ice-corpora.net/ice/index.htm Accessed 7 September 2010.

Kenny, Dorothy. 2001. Lexis and creativity in translation: a corpus-based study. Manchester: St Jerome.

Laviosa, Sara. 1998. Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, 43(4):557-570.

Mutesayire, Martha. 2004. Apposition markers and explicitation: a corpus-based study. *Language Matters: Studies in the Languages of Africa*, 35(1): 54-69.

Olohan, Maeve & Baker, Mona. 2000. Reporting that in translated English: evidence for subconscious processes of explicitation? *Across Languages and Cultures*, 1(2):141-158.

Olohan, Maeve. 2003. How frequent are the contractions? A study of contracted forms in the Translational English Corpus. *Target*, 15(2):59-89.

Williams, D.A. 2005. Recurrent features of translation in Canada: a corpus-based study. Unpublished PhD thesis, University of Ottawa. http://www.scribd.com/doc/7630845/Recurrent-Features-of-Translation-in-Canada-a-Corpus-Based-Study