Proceedings of the Australasian Language Technology Workshop 2007, pages 134-142

Statistical Machine Translation of Australian Aboriginal Languages: Morphological Analysis with Languages of Differing Morphological Richness

Simon Zwarts
Centre for Language Technology
Macquarie University
Sydney, Australia
szwarts@ics.mq.edu.au

Mark Dras
Centre for Language Technology
Macquarie University
Sydney, Australia
madras@ics.mq.edu.au

Abstract

Morphological analysis is often used during preprocessing in Statistical Machine Translation. Existing work suggests that the benefit would be greater for more highly inflected languages, although to our knowledge this has not been systematically tested on languages with comparable morphology. In this paper, two comparable languages with different amounts of inflection are tested, to see if the benefits of morphology used during the translation process, depends on the morphological richness of the language. For this work we use indigenous Australian languages: most Australian Aboriginal languages are highly inflected, where words can take a considerable number of postfixes when compared to Indo-European languages, and for languages in the same (Pama Nyungan) family, the morphological system works similarly. We show in this preliminary work that morphological analysis clearly benefits the richer of the two languages investigated, but is more equivocal in the case of the other.

1 Introduction

The majority of research in the field of Machine Translation (MT) nowadays takes a statistical approach. Morphologically rich languages have some characteristics which make MT hard, particularly in the statistical MT (SMT) context. In one common language group we want to investigate the effect of applying special morphological treatment within SMT for languages with varying degree of morphological richness. Without any morphological preprocessing, individual word counts can be quite low in highly inflected languages, causing more data sparseness than necessary, and ignoring some information which might be useful in Natural Language Processing.

Preprocessing before SMT has been used as a way of improving results. This ranges from basic tokenisation (e.g. separating possessive *'s* on English before training) to extensive syntax-based reordering (e.g. Collins et al. (2005)). Often, the choice of preprocessing proceeds without consideration of the type of language; consider for example recent work on Arabic (Sadat and Habash, 2006), where the various combinations of different preprocessing strategies are systematically worked through, with no particular attention to the characteristics of Arabic.

In most work, there is an intuitive notion that there is a connection between morphological richness of a language and the usefulness of morphological preprocessing. This is suggested in its use in parsing for Korean (Han and Sarkar, 2002) and Turkish (Eryiğit and Oflazer, 2006), and MT for Czech (Al-Onaizan et al., 1999). But in this body of work, as well as the body of work mentioned in section 3.1, only analysis of one language is performed. Moreover there is no specific measure of richness of morphology; it is not obvious how to compare the morphology of different languages such as English, Arabic, Turkish or Korean with their different combinations of prefixing, suffixing and infixing. In this paper, to examine this idea, we look at two Australian Aboriginal languages sharing a similar morphological system, but with different levels of morphological richness. Australian Aboriginal languages are quite different from most others used in Natural Language Processing. Although indigenous Australian languages individually are quite distinct, some features are shared among many of them. In particular, many indigenous Australian languages are morphologically very rich. As for most languages around the world, heavier inflection usually goes together with a freer word order. The inflection of the different words conveys information which languages like English encode in word order, for example to distinguish subjects from objects. Most indigenous Australian languages are very heavily inflected, where it is not uncommon to have three or more postfixes on the same word. In some of these languages the boundaries between postfixes and words are quite imprecise. The form of a word reflects this, and morphology might be explicitly marked on words, where roots and postfixes are separated by special characters.

This morphologically rich nature of indigenous Australian languages becomes even clearer when set against European languages. In indigenous Australian languages suffixes attached to one word can carry a meaning which in Indo-European languages has to be expressed by separate individual words as opposed to suffixes. The boundary between these suffixes and individual words is starting to become vague as the suffixes do not just add some information to the root word, but can introduce complete new meaning elements.

Our work focuses on the languages Warlpiri (an indigenous language of central Australia) and Wik Mungkan (northern Cape York, Queensland, Australia). To the best of our knowledge, no machine translation on indigenous Australian languages has been attempted before, even though these languages share some quite interesting characteristics which are unique in the world. The major part of work in MT focuses on Indo-European and Asian languages. Applying MT to indigenous Australian languages therefore presents us with a new set of challenges.

The paper is structured as follows: in section 2 we provide the background on two Australian Aboriginal languages, and we describe the available data in these languages. Section 3 starts with some work related to our method, gives some background on the data characteristics of our domain, then describes our approach, experiment and the method for evaluation. Section 4 contains the results from this evaluation and has discusses interpret these results; this leads to a conclusion in section 5.

2 Languages and Data

It is difficult to research the effect of morphological analysis between languages with a different amount of morphological richness. It is very hard to compare different languages from completely different languages families, such as comparing English with Arabic or Czech. In trying to answer the question if morphological treatment is more beneficial for more morphological rich languages, we picked two highly inflected languages from the same language family. The Pama-Nyungan languages are the most widespread family of Australian Aboriginal languages and have in common a morphological system based entirely on suffixation (Austin, 2006). By using two languages from the same family, we can make more valid comparisons between them. Another reason for using Australian Aboriginal languages, is that some of them come with some 'free' morphological analysis: morphology is indicated to a certain degree in the writing system itself. As this is a human analysis, it is therefore more reliable than automatically acquired morphology.

We will first describe the two languages and show how they differ in morphological richness.

2.1 Warlpiri

Warlpiri is an interesting language to investigate because it is often considered the prototypical free word order language, and has a number of unusual characteristics. Morphosyntactic analyses have been proposed that describe these: extensive use of case-marking morphology, syntactic ergativity, PRO-drop (null pronominals), clitic-doubling, free word order (but with tight restrictions on the location of the auxiliary), discontinuous constituents, lack of a copula verb, a grammatical category of preverbs, and so on. In terms of linguistic analysis, there is extensive coverage of the grammar in Laughren and Hoogenraad (1996). Further, it is one of the major Aboriginal languages in Australia: it is spoken natively by roughly 3000 people, with at least another 1000 speaking it as a second language; it is one of the few where children are still learning to speak the language as their first language; and it has a deal of cul-

1 Nyampu yimi, ngulaju kamparru-warnu-juku nyiya-kanti-kantiki. Kamparruju nyurru-wiyiji, ngulaju God-ju nyinajalpa yangarlu-wiyi nyanungu-mipa, yalkiri manu walyaku lawa-juku. Ngula-jangkaju, ngurrju-manu yalkiri manu walya-wiyi. 2 Yalkiri kapu walya kuja ngurrju-manu, ngula-julpa lawa-juku walyaju ngunaja kirlka-juku. Ngulaju nyiya-kanti-kanti-wangu-juku. God-rlu kuja yalkiri manu walya ngurrju-manu, ngula-jangkaju, mangkurdurlulku wuuly-kujurnu, ngulalpa parra-wangu-juku karrija murnma-juku. God-rlu-julpa Pirlirrparlu warru warra-warra-kangu mangkurdu-wanarlu.

Table 1: Warlpiri sample extract, Genesis 1:1, 2

1 Ngay John=ang, ngay wik inanganiy umpang niiyant. Ngay wik inangan Jesus Christ=antam waa'-waa'ang niiyant aak ngeen nathan yaam ke'anaman wampow. Nil piip God=angan waa' nungant Jesus Christ=ant puth than pam wanch yotamang nunangan monkan-wakantan than mee'miy ngul yipam iiyayn. Nil puth Jesus=anganiy-a, ngaantiyongkan kuch nunang nil yipam meenathow ngathar ke' pithang yimanangan, ngay puth piip God nunang monkan-wak-wakang a' puth work nungant iiy-iiyang. 2 Ngay puth latang ump-umpang niiyant ngay pithangan thath-thathanga, wik God=antam anangana niiyant ngul waa' ang, wik anangan kan-kanam nil Jesus Christ=angan waa'-waa' nil God=angan meenath nungant.

Table 2: Wik Mungkan sample extract, Revelations 1:1, 2

tural support, for example through Warlpiri Media[1] and through bilingual teaching at the Northern Territory's Community Education Centres such as Yuendumu. Table 1 gives an impression of what Warlpiri looks like.

Because Warlpiri is a heavily agglutinative language, words can have many suffixes. The result can be very long words. To not confuse speakers, suffixes longer than one syllable are usually explicitly marked with a hyphen. This is an important feature we want to exploit later. Other inflections are not marked with hyphen:[2] *Nyangkajulu* which translates as *Look at me* is built from the blocks *(nyangka, look at) (+ju, me)* and *(+lu, you)*.

Suffixes can indicate many things, like tense, case, prepositions, location and more. Some examples are: *-wangu* which translates as *not, without*; *-pala* which indicates two speakers; *-kari* which means *another*; and *-nawu* which indicates it is that specific one. An extensive lexically based analysis of Warlpiri morphosyntax is given by Simpson (1991).

To have a first indication of which part of the written language consist of explicitly marked suffixes we counted how many hyphens the average word in Warlpiri has in our corpus (section 2.3). In table 3 we can see that over half the words carry at least one suffix, with many words carrying more.

2.2 Wik Mungkan

To investigate the effects of morphological analysis we also look at another Australian Aboriginal language. We chose Wik Mungkan (Gordon, 2005), because of data availability and because it belongs to the same Pama-Nyungan language family as Warlpiri, and shares the highly agglutinative characteristics of Warlpiri. Wik Mungkan is a language which originates in northern Cape York, Queensland, Australia. The language nowadays is spoken by far fewer people (600 speakers, 400 native) and fewer resources are available for this language.

Table 2 gives an example of written Wik Mungkan. Wik Mungkan has less extensively marked morphology than Warlpiri, as can be concluded from table 4. Whereas Warlpiri has 0.615 postfixes on average per token, in Wik Mungkan we only have 0.257.

There are different writing conventions for Wik Mungkan as compared to Warlpiri. While in Warlpiri we only split on the hyphen token (−), in

[1] `http://www.warlpiri.com.au`

[2] We follow the notation convention which is common for Warlpiri to use a + for suffixes which 'glue' to the word without a hyphen and a - for suffixes where the hyphen remains when attached.

Postfixes	count	percentage
0	36389	49.32%
1	30248	41.00%
2	6373	8.63%
3	704	0.95%
4	54	0.07%
5	3	0.00%

Average Postfixes per word: 0.615

Table 3: Warlpiri words carrying postfixes

Postfixes	count	percentage
0	211563	77.80%
1	51505	18.94%
2	8226	3.02%
3	647	0.24%
4	7	0.00%

Average Postfixes per word: 0.257

Table 4: Wik Mungkan words carrying postfixes

Wik Mungkan we split on the at-sign (@), the equal sign (=), the hyphen (−), the tilde (˜) and the apostrophe (′). A token like *Jesus=anganiy-a* is split into 3 individual tokens.

2.3 Bible Corpus

Bilingual data comprising English and an indigenous Australian language is extremely scarce. SMT models usually are data hungry, with performance increasing with availability of training data. Languages like Warlpiri have more texts available, but are either not translated, or do not have a close English translation. In our experiments we used parts of the Bible. Warlpiri and other indigenous Australian languages have Bible translations, which obviously are also available in English. We used a couple of books of the Bible which are translated into Warlpiri and the complete New Testament for Wik Mungkan.[3] We verse-aligned the texts in the Aboriginal language with an English Bible translation, the World English Bible (WEB) version. In English we had the opportunity to pick between several translations. We chose for the WEB translation because of the literalness of translations and, because the language is reasonably modern English, unlike the even more literal King James version.

Overall our corpus is very small for SMT models, and we are trying to obtain more data. For the moment we are interested in relative machine translation quality, and hope that translation quality will improve when provided with more bilingual data.

3 Method

3.1 Related approaches

To treat morphologically rich indigenous languages we want to do morphological analyses before translating. We do this as a preprocessing step in Phrase Based SMT (PSMT), leaving all the other PSMT steps untouched.

Preprocessing before applying PSMT has shown to be able to improve overall MT quality. As examples, Xia and McCord (2004), Collins et al. (2005) and Zwarts and Dras (2006) present an PSMT approach with word reordering as a preprocessing step, and demonstrate improved results in translation quality.

Work in Czech, done during the 1999 Summer Workshop at John Hopkins University (Al-Onaizan et al., 1999), describes an approach where Czech was turned into ‘Czech-prime’ as a preprocessing step. For Indo-European languages, Czech is highly inflected and has a relatively free word order. In their approach they first completely discarded inflective information like number, tense and gender. Later they used this information to artificially enhance their statistical model, by enriching the vocabulary of their statistical look-up table by adding new tokens based on seen roots of words with known morphology. Note that this work was not done in the PSMT paradigm, but using the original IBM statistical models (Brown et al., 1993) for MT.

An example of a fairly comprehensive analysis of the use of morphological analysis as a preprocessing step has been done on Arabic (Sadat and Habash, 2006). An Arabic morphological analyser was used to obtain an analysis of the build-up of

[3]These texts were made available to us by the Aboriginal Studies Electronic Data Archive (ASEDA).

Arabic words. Several models were presented which preprocessed the Arabic text. The key idea was to split off word parts based on specific analysis of the word. For example, pronominal clitics are split into several words. However, Arabic morphology is not as extensive as in languages like Warlpiri. Riesa et al. (2006) is another example where the use of morphological information boosts SMT quality. In this approach the tokens are separated from prefixes and postfixes based on a predefined list, derived from a grammar book. Lee (2004) similarly works on Arabic to English translation and separates prefixes and suffixes from the word stem. In contrast with our data, where we do not need to differentiate between different affixes. We only have postfixes, although stacked on each other and playing different roles, so we treat all morphology uniformly.

3.2 Data characteristics

We want to apply morphological preprocessing to Aboriginal languages to investigate its effect on morphologically rich languages as opposed to morphologically poorer ones. In Warlpiri it is possible to explicitly mark suffixes. We separate the suffixes from the main word and treat them as individual tokens. If we have the example sentence *Pina wangkaya yimi-kari* (*Say it again another way*) where we observe *yimi-kari* with *yimi* is *word, sentence* and *-kari* is *another*; we thus separate this to *Pina wangkaya yimi -kari*. Now the SMT models can pick up the individual meaning for *yimi* and *-kari* where this previously could not have been done. In situations where we find *-kari* without the original root word, we assume the SMT model can still translate it.

As a first step to see if our intuition is right we have done a word count for both the original tokens as for the tokens when split on hyphen, to get an idea of the frequency distribution. Some words which are not frequent when counted by string match become frequent if split on suffixes. Table 5 gives an overview of this distribution.

The most frequent word when split on suffixes appears less than ten times only by itself without splitting. Also, some suffixes suddenly appear very high in the frequency list when counting them as separate tokens, while it is impossible for them to feature in the top when we do not apply splitting. The third most common token after splitting, for example, is already a suffix, beating normal root words. In the top 100, we observe 46 suffixes.

rank	count	normal	count	normalised
1	2204	manu	3018	ngula
2	1330	ngulaju	2342	manu
3	934	yangka	2192	-jana
4	773	kuja	1748	God
5	538	ngula-gankaj	1656	-kari
6	529	Jesus	1619	kuja
7	453	wankaja	1557	-juku
8	438	nyina	1508	ngulaju
9	421	nyinaja	1479	-kurra
10	420	junga-juku	1314	yangka
17	226	God-rlu	859	-nyangu
18	256	God-kurlangu	807	-nyayirni
19	255	God-ku	804	-wangu

Table 5: Warlpiri word count and postfix normalised token count

Furthermore if we look at the positions 17, 18 and 19 in the top 100 we see the same root word. If we treat these tokens literally for the PSMT machinery they are three completely separate tokens, but surely they share some meaning. If we split them on hyphen, this partially reduces the data sparseness problem.

Phrase-based translation still allows to treat the split words with morphemes together and even map them to a single English token. Because both the root token and the separated suffixes are still in the same phrasal window, as far as the PSMT machinery is concerned it can still handle them together as if they are one token.

In that case on the Warlpiri side the phrase has several tokens. The difference is that it is now up to the phrasal model to decide how to treat them, individually or as a root suffix combination. Also the individual components have been observed more often in training, so the statistical accuracy for them individually should be higher. The model can choose to use the phrase or the individual components.

3.3 Experiment

For our baseline, we use the original corpus; we compare this against the corpus where the words are split on morphology. We verse-align them, because

1	B	I am most God jaru-kari so nyurrurla-kari-piya-wangu therefore concerned the marnkurrpaku-mipa working because christ jungangku out fruit
	S	I say this to be with you as in the three other wangkamirra because Christ speech give you for an prayed with you
	R	I thank my God I speak with other languages more than you all
2	B	but pina-yanta samaria ngajuku-palangu-kurlangu-kurra does
	S	but back -yanta from my father's of to right away
	R	but you shall go to my father's house
3	B	he ngula-warnurluju-jana Peter-rluju met all the Cornelius-kurlu and to all
	S	thus in all the Peter he told all the Cornelius life and of his life
	R	but Peter began and explained to them in order saying

Table 6: Warlpiri improvement: example translation set: (B)aseline, (S)plit, (R)eference

1	B	when he said to the house will and the assembly jews-antamakan hades
	S	so when he was lost and to enter into the synagogue
	R	he departed there and went into their synagogue
2	B	but he began again to the uuyamam he said to assuredly I tell this as I would like to know by inaniu I uuyaminga
	S	but when Peter to uuama he said truly I head a uuyaminga man was not know no again I is speaks
	R	Again he denied it with an oath I don't know the man

Table 7: Wik Mungkan improvement: example translation set: (B)aseline, (S)plit, (R)eference

both corpora come with verse information. Aligning them on a sentence level within verses was found to be extremely hard, especially since the same information was probably distributed over different sentence in a way problematic for the statistical machinery.

We use the normal tools to for PSMT: GIZA++ (Och and Ney, 2003) to statistically derive a sentence alignment on token level; and the decoder Pharaoh (Koehn, 2004), a beam search decoder for PSMT. Phrases are extracted by our own Phrase Builder, which extracts phrases based on the GIZA++ alignment as described in the Pharaoh manual. We used a trigram model with interpolated Kneser-Ney discounting as a language model. The language model was built using Biblical text and was enriched with extracts from the European Parliament in order to reduce data sparseness. The SRILM (Stolcke, 2002) toolkit was used to build this language model.

Our system still suffers from quite some considerable noise. This is not uncommon for a statistical approach, but particularly hits the system hard in data-poor environments. In an abundant data scenario, noise tends to get averaged out. Some of the noise we experienced in our domain was due to poor verse alignment. There is strong indication in the test set that the PSMT system is actually translating a different sentence than the reference. Since the test and training data are obtained via the same means we assume this is also the case in the training set.

3.4 Evaluation

Often the BLEU metric is used in MT next to human evaluation, to assess translation quality. We did not perform BLEU evaluation, since our overall translation quality is quite poor. This means that in translation often synonyms are selected which BLEU does not pick up. In data sparse environments this might increase the randomness in BLEU results unfavourably and the test set already is small. More importantly however, it might favour one of the systems unfairly above the others. The PSMT system used leaves unknown words untranslated.

Proper nouns are quite likely to have morphology attached in the baseline system. When untranslated this means the proper noun is not matched against a possible English proper noun. Our system splits on morphology, leaving the proper noun by itself, which is identical to English and can be matched without translation. So although nothing is translated our system would score higher using the BLEU metric. Therefore we decided to do a human evaluation only.

To evaluate our model properly we asked human annotators to evaluate the new model against the baseline model. We used three human annotators to evaluate the Warlpiri set, and two to evaluate the Wik Mungkan set. In a blind evaluation we provided them with two alternatives for a translation and a reference verse. For each verse the ordering of baseline and 'split' version was random. So as not to overload our human annotators we drew 50 sentences from our test set, which was based on translation of unseen verses during the training period for the PSMT system. We asked the annotator to indicate which option was a better translation. They were also allowed to leave sentence-options undecided if they could not distinguish quality or if translation quality was too poor to make a good indication. They were provided with a reference translation in English.

4 Results and Discussion

The results of the human evaluation are presented in table 8. For both languages the sum over different annotates is presented for each time they chose that system. To test for statistical significance we used the non-parametric Sign Test. For the Wik Mungkan language the improvement is not statistically significant at the 5% level. With a probability of 7.5% it is possible that our system was chosen more often by random chance, and not because of improved translation. The largest frequently used threshold for statistical significance is 5%, although occasionally 10% is used, so this gives at best weak support to the rejection of the null hypothesis. For Warlpiri, however, there is overwhelming support to indicate we indeed achieved translation quality. This provides some initial support for the intuition that more highly inflected Aboriginal languages will benefit more from morphological analysis.

In absolute terms the quality of translation is quite poor, because we operate in an extremely data-poor scenario. We give some examples of translations for which the authors thought there was a clear improvement of translation quality. This also gives an indication of overall translation quality and shows the clear need for more training data for PSMT. For Warlpiri the examples can be found in Table 6, and for Wik Mungkan in Table 7.

If we take the first translation we see that the baseline has four times as many untranslated words as our system based on splitting. Furthermore we can recognise some more words, like *language* and *speech* which presumable link to each other. Now many more steps need to be performed to build a decent translation out of it, but at least there is a strong indication for a relative improvement.

Many suffixes are not captured yet. At the moment we only treated the explicitly marked suffixes because here we can be sure they are suffixes. Warlpiri knows many suffixes which are not separated with a hyphen, usually one syllable suffixes. To recognise these suffixes we need a morphological analyser. Since we have shown that splitting words contributes positively towards translation quality this seems like a logical step to extend this project in the future. Further experiments need to be carried out to see if these not explicitly marked suffixes can also improve overall quality when they are separated from their root word.

We assume our model performs better for different reasons. First of all, because we have a PSMT system, we can still pick the word with morphology if the system prefers it (word and morphology still fits the phrase window), removing most of the drawbacks a morphological prepocessing step would have without the ability to group things together in phrases. Also, the system can actually use words in cases where the individual words with that morphology attached have never been encountered before, in cases where we have seen it with different morphology. Secondly, because more words are translated, the language model starts to kick in. When words remain untranslated the language model cannot differentiate; when more words are translated we get a positive feedback. Most of all, many suffixes do not only carry morphosyntactic in-

	Wik Mungkan			Warlpiri			
	Ann. I	Ann. II	**Total**	Ann. I	Ann. II	Ann. III	**Total**
System							
Split	24	30	54	26	35	45	106
Baseline	23	18	41	7	4	3	14
Undecided	3	2	5	17	11	2	30
Sign Test							
Likelihood			$7.5 \cdot 10^{-2}$				$7.5 \cdot 10^{-20}$

Table 8: Assessment of human annotators

formation, but are actual meaning elements. Unlike English where inflection represents only a small amount of information such as tense or number, in Aboriginal languages the morphology is so extensive that to translate this morphology itself, we might need (multiple) separate words in English. By separating them, our model gives the PSMT machinery the option to exploit this.

5 Conclusion

Previous work indicates that preprocessing of Natural Language helps achieving overall quality in different Natural Language applications. Our focus is the Phrasal Statistical Machine Translation paradigm in the highly inflected indigenous Australian languages. We show a clear relative improvement of overall Machine Translation quality by separating explicitly marked suffixes when we preprocess languages like Warlpiri, which is the language with the heavier explicitly marked morphology of the two. In Wik Mungkan we observe only a possible but not statistically significant improvement.

A clear improvement of translation quality is achieved by targeting explicitly marked morphology only. However there is more morphological analysis possible in these languages. In future work we would like to included other morphology by using morphological analysers and measure their impact on machine translation quality: looking at a wider range of languages will let us test more extensively the relationship between morphological richness and the usefulness of morphological preprocessing.

6 Acknowledgement

We would like to acknowledge the translators of the Bible parts, Steve Schwarz for Walpiri and Christine Kilham for Wik Mungkan and the institute which granted us the digital resources and the right to use this material: Aboriginal Studies Electronic Data Archive (ASEDA) and the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS).

We would also like to thank the individual annotators who were willing to annotate translation quality over the evaluation texts we provided them.

References

Yaser Al-Onaizan, Jan Cuřín, Michael Jahr, Kevin Knight, John Lafferty, Dan Melam ed, Franz-Josef Och, David Purdy, Noah A. Smith, and David Yarowsky. 1999. *Statistical Machine Translation, Final Report, JHU Workshop*. URL `http://www.clsp.jhu.edu/ws99/projects/mt/final_report/mt-final-report.ps`.

Peter Austin. 2006. Countries and Language – Australia. In *International Encyclopedia of Language and Linguistics 2nd edition*, Article 1711. Oxford: Elsevier.

P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics 19(2)*, pages 263–311.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 531–540. Ann Arbor, Michigan. URL `http://www.aclweb.org/anthology/P/P05/P05-1066`.

Gülsen Eryiğit and Kemal Oflazer. 2006. Statistical

Dependency Parsing for Turkish. In *Proceedings of EACL 2006 - The 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 89–96.

Raymond G. Gordon. 2005. *Ethnologue : Languages of the World*, Fifteenth edition. URL `http://www.ethnologue.com/show_language.asp?code=wim`.

Chung-hye Han and Anoop Sarkar. 2002. Statistical Morphological Tagging and Parsing of Korean with an LTAG Grammar. In *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Formalisms*.

Philip Koehn. 2004. Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models, Philipp Koehn, Association for Machine Translation in the Americas.

Mary Laughren and Robert Hoogenraad. 1996. *A Learner's Guide to Warlpiri. Tape Course for Beginners. Wangkamirlipa Warlpirilki*. IAD Press, Alice Springs, Australia.

Young-Suk Lee. 2004. Morphological Analysis for Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 57–60.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Jason Riesa, Behrang Mohit, Kevin Knight, and Daniel Marcu. 2006. Building an English-Iraqi Arabic Machine Translation System for Spoken Utterances with Limited Resources. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech'2006)*, pages 17–21.

Fatiha Sadat and Nizar Habash. 2006. Combination of Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, pages 1–8. Association for Computational Linguistics.

Jane Simpson. 1991. *Warlpiri morphosyntax: a lexicalist approach*. Kluwer, Dordrecht.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings Intl. Conf. on Spoken Language Processing*, volume 2, pages 901–904.

Fei Xia and Michael McCord. 2004. Improving a Statistical MT System with Automatically Learned Rewrite Patterns. In *Proceedings of the 20th International Conference on Computational Linguistics (Coling)*, pages 508–514.

Simon Zwarts and Mark Dras. 2006. This Phrase-Based SMT System is Out of Order: Generalised Word Reordering in Machine Translation. In *Proceedings of the Australasian Language Technology Workshop*, pages 149–156.