Empirical Lower Bounds on the Complexity of Translational Equivalence [*]

Benjamin Wellington
Computer Science Dept.
New York University
New York, NY 10003
{lastname}@cs.nyu.edu

Sonjia Waxmonsky
Computer Science Dept.
University of Chicago[†]
Chicago, IL, 60637
wax@cs.uchicago.edu

I. Dan Melamed
Computer Science Dept.
New York University
New York, NY, 10003
{lastname}@cs.nyu.edu

Abstract

This paper describes a study of the patterns of translational equivalence exhibited by a variety of bitexts. The study found that the complexity of these patterns in every bitext was higher than suggested in the literature. These findings shed new light on why "syntactic" constraints have not helped to improve statistical translation models, including finite-state phrase-based models, tree-to-string models, and tree-to-tree models. The paper also presents evidence that inversion transduction grammars cannot generate some translational equivalence relations, even in relatively simple real bitexts in syntactically similar languages with rigid word order. Instructions for replicating our experiments are at http://nlp.cs.nyu.edu/GenPar/ACL06

1 Introduction

Translational equivalence is a mathematical relation that holds between linguistic expressions with the same meaning. The most common explicit representations of this relation are word alignments between sentences that are translations of each other. The complexity of a given word alignment can be measured by the difficulty of decomposing it into its atomic units under certain constraints detailed in Section 2. This paper describes a study of the distribution of alignment complexity in a variety of bitexts. The study considered word alignments both in isolation and in combination with independently generated parse trees for one or both sentences in each pair. Thus, the study is relevant to finite-state phrase-based models that use no parse trees (Koehn et al., 2003), tree-to-string models that rely on one parse tree (Yamada and Knight, 2001), and tree-to-tree models that rely on two parse trees (Groves et al., 2004, e.g.).

The word alignments that are the least complex on our measure coincide with those that can be generated by an inversion transduction grammar (ITG). Following Wu (1997), the prevailing opinion in the research community has been that more complex patterns of word alignment in real bitexts are mostly attributable to alignment errors. However, the experiments in Section 3 show that more complex patterns occur surprisingly often even in highly reliable alignments in relatively simple bitexts. As discussed in Section 4, these findings shed new light on why "syntactic" constraints have not yet helped to improve the accuracy of statistical machine translation.

Our study used two kinds of data, each controlling a different confounding variable. First, we wanted to study alignments that contained as few errors as possible. So unlike some other studies (Zens and Ney, 2003; Zhang et al., 2006), we used manually annotated alignments instead of automatically generated ones. The results of our experiments on these data will remain relevant regardless of improvements in technology for automatic word alignment.

Second, we wanted to measure how much of the complexity is not attributable to systematic translation divergences, both in the languages as a whole (SVO vs. SOV), and in specific constructions (English *not* vs. French *ne...pas*). To eliminate this source of complexity of translational equivalence, we used English/English bitexts. We are not aware of any previous studies of word alignments in monolingual bitexts.

Even manually annotated word alignments vary in their reliability. For example, annotators sometimes link many words in one sentence to many

[*] Thanks to David Chiang, Liang Huang, the anonymous reviewers, and members of the NYU Proteus Project for helpful feedback. This research was supported by NSF grant #'s 0238406 and 0415933.

[†] SW made most of her contribution while at NYU.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 977–984,
Sydney, July 2006. ©2006 Association for Computational Linguistics

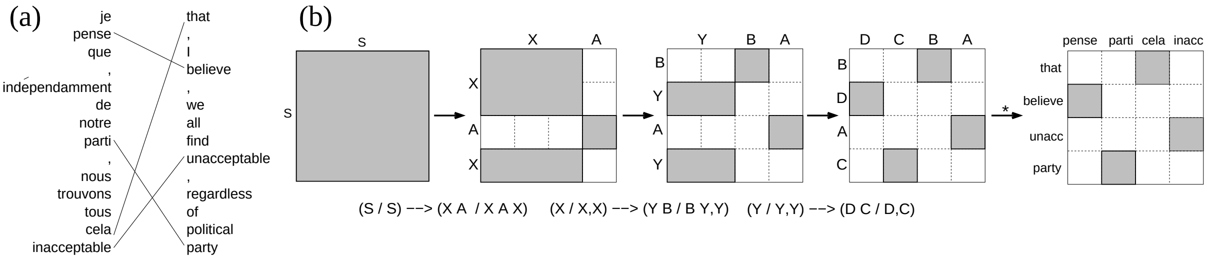

Figure 1: (a) Part of a word alignment. (b) Derivation of this word alignment using only binary and nullary productions requires one gap per nonterminal, indicated by commas in the production rules.

words in the other, instead of making the effort to tease apart more fine-grained distinctions. A study of such word alignments might say more about the annotation process than about the translational equivalence relation in the data. The inevitable noise in the data motivated us to focus on lower bounds, complementary to Fox (2002), who wrote that her results "should be looked on as more of an upper bound." (p. 307) As explained in Section 3, we modified all unreliable alignments so that they cannot increase the complexity measure. Thus, we arrived at complexity measurements that were underestimates, but reliably so. It is almost certain that the true complexity of translational equivalence is higher than what we report.

2 A Measure of Alignment Complexity

Any translation model can memorize a training sentence pair as a unit. For example, given a sentence pair like *(he left slowly / slowly he left)* with the correct word alignment, a phrase-based translation model can add a single 3-word biphrase to its phrase table. However, this biphrase would not help the model predict translations of the individual words in it. That's why phrase-based models typically decompose such training examples into their sub-biphrases and remember them too. Decomposing the translational equivalence relations in the training data into smaller units of knowledge can improve a model's ability to generalize (Zhang et al., 2006). In the limit, to maximize the chances of covering arbitrary new data, a model should decompose the training data into the smallest possible units, and learn from them.[1] For phrase-based models, this stipulation implies phrases of length one. If the model is a synchronous rewriting system, then it should be able to generate every training sentence pair as the yield of a *binary*-branching synchronous derivation tree, where every word-to-word link is generated by a different derivation step. For example, a model that uses production rules could generate the previous example using the synchronous productions
(S, S) → (X Y / Y X); (X, X) → (U V / U V);
(Y, Y) → (slowly, slowly); (U, U) → (he, he);
and (V, V) → (left, left).

A problem arises when this kind of decomposition is attempted for the alignment in Figure 1(a). If each link is represented by its own nonterminal, and production rules must be binary-branching, then some of the nonterminals involved in generating this alignment need discontinuities, or **gaps**. Figure 1(b) illustrates how to generate the sentence pair and its word alignment in this manner. The nonterminals X and Y have one discontinuity each.

More generally, for any positive integer k, it is possible to construct a word alignment that cannot be generated using binary production rules whose nonterminals all have fewer than k gaps (Satta and Peserico, 2005). Our study measured the complexity of a word alignment as the minimum number of gaps needed to generate it under the following constraints:

1. Each step of the derivation generates no more than two different nonterminals.

2. Each word-to-word link is generated from a separate nonterminal.[2]

Our measure of alignment complexity is analogous to what Melamed et al. (2004) call "fan-out."[3] The least complex alignments on this measure — those that can be generated with zero gaps — are precisely those that can be generated by an

[1] Many popular models learn from larger units at the same time, but the size of the smallest learnable unit is what's important for our purposes.

[2] If we imagine that each word is generated from a separate nonterminal as in GCNF (Melamed et al., 2004), then constraint 2 becomes a special case of constraint 1.

[3] For grammars that generate bitexts, fan-out is equal to the maximum number of allowed gaps plus two.

bitext	# SPs	min	median	max	95% C.I.
Chinese/English	491	4	24	52	.02
Romanian/English	200	2	19	76	.03
Hindi/English	90	1	10	40	.04
Spanish/English	199	4	23	49	.03
French/English	447	2	15	29	.01
Eng/Eng MTEval	5253	2	26	92	.01
Eng/Eng fiction	6263	2	15	97	.01

Table 1: Number of sentence pairs and minimum/median/maximum sentence lengths in each bitext. All failure rates reported later have a 95% confidence interval that is no wider than the value shown for each bitext.

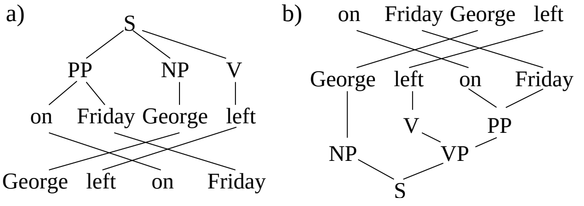

Figure 2: a) With a parse tree constraining the top sentence, a hierarchical alignment is possible without gaps. b) With a parse tree constraining the bottom sentence, no such alignment exists.

ITG. For the rest of the paper, we restrict our attention to binary derivations, except where explicitly noted otherwise.

To measure the number of gaps needed to generate a given word alignment, we used a bottom-up hierarchical alignment algorithm to infer a binary synchronous parse tree that was consistent with the alignment, using as few gaps as possible. A hierarchical alignment algorithm is a type of synchronous parser where, instead of constraining inferences by the production rules of a grammar, the constraints come from word alignments and possibly other sources (Wu, 1997; Melamed and Wang, 2005). A bottom-up hierarchical aligner begins with word-to-word links as constituents, where some of the links might be to nothing ("NULL"). It then repeatedly composes constituents with other constituents to make larger ones, trying to find a constituent that covers the entire input.

One of the important design choices in this kind of study is how to treat multiple links attached to the same word token. Word aligners, both human and automatic, are often inconsistent about whether they intend such sets of links to be disjunctive or conjunctive. In accordance with its focus on lower bounds, the present study treated them as disjunctive, to give the hierarchical alignment algorithm more opportunities to use fewer gaps. This design decision is one of the main differences between our study and that of Fox (2002), who treated links to the same word conjunctively.

By treating many-to-one links disjunctively, our measure of complexity ignored a large class of discontinuities. Many types of discontinuous constituents exist in text independently of any translation. Simard et al. (2005) give examples such as English verb-particle constructions, and the French negation *ne. . . pas*. The disparate elements of such constituents would usually be aligned to the same word in a translation. However, when our hierarchical aligner saw two words linked to one word, it ignored one of the two links. Our lower bounds would be higher if they accounted for this kind of discontinuity.

3 Experiments

3.1 Data

We used two monolingual bitexts and five bilingual bitexts. The Romanian/English and Hindi/English data came from Martin et al. (2005). For Chinese/English and Spanish/English, we used the data from Ayan et al. (2005). The French/English data were those used by Mihalcea and Pedersen (2003). The monolingual bitext labeled "MTEval" in the tables consists of multiple independent translations from Chinese to English (LDC, 2002). The other monolingual bitext, labeled "fiction," consists of two independent translations from French to English of Jules Verne's novel *20,000 Leagues Under the Sea*, sentence-aligned by Barzilay and McKeown (2001).

From the monolingual bitexts, we removed all sentence pairs where either sentence was longer than 100 words. Table 1 gives descriptive statistics for the remaining data. The table also shows the upper bound of the 95% confidence intervals for the coverage rates reported later. The results of experiments on different bitexts are not directly comparable, due to the varying genres and sentence lengths.

3.2 Constraining Parse Trees

One of the main independent variables in our experiments was the number of monolingual parse trees used to constrain the hierarchical alignments. To induce models of translational equivalence, some researchers have tried to use such trees to constrain bilingual constituents: The span of every node in the constraining parse tree must coincide with the relevant monolingual span of some

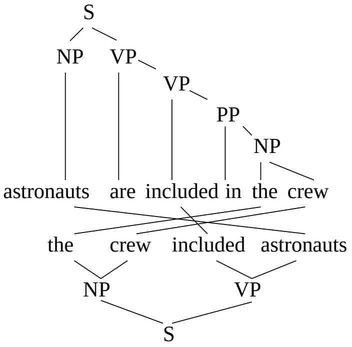

Figure 3: A word alignment that cannot be generated without gaps in a manner consistent with both parse trees.

node in the bilingual derivation tree. These additional constraints can thwart attempts at hierarchical alignment that might have succeeded otherwise. Figure 2a shows a word alignment and a parse tree that can be hierarchically aligned without gaps. *George* and *left* can be composed in both sentences into a constituent without crossing any phrase boundaries in the tree, as can *on* and *Friday*. These two constituents can then be composed to cover the entire sentence pair. On the other hand, if a constraining tree is applied to the other sentence as shown in Figure 2b, then the word alignment and tree constraint conflict. The projection of the VP is discontinuous in the top sentence, so the links that it covers cannot be composed into a constituent without gaps. On the other hand, if a gap is allowed, then the VP can compose as *on Friday* ... *left* in the top sentence, where the ellipsis represents a gap. This VP can then compose with the NP complete a synchronous parse tree. Some authors have applied constraining parse trees to both sides of the bitext. The example in Figure 3 can be hierarchically aligned using either one of the two constraining trees, but gaps are necessary to align it with both trees.

3.3 Methods

We parsed the English side of each bilingual bitext and both sides of each English/English bitext using an off-the-shelf syntactic parser (Bikel, 2004), which was trained on sections 02-21 of the Penn English Treebank (Marcus et al., 1993).

Our bilingual bitexts came with manually annotated word alignments. For the monolingual bitexts, we used an automatic word aligner based on a cognate heuristic and a list of 282 function words compiled by hand. The aligner linked two words to each other only if neither of them was on the function word list and their longest common subsequence ratio (Melamed, 1995) was at least 0.75. Words that were not linked to another word in this manner were linked to NULL. For the purposes of this study, a word aligned to NULL is a non-constraint, because it can always be composed without a gap with some constituent that is adjacent to it on just one side of the bitext. The number of automatically induced non-NULL links was lower than what would be drawn by hand.

We modified the word alignments in all bitexts to minimize the chances that alignment errors would lead to an over-estimate of alignment complexity. All of the modifications involved adding links to NULL. Due to our disjunctive treatment of conflicting links, the addition of a link to NULL can decrease but cannot increase the complexity of an alignment. For example, if we added the links *(cela, NULL)* and *(NULL, that)* to the alignment in Figure 1, the hierarchical alignment algorithm could use them instead of the link between *cela* and *that*. It could thus generate the modified alignment without using a gap. We added NULL links in two situations. First, if a subset of the links in an alignment formed a many-to-many mapping but did not form a bipartite clique (i.e. every word on one side linked to every word on the other side), then we added links from each of these words to NULL. Second, if n words on one side of the bitext aligned to m words on the other side with $m > n$ then we added NULL links for each of the words on the side with m words.

After modifying the alignments and obtaining monolingual parse trees, we measured the alignment complexity of each bitext using a hierarchical alignment algorithm, as described in Section 2. Separate measurements were taken with zero, one, and two constraining parse trees. The synchronous parser in the GenPar toolkit[4] can be configured for all of these cases (Burbank et al., 2005).

Unlike Fox (2002) and Galley et al. (2004), we measured failure rates per corpus rather than per sentence pair or per node in a constraining tree. This design was motivated by the observation that if a translation model cannot correctly model a certain word alignment, then it is liable to make incorrect inferences about arbitrary parts of that alignment, not just the particular word links involved in a complex pattern. The failure rates we report represent lower bounds on the fraction of training data

[4] `http://nlp.cs.nyu.edu/GenPar`

# of gaps allowed →	0/0	0/1 or 1/0
Chinese/English	26 = 5%	0 = 0%
Romanian/English	1 = 0%	0 = 0%
Hindi/English	2 = 2%	0 = 0%
Spanish/English	3 = 2%	0 = 0%
French/English	3 = 1%	0 = 0%

Table 2: Failure rates for hierarchical alignment of bilingual bitexts under word alignment constraints only.

# of gaps allowed on non-English side →	0	1	2
Chinese/English	298 = 61%	28 = 6%	0 = 0%
Romanian/English	82 = 41%	6 = 3%	1 = 0%
Hindi/English	33 = 37%	1 = 1%	0 = 0%
Spanish/English	75 = 38%	4 = 2%	0 = 0%
French/English	67 = 15%	2 = 0%	0 = 0%

Table 3: Failure rates for hierarchical alignment of bilingual bitexts under the constraints of a word alignment and a monolingual parse tree on the English side.

# of gaps →	0/0	0/1	0/2
0 CTs	171 = 3%	0 = 0%	0 = 0%
1 CTs	1792 = 34%	143 = 3%	7 = 0%
2 CTs	3227 = 61%	3227 = 61%	3227 = 61%

Table 4: Failure rates for hierarchical alignment of the MTEval bitext, over varying numbers of gaps and constraining trees (CTs).

# of gaps →	0/0	0/1	0/2
0 CTs	23 = 0%	0 = 0%	0 = 0%
1 CTs	655 = 10%	22 = 0%	1 = 0%
2 CTs	1559 = 25%	1559 = 25%	1559 = 25%

Table 5: Failure rates for hierarchical alignment of the fiction bitext, over varying numbers of gaps and constraining trees (CTs).

that is susceptible to misinterpretation by overconstrained translation models.

3.4 Summary Results

Table 2 shows the lower bound on alignment failure rates with and without gaps for five languages paired with English. This table represents the case where the only constraints are from word alignments. Wu (1997) has "been unable to find real examples" of cases where hierarchical alignment would fail under these conditions, at least in "fixed-word-order languages that are lightly inflected, such as English and Chinese." (p. 385). In contrast, we found examples in all bitexts that could not be hierarchically aligned without gaps, including at least 5% of the Chinese/English sentence pairs. Allowing constituents with a single gap on one side of the bitext decreased the observed failure rate to zero for all five bitexts.

Table 3 shows what happened when we used monolingual parse trees to restrict the compositions on the English side. The failure rates were above 35% for four of the five language pairs, and 61% for Chinese/English! Again, the failure rate fell dramatically when one gap was allowed on the unconstrained (non-English) side of the bitext. Allowing two gaps on the non-English side led to almost complete coverage of these word alignments.

Table 3 does not specify the number of gaps allowed on the English side, because varying this parameter never changed the outcome. The only way that a gap on that side could increase coverage is if there was a node in the constraining parse tree that had at least four children whose translations were in one of the complex permutations. The absence of such cases in the data implies that the failure rates under the constraints of one parse tree would be identical even if we allowed production rules of rank higher than two.

Table 4 shows the alignment failure rates for the MTEval bitext. With word alignment constraints only, 3% of the sentence pairs could not be hierarchically aligned without gaps. Allowing a single gap on one side decreased this failure rate to zero. With a parse tree constraining constituents on one side of the bitext and with no gaps, alignment failure rates rose from 3% to 34%, but allowing a single gap on the side of the bitext that was not constrained by a parse tree brought the failure rate back down to 3%. With two constraining trees the failure rate was 61%, and allowing gaps did not lower it, for the same reasons that allowing gaps on the tree-constrained side made no difference in Table 3.

The trends in the fiction bitext (Table 5) were similar to those in the MTEval bitext, but the coverage was always higher, for two reasons. First, the median sentence size was lower in the fiction bitext. Second, the MTEval translators were instructed to translate as literally as possible, but the fiction translators paraphrased to make the fiction more interesting. This freedom in word choice reduced the frequency of cognates and thus imposed fewer constraints on the hierarchical alignment, which resulted in looser estimates of the lower bounds. We would expect the opposite effect with hand-aligned data (Galley et al., 2004).

To study how sentence length correlates with the complexity of translational equivalence, we took subsets of each bitext while varying the max-

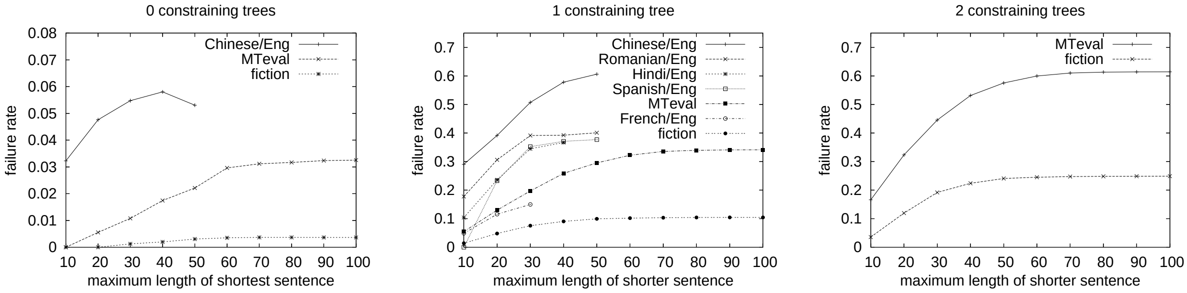

Figure 4: Failure rates for hierarchical alignment without gaps vs. maximum length of shorter sentence.

category →	1	2	3
valid reordering	12	10	5
parser error	n/a	16	25
same word used differently	15	4	0
erroneous cognates	3	0	0
total sample size	30	30	30
initial failure rate (%)	3.25	31.9	38.4
% false negatives	60±7	66±7	84±3
adjusted failure rate (%)	1.3±.22	11±2.2	6±1.1

Table 6: Detailed analysis of hierarchical alignment failures in MTEval bitext.

imum length of the shorter sentence in each pair.[5] Figure 4 plots the resulting alignment failure rates with and without constraining parse trees. The lines in these graphs are not comparable to each other because of the variety of genres involved.

3.5 Detailed Failure Analysis

We examined by hand 30 random sentence pairs from the MTEval bitext in each of three different categories: (1) the set of sentence pairs that could not be hierarchically aligned without gaps, even without constraining parse trees; (2) the set of sentence pairs that could not be hierarchically aligned without gaps with one constraining parse tree, but that did not fall into category 1; and (3) the set of sentence pairs that could not be hierarchically aligned without gaps with two constraining parse trees, but that did not fall into category 1 or 2. Table 6 shows the results of this analysis.

In category 1, 60% of the word alignments that could not be hierarchically aligned without gaps were caused by word alignment errors. E.g.:

1a GlaxoSmithKline's second-best selling **drug** may have to face competition.

1b **Drug** maker GlaxoSmithKline may have to face competition on its second best selling product.

The word *drug* appears in both sentences, but for different purposes, so *drug* and *drug* should not have been linked.[6] Three errors were caused by words like *targeted* and *started*, which our word alignment algorithm deemed cognates. 12 of the hierarchical alignment failures in this category were true failures. For example:

2a **Cheney** denied **yesterday** that the **mission** of his trip was to organize an assault on Iraq, while in **Manama**.

2b **Yesterday** in **Manama**, **Cheney** denied that the **mission** of his trip was to organize an assault on Iraq.

The alignment pattern of the words in bold is the familiar (3,1,4,2) permutation, as in Figure 1. Most of the 12 true failures were due to movement of prepositional phrases. The freedom of movement for such modifiers would be greater in bitexts that involve languages with less rigid word order than English.

Of the 30 sentence pairs in category 2, 16 could not be hierarchically aligned due to parser errors and 4 due to faulty word alignments. 10 were due to valid word reordering. In the following example, a co-referring pronoun causes the word alignment to fail with a constraining tree on the second sentence:

3a But **Chretien** appears to have changed his stance after meeting with Bush in Washington last Thursday.

3b But after **Chretien** talked to Bush last Thursday in Washington, **he** seemed to change his original stance.

25 of the 30 sentence pairs in category 3 failed to align due to parser error. 5 examples failed because of valid word reordering. 1 of the 5 reorderings was due to a difference between active voice and passive voice, as in Figure 3.

The last row of Table 6 takes the various reasons for alignment failure into account. It estimates what the failure rates would be if the monolingual parses and word alignments were perfect, with 95% confidence intervals. These revised rates emphasize the importance of reliable word alignments for this kind of study.

[5] The length of the shorter sentence is the upper bound on the number of non-NULL word alignments.

[6] This sort of error is likely to happen with other word alignment algorithms too, because words and their common translations are likely to be linked even if they're not translationally equivalent in the given sentence.

4 Discussion

Figure 1 came from a real bilingual bitext, and Example 2 in Section 3.5 came from a real monolingual bitext.[7] Neither of these examples can be hierarchically aligned correctly without gaps, even without constraining parse trees. The received wisdom in the literature led us to expect no such examples in bilingual bitexts, let alone in monolingual bitexts. See `http://nlp.cs.nyu.edu/GenPar/ACL06` for more examples. The English/English lower bounds are very loose, because the automatic word aligner would not link words that were not cognates. Alignment failure rates on a hand aligned bitext would be higher. We conclude that the ITG formalism cannot account for the "natural" complexity of translational equivalence, even when translation divergences are factored out.

Perhaps our most surprising results were those involving one constraining parse tree. These results explain why constraints from independently generated monolingual parse trees have not improved statistical translation models. For example, Koehn et al. (2003) reported that "requiring constituents to be syntactically motivated does not lead to better constituent pairs, but only fewer constituent pairs, with loss of a good amount of valuable knowledge." This statement is consistent with our findings. However, most of the knowledge loss could be prevented by allowing a gap. With a parse tree constraining constituents on the English side, the coverage failure rate was 61% for the Chinese/English bitext (top row of Table 3), but allowing a gap decreased it to 6%. Zhang and Gildea (2004) found that their alignment method, which did not use external syntactic constraints, outperformed the model of Yamada and Knight (2001). However, Yamada and Knight's model could explain only the data that would pass the no-gap test in our experiments with one constraining tree (first column of Table 3). Zhang and Gildea's conclusions might have been different if Yamada and Knight's model were allowed to use discontinuous constituents. The second row of Table 4 suggests that when constraining parse trees are used without gaps, at least 34% of training sentence pairs are likely to introduce noise into the model, even if systematic syntactic differences between languages are factored out. We should not be surprised when such constraints do more harm than good.

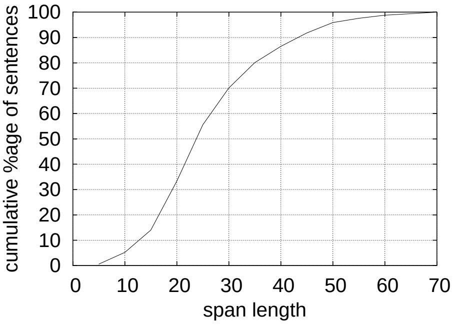

Figure 5: Lengths of spans covering words in (3,1,4,2) permutations.

To increase the chances that a translation model can explain complex word alignments, some authors have proposed various ways of extending a model's domain of locality. For example, Callison-Burch et al. (2005) have advocated for longer phrases in finite-state phrase-based translation models. We computed the phrase length that would be necessary to cover the words involved in each (3,1,4,2) permutation in the MTEval bitext. Figure 5 shows the cumulative percentage of these cases that would be covered by phrases up to a certain length. Only 9 of the 171 cases (5.2%) could be covered by phrases of length 10 or less. Analogous techniques for tree-structured translation models involve either allowing each nonterminal to generate both terminals and other nonterminals (Groves et al., 2004; Chiang, 2005), or, given a constraining parse tree, to "flatten" it (Fox, 2002; Zens and Ney, 2003; Galley et al., 2004). Both of these approaches can increase coverage of the training data, but, as explained in Section 2, they risk losing generalization ability.

Our study suggests that there might be some benefits to an alternative approach using discontinuous constituents, as proposed, e.g., by Melamed et al. (2004) and Simard et al. (2005). The large differences in failure rates between the first and second columns of Table 3 are largely independent of the tightness of our lower bounds. Synchronous parsing with discontinuities is computationally expensive in the worst case, but recently invented data structures make it feasible for typical inputs, as long as the number of gaps allowed per constituent is fixed at a small maximum (Waxmonsky and Melamed, 2006). More research is needed to investigate the trade-off between these costs and benefits.

[7] The examples were shortened for the sake of space and clarity.

5 Conclusions

This paper presented evidence of phenomena that can lead to complex patterns of translational equivalence in bitexts of any language pair. There were surprisingly many examples of such patterns that could not be analyzed using binary-branching structures without discontinuities. Regardless of the languages involved, the translational equivalence relations in most real bitexts of non-trivial size cannot be generated by an inversion transduction grammar. The low coverage rates without gaps under the constraints of independently generated monolingual parse trees might be the main reason why "syntactic" constraints have not yet increased the accuracy of SMT systems. Allowing a single gap in bilingual phrases or other types of constituent can improve coverage dramatically.

References

Necip Ayan, Bonnie J. Dorr, and Christof Monz. 2005. Alignment link projection using transformation-based learning. In *EMNLP*.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *ACL*.

Andrea Burbank, Marine Carpuat, Stephen Clark, Markus Dreyer and Pamela Fox, Declan Groves, Keith Hall, Mary Hearne, I. Dan Melamed, Yihai Shen, Andy Way, Ben Wellington, and Dekai Wu. 2005. Final Report on Statistical Machine Translation by Parsing. JHU CLSP. http://www.clsp.jhu.edu/ws2005/groups/statistical/report.html

Dan Bikel. 2004. A distributional analysis of a lexicalized statistical parsing model. In *EMNLP*.

Chris Callison-Burch, Colin Bannard, and Josh Scroeder. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *ACL*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *ACL*.

Bonnie Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics 20(4)*:597–633.

Heidi Fox. 2002. Phrasal cohesion and statistical machine translation. In *EMNLP*.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *HLT-NAACL*.

Declan Groves, Mary Hearne, and Andy Way. 2004. Robust sub-sentential alignment of phrase-structure trees. In *COLING*.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL*.

Mitchell Marcus, Beatrice Santorini, and Mary-Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Joel Martin, Rada Mihalcea, and Ted Pedersen. 2005. Word alignments for languages with scarce resources. In *ACL Workshop on Building and Using Parallel Texts*.

I. Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons. In *ACL Workshop on Very Large Corpora*.

I. Dan Melamed, Giorgio Satta, and Benjamin Wellington. 2004. Generalized multitext grammars. In *ACL*.

I. Dan Melamed and Wei Wang. 2005. Generalized Parsers for Machine Translation. NYU Proteus Project Technical Report 05-001 http://nlp.cs.nyu.edu/pubs/.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *HLT-NAACL Workshop on Building and Using Parallel Texts*.

LDC. 2002. NIST MT evaluation data, Linguistic Data Consortium catalogue # LDC2002E53. http://projects.ldc.upenn.edu/TIDES/mt2003.html.

Giorgio Satta and Enoch Peserico. 2005. Some computational complexity results for synchronous context-free grammars. In *EMNLP*.

Michel Simard, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Guassier, Cyril Goutte, and Kenji Yamada. 2005. Translating with non-contiguous phrases. In *EMNLP*.

Sonjia Waxmonsky and I. Dan Melamed. 2006. A dynamic data structure for parsing with discontinuous constituents. NYU Proteus Project Technical Report 06-001 http://nlp.cs.nyu.edu/pubs/.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *ACL*.

Richard Zens and Hermann Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *ACL*.

Hao Zhang and Daniel Gildea. 2004. Syntax-based alignment: Supervised or unsupervised? In *COLING*.

Hao Zhang, Liang Huang, Daniel Gildea, and Kevin Knight. 2006. Synchronous binarization for machine translation. In *HLT-NAACL*.