Initial Explorations in English to Turkish Statistical Machine Translation

İlknur Durgar El-Kahlout
Faculty of Enginering
and
Natural Sciences
Sabancı University
Istanbul, 34956, Turkey
ilknurdurgar@su.sabanciuniv.edu

Kemal Oflazer
Faculty of Engineering
and
Natural Sciences
Sabancı University
Istanbul, 34956, Turkey
oflazer@sabanciuniv.edu

Abstract

This paper presents some very preliminary results for and problems in developing a statistical machine translation system from English to Turkish. Starting with a baseline word model trained from about 20K aligned sentences, we explore various ways of exploiting morphological structure to improve upon the baseline system. As Turkish is a language with complex agglutinative word structures, we experiment with morphologically segmented and disambiguated versions of the parallel texts in order to also uncover relations between morphemes and function words in one language with morphemes and functions words in the other, in addition to relations between open class content words. Morphological segmentation on the Turkish side also conflates the statistics from allomorphs so that sparseness can be alleviated to a certain extent. We find that this approach coupled with a simple grouping of most frequent morphemes and function words on both sides improve the BLEU score from the baseline of 0.0752 to 0.0913 with the small training data. We close with a discussion on why one should not expect distortion parameters to model word-local morpheme ordering and that a new approach to handling complex morphotactics is needed.

1 Introduction

The availability of large amounts of so-called parallel texts has motivated the application of statistical techniques to the problem of machine translation starting with the seminal work at IBM in the early 90's (Brown et al., 1992; Brown et al., 1993). Statistical machine translation views the translation process as a noisy-channel signal recovery process in which one tries to recover the input "signal" *e*, from the observed output signal *f*.[1]

Early statistical machine translation systems used a purely word-based approach without taking into account any of the morphological or syntactic properties of the languages (Brown et al., 1993). Limitations of basic word-based models prompted researchers to exploit morphological and/or syntactic/phrasal structure (Niessen and Ney, (2004), Lee,(2004), Yamada and Knight (2001), Marcu and Wong (2002), Och and Ney (2004),Koehn et al. (2003), among others.)

In the context of the agglutinative languages similar to Turkish (in at least morphological aspects) , there has been some recent work on translating from and to Finnish with the significant amount of data in the Europarl corpus. Although the BLEU (Papineni et al., 2002) score from Finnish to English is 21.8, the score in the reverse direction is reported as 13.0 which is one of the lowest scores in 11 European languages scores (Koehn, 2005). Also, reported *from* and *to* translation scores for Finnish are the lowest on average, even with the large number of

[1] Denoting *English* and *French* as used in the original IBM Project which translated from French to English using the parallel text of the Hansards, the Canadian Parliament Proceedings.

Proceedings of the Workshop on Statistical Machine Translation, pages 7–14,
New York City, June 2006. ©2006 Association for Computational Linguistics

sentences available. These may hint at the fact that standard alignment models may be poorly equipped to deal with translation from a poor morphology language like English to an complex morphology language like Finnish or Turkish.

This paper presents results from some very preliminary explorations into developing an English-to-Turkish statistical machine translation system and discusses the various problems encountered. Starting with a baseline word model trained from about 20K aligned sentences, we explore various ways of exploiting morphological structure to improve upon the baseline system. As Turkish is a language with agglutinative word structures, we experiment with morphologically segmented and disambiguated versions of the parallel text, in order to also uncover relations between morphemes and function words in one language with morphemes and functions words in the other, in addition to relations between open class content words; as a cursory analysis of sentence aligned Turkish and English texts indicates that translations of certain English words are actually morphemes embedded into Turkish words. We choose a morphological segmentation representation on the Turkish side which abstracts from word-internal morphological variations and conflates the statistics from allomorphs so that data sparseness can be alleviated to a certain extent.

This paper is organized as follows: we start with the some of the issues of building an SMT system into Turkish followed by a short overview Turkish morphology to motivate its effect on the word alignment problem with English. We then present results from our explorations with a baseline system and with morphologically segmented parallel aligned texts, and conclude after a short discussion.

2 Issues in building a SMT system for Turkish

The first step of building an SMT system is the compilation of a large amount of parallel texts which turns out to be a significant problem for the Turkish and English pair. There are not many sources of such texts and most of what is electronically available are parallel texts diplomatic or legal domains from NATO, EU, and foreign ministry sources. There is also a limited amount data parallel news corpus available from certain news sources. Although we have collected about 300K sentence parallel texts, most of these require significant clean-up (from HTML/PDF sources) and we have limited our training data in this paper to about 22,500 sentence subset of these parallel texts which comprises the subset of sentences of 40 words or less from the 30K sentences that have been cleaned-up and sentence aligned.[2,3]

The main aspect that would have to be seriously considered first for Turkish in SMT is the productive inflectional and derivational morphology. Turkish word forms consist of morphemes concatenated to a root morpheme or to other morphemes, much like "beads on a string" (Oflazer, 1994). Except for a very few exceptional cases, the surface realizations of the morphemes are conditioned by various local regular morphophonemic processes such as vowel harmony, consonant assimilation and elisions. Further, most morphemes have phrasal scopes: although they attach to a particular stem, their syntactic roles extend beyond the stems. The morphotactics of word forms can be quite complex especially when multiple derivations are involved. For instance, the derived modifier `sağlamlaştırdığımızdaki` [4] would be broken into surface morphemes as follows:

`sağlam+laş+tır+dığ+ımız+da+ki`

Starting from an adjectival root *sağlam*, this word form first derives a verbal stem *sağlamlaş*, meaning "to become strong". A second suffix, the causative surface morpheme *+tır* which we treat as a verbal derivation, forms yet another verbal stem meaning "to cause to become strong" or "to make strong (fortify)". The immediately following participle suffix

[2] We are rapidly increasing our cleaned-up text and expect to clean up and sentence align all within a few months.

[3] As the average Turkish word in running text has between 2 and 3 morphemes we limited ourselves to 40 words in the parallel texts in order not to exceed the maximum number of words recommended for GIZA++ training.

[4] Literally, "(the thing existing) at the time we caused (something) to become strong". Obviously this is not a word that one would use everyday, but already illustrates the difficulty as one Turkish "word" would have to be aligned to a possible discontinues sequence of English words if we were to attempt a word level alignment. Turkish words (excluding noninflecting frequent words such as conjunctions, clitics, etc.) found in typical running text average about 10 letters in length. The average number of bound morphemes in such words is about 2.

+dığ, produces a participial nominal, which inflects in the normal pattern for nouns (here, for 1^{st} person plural possessor which marks agreement with the subject of the verb, and locative case). The final suffix, *+ki*, is a relativizer, producing a word which functions as a modifier in a sentence, modifying a noun somewhere to the right.

However, if one further abstracts from the morphophonological processes involved one could get a lexical form

`sağlam+lAş+DHr+DHk+HmHz+DA+ki`

In this representation, the lexical morphemes except the lexical root utilize meta-symbols that stand for a set of graphemes which are selected on the surface by a series of morphographemic processes which are rooted in morphophonological processes some of which are discussed below, but have nothing whatsoever with any of the syntactic and semantic relationship that word is involved in. For instance, `A` stands for back and unrounded vowels *a* and *e*, in orthography, `H` stands for high vowels *ı, i, u* and *ü*, and `D` stands for *d* and *t*, representing alveolar consonants. Thus, a lexical morpheme represented as `+DHr` actually represents 8 possible allomorphs, which appear as one of *+dır, +dir, +dur, +dür, +tır, +tir, +tur, +tür* depending on the local morphophonemic context. Thus at this level of representation words that look very different on the surface, look very similar. For instance, although the words *masasında* 'on his table' and *defterinde* 'in his notebook' in Turkish look quite different, the lexical morphemes except for the root are the same: *masasında* has the lexical structure `masa+sH+ndA`, while *defterinde* has the lexical structure `defter+sH+ndA`.

The use of this representation is particularly important for Turkish for the following reason. Allomorphs which differ because of local word-internal morphographemic and morphotactical constraints almost always correspond to the same words or units in English when translated. When such units are considered by themselves as the units in alignment, statistics get fragmented and the model quality suffers. On the other hand, this representation if directly used in a standard SMT model such as IBM Model 4, will most likely cause problems, since now, the distortion parameters will have to take on the task of generating the correct sequence of morphemes in a word (which is really a local word-internal problem to be solved) in addition to generating the correct sequence of words.

3 Aligning English–Turkish Sentences

If an alignment between the components of parallel Turkish and English sentences is computed, one obtains an alignment like the one shown in Figure 1, where it is clear that Turkish words may actually correspond to whole phrases in the English sentence.

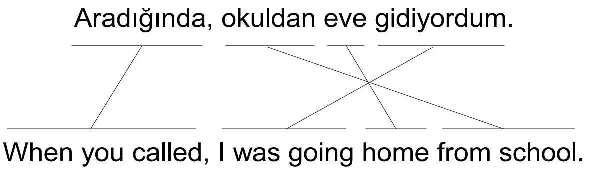
Aradığında, okuldan eve gidiyordum.
When you called, I was going home from school.

Figure 1: Word level alignment between a Turkish and an English sentence

One major problem with this situation is that even if a word occurs many times in the English side, the actual Turkish equivalent could be either missing from the Turkish part, or occur with a very low frequency, but many inflected variants of the form could be present. For example, Table 1 shows the occurrences of different forms for the root word *faaliyet* 'activity' in the parallel texts we experimented with. Although, many forms of the root word appear, none of the forms appear very frequently and one may even have to drop occurrences of frequency 1 depending on the word-level alignment model used, further worsening the sparseness problem.[5]

To overcome this problem and to get the maximum benefit from the limited amount of parallel texts, we decided to perform morphological analysis of both the Turkish and the English texts to be able to uncover relationships between root words, suffixes and function words while aligning them.

[5] A noun root in Turkish may have about hundred inflected forms and substantially more if productive derivations are considered, meanwhile verbs can have thousands of inflected and derived forms if not more.

Table 1: Forms of the word *faaliyet* ’activity’

Wordform	Count	Gloss
faaliyet	3	’activity’
faaliyete	1	’to the activity’
faaliyetinde	1	’in its activity’
faaliyetler	3	’activities’
faaliyetlere	6	’to the activities’
faaliyetleri	7	’their activities’
faaliyetlerin	7	’of the activities’
faaliyetlerinde	1	’in their activities’
faaliyetlerine	5	’to their activities’
faaliyetlerini	1	’their activities (acc.)’
faaliyetlerinin	2	’of their activities’
faaliyetleriyle	1	’with their activities’
faaliyette	2	’in (the) activity’
faaliyetteki	1	’that which is in activity’
Total	41	

On the Turkish side, we extracted the lexical morphemes of each word using a version of the morphological analyzer (Oflazer, 1994) that segmented the Turkish words along morpheme boundaries and normalized the root words in cases they were deformed due to a morphographemic process. So the word *faaliyetleriyle* when segmented into lexical morphemes becomes *faaliyet +lAr +sH +ylA*. Ambiguous instances were disambiguated statistically (Külekçi and Oflazer, 2005).

Similarly, the English text was tagged using TreeTagger (Schmid, 1994), which provides a lemma and a POS for each word. We augmented this process with some additional processing for handling derivational morphology. We then dropped any tags which did not imply an explicit morpheme (or an exceptional form). The complete set of tags that are used from the Penn-Treebank tagset is shown in Table 2.[6] To make the representation of the Turkish texts and English texts similar, tags are marked with a ’+’ at the beginning of all tags to indicate that such tokens are treated as “morphemes.” For instance, the English word *activities* was segmented as *activity +NNS*. The alignments we expected to obtain are depicted in Figure 2 for the example sentence given earlier in Figure 1.

Table 2: The set of tags used to mark explicit morphemes in English

Tag	Meaning
JJR	Adjective, comparative
JJS	Adjective, superlative
NNS	Noun, plural
POS	Possessive ending
RBR	Adverb, comparative
RBS	Adverb, superlative
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present

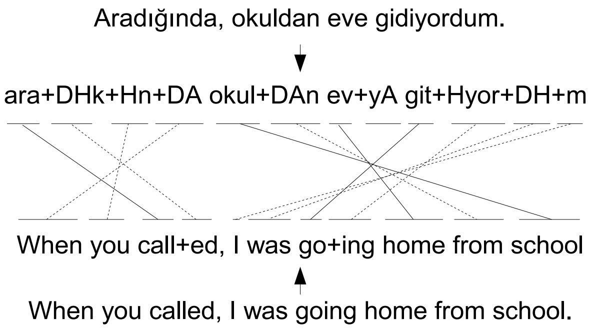
Aradığında, okuldan eve gidiyordum.
ara+DHk+Hn+DA okul+DAn ev+yA git+Hyor+DH+m
When you call+ed, I was go+ing home from school
When you called, I was going home from school.

Figure 2: “Morpheme” alignment between a Turkish and an English sentence

4 Experiments

We proceeded with the following sequence of experiments:

(1) **Baseline:** As a baseline system, we used a pure word-based approach and used Pharaoh Training tool (2004), to train on the 22,500 sentences, and decoded using Pharaoh (Koehn et al., 2003) to obtain translations for a test set of 50 sentences. This gave us a baseline BLEU score of 0.0752.

(2) **Morpheme Concatenation:** We then trained the same system with the morphemic representation

[6] The tagset used by the TreeTagger is a refinement of Penn-Treebank tagset where the second letter of the verb part-of-speech tags distinguishes between ”be” verbs (B), ”have” verbs (H) and other verbs (V),(Schmid, 1994).

of the parallel texts as discussed above. The decoder now produced the translations as a sequence of root words and morphemes. The surface words were then obtained by just concatenating all the morphemes following a root word (until the next root word) taking into just morphographemic rules but not any morphotactic constraints. As expected this "morpheme-salad" produces a "word-salad", as most of the time wrong morphemes are associated with incompatible root words violating many morphotactic constraints. The BLEU score here was 0.0281, substantially worse than the baseline in (1) above.

(3) **Selective Morpheme Concatenation:** With a small script we injected a bit of morphotactical knowledge into the surface form generation process and only combined those morphemes following a root word (in the given sequence), that gave rise to a valid Turkish word form as checked by a morphological analyzer. Any unused morphemes were ignored. This improved the BLEU score to 0.0424 which was still below the baseline.

(4) **Morpheme Grouping:** Observing that certain sequence of morphemes in Turkish texts are translations of some continuous sequence of functional words and tags in English texts, and that some morphemes should be aligned differently depending on the other morphemes in their context, we attempted a morpheme grouping. For example the morpheme sequence *+DHr +mA* marks infinitive form of a causative verb which in Turkish inflects like a noun; or the lexical morpheme sequence *+yAcAk +DHr* usually maps to "it/he/she will". To find such groups of morphemes and functional words, we applied a sequence of morpheme groupings by extracting frequently occuring n-grams of morphemes as follows (much like the grouping used by Chiang (2005): in a series of iterations, we obtained high-frequency bi-grams from the morphemic representation of parallel texts, of either morphemes, or of previously such identified morpheme groups and neighboring morphemes until up to four morphemes or one root 3 morpheme could be combined. During this process we ignored those combinations that contain punctuation or a morpheme preceding a root word. A similar grouping was done on the English side grouping function words and morphemes before and after root words.

The aim of this process was two-fold: it let frequent morphemes to behave as a single token and help Pharaoh with identification of some of the phrases. Also since the number of tokens on both sides were reduced, this enabled GIZA++ to produce somewhat better alignments.

The morpheme level translations that were obtained from training with this parallel texts were then converted into surface forms by concatenating the morphemes in the sequence produced. This resulted in a BLEU score of 0.0644.

(5) **Morpheme Grouping with Selective Morpheme Concatenation:** This was the same as (4) with the morphemes selectively combined as in (3). The BLEU score of 0.0913 with this approach was now above the baseline.

Table 3 summarizes the results in these five experiments:

Table 3: BLEU scores for experiments (1) to (4)

Exp.	*System*	*BLEU*
(1)	Baseline	0.0752
(2)	Morph. Concatenation.	0.0281
(3)	Selective Morph. Concat.	0.0424
(4)	Morph. Grouping and Concat.	0.0644
(5)	Morph. Grouping + (3)	**0.0913**

In an attempt to factor out and see if the translations were at all successful in getting the root words in the translations we performed the following: We morphologically analyzed and disambiguated the reference texts, and reduced all to sequences of root words by eliminating all the morphemes. We performed the same for the outputs of (1) (after morphological analysis and disambiguation), (2) and (4) above, that is, threw away the morphemes ((3) is the same as (2) and (5) same as (4) here). The translation root word sequences and the reference root word sequences were then evaluated using the BLEU (which would like to label here as BLEU-r for *BLEU root*, to avoid any comparison to previous results, which will be misleading. These scores are shown in Figure 4.

The results in Tables 3 and 4 indicate that with the standard models for SMT, we are still quite far from even identifying the correct root words in the trans-

Table 4: BLEU-r scores for experiments (1), (2) and (4)

Exp.	*System*	*BLEU*
(1)	Baseline	0.0955
(2)	Morph. Concatenation.	0.0787
(4)	Morph. Grouping	**0.1224**

Table 5: Some good SMT results

Input: international terrorism also remains to be an important issue .
Baseline: ulus+lararası terörizm de önem+li kal+mış+tır . bir konu ol+acak+tır
Selective Morpheme Concatenation: ulus+lararası terörizm de ol+ma+ya devam et+mek+te+dir önem+li bir sorun+dur .
Morpheme Grouping: ulus+lararası terörizm de önem+li bir sorun ol+ma+ya devam et+mek+te+dir .
Reference Translation: ulus+lararası terörizm de önem+li bir sorun ol+ma+ya devam et+mek+te+dir .

Input: the initiation of negotiations will represent the beginning of a next phase in the process of accession
Baseline: müzakere+ler+in gör+üş+me+ler yap+ıl+acak bir der+ken aşama+nın hasar+ı sürec+i başlangıc+ı+nı 15+'i
Selective Morpheme Concatenation: initiation müzakere+ler temsil ed+il+me+si+nin başlangıc+ı bir aşama+sı+nda katılım sürec+i+nin ertesi
Morpheme Grouping: müzakere+ler+in başla+ma+sı+nın başlangıc+ı+nı temsil ed+ecek+tir katılım sürec+i+nin bir sonra+ki aşama
Reference Translation: müzakere+ler+in başla+ma+sı , katılım sürec+i+nin bir sonra+ki aşama+sı+nın başlangıc+ı+nı temsil ed+ecek+tir

lations into Turkish, let alone getting the morphemes and their sequences right. Although some of this may be due to the (relatively) small amount of parallel texts we used, it may also be the case that splitting the sentences into morphemes can play havoc with the alignment process by significantly increasing the number of tokens per sentence especially when such tokens align to tokens on the other side that is quite far away.

Nevertheless the models we used produce some quite reasonable translations for a small number of test sentences. Table 5 shows the two examples of translations that were obtained using the standard models (containing no Turkish specific manipulation except for selective morpheme concatenation). We have marked the *surface* morpheme boundaries in the translated and reference Turkish texts to indicate where morphemes are joined for exposition purposes here – they neither appear in the reference translations nor in the produced translations!

5 Discussion

Although our work is only an initial exploration into developing a statistical machine translation system from English to Turkish, our experiments at least point out that using standard models to determine the correct sequence of morphemes within the words, using more powerful mechanisms meant to determine the (longer) sequence of words in sentences, is probably not a good idea. Morpheme ordering is a very local process and the correct sequence should be determined locally though the existence of morphemes could be postulated from sentence level features during the translation process. Furthermore, insisting on generating the exact sequence of morphemes could be an overkill. On the other hand, a morphological generator could take a *root word* and a *bag of morphemes* and generate possible legitimate surface words by taking into account morphotactic constraints and morphographemic constraints, possibly (and ambiguously) filling in any morphemes missing in the translation but actually required by the morphotactic paradigm. Any ambiguities from the morphological generation could then be filtered by a language model.

Such a bag-of-morphemes approach suggests that we do not actually try to determine exactly where the morphemes actually go in the translation but rather determine the root words (including any function words) and then *associate* translated morphemes with the (bag of the) right root word. The resulting sequence of root words and their bags-of-morpheme can be run through a morphological generator which can handle all the word-internal phenomena such as proper morpheme ordering, filling in morphemes or even ignoring spurious morphemes, handling local morphographemic phenomena such as vowel harmony, etc. However, this approach of not placing morphemes into specific position in the translated output but just associating them with certain root words requires that a significantly different alignment and decoding models be developed.

Another representation option that could be employed is to do away completely with morphemes on the Turkish side and just replace them with morphological feature symbols (much like we did here for English). This has the advantage of better handling allomorphy – all allomorphs including those that are not just phonological variants map to the same feature, and homograph morphemes which signal different features map to different features. So in a sense, this would provide a more accurate decomposition of the words on the Turkish side, but at the same time introduce a larger set of features since default feature symbols are produced for any morphemes that do not exist on the surface. Removing such redundant features from such a representation and then using reduced features could be an interesting avenue to pursue. Generation of surface words would not be a problem since, our morphological generator does not care if it is input morphemes or features.

6 Conclusions

We have presented the results of our initial explorations into statistical machine translation from English to Turkish. Using a relatively small parallel corpus of about 22,500 sentences, we have experimented with a baseline word-to-word translation model using the Pharaoh decoder. We have also experimented with a morphemic representation of the parallel texts and have aligned the sentences at the morpheme level. The decoder in this cases produces root word and morpheme sequences which are then selectively concatenated into surface words by possibly ignoring some morphemes which are redundant or wrong. We have also attempted a simple grouping of root words and morphemes to both help the alignment by reducing the number of tokens in the sentences and by already identifying some possible phrases. This grouping of morphemes and the use of selective morpheme concatenation in producing surface words has increased the BLEU score for our test set from 0.0752 to 0.0913. Current ongoing work involves increasing the parallel corpus size and the development of bag-of-morphemes modeling approach to translation to separate the sentence level word sequencing from word-internal morpheme sequencing.

7 Acknowledgements

This work was supported by TÜBİTAK (Turkish Scientific and Technological Research Foundation) project 105E020 ”Building a Statistical Machine Translation for Turkish and English”.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John D. Lafferty, and Robert L. Mercer. 1992. Analysis, statistical transfer, and synthesis in machine translation. In *Proceeding of TMI: Fourth International Conference on Theoretical and Methodological Issues in MT*, pages 83–100.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*.

Philip Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*, Phuket, Thailand.

M. Oguzhan Külekçi and Kemal Oflazer. 2005. Pronunciation disambiguation in turkish. In Pinar Yolum, Tunga Güngör, Fikret S. Gürgen, and Can C. Özturan, editors, *ISCIS*, volume 3733 of *Lecture Notes in Computer Science*, pages 636–645. Springer.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL 2004 - Companion Volume*, pages 57–60.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-02)*, Philadelphia.

Sonja Niessen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntatic information. *Computational Linguistics*, 30(2):181–204.

Franz J. Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Kemal Oflazer. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148.

Kishore Papineni, Todd Ward Salim Roukos, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 311–318, Philadelphia, July. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

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 00–00, Toulouse.