

Phrasal Cohesion and Statistical Machine Translation

Heidi J. Fox

Brown Laboratory for Linguistic Information Processing
Department of Computer Science
Brown University, Box 1910, Providence, RI 02912
hj@cs.brown.edu

Abstract

There has been much interest in using phrasal movement to improve statistical machine translation. We explore how well phrases cohere across two languages, specifically English and French, and examine the particular conditions under which they do not. We demonstrate that while there are cases where coherence is poor, there are many regularities which can be exploited by a statistical machine translation system. We also compare three variant syntactic representations to determine which one has the best properties with respect to cohesion.

1 Introduction

Statistical machine translation (SMT) seeks to develop mathematical models of the translation process whose parameters can be automatically estimated from a parallel corpus. The first work in SMT, done at IBM (Brown et al., 1993), developed a noisy-channel model, factoring the translation process into two portions: the translation model and the language model. The translation model captures the translation of source language words into the target language and the reordering of those words. The language model ranks the outputs of the translation model by how well they adhere to the syntactic constraints of the target language.¹

The prime deficiency of the IBM model is the reordering component. Even in the most complex of

the five IBM models, the reordering operation pays little attention to context and none at all to higher-level syntactic structures. Many attempts have been made to remedy this by incorporating syntactic information into translation models. These have taken several different forms, but all share the basic assumption that phrases in one language tend to stay together (i.e. cohere) during translation and thus the word-reordering operation can move entire phrases, rather than moving each word independently.

(Yarowsky et al., 2001) states that during their work on noun phrase bracketing they found a strong cohesion among noun phrases, even when comparing English to Czech, a relatively free word order language. Other than this, there is little in the SMT literature to validate the coherence assumption. Several studies have reported alignment or translation performance for syntactically augmented translation models (Wu, 1997; Wang, 1998; Alshawi et al., 2000; Yamada and Knight, 2001; Jones and Havrilla, 1998) and these results have been promising. However, without a focused study of the behavior of phrases across languages, we cannot know how far these models can take us and what specific pitfalls they face.

The particulars of cohesion will clearly depend upon the pair of languages being compared. Intuitively, we expect that while French and Spanish will have a high degree of cohesion, French and Japanese may not. It is also clear that if the cohesion between two closely related languages is not high enough to be useful, then there is no hope for these methods when applied to distantly related languages. For this reason, we have examined phrasal cohesion for French and English, two languages which are fairly close syntactically but have enough differences to be

¹Though usually a simple word n-gram model is used for the language model.

interesting.

2 Alignments, Spans and Crossings

An *alignment* is a mapping between the words in a string in one language and the translations of those words in a string in another language. Given an English string, $\mathbf{e} = e_1^l \equiv e_1 e_2 \dots e_l$, and a French string, $\mathbf{f} = f_1^m \equiv f_1 f_2 \dots f_m$, an alignment \mathbf{a} can be represented by $\mathbf{a} = a_1^l \equiv a_1 a_2 \dots a_l$. Each a_i is a set of indices into \mathbf{e} where $j \in a_i$; $1 \leq j \leq m$; $0 \leq i \leq l$ indicates that word j in the French sentence is aligned with word i in the English sentence. $a_i = \emptyset$ indicates that English word i has no corresponding French word.

Given an alignment \mathbf{a} and an English phrase covering words $e_i \dots e_j$, the *span* is a pair where the first element is $\min(a_i, \dots, a_j)$ and the second element is $\max(a_i, \dots, a_j)$. Thus, the span includes all words between the two extrema of the alignment, whether or not they too are part of the translation. If phrases cohere perfectly across languages, the span of one phrase will never overlap the span of another. If two spans do overlap, we call this a *crossing*.

Figure 1 shows an example of an English parse along with the alignment between the English and French words (shown with dotted lines). The English word “not” is aligned to the two French words “ne” and “pas” and thus has a span of [1,3]. The main English verb “change” is aligned to the French “modifie” and has a span of [2,2]. The two spans overlap and thus there is a crossing. This definition is asymmetric (i.e. what is a crossing when moving from English to French is not guaranteed to be a crossing when moving from French to English). However, we only pursue translation direction since that is the one for which we have parsed data.

3 Experiments

3.1 Data

To calculate spans, we need aligned pairs of English and French sentences along with parses for the English sentences. Our aligned data comes from a corpus described in (Och and Ney, 2000) which contains 500 sentence pairs randomly selected from the Canadian Hansard corpus and manually aligned. The alignments are of two types: *sure* (S) and *possible* (P). S alignments are those which are unam-

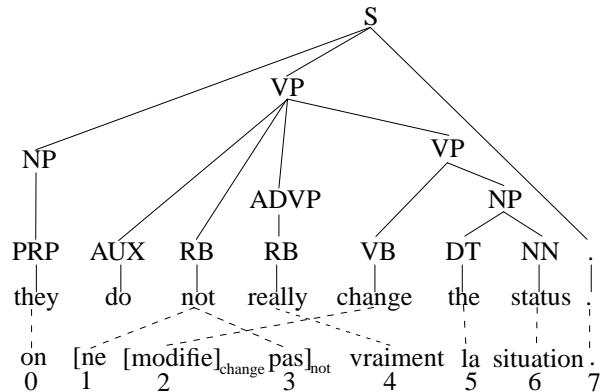


Figure 1: Alignment Example with Crossing

biguous while P alignments are those which are less certain. P alignments often appear when a phrase in one language translates as a unit into a phrase in the other language (e.g. idioms, free translations, missing function words) but can also be the result of genuine ambiguity. When two annotators disagree, the union of the P alignments produced by each annotator is recorded as the P alignment in the corpus. When an S alignment exists, there will always also exist a P alignment such that $P \supseteq S$. The English sentences were parsed using a state-of-the-art statistical parser (Charniak, 2000) trained on the University of Pennsylvania Treebank (Marcus et al., 1993).

3.2 Phrasal Translation Filtering

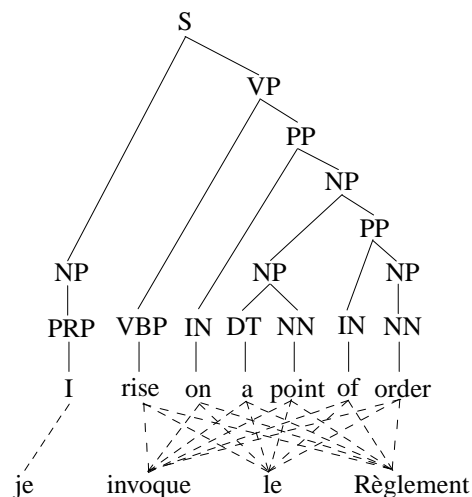


Figure 2: Phrasal Translation Example

Since P alignments often align phrasal transla-

Alignment Type	Phrasal Filter Off			Phrasal Filter On		
	S	S→P	P	S	S→P	P
Head Crossings	0.236	4.790	5.284	0.172	2.772	2.492
Modifier Crossings	0.056	0.880	0.988	0.048	0.516	0.362
Phrasal Translations	–	–	–	0.072	2.382	3.418

Table 1: Average Number of Crossings per Sentence

tions, the number of crossings when P alignments are used will be artificially inflated. For example, in Figure 2 note that every pair of English and French words under the verb phrase is aligned. This will generate five crossings, one each between the pairs VBP-PP, IN-NP₃, NP₂-PP, NN-DT, and IN-NP₁. However, what is really happening is that the whole verb phrase is first being moved without crossing anything else and then being translated as a unit. For our purposes we want to count this example as producing zero crossings. To accomplish this, we defined a simple heuristic to detect phrasal translations so we can filter them if desired.

3.3 Calculating Crossings

After calculating the French spans from the English parses and alignment information, we counted crossings for all pairs of child constituents in each constituent in the sentence, maintaining separate counts for those involving the head constituent of the phrase and for crossings involving modifiers only. We did this while varying conditions along two axes: alignment type and phrasal translation filtering. Recalling the two different types of alignments, S and P, we examined three different conditions: S alignments only, P alignments only, or S alignments where present falling back to P alignments (S→P). For each of these conditions, we counted crossings both with and without using the phrasal translation filter.

For a given alignment type $A \in \{S, S \rightarrow P, P\}$, let $C_A(p_1, p_2) = 1$ if phrases p_1 and p_2 cross each other and 0 otherwise. Let $F_A(p_1, p_2) \equiv 1$ if the phrasal translation filter is turned off. If the filter is on,

$$F_A(p_1, p_2) = \begin{cases} 0 & \text{if } p_1 \text{ and } p_2 \text{ are part} \\ & \text{of a phrasal translation} \\ & \text{in alignment } A \\ 1 & \text{otherwise} \end{cases}$$

Then, for a given phrase p with head constituent

h , modifier constituents M , and child constituents $C = M \cup \{h\}$ and for a particular alignment type A , the number of head crossings C_A^h and modifier crossings C_A^m can be calculated recursively:

$$C_A^h(p) = \sum_{c \in C} C_a^h(c) + \sum_{m \in M} C_A(h, m) F_A(h, m)$$

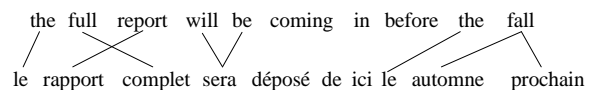
$$C_A^m(p) = \sum_{c \in C} C_a^m(c) + \sum_{\substack{m \in M \\ m' \in M, m' \neq m}} C_A(m', m) F_A(m', m)$$

4 Results

4.1 Average Crossings

Table 1 shows the average number of crossings per sentence. The table is split into two sections, one for results when the phrasal filter was used and one for when it was not. “Alignment Type” refers to whether we used S, P or S→P as the alignment data. The “Head Crossings” line shows the results when comparing the span of the head constituent of a phrase with the spans of its modifier constituents, and “Modifier Crossings” refers to the case where we compare the spans of pairs of modifiers. The “Phrasal Translations” line shows the average number of phrasal translations detected per sentence.

For S alignments, the results are quite promising, with an average of only 0.236 head crossings per sentence and an even smaller average for modifier crossings (0.056). However, these results are overly optimistic since often many words in a sentence will not have an S alignment at all, such as “coming”, “in”, and “before” in following example:



When we use P alignments for these unaligned words (the S→P case), we get a more meaningful result. Both types of crossings are much more frequent (4.790 for heads and 0.88 for modifiers) and

phrasal translation filtering has a much larger effect (reducing head average to 2.772 and modifier average to 0.516). Phrasal translations account for almost half of all crossings, on average. This effect is even more pronounced in the case where we use P alignments only. This reinforces the importance of phrasal translation in the development of any translation system.

Even after filtering, the number of crossings in the S→P case is quite large. This is discouraging, however there are reasons why this result should be looked on as more of an upper bound than anything precise. For one thing, there are cases of phrasal translation which our heuristic fails to recognize, an example of which is shown in Figure 3. The alignment of “explorer” with “this” and “matter” seems to indicate that the intention of the annotator was to align the phrase “work this matter out”, as a unit, to “de explorer la question”. However, possibly due to an error during the coding of the alignment, “work” and “out” align with “de” (indicated by the solid lines) while “this” and “matter” do not. This causes the phrasal translation heuristic to fail resulting in a crossing where there should be none.

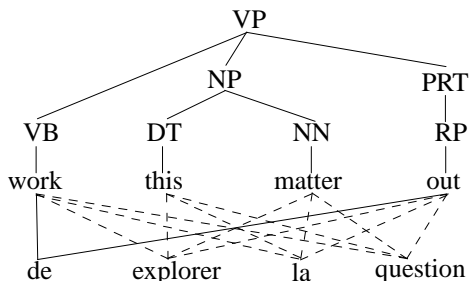


Figure 3: Phrasal Translation Heuristic Failure

Also, due to the annotation guidelines, P alignments are not as consistent as would be ideal. Recall that in cases of annotator disagreement, the P alignment is taken to be the union of the P alignments of both annotators. Thus, it is possible for the P alignment to contain two mutually conflicting alignments. These composite alignments will likely generate crossings even where the alignments of each individual annotator would not. While reflecting genuine ambiguity, an SMT system would likely pursue only one of the alternatives and only a portion of the crossings would come into play.

4.2 Percentage Crossings

Our results show a significantly larger number of head crossings than modifier crossings. One possibility is that this is due to most phrases having a head and modifier pair to test, while many do not have multiple modifiers and therefore there are fewer opportunities for modifier crossings. Thus, it is informative to examine how many potential crossings actually turn out to be crossings. Table 2 provides this result in the form of the percentage of crossing tests which result in detection of a crossing.

To calculate this, we kept totals for the number of head (T_A^h) and modifier (T_A^m) crossing tests performed as well as the number of phrasal translations detected (T_A^{ph}). Note that when the phrasal translation filter is turned on, these totals differ for each of the different alignment types (S, S→P, and P).

$$T_A^h(p) = \sum_{c \in C} T_A^h(c) + \sum_{m \in M} F_A(h, m)$$

$$T_A^m(p) = \sum_{c \in C} T_A^m(c) + \sum_{\substack{m \in M \\ m' \in M, m' \neq m}} F_A(m', m)$$

$$T(p) = \binom{|C|}{2} + \sum_{c \in C} T(c)$$

$$T_A^{ph}(p) = T(p) - T_A^h(p) - T_A^m(p)$$

The percentages are calculated after summing over all sentences s in the corpus:

$$\%_A^h = 100 * \frac{\sum_s C_A^h(s)}{\sum_s T_A^h(s)} \quad \%_A^m = 100 * \frac{\sum_s C_A^m(s)}{\sum_s T_A^m(s)}$$

$$\%_A^{ph} = 100 * \frac{\sum_s T_A^{ph}(s)}{\sum_s T(s)}$$

There are still many more crossings in the S→P and P alignments than in the S alignments. The S alignment has 1.58% head crossings while the S→P and P alignments have 32.16% and 35.47% respectively, with similar relative percentages for modifier crossings. Also as before, half to two-thirds of crossings in the S→P and P alignments are due to phrasal translations. More interestingly, we see that modifier crossings remain significantly less prevalent than head crossings (e.g. 14.45% vs. 32.16% for the S→P case) and that this is true uniformly across all parameter settings. This indicates that heads are more intimately involved with their modifiers than

Alignment Type	Phrasal Filter Off			Phrasal Filter On		
	S	S→P	P	S	S→P	P
Head Crossings	1.58%	32.16%	35.47%	1.15%	18.61%	16.73%
Modifier Crossings	0.92%	14.45%	16.23%	0.78%	8.47%	5.94%
Phrasal Translations	–	–	–	0.34%	11.35%	16.29%

Table 2: Percent Crossings per Chance

Cause	Count
Ne Pas	13
Modal	9
Adverb	8
Possessive	2
Pronoun	2
Adjective	1
Parser Error	16
Reword	16
Reorder	13
Translation Error	5
Total	86

Table 3: Causes of Head Crossings

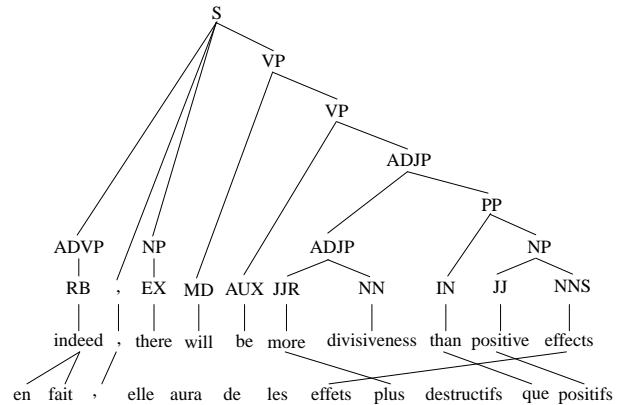


Figure 4: Crossing Due to Rewording

modifiers are with each other and therefore are more likely to be involved in semi-phrasal constructions.

5 Analysis of Causes

Since it is clear that crossings are too prevalent to ignore, it is informative to try to understand exactly what constructions give rise to them. To that end, we examined by hand all of the head crossings produced using the S alignments with phrasal filtering. Table 3 shows the results of this analysis.

The first thing to note is that by far most of the crossings do not reflect lack of phrasal cohesion between the two languages. Instead, they are caused either by errors in the syntactic analysis or the fact that translation as done by humans is a much richer process than just replication of the source sentence in another language. Sentences are reworded, clauses are reordered, and sometimes human translators even make mistakes.

Errors in syntactic analysis consist mostly of attachment errors. Rewording and reordering accounted for a large number of crossings as well. In most of the cases of rewording (see Figure 4) or re-

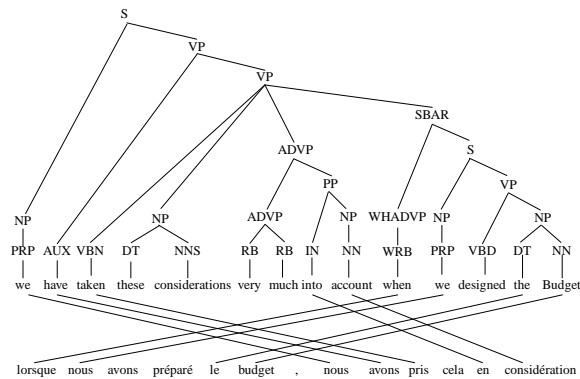


Figure 5: Crossing Due to Reordering of Clauses

ordering (see Figure 5) a more “parallel” translation would also be valid. Thus, while it would be difficult for a statistical model to learn from these examples, there is nothing to preclude production of a valid translation from a system using phrasal movement in the reordering phase. The rewording and reordering examples were so varied that we were unable to find any regularities which might be exploited by a translation model.

Among the cases which do result from language differences, the most common is the “ne ... pas” construction (e.g. Figure 1). Fifteen percent of the 86 total crossings are due to this construction. Because “ne ... pas” wraps around the verb, it will always result in a crossing. However, the types of syntactic structures (categorized as context-free grammar rules) which are present in cases of negation are rather restricted. Of the 47 total distinct syntactic structures which resulted in crossings, only three of them involved negation. In addition, the crossings associated with these particular structures were unambiguously caused by negation (i.e. for each structure, only negation-related crossings were present).

Next most common is the case where the English contains a modal verb which is aligned with the main verb in the French. In the example in Figure 6, “will be” is aligned to “sera” (indicated by the solid lines) and because of the constituent structure of the English parse there is a crossing. As with negation, this type of crossing is quite regular, resulting uniquely from only two different syntactic structures.

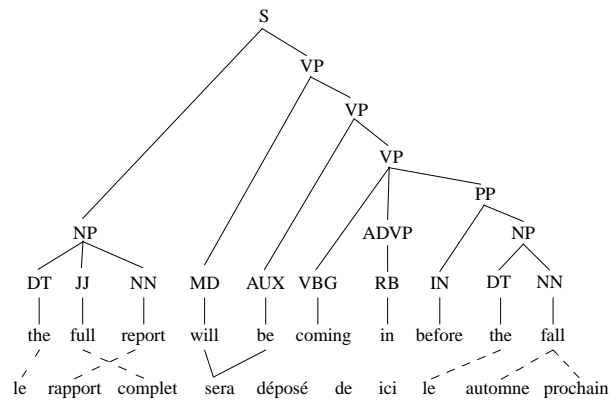


Figure 6: Crossing Due to Modal

Adverbs are a third common cause, as they typically follow the verb in French while preceding it in English. Figure 7 shows an example where the span of “simplement” overlaps with the span of the verb phrase beginning with “tells” (indicated by the solid lines). Unlike negation and modals, this case is far less regular. It arises from six different syntactic constructions and two of those constructions are implicated in other types of crossings as well.

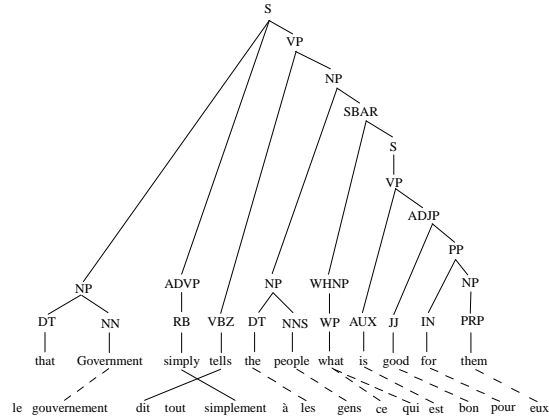


Figure 7: Crossing Due to Adverb

6 Further Experiments

6.1 Flattening Verb Phrases

Many of the causes listed above are related to verb phrases. In particular, some of the adverb-related crossings (e.g. Figure 1) and all of the modal-related crossings (e.g. Figure 6) are artifacts of the nested verb phrase structure of our parser. This nesting usually does not provide any extra information beyond what could be gleaned from word order. Therefore, we surmised that flattening verb phrases would eliminate some types of crossings without reducing the utility of the parse.

The flattening operation consists of identifying all nested verb phrases and splicing the children of the nested phrase into the parent phrase in its place. This procedure is applied recursively until there are no nested verb phrases. An example is shown in Figure 8. Crossings can be calculated as before.

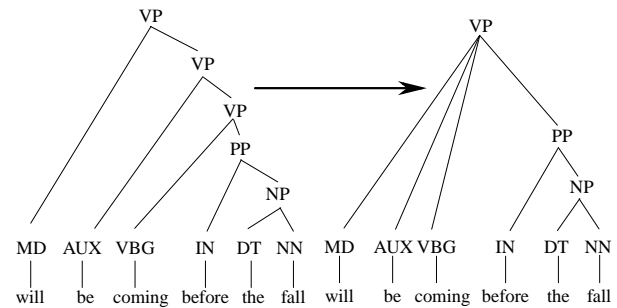


Figure 8: Verb Phrase Flattening

Alignment Type	S	S→P	P
Baseline	0.172	2.772	2.492
Flattened VPs	0.136	2.252	1.91
Dependencies	0.078	1.88	1.476

Table 4: Average Head Crossings per Sentence (Phrasal Filter On)

Alignment Type	S	S→P	P
Baseline	0.048	0.516	0.362
Flattened VPs	0.06	0.86	0.694
Dependencies	0.1	1.498	1.238

Table 5: Average Modifier Crossings per Sentence (Phrasal Filter On)

Flattening reduces the number of potential head crossings while increasing the number of potential modifier crossings. Therefore, we would expect to see a comparable change to the number of crossings measured, and this is exactly what we find, as shown in Tables 4 and 5. For example, for S→P alignments, the average number of head crossings decreases from 2.772 to 2.252, while the average number of modifier crossings increases from 0.516 to 0.86. We see similar behavior when we look at the percentage of crossings per chance (Tables 6 and 7). For the same alignment type, the percentage of head crossings decreases from 18.61% to 15.12%, while the percentage of modifier crossings increases from 8.47% to 10.59%. One thing to note, however, is that the total number of crossings of both types detected in the corpus decreases as compared to the baseline, and thus the benefits to head crossings outweigh the detriments to modifier crossings.

Alignment Type	S	S→P	P
Baseline	1.15%	18.61%	16.73%
Flattened VPs	0.91%	15.12%	12.82%
Dependencies	0.52%	12.62%	9.91%

Table 6: Percent Head Crossings per Chance (Phrasal Filter On)

Alignment Type	S	S→P	P
Baseline	0.78%	8.47%	5.94%
Flattened VPs	0.73%	10.59%	8.55%
Dependencies	0.61%	9.22%	7.62%

Table 7: Percent Modifier Crossings per Chance (Phrasal Filter On)

6.2 Dependencies

Our intuitions about the cohesion of syntactic structures follow from the notion that translation, as a meaning-preserving operation, preserves the dependencies between words, and that syntactic structures encode these dependencies. Therefore, dependency structures should cohere as well as, or better than, their corresponding syntactic structures. To examine the validity of this, we extracted dependency structures from the parse trees (with flattened verb phrases) and calculated crossings for them. Figure 9 shows a parse tree and its corresponding dependency structure.

The procedure for counting modifier crossings in a dependency structure is identical to the procedure for parse trees. For head crossings, the only difference is that rather than comparing spans of two siblings, we compare the spans of a child and its parent.

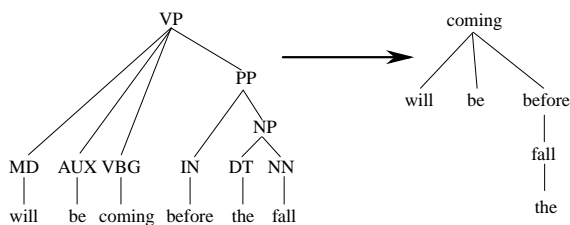


Figure 9: Extracting Dependencies

Again focusing on the S→P alignment case, we see that the average number of head crossings (see Table 4) continues to decrease compared to the previous case (from 2.252 to 1.88), and that the average number of modifier crossings (see Table 5) continues to increase (from 0.86 to 1.498). This time, however, the percentages for both types of crossings (see Tables 6 and 7) decrease relative to the case of flattened verb phrases (from 15.12% to 12.62% for heads and from 10.59% to 9.22% for modifiers). The percentage of modifier crossings is still higher

than in the base case (9.22% vs. 8.47%). Overall, however, the dependency representation has the best cohesion properties.

7 Conclusions

We have examined the issue of phrasal cohesion between English and French and discovered that while there is less cohesion than we might desire, there is still a large amount of regularity in the constructions where breakdowns occur. This reassures us that reordering words by phrasal movement is a reasonable strategy. Many of the initially daunting number of crossings were due to non-linguistic reasons, such as rewording during translation or errors in syntactic analysis. Among the rest, there are a small number of syntactic constructions which result in the majority of the crossings examined in our analysis. One practical result of this skewed distribution is that one could hope to discover the major problem areas for a new language pair by manually aligning a small number of sentences. This information could be used to filter a training corpus to remove sentences which would cause problems in training the translation model, or for identifying areas to focus on when working to improve the model itself. We are interested in examining different language pairs as the opportunity arises.

We have also examined the differences in cohesion between Treebank-style parse trees, trees with flattened verb phrases, and dependency structures. Our results indicate that the highest degree of cohesion is present in dependency structures. Therefore, in an SMT system which is using some type of phrasal movement during reordering, dependency structures should produce better results than raw parse trees. In the future, we plan to explore this hypothesis in an actual translation system.

8 Acknowledgments

The work reported here was supported in part by the Defense Advanced Research Projects Agency under contract number N66001-00-C-8008. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the United States Gov-

ernment.

We would like to thank Franz Och for providing us with the manually annotated data used in these experiments.

References

- Hiyan Alshawi, Srinivas Bangalore, and Shona Douglas. 2000. Learning dependency translation models as collections of finite-state head transducers. *Computational Linguistics*, 26(1):45–60, March.
- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.
- Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*.
- Douglas Jones and Rick Havrilla. 1998. Twisted pair grammar: Support for rapid development of machine translation for low density languages. In *Proceedings of the Conference of the Association for Machine Translation in the Americas*.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 13(2):313–330, June.
- Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447.
- Ye-Yi Wang. 1998. *Grammar Inference and Statistical Machine Translation*. Ph.D. thesis, Carnegie Mellon University.
- Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403, September.
- Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*.
- David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 109–116.