Proceedings of the Second Workshop on African Language Technology
- AFLAT 2010 -

A Knowledge-Light Approach to Luo Machine Translation and Part-of-Speech Tagging

Guy De Pauw[1,2], **Naomi Maajabu**[1], **Peter Waiganjo Wagacha**[2]

[1]CLiPS - Computational Linguistics Group
University of Antwerp, Belgium
{firstname.lastname}@ua.ac.be

[2]School of Computing & Informatics
University of Nairobi, Kenya
waiganjo@uonbi.ac.ke

Abstract

This paper describes the collection and exploitation of a small trilingual corpus English - Swahili - Luo (Dholuo). Taking advantage of existing morphosyntactic annotation tools for English and Swahili and the unsupervised induction of Luo morphological segmentation patterns, we are able to perform fairly accurate word alignment on factored data. This not only enables the development of a workable bidirectional statistical machine translation system English - Luo, but also allows us to part-of-speech tag the Luo data using the projection of annotation technique. The experiments described in this paper demonstrate how this knowledge-light and language-independent approach to machine translation and part-of-speech tagging can result in the fast development of language technology components for a resource-scarce language.

1. Introduction

In recent years quite a few research efforts have investigated the applicability of statistical and machine learning approaches to African language technology. Digitally available corpora, many of which are compiled through web mining (de Schryver, 2002; Scannell, 2007), have proved to be the key component in the development of accurate and robust *performance* models of language.

Unfortunately, linguistically annotated gold-standard corpora, publicly available digital lexicons and basic language technology components, such as morphological analyzers and part-of-speech taggers for African languages are still few and far between. While great advances have been made for many of Africa's larger languages, such as Swahili, Zulu and Hausa, very few research efforts are ongoing for the majority of the continent's 2000+ languages.

This is particularly problematic for many vernacular languages that do not have official status, as they run the risk of being politically and technologically marginalized. An example of such a language is Luo (Dholuo), spoken by about three million people in Kenya, Tanzania and Uganda. This language was recently met with renewed interest, as it is the language of the Kenyan Luo tribe, which constitutes half of United States President Barack Obama's heritage.

This Nilo-Saharan language can undoubtedly be considered as *resource scarce*, as digital text material, as well as the commercial interest to develop it, is largely absent. Apart from the morphological clustering tool described in De Pauw et al. (2007), we are not aware of any published research on Luo in the field of computational linguistics.

This paper explores the applicability of a *knowledge light* approach to the development of Luo language technology. By compiling and exploiting a small corpus of translated texts in Luo, English and Swahili, we will show how a basic machine translation system for the language can be built and how linguistic annotation can be transferred from a *resource-rich* language to the resource-scarce language in question.

In Section 2 we describe the trilingual parallel corpus that forms the building blocks of our approach. We then describe how we can use standard techniques to develop a basic statistical machine translation system for the language pairs in question (Section 3). By further exploiting the automatically induced word alignment patterns, we show in Section 4 how we can project part-of-speech tag annotation from English and Swahili onto Luo. We conclude the paper with a discussion of the advantages and limitations of the approach and outline some options for future research.

2. A Trilingual Corpus English - Swahili - Luo

Unlike many other smaller vernacular languages, such as Gĩkũyũ, Luo has relatively little data available on the Internet. A small web-mined corpus is nevertheless available in the archive of the Crúbadán project (Scannell, 2007). We used this corpus as a seed to perform further web mining, resulting in an updated monolingual Luo corpus of about 200,000 words.

2.1. Parallel Data

A modernized translation of the New Testament in Luo was recently made digitally available on-line by the International Bible Society (2005). Not only does this document effectively double the size of our monolingual Luo corpus, it also enables the compilation of a parallel corpus, by aligning the text with the same documents in other languages. This parallel corpus can consequently be used to power a statistical machine translation system[1].

We used the New Testament data of the SAWA corpus (De Pauw et al., 2009) to construct a small trilingual parallel corpus English - Luo - Swahili. The chapter and verse indications that are inherently present in the data hereby function as paragraph alignment markers, further facilitating automatic sentence alignment, using the approach described in Moore (2002).

[1]Interestingly, an audio-book version of the New Testament is also available from *Faith Comes by Hearing*, opening up the possibility of investigating data-driven speech recognition for Luo as well.

Token counts for the trilingual parallel corpus can be found in Table 1. The corpus was randomly divided into an 80% training set, a 10% validation set and a 10% evaluation set. The validation set allows us to tweak algorithm settings for language modeling, while the evaluation set can be used to measure the accuracy of the machine translation system on previously unseen data.

	English	Swahili	Luo
New Testament	192k	156k	170k

Table 1: Token counts for trilingual parallel NT corpus

English	You/PP/you have/VBP/have let/VBN/let go/VB/go of/IN/of the/DT/the commands/NNS/command ...
Swahili	Ninyi/PRON/ninyi mnaiacha/V/mnai amri/N/amri ya/GEN-CON/ya Mungu/PROPNAME/Mungu ...
Luo	Useweyo/useweyo Chike/chike Nyasaye/nyasaye mi/mi koro /koro umako/mako ...

Table 2: Factored Data (Mark 7:8)

2.2. Annotation

Standard statistical machine translation tools can be used to *align* the words of the source and target language, by looking for translation pairs in the sentences of the parallel corpus. This typically requires a large amount of data, as it involves scanning for statistically significant collocation patterns and cognates in the orthography. For the language pairs under investigation, the latter information source is limited, as named entities are often transliterated to match the Swahili and Luo pronunciation patterns.

Introducing a layer of linguistic abstraction, however, can aid the alignment process: knowing that a word in the source language is a noun, can help connecting it with a corresponding noun in the target language. Similarly, lemmatization can aid word alignment, as it allows scanning for word types, rather than tokens.

We therefore part-of-speech tagged and lemmatized the English part of the corpus using the TreeTagger (Schmid, 1994). We used the systems described in De Pauw et al. (2006) and De Pauw and de Schryver (2008) to tag and stem the Swahili data. The result is illustrated in the first two rows of Table 2. Each English token consists of three parts: the word form itself, its part-of-speech tag and its lemma. Each Swahili token likewise consists of three parts: the word itself, its part-of-speech tag and its stem.

For Luo, no such tools are available. We therefore made use of the MORFESSOR algorithm (Creutz and Lagus, 2005), which attempts to automatically induce the morphotactics of a language on the basis of a word list. While this fully automated approach obviously generates a flawed morphological description of the language in question, both Koehn et al. (2007) and Virpioja et al. (2007) indicate that even a very rough type of morphological normalization can still significantly aid word alignment of highly inflecting languages.

A lexicon compiled from the monolingual Luo corpus and the training and validation sets of the Luo portion of the parallel corpus was used to train the MORFESSOR algorithm. The output contains a segmented word list, indicating prefixes, stems and suffixes, as well as a segmentation model that allows for the segmentation of previously unseen words. This model was consequently used to morphologically segment the words of the evaluation set. The automatically induced stemming information is then added to the Luo part of the parallel corpus, giving us the type of factored data, illustrated in the last row of Table 2.

A small portion of the evaluation set (about 2,000 words) was also manually annotated for part-of-speech tags. We restricted ourselves to a very small set of 13 part-of-speech tags[2]. This gold-standard data allows us to evaluate the accuracy with which the part-of-speech tags are projected from English and Swahili (Section 4).

3. Machine Translation

In a first set of experiments, we explore the feasibility of statistical machine translation between the two language pairs under investigation: English $\leftrightarrow$ Luo and Swahili $\leftrightarrow$ Luo. We use MOSES, the standard toolkit for statistical machine translation (Koehn et al., 2007), which incorporates word-alignment using GIZA++ (Och and Ney, 2003) and a phrase-based decoder. The big advantage of MOSES is its ability to efficiently process *factored* data (Table 2) during word alignment, the building of phrase tables and final decoding.

The four translation models were trained using the factored data, described in Section 2.2. For each of the three target languages we built an n-gram language model using the SRILM toolkit (Stolcke, 2002). The value for n was tuned by optimizing perplexity values on the respective validation sets. The Gigaword corpus (Graff, 2003) served as source material for the English language model. For the Swahili language model, the twenty million word *TshwaneDJe Kiswahili Internet Corpus* (de Schryver and Joffe, 2009) was used. To construct the Luo language model, we used the 200,000 word monolingual Luo corpus (Section 2), complemented by the training and validation sets of the parallel corpus.

After training, MOSES used the resulting translation model to translate the evaluation set. By comparing the generated translations to the reference translations, we can estimate the quality of the translation systems using the standard BLEU and NIST metrics (Papineni et al., 2002; Doddington, 2002). The experimental results can be found in Table 3. For the sake of comparison, we have performed each experiment twice: once on just word forms and once on the factored data (indicated by **[F]** in Table3). This allows us to quantitatively illustrate the advantage of using factored data.

[2]The Luo tag set is listed in Table 6 (Addendum).

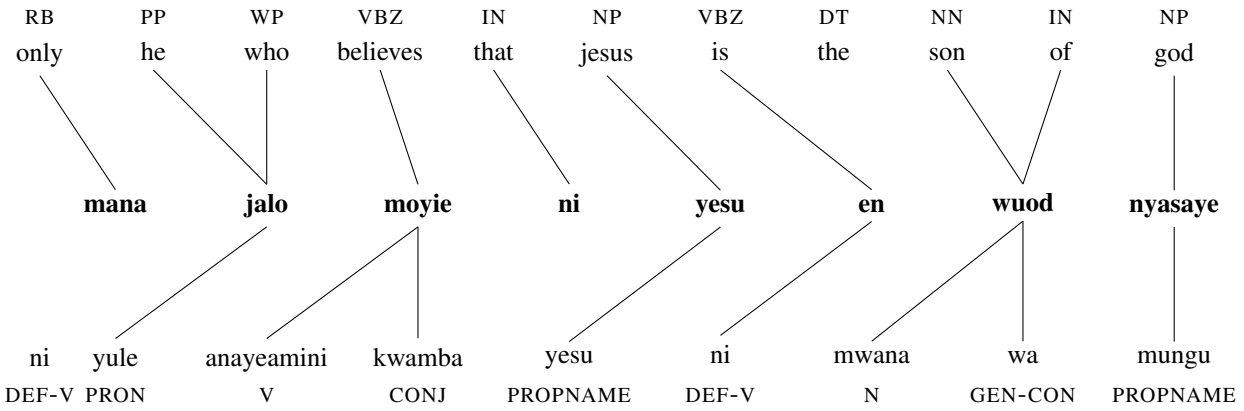

Figure 1: Projection of Part-of-Speech Tag Annotation Using Automatically Induced Word Alignment

The OOV rate expresses the percentage of out-of-vocabulary words, i.e. how many words of the evaluation set of the target language are unknown to the language model. Unsurprisingly, Luo has the highest OOV rate, which can be attributed to the limited amount of corpus data and the morphological complexity of the language.

	OOV rate	NIST	BLUE
Luo → English	4.4%	5.39	0.23
Luo → English [F]		6.52	0.29
English → Luo	11.4%	4.12	0.18
English → Luo [F]		5.31	0.22
Luo → Swahili	6.1%	2.91	0.11
Luo → Swahili [F]		3.17	0.15
Swahili → Luo	11.4%	2.96	0.10
Swahili → Luo [F]		3.36	0.15

Table 3: Results of Machine Translation Experiments

For the language pair English ↔ Luo, NIST and BLEU scores are quite encouraging. While not in the same league as those reported for Indo-European language pairs, the scores are fairly reasonable, considering the very limited amount of training data. We also digitized a translation dictionary English - Luo (Bole, 1997), but adding this information source during word alignment, unfortunately did not lead to significantly higher BLUE or NIST scores during decoding.

The use of factored data seems to have significantly aided word alignment. While it can do nothing to remedy the problem of out-of-vocabulary words, paired T-tests show that the difference in accuracy between the unfactored and factored translation models is statistically significant.

It is important to point out, however, that the encouraging results can be largely attributed to the fact that the system was trained and evaluated on texts from the same domain, i.e. biblical data. Trials on secular data revealed a much higher OOV rate (up to 40% for Luo as a target language) and rather sketchy translations, as illustrated in the translation examples displayed in Table 4.

The language pair Swahili ↔ Luo fairs much worse. While this may seem surprising given the geographical proximity of the two languages, they are genealogically very different, with Swahili being a Bantu language, as opposed to the Nilotic language of Luo.

(1)	**Source**	*en ng'a moloyo piny ? mana jalo moyie ni yesu en wuod nyasaye*
	Reference	*who is it that overcomes the world ? Only he who believes that jesus is the son of god*
	Translation	*who is more than the earth ? only he who believes that he is the son of god*
(2)	**Source**	*atimo erokamano kuom thuoloni*
	Reference	*I am thankful for your leadership*
	Translation	*do thanks about this time*

Table 4: Translation examples for religious data (1) and secular data (2)

The low BLEU and NIST scores can be attributed to the highly inflectional nature of both the source and target language in this case. Despite the fact that word alignment was performed using factored data, there is no morphological generation component for either language in the translation models, so that often an erroneously inflected word form will be generated during decoding, thereby adversely affecting translation quality.

4. Projection of Part-of-Speech Tags

The idea behind projection of annotation is to use the automatically induced word alignment patterns to project the part-of-speech tags of the words in the source language to the corresponding words in the target language. This is illustrated in Figure 1. The Luo sentence in the middle is word aligned with the English sentence at the top and the Swahili sentence at the bottom.

The direct correspondence assumption (Hwa et al., 2002) suggests that the part-of-speech tag of a source language word can be safely projected onto the corresponding word in the target language. The Luo word *moyie* for example can receive the English verbal part-of-speech tag VBZ. Likewise, the word *nyasaye* can be tagged using the Swahili

part-of-speech tag PROPNAME.

In many cases, there is a one-to-many pattern, for example in the alignment of the Luo word *moyie* and the Swahili phrase *anayeamini kwamba*. In such cases, we refer to a predefined tag priority list (see Addendum), which would discard the CONJ in favor of the verbal V tag.

After transferring the source language part-of-speech tags to the target language, a look-up table (see Addendum) maps the projected part-of-speech tags to those of the target tag set (see Footnote 1). We can then compare the projected and converted tag to that of the small gold-standard evaluation set to estimate the feasibility of the approach.

Table 5 displays the results of the experiment. The *Rate* score expresses how many words in the target language received a tag. *Precision* expresses how many of these projected tags are correct. Finally, the *accuracy* score expresses the overall tagging accuracy on the evaluation set.

	Rate	**Precision**	**Accuracy**
English → Luo	73.6%	69.7%	51.3%
Swahili → Luo	71.5%	68.4%	48.9%
Exclusive → Luo	66.5%	78.5%	52.2%
Inclusive → Luo	75.4 %	69.5%	52.4%

Table 5: Results of Projection of Part-of-Speech Tag Annotation Experiments

Projecting part-of-speech tags from English onto Luo works reasonably well. Almost 75% of the words receive a tag. Quite a few words do not receive a part-of-speech tag, either through erroneous word alignment patterns or because there simply was no corresponding word in the target language. Tags are projected with a precision of close to 70%. Overall, more than half of the words in the evaluation set received the correct tag.

The BLEU and NIST scores for the language pair Swahili - Luo (Table 3) were significantly lower than those for the language pair English - Luo. Interestingly, this performance drop is much less significant in this experiment. This further indicates that the problem in translating between these two languages is mostly due to the absence of a morphological generation component during decoding, rather than to fundamental issues in the word alignment phase.

The last two rows of Table 5 show the scores of combined models. The *Exclusive* model only retains a part-of-speech tag for the target word, if both source languages project it. This obviously results in a lower tagging rate, but significantly improves the precision score. The *Inclusive* model likewise combines the two projections, but does not require the two projected tags to be the same. In the case of a projection conflict, the English part-of-speech tag is preferred. This further improves on the Rate and Accuracy scores, but looses out on Precision.

Error analysis shows that many of the tagging errors are being made on the highly frequent function words. This is encouraging, since these constitute a closed-class and mostly unambiguous set of words which can be tagged using a simple look-up table. The translation dictionary (Bole, 1997) can also be used to further complement the part-of-speech tagging database, hopefully further increasing rate and possibly accuracy of the tagger.

Furthermore, the automatically part-of-speech tagged corpus can now be used as training material for a data-driven part-of-speech tagger. Particularly morphological clues can be automatically extracted from this data that can help in tagging previously unseen words.

5. Discussion

To the best of our knowledge, this paper presents the first published results on statistical machine translation for a Nilotic language. A very small trilingual parallel corpus English - Swahili - Luo, consisting of biblical data, was compiled. Morphosyntactic information was added, either by using existing annotation techniques, or by using the unsupervised induction of morphological segmentation patterns.

The results show that using factored data enables the development of a basic machine translation system English - Luo, as well as the projection of part-of-speech tag information for the resource-scarce language of Luo. We hope to replicate this experiment for other vernacular languages as well, for example Gĩkũyũ, which may benefit from its genealogical proximity to Swahili (Wagacha et al., 2006).

One of the bottlenecks in the current approach is the limited morphological analysis and generation capabilities for Swahili and particularly Luo. The unsupervised approach used to stem the Luo data can only serve as a stop-gap measure and a more intricate morphological component will be needed to improve on the current BLEU and NIST scores.

For part-of-speech tagging we will investigate how the Luo corpus, annotated through projection, can be used as training material for a morphologically aware data-driven tagger. Such a part-of-speech tagger may significantly outperform the current approach, as it will be able to process out-of-vocabulary words and smooth over errors introduced by erroneous word alignment patterns or errors made in the annotation of the source language.

The experiments described in this paper serve as a proof-of-concept that language technology components and applications for African languages can be developed quickly without using manually compiled linguistic resources for the target language in question. While we certainly do not claim that the resulting part-of-speech tagger or machine translation system can serve as an end product, we are confident that they can aid in the further development of linguistically annotated Luo corpora.

Acknowledgments and Demonstration

The first author is funded as a Postdoctoral Fellow of the Research Foundation - Flanders (FWO).

The authors wish to thank the reviewers for their constructive and insightful comments. We are also indebted to Kevin Scannell for making his Dholuo data available for research purposes.

A demonstration system and the trilingual parallel corpus is available at `http://aflat.org/luomt`.

6. References

Bole, O.A. (1997). *English-Dholuo dictionary*. Kisumu, Kenya: Lake Publishers & Enterprises.

Creutz, M. & Lagus, K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. In T. Honkela, V. Könönen, M. Pöllä & O. Simula (Eds.), *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*. Espoo, Finland: Helsinki University of Technology, Laboratory of Computer and Information Science, pp. 106–113.

De Pauw, G. & de Schryver, G-M. (2008). Improving the computational morphological analysis of a Swahili corpus for lexicographic purposes. *Lexikos*, 18, pp. 303–318.

De Pauw, G., de Schryver, G-M. & Wagacha, P.W. (2006). Data-driven part-of-speech tagging of Kiswahili. In P. Sojka, I. Kopeček & K. Pala (Eds.), *Proceedings of Text, Speech and Dialogue, 9th International Conference*. Berlin, Germany: Springer Verlag, pp. 197–204.

De Pauw, G., Wagacha, P.W. & Abade, D.A. (2007). Unsupervised induction of Dholuo word classes using maximum entropy learning. In K. Getao & E. Omwenga (Eds.), *Proceedings of the First International Computer Science and ICT Conference*. Nairobi, Kenya: University of Nairobi, pp. 139–143.

De Pauw, G., Wagacha, P.W. & de Schryver, G-M. (2009). The SAWA corpus: a parallel corpus English - Swahili. In G. De Pauw, G-M. de Schryver & L. Levin (Eds.), *Proceedings of the First Workshop on Language Technologies for African Languages (AfLaT 2009)*. Athens, Greece: Association for Computational Linguistics, pp. 9–16.

de Schryver, G-M. & Joffe, D. (2009). *TshwaneDJe Kiswahili Internet Corpus*. Pretoria, South Africa: TshwaneDJe HLT.

de Schryver, G-M. (2002). Web for/as corpus: A perspective for the African languages. *Nordic Journal of African Studies*, 11(2), pp. 266–282.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In M. Marcus (Ed.), *Proceedings of the second international conference on Human Language Technology Research*. San Francisco, USA: Morgan Kaufmann Publishers Inc., pp. 138–145.

Faith Comes by Hearing. (2010). *Luo Audio Bible*. [Online]. Available: http://www.faithcomesbyhearing.com (accessed March 2010).

Graff, D. (2003). *English Gigaword*. [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05 (accessed March 2010).

Hurskainen, A. (2004). HCS 2004 – Helsinki Corpus of Swahili. Technical report, Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC.

Hwa, R., Resnik, Ph., Weinberg, A. & Kolak, O. (2002). Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, USA: Association for Computational Linguistics, pp. 392–399.

International Bible Society. (2005). *Luo New Testament*. [Online]. Available: http://www.biblica.com/bibles/luo (accessed March 2010).

Koehn, Ph., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. & Herbst, E. (2007). MOSES: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*. Prague, Czech Republic: Association for Computational Linguistics, pp. 177–180.

Marcus, M.P., Santorini, B. & Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), pp. 313–330.

Moore, R.C. (2002). Fast and accurate sentence alignment of bilingual corpora. In S.D. Richardson (Ed.), *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*. Berlin, Germany: Springer Verlag, pp. 135–144.

Och, F.J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), pp. 19–51.

Papineni, K., Roukos, S., Ward, T. & Zhu, W-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Morristown, USA: Association for Computational Linguistics, pp. 311–318.

Scannell, K. (2007). The Crúbadán project: Corpus building for under-resourced languages. In C. Fairon, H. Naets, A. Kilgarriff & G-M Schryver (Eds.), *Building and Exploring Web Corpora - Proceedings of the 3rd Web as Corpus Workshop*. volume 4 of *Cahiers du Cental*, Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain, pp. 5–15.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In D. Jones (Ed.), *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK: UMIST, pp. 44–49.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In J.H.L. Hansen & B. Pellom (Eds.), *Proceedings of the International Conference on Spoken Language Processing*. Denver, USA: International Speech Communication Association, pp. 901–904.

Virpioja, S., Väyrynen, J.J., Creutz, M. & Sadeniemi, M. (2007). Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In B. Maegaard (Ed.), *Proceedings of the Machine Translation Summit XI*. Copenhagen, Denmark: European Association for Machine Translation, pp. 491–498.

Wagacha, P.W., De Pauw, G. & Getao, K. (2006). Development of a corpus for Gĩkũyũ using machine learning techniques. In J.C. Roux (Ed.), *Networking the development of language resources for African languages*. Genoa, Italy: ELRA, pp. 27–30.

Addendum: Tag Projection Table

Table 6 describes how English part-of-speech tags (Marcus et al., 1993) and Swahili part-of-speech tags (Hurskainen, 2004; De Pauw et al., 2006) are projected onto the Luo tag set. The Luo tag lists expresses tag priority (from top to bottom) in the event of projection conflicts. The Luo tag `particle` was used as a retainer for particles and language specific function words in English and Swahili.

English	Luo	Swahili
NN NNS SYM	noun	ABBR CAP IDIOM N (all concords)
MD VB VBD VBG VBN VBP VBZ	verb	DEF-V V (all concords) IMP
JJ JJR JJS	adjective	A-INFL A-UINFL AD-ADJ ADJ
RB RBR RBS WRB	adverb	ADV NEG
PP PP$ WP$	pronoun	DEM PRON
CD	number	NUM
FW	loan word	AR ENG
IN TO	preposition	PREP
NP NPS	proper name	PROPNAME
UH	exclamation	EMPH EXCLAM RHET
CC	conjunction	CC CONJ
DT EX PDT POS RP WDT WP	particle	AG-PART GEN-CON INTERROG NA-POSS REL SELFSTANDING
LS	punctuation	PUNC

Table 6: Tag Projection Table