Divergences: Exploring One Rabbit-Hole in Word-Level Alignment

Lisa Pearl
Department of Computer Science
University of Maryland, College Park
llsp@clip.umiacs.umd.edu

April 3, 2002

Abstract

Word-level alignments of bilingual text (bitexts) are not only an integral part of statistical machine translation models, but also useful for lexical acquisition, treebank construction, and part-of-speech tagging. The frequent occurrence of *divergences*, structural differences between languages, presents a great challenge to the alignment task. I propose that syntactic parse information can be used to transform the sentence structure of one language to bear a closer resemblance to that of the other language. This, in turn, will produce more one-to-one correspondences between the languages and thus ease the word-level alignment task.

In this paper, I perform a case-study on English-Spanish to show that common divergence types can be found and systematically identified. I describe specifically the transformation technique and also present the results of an empirical analysis comparing the complexities of performing word-level alignment with and without divergence handling. These results suggest that divergence handling can improve word-level alignment[1].

This work is submitted in partial fulfillment of the requirements for the degree of **Bachelor of Science in Computer Science with High Honors** at the **University of Maryland, College Park**.

Thesis Advisor: Professor Bonnie J. Dorr, Department of Computer Science and Institute for Advanced Computer Study.

[1]The research presented here was supported in part by ONR MURI Contract FCPO.810548265.

Contents

List of Tables

Chapter 1

Introduction

Word-level alignments of bilingual text (bitexts) are a key component of statistical machine translation models, as well as tools for lexical acquisition, treebank construction, and part-of-speech tagging [Yarowsky and Ngai 2001]. However, *divergences*, which are structural differences between languages, are a significant stumbling block in the alignment task. The ideas driving the current research effort were the following:

1. Every language pair has divergences that are easy to recognize.
2. Knowing what these divergence types are and how to systematically accommodate them would aid word-level alignment immensely.

With improved word-level alignment, we can obtain an improved projection of structural information from one language to another - which, in turn, can be used for such useful things as inducing foreign language parse trees and training foreign language parsers to be used in statistical machine translation. Thus, improving word-level alignment is a practical goal, providing a means to several ends.

Software which specifically addresses the issue of systematically accommodating divergence types is available at **ftp.umiacs.umd.edu/pub/bonnie/divergence.tar.gz**.

Chapter 2

Divergences Overview

A *divergence* occurs when the underlying concepts or gist of a sentence et spread into different words for different languages. For example, the notion of floating across a river is expressed as *it floats across a river* in English and *it crosses a river floating*(*atravesó el río flotando*) in Spanish.

While this transformation from one expression of the idea to the other is seemingly transparent for human readers, it throws word-level aligners for a serious loop. Far from being a rare occurrence, my preliminary investigation revealed that divergences occurred in approximately 1 out of every 3 sentences in a sample size of 19K sentences from the TREC El Norte Newspaper Corpus[1] (using automatic detection techniques followed by human confirmation). Thus, finding a way to deal effectively with these divergences and repair them would be a massive advance for bilingual alignment.

The current avenue of research involves transforming English into a pseudo-English form (which I call E') that more closely matches the physical form of the foreign language, e.g., "float across a river" becomes "cross a river floating" if the foreign language is Spanish. Ideally, this rewriting of the English sentence creates more one-to-one correspondences which, in turn, facilitates our statistical alignment process. The key is to identify possible rewritings for known examples (my training set), which then generalize to rewritings for examples not yet covered in the training set. In theory, the transformation approach applies to all divergence types. Thus, given a corpus, divergences are identified, rewritten, and then run through the statistical aligner of choice.

[1]This analysis was done on the TREC Spanish Data, LDC catalog no LDC2000T51, ISBN 1-58563-177-9, 2000.

Chapter 3

Related Work

Recently, researchers have extended traditional statistical machine translation (MT) models [Brown et al. 1990, Brown et al. 1993] to include the syntactic structures of the languages [Alshawi et al. 2000, Alshawi and Douglas 2000, Wu 1997]. Furthermore, [Yamada and Knight 2001] have shown that MT models are significantly improved when trained on syntactically *annotated* data [Yamada and Knight 2001]. Unfortunately, the cost of human labor in producing annotated treebanks is often prohibitive, thus making the construction of such data for new languages infeasible. Some researchers have developed techniques for fast acquisition of hand-annotated treebanks [Fellbaum et al. 2001]. Others have appealed to machine learning techniques. For example, the work of [Hermjakob and Mooney 1997] and [Hwa 2000] aimed to minimize the amount of annotated data needed to induce a parser.

In a cross-language setting, some researchers have taken approaches that circumvent the construction of annotated treebanks, producing MT systems without parsers and dictionaries. One example is the Expedition effort, an enterprise to machine translation capability for low-density languages in short periods of time [Amtrup et al. 1999]. Others have proposed to induce a parser from a noisy treebank of foreign-language dependency trees that are automatically projected from English [Hwa et al. 2002]. The approach here is the most relevant to the latter in that it provides a significantly noise-reduced foreign-language dependency treebank for inducing a foreign language parser. The transformation process repairs misalignments resulting from cross-language divergences, thus effecting a more accurate dependency-tree projection from English into the foreign language.

Chapter 4

Divergence Types

Working with the idea that every language pair has divergences that are easy to recognize, I conducted a case study on English-Spanish and discovered six canonical types. Each one, in turn, is described below:

4.1 Light Verb Construction

A *light verb construction* involves a single verb in one language being translated into the other with a semantically "light" verb, i.e. one that carries little or no specific meaning in its own right, and some other meaning unit (perhaps a noun) to convey the appropriate meaning. Canonical light verbs include *give*, *make*, *do*, *take*, and *have*.

As an example of this type of divergence, we have the English verb *kick* and its Spanish equivalent:

to kick $\Leftrightarrow$ dar una patada (give a kick)

Other examples are given in Table 4.2.

4.2 Manner Conflation

This divergence involves translating a single manner verb (e.g., *float*) (and sometimes an adverbial describing direction) as a light verb of motion (e.g. *go*, *stay*) and a manner-indicating content word (e.g. *floating*). For Spanish, the content word tends to be a progressive manner verb. Other languages, however, may express the content word in different ways.

As an example of this divergence type, we have the English verb *float* and its Spanish equivalent:

to float $\Leftrightarrow$ ir flotando (go (via) floating)

Other Examples are given in Table 4.3.

A subtype of this divergence involves specifically the idea of direction or path - a single manner verb and a preposition, indicating direction, are translated with a light verb of motion and a manner-indicating content word. Here is where we can see quite well how the same gist is spread differently in the two languages:

run in ⇔ entrar corriendo
(run in) ⇔ (enter running)
[move-running in] ⇔ [move-in running]

Other Examples are given in Table 4.4.

4.3 Head Swapping

A head-swapping divergence banks on the idea of phrases in tree structure, i.e. an X is the head of an XP and the XP has auxiliary branches for the modifiers. Head-swapping occurs when the head of the XP in one language is the modifier of the XP in the other language. In other words, a permutation of semnatically equivalent words is necessary to go from one language to the other.

Spanish's *soler* demonstrates this quite nicely:

usually go ⇔ soler ir (to-be-used-to going)

$$[usually_{(Mod)} go_{(X)}]_{XP} \Leftrightarrow [to - be - used - to_{(X)} going_{(Mod)}]_{(XP)}$$

While there are not, in Spanish, many different examples, the number of *soler* instances is significant.

4.4 Thematic Divergence

A thematic divergence occurs when the verb's arguments switch thematic roles from one language to another. *Gustar* is an excellent example of this.

I_{Agent} like $grapes_{Theme}$ ⇔ Me gustan uvas ($to - me_{Theme}$ please $grapes_{Agent}$)

This type of divergence is very common in Spanish and is one of the most abundant divergence types in the TREC El Norte Corpus.

Other examples are shown in Table 4.5.

4.5 Categorial Divergence

A categorial divergence involves using different parts of speech to express the same idea. For instance, from English to Spanish, an adjectival phrase is translated as a light verb, accompanied by a nominal form of the adjective.

to be jealous ⇔ tener celos (to have jealousy)

Other examples are shown in Table 4.6.

4.6 Structural Divergence

A structural divergence involves the realization of incorporated arguments such as subject and object as obliques (i.e. headed by a preposition in a PP). The following are examples in English-Spanish:

Table 4.1: **Translation from English Adjectival Phrases into Spanish Nominalized Phrases**

to enter the house $\Leftrightarrow$ entrar en la casa (enter in the house)
ask for a referendum $\Leftrightarrow$ pedir un referendum (ask-for a referendum)

Table 4.2: **Translation from English Verbs into Spanish Light Verb Phrases**

use $\Leftrightarrow$ hacer uso de (make use of)
welcome $\Leftrightarrow$ dar la bienvenida (give the welcome)
jump $\Leftrightarrow$ dar un salto (give a jump)
please $\Leftrightarrow$ dar gusto (give pleasure)
aid $\Leftrightarrow$ dar apoyo (give aid)
face $\Leftrightarrow$ dar la cara (give the face)
initiate $\Leftrightarrow$ dar inicio (give the beginning)
surprise $\Leftrightarrow$ dar la sorpresa (give the surprise)
beat $\Leftrightarrow$ dar una paliza (give the beating)
borrow $\Leftrightarrow$ tomar prestado (take lent-thing)
speak $\Leftrightarrow$ tomar las palabras (take the words)
breathe $\Leftrightarrow$ tomar un respiro (take a breath)
note $\Leftrightarrow$ tomar nota (take note)
hope $\Leftrightarrow$ tener aspiración (have hope)
fear $\Leftrightarrow$ tener miedo (have fear)
dread $\Leftrightarrow$ tener pavor (have dread)
remember $\Leftrightarrow$ tener memoria (have memory)
end $\Leftrightarrow$ poner fin (put end)
test $\Leftrightarrow$ poner a prueba (put to test)

Table 4.3: **Translation from English Manner Verbs into Spanish Light Verb Phrases**

reduce ⇔ ir reduciendo (go reducing)
remain ⇔ ir quedando (go remaining)
develop ⇔ ir desarrollando (go developing)
give ⇔ ir dando (go giving)
increment ⇔ ir incrementando (go incrementing)
take ⇔ ir tomando (go taking)
advance ⇔ ir avanzando (go advancing)
walk ⇔ ir caminando (go walking)
pass ⇔ ir pasando (go passing)
grow ⇔ ir creciendo (go growing)

Table 4.4: **Translation from English Manner Verbs into Spanish Light Verb Phrases**

bounce out ⇔ irse botando (go-out bouncing)
fly about ⇔ andar volando (go-about flying)
run across ⇔ cruzar corriendo (go-across running)
run down ⇔ bajar corriendo (go-down running)
run out ⇔ salir corriendo (go-out running)
sail about ⇔ andar navegando (go-about sailing)
search about for ⇔ andar buscando (go-about searching-for)
swim by ⇔ pasar nadando (pass swimming)
walk in ⇔ entrar caminando (go-in running)
walk through ⇔ pasar caminando por (pass-through walking)

Table 4.5: **Translation from English Manner Verbs into Spanish Phrases**

I dislike ⇔ me disgusta (to-me displease)
I have...left ⇔ me quedan (to-me remain)
I have a headache ⇔ me duele la cabeza (to-me hurts the head)
I miss ⇔ me falta (to-me misses)
I am interested in ⇔ me interesa (to-me interests)
I am bothered by ⇔ me molesta (to-me bothers)

Table 4.6: **Translation from English Adjectival Phrases into Spanish Nominalized Phrases**

to be fully aware $\Leftrightarrow$ tener plena conciencia (to have full awareness)
to have handy $\Leftrightarrow$ tener a mano (to have at hand)
to be successful $\Leftrightarrow$ tener éxito (to have success)
to be interested in $\Leftrightarrow$ tener interés (to have interest)

Chapter 5

Occurrence of Divergences in Large Corpora

Since we're investing so much time identifying divergences (and later accomodating them in word-level alignment), it would be a good idea to know how often they appear. Specifically, we want to make sure that this investigation is worth the time put into it.

In my preliminary investigation, I conducted an in-depth linguistic analysis on part of the Trec El Norte Corpus, searching for some commonalities among the divergences I was finding. The first pass resulted in finding very broad groupings - "light" verbs, conflation with "light" verbs of motion, instances of head swapping, and categorial/structural shifts. Using these gross distinctions, I crafted a set of regular expressions to use as a sieve on 19K sentences of the TREC El Norte Corpus. Because the regular expressions were, indeed, very broad, they pulled just over 18K of sentences as potentially containing a divergence. In order to determine whether or not these sentences actually did contain a divergence, I verified (or discarded) them by hand, a very time-consuming and monotonous undertaking. However, the end result was around 6.8K of the 19K sentences human-verified as containing a divergence. Ths was an amazing amount - about 1 out of 3 sentences contained a divergence of one of the six types outlines in the previous section.

With this high occurrence rate, I reasoned that expending effort to accomodate divergences would be a worthwhile endeavor. Moreover, it would behoove me to come up with a quicker, more automatic way to identify them in large corpora. However, the first step would be to at least shrink the amount of potential divergences which required human-verification.

To this end, I crafted a set of very refined regular expressions which had a higher precision at the expense of a rather low recall. In a sample size of 19K, these regular expressions identified 3,628 potential divergences, of which 1750 were verified as actual divergences. While still identifying a fairly significant portion of sentences containing divergences (9%), the method of regular expressions seems rather insufficient for the rather complicated task of divergence identification. More sophisticated identification methods are the subject of future research.

However, having confirmed that divergences occurred in significant numbers, a way of dealing with them was then required.

Chapter 6

Overview of the Transformation Process

As mentioned previously, the current avenue of research focuses on transforming the syntactic structure of English into an E' which more closely matches the syntactic structure of the foreign lanaguage. Again, the thrust of this transformation is to create more one-to-one correspondences between the two languages and thus ease the work of word-alignment. The syntactic structures used in this approach are modeled after the dependency-tree representations used in the Minipar system [Lin 1995, Lin 1998].. A sample ideal transformation would take the English sentence *John kicked Mary*, depicted in the simplified dependency tree below:

```
(kick V_N_N root (john N subj) (mary N obj))
```

and turn it into *John gave kicks to Mary*, depicted below (again, simplified format):

```
(give V_N_N_P root (john N subj)
      (kicks N obj) (to P mod (mary N pcomp-n)))
```

A more detailed explanation of the transformation process is given in the next section.

Chapter 7

The Transformation Process in Detail

First, we use minipar dependency trees to model the syntactic structure of the English sentence. Then, accomodating divergences is simply a matter of transforming the English dependency tree to an E' dependency tree which resembles the syntactic structure of the Spanish sentence. As an example, suppose we again have the sentence *John kicked Mary.* As this is an example of a light verb divergence for Spanish, suppose the transformation involved turning the syntactic structure into *John gave kicks to Mary.* The dependency trees might then look like this:

```
E0     *      C       *       *root*  [ ]
E2     *      N       2       subj    [ antecedent:1 gov:kick ]
1   John      N       2       s       [ gov:kick ]
2 kicked      V_N_N   E0      i       [ gov:fin ]
3   Mary      N       2       obj     [ gov:kick ]
```

John kicked Mary,
which becomes

```
1  John      N       2       s       [ gov:give ]
2  gave      V_N_N_P *       *root*  [ ]
3 kicks      N       2       obj     [ gov:give ]
4    to      Prep    2       mod     [ gov:give ]
5  Mary      N       4       pcomp-n [ gov:to ]
```

John gave kicks to Mary

Given, then, that out goal is to transform the dependency trees depending on what kind of divergence is being examined, the first thing was to identify the correct transformation for each divergence type. To do this, we needed to get a canonical example of each divergence type and then tranform it. Then, the transformation itself involved writing specific rules to do a gross syntactic transformation and then a lexical massaging afterwards. The software to do this can be downloaded from **ftp.umiacs.umd.edu/pub/bonnie/divergence.tar.gz**.

To make the details of the transformation process more clear, we illustrate with *John kicked Mary* - a light verb divergence of type 1B.

The English dependency tree:

```
E0      *      C       *       *root*  [ ]
E2      *      N       2       subj    [ antecedent:1 gov:kick ]
1   John       N       2       s       [ gov:kick ]
2 kicked       V_N_N   E0      i       [ gov:fin ]
3   Mary       N       2       obj     [ gov:kick ]
```

After the gross syntactic transformation:

```
1     John      N              2        s       [ gov:give ]
2 LightVB      V_N_N_P        *       *root*  [ ]
3  kicked      N-nounifyV     2       obj     [ gov:give ]
4      to      Prep           2       mod     [ gov:give ]
5    Mary      N              4       pcomp-n [ gov:to ]
```

After the lexical massaging:

```
1    John      N              2       s       [ gov:give ]
2 LightVB      V              *       *root*  [ ]
3    kick      N              2       obj     [ gov:give ]
4      to      Prep           2       mod     [ gov:give ]
5    Mary      N              4       pcomp-n [ gov:to ]
```

As we can see, this lexical massaging process is not quite sufficient - a future extension might be, for instance, to statistically learn which light verbs go with which nouns, thus resulting in:

```
1 John      N              2       s       [ gov:give ]
2 gave      V              *       *root*  [ ]
3 kick      N              2       obj     [ gov:give ]
4   to      Prep           2       mod     [ gov:give ]
5 Mary      N              4       pcomp-n [ gov:to ]
```

Chapter 8

Experiment: Impact of Divergence Correction on Alignment

To evaluate the hypothesis that transformations of divergent cases can, indeed, facilitate the word-level alignment process, a human alignment study was conducted for English-Spanish.

The experiment involves four steps:

i. Identify 6 canonical rewritten structures—one for each divergence category.

ii. Automatically categorize English sentences into one of the 6 divergence categories (or "none") based on the foreign language.

iii. Apply the appropriate canonical rewriting to each divergence-categorized English sentence, renaming it English Prime (E').

iv. Align and Compare:

- Human align the true English sentence and the foreign-language sentence.
- Human align the rewritten E' sentence and the foreign-language sentence.
- Compare inter-annotator agreement between the first and second sets.

Table 8.1 shows examples of the types of sentences that were aligned with the Spanish in our experiment (including both English and E').

8.1 Experimental Setup

In the experiment, four fluently bilingual human subjects were asked to perform word-level alignments on the same set of sentences selected from the Bible. They were all provided the same instructions and software, similar to the methodology and system described by [Melamed 1998]. Two of the four subjects were given

Type	English	E'	Spanish
	fear	have fear	tiene miedo
	try	put to trying	poner a prueba
Manner/Path	teaches	walks teaching	anda enseñando
	is spent	self goes spending	se va gastando
	walked out	lmove-out walking	salió caminando
Thematic	I am pained	me pain they	me duelen
	He loves it	to him be-loved it	le gusta
Categorial	I am jealous	I have jealousy	tengo celos
	I require of you	I require-of you	te pido
Structural	after six years	after of six years	después de seis años
	and because of that in other parts	and for that in other parts	y por ello en otras partes

Table 8.1: Examples of True English, E', and Spanish Equivalent

the original English and foreign language sentences; they served as the control for the experiment. The sentence given to the other two consisted of the original foreign language sentences paired with the transformed English (denoted as E'). The inter-annotator agreement rates (as well as other relevant statistics, such as number of unaligned and multiply-aligned words) were then compared. If the divergence transformations had successfully modified English structures to match those of the foreign language, we would expect the inter-annotator agreement rate between the subjects aligning the E' set to be higher than the control set. We would also expect that the E' set would have fewer unaligned and multiply-aligned words.

8.2 Experiment Details

For this experiment, the subjects were presented with 150 English-Spanish sentence pairs from the English and Spanish Bibles. These sentences were first selected as potential divergences, using the refined hand-crafted regular expressions referred to in Section 5; they were then verified by the experimenter as belonging to a particular divergence type. The average length of the English sentences is 25.6 words; the average length of the Spanish sentences is 24.7 words. Of the four human subjects, two are native Spanish speakers, and two are native English speaker studying Spanish literature. The backgrounds of the four human subjects are summarized in Table 8.2.

	data set	native-tongue	linguistic knowledge?	ease with computers
Subject 1	control	Spanish	yes	high
Subject 2	control	Spanish	no	low
Subject 3	divergence	English	no	high
Subject 4	divergence	English	no	low

Table 8.2: A summary of the backgrounds of the English-Spanish subjects

	(F-score)	% of unaligned words	Avg. alignments per word
E-S	79.3	17.2	1.35
E'-S	82.1	14.0	1.16

Table 8.3: The results of the experiment. Note: In computing the average number of alignments per word, unaligned words were not computed.

The averaged precision and recall figures (F-scores)[1] and other relevant statistics are summarized below:

Inter-annotator agreement rate is quantified for each pair of subjects who viewed the same set of data. We hold one subject's alignments as the "ideal" and compute the precision and recall figures for the other subject based on how many alignment links were made by both people.

For this experiment, the inter-annotator agreement, indeed, is higher for the bitext in which the divergent portions of the English sentences have been transformed. Specifically, the agreement rate increased from 79.3% to 82.1%, resulting in an error reduction of 13.5%.

Additional statistics also support the hypothesis that transforming divergent English sentences eases word-level alignment by reducing the number of unaligned and multiply-aligned words. In this experiment, both the appearances of unaligned words and multiply-aligned words decreased when aligning to the modified English sentences. The percentage of unaligned words decreased from 17% to 14%, and the average number of links to a word is lowered from 1.35 to 1.16.[2]

[1] $F = \frac{2 \times Precision \times Recall}{Precision + Recall}$

[2] The relatively high overall percentage of unaligned words is due to the fact that the subjects did not align punctuations.

Chapter 9

Conclusion and Future Work

In this paper, I have described a case study of divergences in the language pair English-Spanish, resulting in the identification of six canonical divergence types. By examining bitext corpora, I have established a lower bound for divergence occurrence - approximately 35%. In addition, I have shown that divergence cases can be systematically handled by transforming the syntactic structures of English to bear a closer resemblance to those of Spanish, using a template based on the six canonical divergence types. The validity of divergence handling has been verified via a word-level alignment experiment - human subjects consistently had a higher agreement rate with each other on the word-level alignment task when the divergence English phrases were transformed. This result suggests that divergence handling will significantly improve automatic methods of word-level alignment and facilitate cross-language processing research such as creating foreign language treebanks from projected English syntactic structures.

Future work will involve a refinement of the transformation process (the additional lexical massaging mentioned in Section 7) as well as automatic identification of both divergences and automatic specification of which divergence type a potential divergence is.

Bibliography

[Alshawi and Douglas 2000] H. Alshawi and S. Douglas. 2000. Learning Dependency Transduction Models from Unannotated Examples. *Philosophical Transactions, Series A: Mathematical, Physical and Engineering Sciences.*

[Alshawi et al. 2000] Hiyan Alshawi, Srinivas Bangalore, and Shona Douglas. 2000. Learning Dependency Translation Models as Collections of Finite State Head Transducers. *Computational Linguistics*, 26.

[Amtrup et al. 1999] Jan W. Amtrup, Karine Megerdoomian, and Remi Zajac. 1999. Rapid Development of Translation Tools. In *Proceedings of the Machine Translation Summit VII*, pages 385–389, Singapore.

[Brown et al. 1990] Peter F. Brown, John Cocke, Stephen Della-Pietra, Vincent J. Della-Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85, June.

[Brown et al. 1993] Peter F. Brown, Stephen A. DellaPietra, Vincent J. DellaPietra, and Robert L. Mercer. 1993. The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics.*

[Dorr 1993] Bonnie J. Dorr. 1993. *Machine Translation: A View from the Lexicon.* The MIT Press, Cambridge, MA.

[Fellbaum et al. 2001] Christiane Fellbaum, Martha Palmer, Hoa Trang Dang, Lauren Delfs, and Susanne Wolff. 2001. Manual and Automatic Semantic Annotation with WordNet. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Customizations*, Carnegie Mellon University, Pittsburg, PA.

[Hermjakob and Mooney 1997] Ulf Hermjakob and Raymond J. Mooney. 1997. Learning Parse and Translation Decisions from Examples with Rich Context. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–489.

[Hwa et al. 2002] Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating Translational Correspondence using Annotation Projection. In *submitted to the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA.

[Hwa 2000] Rebecca Hwa. 2000. Sample selection for statistical grammar induction. In *Proceedings of the 2000 Joint SIGDAT Conference on EMNLP and VLC*, pages 45–52, Hong Kong, China, October.

[Lin 1995] Dekang Lin. 1995. Government-Binding Theory and Principle-Based Parsing. Technical report, University of Maryland. Submitted to Computational Linguistics.

[Lin 1998] Dekang Lin. 1998. Dependency-Based Evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*, Granada, Spain, May.

[Melamed 1998] I. Dan Melamed. 1998. Empirical Methods for MT Lexicon Development. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529*, pages 18–30, Langhorne, PA, October 28–31.

[Wu 1997] Dekai Wu. 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23(3):377–400.

[Yamada and Knight 2001] 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*, pages 523–529, Toulouse, France.

[Yarowsky and Ngai 2001] David Yarowsky and Grace Ngai. 2001. Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. In *Proc. of NAACL-2001*, pages 200–207.