Comparing source and target texts in a translation corpus

Magnus Merkel
Department of Computer and Information Science
Linköping University
magme@ida.liu.se

1 Introduction

In this paper the Linköping Translation Corpus (LTC) is used as an example on how simple methods and tools can be applied to investigate the relationships between source and target texts in a translation corpus.

The Linköping Translation Corpus consists of English source texts linked to Swedish target texts. The text material comes from two major text types: user’s guides to computer programs and fiction. There is also a shorter machine-translated text consisting of dialogue included in the corpus. LTC consists of 805,277 words in the source text and 732,628 words in the target texts, making the total word size of just over 1,500,000 words.

Table 1 below shows an overview of the translation corpus:

Table 1. The Linköping Translation Corpus - an overview

Text type	Title	No. of source words	No. of target words	Transl. method
User’s Guide	Microsoft Access UG	179,631	157,302	Human
User’s Guide	Microsoft Excel UG	141,381	127,436	Human
User’s Guide	IBM OS2 UG	127,499	99,853	TM
User’s Guide	IBM InfoWindows UG	69,428	53,619	TM
User’s Guide	IBM Client Access UG	21,321	16,752	TM
Novel	Gordimer: A Guest of Honour	197,078	210,350	Human
Novel	Bellow: To Jerusalem and Back	66,760	65,268	Human
Dialog	ATIS dialogues	2,179	2,048	MT
Total		**805,277**	**732,628**	

Three of the translations were translated with the aid of IBM’s translation memory tool (TM), which gave an extra dimension to the corpus.

Given a translation corpus such as LTC, one task is to uncover the characteristics of the translation as a whole, or to see whether the translations could be characterised as source-oriented or target-oriented translations (Newmark 1988). The analysis of translation corpus can be made in several steps, going from the simplest case, namely just to compare surface data from the source and target texts independently to a full-blown analysis of how the translator(s) have chosen to render the target text given lexical, syntactic and semantic constraints. In this paper, the first steps of such an analysis and later are shown as well as a sketch on how these analyses correspond with a thorough linguistic analysis of the relationships between the source and target texts.

2 Step 1: Source and target texts independently

The majority of the translation analyses comes from using the DAVE toolbox, developed at Linköping University. With the DAVE tools we extracted data on the Linköping Translation Corpus first as separate texts, i.e. the source texts and target texts independently, including data for

- Word type/token ratio
- Sentence type/token ratio
- Average number of words per sentence
- Number of repeated sentences
- Recurrent sentence rate

These data for the source and target texts, respectively, are listed in Table 2 and 3 below.

There is nothing strikingly unexpected in the source text data in Table 2, although it should be noted that the two novels (Gord and Bellow) contain the largest number of word types, longest sentences and lowest recurrent sentence rates. The fact that there are 184 sentences which are repeated in the Gordimer novel, may even seem somewhat high as repetitiveness is not a common characteristic for fiction, but at closer scrutiny, it turns out that almost all the repeated sentences is contained in the dialogue part of the novel. For example, utterances like “I know.”, “Yes.” and “All right.” occur several times in the novel and constitute to a large extent these 184 repetitive sentences. Furthermore, it is worth mentioning the relative similarites between the two Microsoft texts (Access and Excel); the number of words per sentence is comparable as well as the recurrence rates. Of the IBM texts, the InfoWin text has a considerably higher recurrence rate than the others (31.1 per cent).

Table 2. Source texts - general data

	Access	Excel	OS2	InfoWin	Client	Gord	Bellow	ATIS
Word tokens	179631	141381	127499	69428	21321	197078	66760	2179
Word types	4370	4483	7537	3276	1680	17539	10139	245
Word type/ token	41.11	31.54	16.92	21.19	12.69	11.24	6.58	8.89
Sentences	14829	12610	12242	7834	2427	12310	4215	263
Words/ sentence	12.11	11.21	10.41	8.86	8.78	16.01	15.84	8.29
Repeated sentences	5361	3807	3333	4116	904	184	4	0
Recurrent sentence rate	14.7%	13.62%	13.93%	31.10%	17.55%	0.18%	0.01%	0.00%

Table 3. Target texts - general data

	Access	Excel	OS2	InfoWin	Client	Gord	Bellow	ATIS
Word tokens	157302	127436	99853	53619	16752	210350	65268	2048
Word types	6703	7246	10152	4308	2266	23599	13026	255
Word type/ token	23.47	17.59	9.84	12.45	7.39	8.91	5.01	8.03
Sentences	15079	13020	11943	7735	2457	13427	4285	263
Words/ sentence	10.43	9.79	8.36	6.93	6.82	15.67	15.23	7.79
Repeated sentences	5040	3853	3066	4351	933	291	8	0
Recurrent sentence rate	11.37%	13.06%	9.84	39.26%	18.70%	0.31%	0.02%	0.00%

Although the figures vary slightly, the same pattern is discernible for the target texts as for the source texts, namely that the novels contain the highest numbers of word types, longest sentences and lowest recurrent sentence rates. The Microsoft texts and IBM texts also seem to be relatively similar.

The next step is then to compare the general data from the source and target texts and see if we can conclude something about the translations. We do this by comparing the relative proportions of number of sentences, number of word tokens and recurrent sentence rates, by using the following simple measures (the figures summarized in Table 4 below):

- ST-Sentence = the number of source sentences/number of target sentences
- ST-Word ratio = number of source words/number of target words,
- ST-Recurrent sentence ratio = Recurrent sentence rate(Source text)/Recurrent sentence rate(Target text).

The figures tells us that only two of the texts have more source sentences than target sentences (namely OS2 and InfoWin). This could indicate that most of the texts contain more deletions or that the sentence pairs have a high degree of n-1 sentence correspondences, but at this point this is mere speculation. Only one text, the Gordimer novel, contain more running words in the translation than in the original text. Due to fact that Swedish contain more compounds (written as single words) than English, and that at a large proportion of the definite article "the" and the verb "do" do not have Swedish counterparts, it would be reasonable to expect that the number of words be smaller in the Swedish text. But, again we can only speculate that the text type, in this case fiction, seems to give rise to a relatively higher number of target words.

Table 4. Relative comparisons between source texts and target texts

	Acc.	Excel	OS2	Info	Client	Gord	Bellow	ATIS
ST-Sentence ratio	0.98	0.97	1.02	1.01	0.99.	0.92	0.98	1.00
ST-Word ratio	1.14	1.11	1,28	1.29	1,27	0.94	1.02	1,06
ST-Recurrent sentence ratio	1.29	1.04	1.09	0.79	0.94	0.58	0.50	N/A

This is apparent if we also look at the word ratio for the other novel by Bellow which contain fewer words in the translation compared to the original, but the figure (1.02) is still considerably lower than for the translations of the computer manuals.

The English-Swedish Parallel Corpus (ESPC) from Lund contain comparable ST-word ratios for the translations of fiction from English to Swedish (0.98). Looking at the total material (English to Swedish) including non-fiction, gives a ST-word ratio of 1.003 in ESPC.[1] This means that the non-fiction part of the ESPC corpus contain more source words than target words (ST-word ratio 1.028). The computer manuals in the Linköping Translation Corpus do seem to be different in this respect as the ST-word ratios range from 1.11 to 1.29. In relative terms it is reasonable to expect that more information is preserved or added in the fiction translations compared to the translations of manuals.

When we compare the values for sentence recurrence in the texts, we can see that two of the IBM texts (InfoWin and Client) actually have higher sentence recurrence rates in the target than in the source, which is in line with the first hypothesis as these texts were translated with the aid of translation memories.

The two Microsoft texts have higher recurrence rates for the source text than the target text which is in accordance with the second hypothesis, namely that consistency on the sentence level would be more difficult in traditional translation.

The text that does not fit the pattern then is the OS2 text, which has a higher recurrence rate in the source text than in the target text even though the translation was produced with translation memory.

3 Step 2: Source and target texts jointly (as an aligned corpus)

To be able to make more detailed observations on the relationships between the source and the arget text, it is necessary to investigate the source text and target text as a whole, that is, as parallel texts and this is done by using the DAVE tool box (see Merkel 1999).The DAVE toolbox contains some bitext-tailored tools, for example, a module for analysing discrepancies (or inconsistencies) in the translation as well as a module for bilingual concordancing. The Discrepancy analysis provides information on how consistent or inconsistent the translations of the recurrent sentences are. The Bilingual concordance module lets the user browse and search the parallel text for any combination of source and target words and multi-word units and collect data for the co-occurrence of certain items.

In particular, the focus is on whether there are any observable differences in the translations as far as text type and method of translation are concerned, especially for the distinctions of source- vs. target-orientation, consistency and correspondence. The following types of extracted data from the Linköping Translation Corpus (LTC) are in focus: (i) sentence mappings, (ii) consistency and variation and (iii) co-occurrence data for a sample of lexical items.

2.1 Sentence mapping

The majority of the translations in the Linköping translation corpus contain 1-1 sentence mappings to the degree of a 96-98.35 per cent interval, as can be seen in Table 5. The OS2 text has a strikingly high proportion of deletions (1-0) and insertions (0-1) which indicate that the translation is not particularly close to the original, but is rather a kind of communicative, more target-oriented translation, cf. Newmark (1988). However, the translation of this text has been made with the aid of a translation memory tool, which contradicts a target-oriented translation as translation memories should actually steer translators towards source-oriented translation. For the time being we can only note that there seems to be something strange about the OS2 translation, given the use of translation tools and the fact that data from the sentence mappings give us another message.

The second translation that sticks out is the Gordimer text. Here there is only one aspect that seems peculiar, and that is the relatively high proportion of 1-2 mappings. Over 8 per cent of the pairs are instances of when one English sentence has been translated with two Swedish sentences.

Table 5. Sentence mappings from the parallel texts (excluding ATIS)

	Access		Excel		OS2		InfoWin		Client		Gord		Bellow	
	No.	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Pairs	14704		12589		11932		7771		2426		12254		4209	
1-1	14169	*96.36*	12107	*96.17*	10444	*87.53*	7519	*96.76*	2386	*98.35*	11112	*90.68*	4122	*97.93*
1-0	21	*0.14*	15	*0.12*	408	*3.42*	107	*1.38*	4	*0.16*	4	*0.03*	2	*0.05*
0-1	4	*0.03*	6	*0.05*	253	*2.12*	7	*0.09*	0	*0.00*	0	*0.00*	0	*0.00*
2-1	121	*0.82*	26	*0.21*	390	*3.27*	66	*0.85*	1	*0.04*	41	*0.33*	6	*0.14*
1-2	376	*2.56*	426	*3.38*	308	*2.58*	70	*0.90*	35	*1.44*	999	*8.15*	77	*1.83*
Rest	14	*0.09*	9	*0.08*	129	*1.08*	2	*0.02*	0	*0.00*	98	*0.8*	2	*0.05*

[1] Data from ESPC was kindly provided to me by Bengt Altenberg (personal communication October, 7 1999). The fiction ST-word ratio is based on 276,591 source words/281,127 target words and the total ST-word-ratio on 494,374 source words/492,885 target words.

The explanation for this has to do with at least two different uses of punctuation characters in English and Swedish. First, in English the semicolon is used more often than in Swedish as a delimiter between main clauses, which means that perhaps we should have classified the semicolon as a possible sentence delimiter during the alignment process. Secondly, in English a comma may precede an utterance (within quotation characters) whereas in Swedish the line will commonly be indicated with a colon. The two different uses of semicolons and commas are shown in the two examples from the Gordimer novel in Table 6. The actual positions that were discussed above are underlined in the source and target texts.

Table 6. Different uses of the semicolon and the comma in English and Swedish

Source	Target
It was she who had given her glass to him that night at the Independence party; the Pole who had danced the gazatska became the man with ...	Det var hon som hade låtit honom överta sitt glas under självständighetsfesten. Polacken, som hade dansat en gazatska, blev den han....
A youthful black official at passport control said uncertainly, "Just a minute.	En ungdomlig, svart tjänsteman i passkontrollen svarade litet osäkert: "Ett ögonblick.

If we regard semicolons as sentence delimiters as well as commas in sequences of <*comma-space-quotation mark-uppercase letter*>, and then recalculate the proportions, it turns out that the proportion of 1-1 mappings increases to 95.94 per cent, which takes the Gordimer text up to roughly the same relative proportions as the other texts (except the OS2 text). A flexible method to handle "unusual" sentence boundaries has been suggested by Palmer and Hearst (1994) which may help to improve sentence alignment.

2.2 Discrepancy Analysis

By using discrepancy analysis it can be shown that all translated texts are more or less inconsistent. For manually translated texts the variations are what can be expected, but we discovered an unexpected high degree of inconsistency in the translation memory translated manuals and found that this was due to a clash between an established translation culture and new technology (see Merkel 1996). The discrepancy analysis also revealed relative differences between the manuals which showed how the Client translation actually was a more consistent translation and therefore probably more source-oriented translations than the other user's guides.

2.3 Bilingual concordancing

Co-occurrence data could be useful in several applications. In contrastive linguistics, it could form the basis for extracting exactly those sentence pairs that contain the word(s) that are of interest to the scholar. Altenberg (1998) has developed a measure, *mutual correspondence*, that aims to capture the degree of correspondence in parallel corpora between pairs of words in English-to-Swedish translations and Swedish–to-English translations jointly. For example, if the English word "however" is always translated into the Swedish "emellertid" and "emellertid" is always translated by "however" in English translations then the Mutual Correspondence (MC) between "however" and "emellertid" is 100 per cent.

As the Linköping translation corpus only contains translations in one direction, namely into Swedish, mutual correspondence cannot be calculated, but it is possible to investigate the relative word co-occurrence rate (WCR) for a source and a target word, as follows:

$$WCR = \frac{2 \times (cooccur(A, B) \times 100)}{freq(A\ ) + freq(B\ )}$$

If a word A occurs 10 times in the source text, a target word B occurs 15 times in the target text and A and B co-occur 8 times, the WCR for A and B is 64 per cent (16/25). The measure actually takes into account the number of times a token of one of the words occurs in a co-occurrence relation in the corpus.

The MC measure will capture the extent to which two words are mutual translations of each other, while the WCR measure will measure the proportion of co-occurrence between one source and one target word given a corpus containing only one translation direction.

Altenberg measures the *translation bias* as the ratio of how many target tokens that are realised from the source word. The formula for translation bias (TB) can be expressed as follows:

$$TB(A, B) = \frac{cooccur(B, A)}{freq(A)}$$

which means a simple ratio between the number of times a target item, B, co-occurs with the source item, A, in relation to the total number of source items A.

Different hypotheses could be tested by investigating co-occurrence data and translation bias; for example, is it possible to conclude how source-oriented a target text is given only co-occurrence data of word pairs from different translations? Text-type specific translation corpora could provide information about what the standard co-occurrence rates would be for a core of word pairs. These pairs and their co-occurrence rates could then be tested on translations from the same text type and perhaps give an indication of how source-oriented the translations are.

Furthermore, co-occurrence rates could be used to investigate what word pairs that are most suitable to use as anchor words (Johansson and Hofland 1994), cognates (Simard et al. 1992) or cue words Wu (1994) in hybrid approaches to sentence.

The use of the bilingual concordance component from DAVE and the application of word co-occurrence rates to the corpus was applied to word pairs from the Linköping translation corpus. The word pairs belong to four different categories:

conjunctions, subjunctions, numerals and proper names/technical terms.

The brief analysis on word correspondences confirms the view that the best candidates for anchoring words can be found among cognates (numbers and proper names) an, to a certain extent, technical terms. Conjunctions and subjunctions can also function as potentially good candidates for most texts. It also shows that the texts that were considered to be more source-oriented in their translation style also exhibit more consistent translations on the word level.

The Bilingual concordance component is a useful tool for the contrastive linguist, the translation scholar and the language engineer. The contrastive linguist can compile statistical data on co-occurrence and extract sentence pairs from parallel corpora that are specifically interesting for a certain contrastive study. The translation scholar may be more inclined to study translation bias; that is, given a certain source object, what are the preferred choices of the translators as they appear in the text. The translation scholar will focus on the translation direction of the text, whereas the contrastive linguist will be more interested in the relationship between two language systems.

3 Step 3: Structural and Semantic Correspondence

In Ahrenberg & Merkel (2000), a descriptive model for measuring the salient traits and tendencies of a translation as compared with the source text were applied to the LTC. Here samples from each translation from the corpus were analyzed in detail to uncover structural and semantic changes in the translation. Many of the traits that we have seen in steps 1 and 2 were verified in this study. In Figure 1, below it is shown graphically how four of the translations are located as regards structural and semantic changes. The Gordimer translation contains more information than its orignal, but exhibits structural changes on the same level as the Access and Client translations. The MT-produced ATIS translation is, not surprisingly, shown to be equal in both structure and specification degree compared to its original. These data correlate with ST-Word ratio presented for these texts in step 2 earlier.

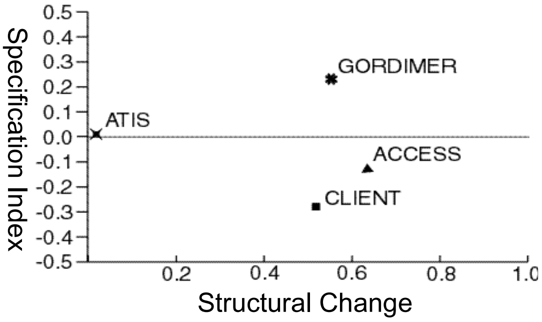

Figure 1. Four different translations displayed according to their tendency for structural and semantic change.

4 Current work: Step 4 - TransMap

At present, the project TransMap, conducted at Linköping University, is developing the analysis methods further by extracting and specifying correspondence data in translation corpora where word and phrase linking is done in combination with a word aligner and a user. The idea is to classify as correspondences on several levels including correspondences of base forms parts-of-speech, syntactic function, and type (such as pronominalization, deletion, addition, convergence and divergence). A shift in direction has here been made compared to the group's earlier work in that the present approach includes using available linguistic resources in the word alignment and phrase extraction tools, such as POS taggers and lemmatizers, and not just string data.

5 Conclusion

By comparing string level data in translation corpora, a great deal of information can be extracted, as have been shown in steps 1 and 2. Many characteristics that were uncovered with these simple methods were actually confirmed when a more thorough linguistic investigation was made on samples from each translation. It remains to be seen if these also holds when the more elaborate techniques of interactive linking and classification of translation units have been finished.

6 References

L. Ahrenberg & M. Merkel. Correspondence measures for MT evaluation. *Proceedings of the LREC 2000 Workshop on Evaluation of Machine Translation*, Athens, Greece 29th May, 2000, pp. 41-46, 2000.

B. Altenberg. Adverbial connectors in English and Swedish. *Out of corpora. Studies in honour of Stig Johansson.* H. Hasselgård & S. Oksefjell (eds.), Rodopi, Amsterdam: 249-268, 1998.

S. Johansson & K. Hofland. Towards an English Norwegian Parallel Corpus. *Creating and Using English Language Corpora.* U Fries, G. Tottie and P. Scheider. Zürich. Rodopi:25:37, 1994.

M. Merkel. Checking Translations for Inconsistency - a Tool for the Editor. In *Proceedings from AMTA-96*, Montreal:157-167, 1996.

M. Merkel. Understanding and Enhancing Translation by Parallel Text Processing. Ph.D. Thesis No. 607. Department of Computer and Information Science, Linköping University, 1999.

P. Newmark, *A Textbook of Translation*, Prentice Hall, London, 1988.

D.D. Palmer & M.A.Hearst. Adaptive Senence Boundary Disambiguation. *Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP-94)*, Stuttgart:78-83, 1994.

M. Simard, G.F. Foster, & P. Isabelle. Using Cognates to Align Sentences in Bilingual Corpora. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal,1992.

D. Wu. Aligning a Parallell English-Chinese Corpus Statistically with Lexical Criteria. Proceedings of the 32n^{d} Annual Meeting of the ACL: 80-87, 1994.