Alignment-based profiling of Europarl data in an English-Swedish parallel corpus

Lars Ahrenberg

Department of Computer and Information Science
Linköpings universitet
SE-58183, Sweden
lah@ida.liu.se

Abstract

This paper profiles the Europarl part of an English-Swedish parallel corpus and compares it with three other subcorpora of the same parallel corpus. We first describe our method for comparison which is based on alignments, both at the token level and the structural level. Although two of the other subcorpora contains fiction, it is found that the Europarl part is the one having the highest proportion of many types of restructurings, including additions, deletions and long distance reorderings. We explain this by the fact that the majority of Europarl segments are parallel translations.

1. Introduction

The Europarl corpus (Koehn, 2005) is the most widely used corpus for training and evaluating statistical machine translation systems for European languages, as evidenced by several recent workshops on the topic. The reasons are not hard to understand; it is very large, it is freely available, and it has data for all pairs of EU languages.

When a parallel corpus is used for statistical machine translation, filters are often applied to the selection of sentence pairs to restrict processing time for training and limit the amount of noise. The filters used are quite simple, however. One way is to put an upper bound on the length of sentences, and another is to remove pairs where one sentence has 5-7 times more tokens than its correspondent.

One can easily observe sentence pairs in the Europarl corpus where the two aligned sentences differ much in structure, and even content. A few examples are given below.

EN: This report is no exception
SE: Så är fallet även i detta betänkande
Gloss: So is case-DEF also in this report

EN: There are several reasons for this
SE: Det beror på flera omständigheter
Gloss: It depends on several circumstances

EN: ...there has been an attempt to put technical make-up on the political face
SE: ...tekniken har försökt dölja politiken
Gloss: ...technology has tried hide politics

There are several possible explanations for these differences: the translation strategy is fairly free, many translators are involved, and, if all available data is used, a majority of the data will be parallel translations rather than original source texts and their translations. Such sentence pairs are likely to introduce noise and could be harmful for the translation models. In any case, since the models of reordering and restructuring that current SMT models employ are fairly crude, it is unlikely that the systems will be able to reproduce translations with the same amount of deletions, additions, reordering and paraphrasing, that are found in such sentence pairs. Thus, it might be a good idea to have them filtered out too, at least from the test sets used for evaluation, if one can find some way to identify them automatically.

This paper, however, has a more restricted goal. It is a study of structural correspondences in a small parallel corpus with English-Swedish Europarl data in focus. All data is taken from the LinES English-Swedish parallel treebank, (Ahrenberg, 2007b), where one of the subcorpora is made up of Europarl data. Comparisons are made with three other subcorpora of the same treebank. The comparisons explore the manual alignments of the corpus and the syntactic annotation based on dependencies. The results indicate that the Europarl part is, in many respects, the most complex one in terms of the frequency of many types of non-isomorphic correspondences and non-local reorderingns, in spite of the fact that two of the other subcorpora are drawn from fiction.

Section 2 presents the data, section 3 presents the method, and section 4 makes a number of comparisons using basic measures at the text, lexical and phrasal levels. Section 5, finally, holds the conclusions and suggestions for further work.

2. Method

Comparisons of monolingual corpora are often made on the basis of frequency profiling of words as well as word categories, such as parts of speech or semantic tags (Rayson and Garside, 2000). The same principles can be applied to the comparison of parallel corpora with respect to translational correspondences, but is not as straight-forward, as it has to deal with two texts. Notions such as deletion, reordering and restructuring are somewhat loose and there are different ways to make them more precise and measure them. Thus, we first need to define types of relations that can be counted.

Second, even if we find that one corpus has more of one type of restructuring than another, this need not be due to choices made by a translator, but to properties of the source

```
        <s id="s10">
            <w id="w150" relpos="1" base="this" pos="DET" msd="DEM-SG"
func="det" fa="2">This</w>
            <w id="w151" relpos="2" base="report" pos="N" msd="SG-NOM"
func="subj" fa="3">report</w>
            <w id="w152" relpos="3" base="be" pos="V" msd="PRES" func="main"
fa="0">is</w>
            <w id="w153" relpos="4" base="no" pos="DET" msd="NEG" func="det"
fa="5">no</w>
            <w id="w154" relpos="5" base="exception" pos="N" msd="SG-NOM"
func="sc" fa="3">exception</w>
            <w id="w155" relpos="6" base="." pos="FE" msd="Period">.</w>
        </s>
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
        <s id="s10">
            <w id="w144" relpos="1" base="så" pos="ADV" msd="" func="sc"
fa="2">Så</w>
            <w id="w145" relpos="2" base="vara" pos="V" msd="PRES" func="main"
fa="0">är</w>
            <w id="w146" relpos="3" base="fall" pos="N" msd="DEF-SG-NOM"
func="subj" fa="2">fallet</w>
            <w id="w147" relpos="4" base="även" pos="ADV" msd="" func="ad"
fa="5">även</w>
            <w id="w148" relpos="5" base="i" pos="PREP" msd="" func="advl"
fa="2">i</w>
            <w id="w149" relpos="6" base="detta" pos="DET" msd="DEM-SG"
func="det" fa="7">detta</w>
            <w id="w150" relpos="7" base="betänkande" pos="N" msd="IND-SG-NOM"
func="pcomp" fa="5">betänkande</w>
            <w id="w151" relpos="8" base="." pos="FE" msd="Period">.</w>
        </s>
```

Figure 1: Excerpts from the Europarl monolingual files with annotation for the English sentence *This report is no exception,* and its Swedish translation *Så är fallet även för detta betänkande.*

text. It may be that the source part of the corpus has a greater number of constructions that, when translated in mostly standard ways, necessitates structural changes because of language differences.

Generally, a Swedish translation of an English text has fewer tokens than the source as a number of very common constructions are usually translated with fewer tokens. Some of the most common cases are listed below:

Compound tenses translated with simple tense forms:
is sleeping $\sim$ *sover*
were given $\sim$ *gavs*

Compound nouns are subject to different orthographic conventions:
file system $\sim$ *filsystem,*
world market prices $\sim$ *världsmarknadspriser*

Definite articles are expressed with a suffix in Swedish when there are no modifiers:
the market $\sim$ *marknaden*
the world market $\sim$ *världsmarknaden*

Do-support is absent in Swedish:[1]
Did she leave? $\sim$ *Gick hon?*
She did not go $\sim$ *Hon gick inte*

[1]Note also that clitics such as n't are tokenized as independent tokens in the treebank used.

Thus, if an English source text has significantly more instances of these constructions than another text, we expect to see a larger drop in the number of tokens of its Swedish translation.

A third factor that needs to be kept in mind is the principles used for word alignment. In LinES, function words are null-aligned when there is no corresponding function word on the other side, even though, as is often the case for the definite article, the same function is expressed by a morpheme. Obviously, any parallel English-Swedish parallel corpus aligned in this way will have a high number of deletions. Hence we must focus the comparisons on relative numbers rather than absolute ones.

The paper reports both absolute and relative frequencies for different types of correspondences. We also provide definitions for the types. The definitions are encoded in a suite of Perl scripts that generate the actual counts. The χ^2 statistic has been used for tests of statistical significance. For the reasons given above, we also quantify properties of the English (source) texts to see to what extent the observed structural differences between the English and the Swedish texts can be accounted for on the grounds of well-known grammatical differences.

3. Data

The parallel corpus used for this study is a fairly small one, but on the other hand all syntactic annotations and word alignments have been manually checked so as to be in ac-

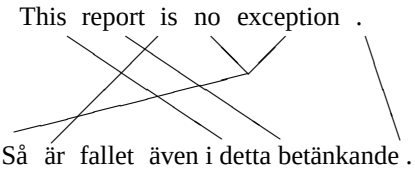
Subcorpus	Text type	Sentences	Src words	Trg words	Ratio
Access	online help texts	595	10,451	8,898	0.85
Europarl	parliamentary debates	594	9,334	8,715	0.93
Bellow	fiction	604	10,310	9,962	0.97
HarryP	fiction	600	10,171	10,501	1.03
Sums:		2393	40,266	38,076	0.95

Table 1: Corpus overview showing text type and size.

cord with specified guidelines (Ahrenberg, 2007a).[2] The corpus comprises four subcorpora as outlined in Table 1: on-line help texts for MS Access for Windows XP (Access), Europarl data (Europarl), excerpts from a novel by Saul Bellow[3] (Bellow), and excerpts from the second Harry Potter book[4] (HarryP). Each subcorpus used for the study has a size of roughly 600 sentence pairs.

The syntactic annotation employs parts-of-speech, morphological properties, and dependency functions. Every sentence is assumed to have a unique head, marked by the function 'main', and all other tokens, except punctuation marks, are direct or indirect dependents of the head. Monolingual files are XML-formatted. An annotated segment pair is shown in Figure 1.

The word alignment is based both on semantic and structural correspondence where many-to-many alignments (as usual) represent corresponding units that cannot be analysed into smaller (1-1, 1-n, or n-1) alignments. Word alignments are complete, i.e., a decision has been made for each token in the corpus if, and how, it corresponds to something in the other language. A word link is represented as a paired list of indices such as (4-5/1) which says that the 4th and 5th words of the source sentence have been linked to the first word of the target sentence. The alignment encoding for the sentences in Figure 1 is shown in Figure 2.



<(1/6),(2/7),(3/2),(4-5/1),(6/8),(0/3),(0/4),(0/5)>

Figure 2: Alignment for the sentences in Figure 1 as graph (top) and position mapping (bottom).

Null links are represented by the number 0. For example, (0/3) means that the third word of the Swedish sentence is judged to have no correspondent in the English sentence.

[2]While I cannot claim that the annotation is completely free of errors and inconsistencies, it is judged to be accurate enough to support global counts of the kind of performed in this study.

[3]Saul Bellow: *To Jerusalem and back: a personal account*, Viking Press, New York, 1976. Swedish translation: *Jerusalem tur och retur* by Caj Lundgren, Stockholm AB, 1977.

[4]J. R. Rowling: *Harry Potter and the Chamber of Secrets*, Arthur Levine Books, 1999. Swedish translation: *Harry Potter och hemligheternas kammare*, by Lena Fries-Gedin, Tiden, 2001.

We refer to the smallest index of an index list λ as $min(\lambda)$ and the largest index as $max(\lambda)$. Thus, with λ = 4-5-7, $min(\lambda)$ = 4, and $max(\lambda)$ = 7.

4. Comparisons

We report absolute and relative frequencies based on different (combinations of) data. First, we discuss the figures from Table 1 and try to explain the differences that can be seen with respect to token counts. Then we make several alignment-based comparisons, first using alignment only, and then combining it with the syntactic annotation.

Subcorpus	En N	En N N	Sw N	Sw N N
Access	2852	695	2196	92
Europarl	1944	193	1795	107
Bellow	1767	152	1736	70
HarryP	1690	154	1622	30
Σ	8253	1194	7349	299

Table 2: Distribution of Noun-Noun sequences in the four subcorpora. Note that not all N N sequences are compounds and that proper nouns are not included in these counts.

While the main focus is on the Europarl subcorpus, we also take note of properties of the other subcorpora when they stick out from the others.

4.1. Comparisons based on monolingual files

Already Table 1 indicates differences between the subcorpora. For HarryP the translation actually has more tokens than the source, while the Access translation has much fewer tokens than the others. The reason for these differences is not obvious without a more fine-grained inspection. We can find some clues, however, by looking at how tokens are distributed on different parts-of-speech and dependency functions. This reveals, for instance, that Access has the largest percentage drop in nouns, something which is explained by the higher numbers of compound nouns in Access. As shown in Table 2, there are more Noun-Noun sequences in Access than in the other three subcorpora combined, and a corresponding drop in the number of nouns in the Swedish version. Further inspection shows that a large part of this reduction (585 instances) is explained by a drop in the number of nouns that are attributes, i.e., nouns that make up the first part of an English nominal compound.

4.2. Comparisons based on word alignments

Alignment-based comparisons can provide more detailed information on restructurings. A basic typing scheme is

Correspondence	Access	Europarl	Bellow	HarryP
1-1 (isomorphism)	6916 (69.7%)	**6236 (64.9%)**	7828 (74.8%)	7585 (71.3%)
1-0 (deletion)	1415 (14.2%)	**1488 (15.4%)**	1150 (10.9%)	1120 (10.5%)
0-1 (addition)	474 (4.7%)	**1001 (10.4%)**	661 (6.3%)	1020 (9.6%)
many-1 (reduction)	**816 (8.1%)**	427 (4.4%)	343 (3.2%)	290 (2.6%)
1-many (expansion)	255 (2.5%)	349 (3.6%)	403 (3.8%)	**501 (4.3%)**
many-many (paraphrase)	40 (0.4%)	100 (1.0%)	74 (0.6%)	108 (1.3%)
Σ	9916	9601	10459	9724

Table 3: Distribution of word alignments on different types. Significant extremes are marked in boldface.

given by the number of tokens in a link, as shown in Table 3.

In Table 3, unlike Tables 1 and 2, Europarl comes out as extreme in certain aspects. It is the subcorpus with the least percentage of 1-to-1 word links, and it is the subcorpus with the highest number of deletions, and is close to the top in terms of additions. Moreover, a smaller share, some 38%, of the deletions are explained by null-links for non-corresponding function words, such as the definite article *the* and copular instances of the verb *be*, than for the other subcorpora, where the share is 40-50%.

If the Europarl corpus is compared to the union of the other three corpora, it is significantly different at the 0.01 level from them in the number of deletions, additions and isomorphisms. It has a larger share of deletions and additions and a smaller share of 1-to-1 correspondences.

We can also see that the large reduction of tokens for Access, as observed in Table 1, is largely explained by a high proportion of many-to-1 word links, the link type that applies to compound nouns.

For HarryP there is a high proportion of both additions and expansions, and a low proportion of deletions. This agrees well with the high ratio of target tokens to source tokens that was observed in Table 1.

Counting only the number of tokens in a link pays no regard to whether the tokens occur in a sequence or not. The types that involve more than one token on either side, i.e., reductions, expansions and paraphrases could be further categorized based on the occurrence of splits. Splits are not very numerous and occur on average in just above 1% of all sentences. In general, there are more splits on the Swedish side although the ratio varies from 1.4 (Access) to 3.9 (Bellow). Total sums are given in Table 4.

Subcorpus	Split src	Split trg	Sum	Ratio
Access	20	27	47	2.43
Europarl	23	52	**75**	**4.15**
Bellow	12	47	59	2.91
HarryP	23	56	**79**	**3.82**

Table 4: Number of links with a split token sequence and ratio per 1000 tokens.

Here, HarryP is the subcorpus with most splits in absolute numbers, whereas Europarl has the highest ratio of splits compared to the number of tokens. The difference between these two is not significant.

4.3. Comparing reordering

Another informative feature of a word alignment is the amount of reordering it contains. (Fox, 2002) made a study of reorderings in the English-French test corpus from the Hansards used by (Och and Ney, 2000) – and many others after them. To measure the amount of reordering she identified instances of crossings, where a crossing was defined as two phrases having overlapping spans on the target side. In addition she took into account the fact that her corpus contains both sure and possible links and distinguished crossings of heads with modifiers and crossings of modifiers.

With a syntactic analysis that is based on dependencies rather than phrase structure we can gain useful information by restricting attention to word links. Also, the relevant feature is permutation rather than overlap, as the word alignment divides the tokens of a sentence pair into clearly separate links. For the metric only link pairs that are adjacent on the source side are considered, ignoring intervening deletions. Thus, a *crossing* occurs if there are two links $< \sigma_1, \tau_1 >, < \sigma_2, \tau_2 >$ such that

$max(\sigma_1) < min(\sigma_2)$, and
$max(\tau_2) < min(\tau_1)$,

and there is no other non-null link $< \sigma_3, \tau_3 >$ with those two properties for which $max(\sigma_1) < min(\sigma_3) < min(\sigma_2)$. In case there is a split, we check positions both before and after the split. Thus, what are counted are instances where a link has a crossing with the nearest non-deleted neighbor. For example, in Figure 1 there are two crossings, one for the pair of links <report:betänkande, is:är> and one for the pair <is:är, no exception:så>.

In addition, we need a measure for the span, or spread, of a crossing. For this purpose we measure the size of a crossing in terms of the difference in target word indices: $min(\tau_1) - max(\tau_2)$. Note that this difference will be greater when there are null-aligned (i.e., added) tokens on the target side in between $max(\tau_2)$ and $min(\tau_1)$. Alternatively, we could count the number of links in the same interval.

Table 5 presents data on crossings. Bellow is the corpus with the highest number of crossings and both Bellow and HarryP have a larger proportion of length 1 crossings than Europarl. This is partly due to a high number of crossings involving a one-word subject and a finite verb. This in turn can be explained by their genre as written narratives with

Measure	Access	Europarl	Bellow	HarryP
Length 1 crossings	247 (46.7%)	257 (43.6%)	**378 (57.5%)**	**275 (55.1%)**
Length 2-5 crossings	224 (42.3%)	**260 (44.1%)**	230 (35.0%)	183 (36.7%)
Longer crossings	58 (11.0%)	**73 (12.4%)**	49 (7.5%)	41 (8.2%)
Sums (No. per sentence):	529 (0.89)	590 (0.99)	**657 (1.09)**	499 (0.83)

Table 5: Number of crossings of different types in each subcorpus.

many source sentences beginning with an adverbial and a subject before the finite verb, where the translator is forced to move one of them after the finite verb in the Swedish translation, as in the following examples (from Bellow):

EN: Then *someone says* that it ca n't be long now...
SE: Då *säger någon* att det inte kan dröja länge nu...
Gloss: Then *says someone* that it not can be long now...

EN: Silent , *I give* his case some thought
SE: Under tystnad *ägnar jag* en smula eftertanke åt fallet
Gloss: Under silence *give I* a bit afterthought to case-DEF

While Europarl is the second subcorpus in order as regard number of crossings, it is the one with the highest proportion of long crossings, i.e., crossings that are not simple swaps. Such crossings are present also in quite short sentences as evidenced by the examples listed in Section 1.

It may be noted that our way of measuring crossings is asymmetric. While absolute numbers are slightly different when crossings are counted in the opposite direction, the general tendencies and relative differences are the same.

4.4. Types of structural correspondence

To estimate the amount of restructuring in a parallel corpus, it is obviously of interest to look at phrasal correspondences. It is not evident, however, how phrasal correspondences should be typed. Most methods for generating sub-sentential correspondences from parse trees rely on some form of wellformedness constraints, whether performed manually (Samuelsson and Volk, 2007) or automatically (Lavie et al., 2008; Tinsley et al., 2007). A common assumption is that head-dependent relations are kept, i.e., if two non-terminal nodes have been aligned, the daughters of one of them can only be aligned to daughters of the other. Head-dependency reversals are not uncommon, though, as in the example below (from Europarl), where the main verb, *have a place*, from the source text has been made the head of an embedded clause in the translation, due to the introduction of a presentation construction *det är naturligt ...* (it is natural) as a translation of the adverbial *Naturally*:

EN: *Naturally,* the Turkish Cypriots will have a place in the representation.
SE: Och *det är naturligt att* i delegationen för Cyperns lagliga och erkända regering även turkcyprioter kommer att kunna ingå.
Gloss: And *it is natural that* in delegation_DEF for Cyprus' legal and recognized government also Turkish Cypriots will be able to take_part

To assess how common this type of restructuring is, the English sentence heads that correspond to Swedish sentence heads have been counted. These counts are shown in Table 6, again putting the Europarl corpus on top. It has significantly more of restructuring in this respect than the other subcorpora.

In another experiment we considered all binary head-dependency relations in the source data and their correspondents in the target data. The taxonomy for these relations is based on the following features: (1) the occurrence of null links for one or both of the source tokens; (2) whether the tokens correspond to the same or different target tokens; (3) whether the dependency direction is kept, reversed, or levelled out; (4) whether the target tokens have an immediate dependency relation, when they have a dependency relation at all. The following types were defined, where D refers to the dependent token, H to its head, and D' and H' to their respective translations:

- **Deletion.** The dependendant D, or the head H, or both of them have been null-aligned.
- **Conflation.** The dependent and the head correspond to the same token on the target side.
- **Isomorphism.** The source dependency D→H corresponds to a dependency D'→H'on the target side.
- **Stretched dependency.** The dependent, D, corresponds to a token D' for which H' is a head, but not an immediate head.
- **Reversal.** The source dependency D→H corresponds to a reversed dependency H'→D'on the target side.
- **Stretched reversal.** The source head, H, corresponds to a token, H', for which D' is a head, but not an immediate head.
- **Levelling.** H' is not a head for D', nor is D' a head for H'.

Table 7 shows the distribution of dependency correspondences for these different types. Only links where both tokens have at most one corresponding token in the target data have been included. Source dependencies that do not meet this criterion are noted as 'Skipped'.

The picture from Table 7 largely corroborates the earlier findings. Europarl has the highest share of deletions, which is not surprising, since it has the highest share of deletions at the word level. It also has the lowest share of pure isomorphisms.

The relative frequency of reversals may seem high and is partly explained by differences in syntactic annotation for

Relation	Access	Europarl	Bellow	HarryP
main - main	527	476	538	530
main - nonmain	68	**118**	66	70
percentage	11.4	**19.8**	10.9	11.7

Table 6: Alignments of sentence heads in the four subcorpora. Note that the category main-main includes 1-many, many-1 and many-many links as well, when both sides contain the main token.

Type	Access	Europarl	Bellow	HarryP
Deletions	1480 (18.0%)	**1766 (26.2%)**	1304 (18.2%)	1125 (16.6%)
Conflation	**890 (10.8%)**	543 (8.0%)	361 (5.0%)	342 (5.0%)
Isomorphisms	4757 (57.9%)	**3458 (51.3%)**	4597 (64.2%)	4254 (62.9%)
Stretched Dependency	407 (4.9%)	366 (5.4%)	372 (5.2%)	397 (5.9%)
Reversals	**201 (2.4%)**	135 (2.0%)	91 (1.3%)	113 (1.7%)
Stretched Reversal	27 (0.3%)	22 (0.3%)	27 (0.3%)	23 (0.3%)
Levelling	457 (5.6%)	452 (6.7%)	413 (5.7%)	512 (7.6%)
Subtotals	8219	6742	7165	6766
Skipped	649	1095	1100	1542
Σ	8868	7837	8265	8308

Table 7: Frequencies for different types of relation between a source text dependency relation and its corresponding target image. Percentages are based on the number of relations that could be typed, stated in the row Subtotals.

the two languages. For example, Swedish passive participles agree in number and gender with their subjects, as do adjectives, and are usually analysed as subject predicatives, while English passive participles are analysed as heads. To illustrate: in an English sentence *X is installed* the participle is the head of the copula, while in the corresponding Swedish sentence *X är installerad*, the direction of the dependency goes in the opposite direction. Another contributing factor is that some common English verbs, such as *want* are analysed as main verbs, while the common Swedish translation *vill* is analysed as an auxiliary. And coordinations that relate two first parts of a compound are also analyzed differently. Normally in a coordinated construction the first conjunct is taken to be the head. When two first parts of a compound are coordinated in Swedish, however, the second part is taken to be the head, as in the following example:

EN: *row and column areas*
SE: *rad- och kolumnområden*

Here, in the English phrase, *column* is a dependent of *row*, while in the Swedish translation, the correspondent of *row*, namely *rad* is a dependent of *kolumnområden*.

However, the relative difference in the number of reversals for the Access subcorpus and the others is still significant. It is explained by a large number of noun phrases consisting of a proper noun, or a noun used as a name, and a descriptive noun, where the Swedish translation reverses both the linear order and the dependency direction, as in the following examples:

EN: *the Orders field*
SE: *fältet Order*
Gloss: field-DEF Order

EN: *Enable system administrator user name check box*
SE: *kryssrutan Aktivera användarnamn för systemadministratör*
Gloss: *check-box-DEF Enable user-name for system-administrator*

Access also shows the highest number of conflations but this is not so surprising since we know from Table 2 that Access has a large number of compound nouns.

5. Conclusions and future work

This is a small study which does not permit very definite conclusions. There is a clear indication, however, that Europarl data contains a high share of structurally complex relations, in particular additions, deletions, and long distance reorderings on a level comparable to those that can be found in fiction. It also has a high share of non-corresponding sentence heads. At least, this seems to be the case for the English-Swedish data in the LinES parallel treebank. We believe that a major reason is that Europarl data includes parallel translations, not just source text and translation.

The study can be extended in several ways. First, more detailed studies can be performed by further categorization in terms of the parts-of-speech and dependency relations involved. We have shown a few examples of such more fine-grained analyses, but the picture could easily be made more complete. Second, we would like to include other corpora that are much used in statistical MT, such as the JRC-Acquis (Steinberger et al., 2006) in the LinES treebank and perform a similar study. The method itself can also be improved by the inclusion of a more developed taxonomy for correspondences at the phrasal level.

Another extension is to see whether profiling can be based on automatic tools. In particular, it would be interesting to compare results from precision-oriented align-

ment methods such as symmetrized Giza++ with intersection (Och and Ney, 2003) or Holmqvist's pattern-based word alignment (Holmqvist,).

Another line of research is the development of appropriate automatic filters on existing training corpora to reduce complexity, and the compilation of alternative parallel and annotated corpora of less complexity. Of course, we cannot say on the basis of this study how instances of complex correspondence relations affect the translation models and phrase tables that are generated from Europarl data and used in statistical machine translation, nor whether they have any adverse effects at all. This should also be a topic for further investigation. However, the restructuring that is found certainly goes beyond what current SMT systems can produce, and rather calls for an example-based approach. Thus, we believe that it is not a good idea to include such translations in reference sets used for testing, since the prime application for statistical systems is gisting rather than translations with publishing-quality.

6. References

Lars Ahrenberg. 2007a. LinES 1.0 annotation: format, contents and guidelines. http://www.ida.liu.se/ lah/transmap/Corpus/guidelines.pdf.

Lars Ahrenberg. 2007b. LinES: An English-Swedish parallel treebank. In *Proceedings of The 16th Nordic Conference of Computational Linguistics*, Tartu, Estonia, May, 24-26.

Heidi J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Maria Holmqvist. 2010. Heuristic word alignment with parallel phrases. In *The 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Malta, May 19-21.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, Phuket, Thailand.

Alon Lavie, Alok Parlikar, and Vamshi Ambati. 2008. Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. In *Proceedings of the Second ACL Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, pages 87–95, Columbus, Ohio, USA, June. Association for Computational Linguistics.

Franz J. Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Coling '00: The 18th International Conference on Computational Linguistics*, pages 1086–1090, Saarbrücken, Germany, August.

Franz J. Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*, pages 1–6, Hong Kong. ACL.

Yvonne Samuelsson and Martin Volk. 2007. Alignment tools for parallel treebanks. In *Proceedings of the GLDV Früjahrestagung*, Tübingen, Germany.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, s Dan Tufi and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genoa, Italy, May, 24-26.

John Tinsley, Mary Hearne, and Andy Way. 2007. Exploiting parallel treebanks to improve phrase-based statistical machine translation. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT-07)*, pages 175–187, Bergen, Norway.