Korean Language Resources for Everyone

Jungyeul Park
Department of Linguistics
University of Arizona
Tucson, AZ 85721
jungyeul@email.arizona.edu

Jeen-Pyo Hong
NAVER LABS
NAVER Corporation
Republic of Korea
jeenpyo.hong@navercorp.com

Jeong-Won Cha
Department of Computer Engineering
Changwon National University
Republic of Korea
jcha@changwon.ac.kr

Abstract

This paper presents open language resources for Korean. It includes several language processing models and systems including morphological analysis, part-of-speech tagging, syntactic parsing for Korean, and standard evaluation Korean-English machine translation data with the Korean-English statistical machine translation baseline system. We make them publicly available to pave the way for further development regarding Korean language processing.

1 Introduction

This paper presents open language resources (LRs) for Korean. We provide necessary data, models, tools, and systems to analyze Korean sentences. It includes the whole working pipeline from part-of-speech (POS) tagging to syntactic parsing for Korean. We also provide the Korean-English statistical machine translation (SMT) baseline system and newly created standard data for MT evaluation. All LRs described in this paper will be publicly available under the MIT License (MIT).

2 Korean Language

Korean is an agglutinative language in which "words typically contain a linear sequence of MORPHS" (Crystal, 2008). Words in Korean (*eojeols*), therefore, can be formed by joining content and functional morphemes to indicate such meaning. These *eojeols* can be interpreted as the basic segmentation unit and they are separated by a blank space in the Korean sentence. Let us consider the sentence in (1). For example, *unggaro* is a content morpheme (a proper noun) and a postposition *-ga* (a nominative case marker) is a functional morpheme. They form together a single word *unggaro-ga* ('Ungaro + NOM'). For convenience sake, we add - at the beginning of functional morphemes, such as *-ga* for NOM to distinguish between content and functional morphemes. The nominative case marker *-ga* or *-i* may vary depending on the previous letter - vowel or consonant. A predicate *naseo-eoss-da* also consists of the content morpheme *naseo* ('become') and its functional morphemes (*-eoss* 'PAST' and *-da* 'DECL').

3 Morphological analysis and POS tagging

Numerous studies pertaining to morphological analysis and POS tagging for Korean have been conducted over the past decades (Cha et al., 1998; Lee and Rim, 2004; Kang et al., 2007; Lee, 2011). Most morphological analysis and POS tagging for Korean have been conducted based on an *eojeol*. In the system of Korean POS taggers, a morphological analysis is generally followed by a POS tagging step. That is, all possible sequences of morphological segmentation for a given word are generated during the morphological analysis and the *possible* (or best) correct sequences are then selected during POS tagging.

ESPRESSO, a Korean POS tagger described in Hong (2009) is publicly available[1]. It greatly improves the accuracy of POS tagging using POS patterns of words in which it obtains up to 95.85% ac-

[1] Note that there is another resource with the same name (Pantel and Pennacchiotti, 2006).

(1) a. 프랑스의 세계적인 의상 디자이너 엠마누엘 웅가로가 실내 장식용 직물 디자이너로 나섰다.

b. *peurangseu-ui segyejeok-in uisang dijaineo emmanuel unggaro-ga silnae jangsikyong*
France-GEN world class-REL fashion designer Emanuel Ungaro-NOM interior decoration
jikmul dijaineo-ro naseo-eoss-da.
textile designer-AJT become-PAST-DECL.

'The world class French fashion designer Emanuel Ungaro became an interior textile designer.'

Figure 1: Example of the Korean sentence

Input:
프랑스의 세계적인 의상 디자이너 엠마누엘 웅가로가 실내 장식용 직물 디자이너로 나섰다.

Output:
```
 프랑스의     BOS  프랑스/NNP+의/JKG
 세계적인          세계/NNG+적/XSN+이/VCP+ㄴ/ETM
 의상              의상/NNG
 디자이너          디자이너/NNG
 엠마누엘          엠마누엘/NNP
 웅가로가          웅가로/NNP+가/JKS
 실내              실내/NNG
 장식용            장식용/NNG
 직물              직물/NNG
 디자이너로        디자이너/NNG+로/JKB
 나섰다.      EOS  나서/VV+었/EP+다/EF+./SF
```

Figure 2: Input and output examples of ESPRESSO for Korean POS tagging

curacy for Korean. Figure 2 shows the input and output formats of ESPRESSO for Korean POS tagging. Even though ESPRESSO can yield several output formats, we only show the Sejong corpus-like format in this paper, in which we use the format for the input of syntactic analysis. While ESPRESSO indicates BOS and EOS (the beginning and the end of a sentence, respectively), the actual Sejong corpus does not contain BOS and EOS labels. The original Sejong morphologically analyzed corpus annotates the sentence boundary using the markup language.

We use Sejong POS tags, the mostly used POS tag information for Korean. Figure 3 shows the summary of the Sejong POS tag set and its mapping to the Universal POS tag (Petrov et al., 2012). We convert the `XR` (non-autonomous lexical root) into the `NOUN` because they are mostly considered as a noun or a part of noun (*e.g. minju*/XR ('democracy')). The current Universal POS tag mapping for Sejong POS tags is based on a handful of POS patterns of Korean words. However, combinations of words in Korean are very productive and exponential. Therefore, the number of POS patterns of the word does not converge as the number of words increases. For example, the Sejong Treebank contains about 450K words and almost 5K POS patterns. We also test with the Sejong morphologically analyzed corpus which contains over 10M words. The number of POS patterns does not converge and it increases up to over 50K. The wide range of POS patterns is mainly due to the fine-grained morphological analysis results, which shows all possible segmentations divided into lexical and functional morphemes. These various POS patterns indicate useful morpho-syntactic information for Korean. For example, Oh et al. (2011) predicted function labels (phrase-level tags) using POS patterns that would improve dependency parsing results.

Sejong POS	description	Universal POS
NNG, NNP, NNB, NR, XR	Noun related	NOUN
NP	Pronoun	PRON
MAG,	Adverb	ADV
MAJ	Conjunctive adverb	CONJ
MM	Determiner	DET
VV, VX, VCN, VCP	Verb related	VERB
VA	Adjective	ADJ
EP, EF, EC, ETN, ETM	Verbal endings	PRT
JKS, JKC, JKG, JKO, JKB, JKV, JKQ, JX, JC	Postpositions (case markers)	ADP
XPN, XSN, XSA, XSV	Suffixes	PRT
SF, SP, SE, SO, SS	Punctuation marks	PUNC (.)
SW	Special characters	X
SH, SL	Foreign characters	X
SN	Number	NUM
NA, NF, NV	Unknown words	X

Figure 3: POS tags in the Sejong corpus and their 1-to-1 mapping to Universal POS tags

4 Syntactic analysis

Statistical parsing trained from an annotated data set has been widespread. However, while there are manually annotated several Korean Treebank corpora such as the Sejong Treebank (SJTree), only a few works on statistical Korean parsing have been conducted.

4.1 Phrase structure parsing

For previous work on constituent parsing, Sarkar and Han (2002) used an early version of the Korean Penn Treebank (KTB) to train lexicalized Tree Adjoining Grammars (TAG). Chung et al. (2010) used context-free grammars and tree-substitution grammars trained on data from the KTB. Choi et al. (2012) proposed a method to transform the word-based SJTree into an entity-based Treebank to improve the parsing accuracy. There exit several phrase structure parsers such as Stanford (Klein and Manning, 2003), Bikel (Bikel, 2004), and Berkeley (Petrov and Klein, 2007) parsers (either lexicalized or unlexicalized) that we can train with the Treebank.

For phrase structure parsing, we provide a parsing model for the Berkely parser.[2] Choi et al. (2012) tested Stanford, Bikel, and Berkeley parsers and the Berkeley parser shows the best results for phrase structure parsing for Korean. The input sentence of phrase structure parsers is generally the tokenized sentence. It can be obtained by performing the segmentation task for a word. Each segmented morpheme becomes a leaf node in the phrase structure. Therefore, we use the tokenization scheme based on POS tagging. Figure 4 shows the input and output formats for the Berkeley parser. As preprocessing tools, we provide `MakeBerkeleyTestIn` and `MakeBerkeleyTestWithPOSIn`. They convert ESPRESSO's output into the Berkely parser's input by tokenizing the Korean sentence with or without POS information, respectively.

4.2 Dependency parsing

For previous work on dependency parsing for Korean, Chung (2004) presented a model for dependency parsing using surface contextual information. Oh and Cha (2010), Choi and Palmer (2011) and Park et al. (2013) independently developed a parsing model from the Korean dependency Treebank. They converted automatically the phrase-structured Sejong Treebank into the dependency Treebank. To convert into dependency grammars, Park et al. (2013) summarized as follows.

We, first, assign an anchor for nonterminal nodes using bottom-up breadth-first search. An anchor is

[2] `https://github.com/slavpetrov/berkeleyparser`

```
Input:
프랑스 의 세계 적 이 ㄴ 의상 디자이너 엠마누엘 웅가로 가 실내 장식 용 직물 디자이너 로 나
서 었 다 .

Output:
(S (NP-SBJ (NP (NP-MOD (NNP 프랑스) (JKG 의)
               (NP (VNP-MOD (NNG 세계) (XSN 적) (VCP 이) (ETM ㄴ))
                    (NP (NP (NNG 의상))
                         (NP (NNG 디자이너)))))
                              (NP-SBJ (NP (NNP 엠마누엘))
                                        (NP-SBJ (NNP 웅가로) (JKS 가))))
   (VP (NP-AJT (NP (NP (NP (NNG 실내))
                              (NP (NNG 장식) (XSN 용)))
                         (NP (NNG 직물)))
                    (NP-AJT (NNG 디자이너) (JKB 로)))
       (VP (VV 나서) (EP 었) (EF 다) (SF .))))
```

Figure 4: Input and output examples for Korean phrase structure parsing

the lexical terminal node where each nonterminal node can have as a head node. We use lexical anchor rules described in Park (2006) for the SJTree. Lexical anchor rules distinguish dependency relations. We assign only the lexical anchor for nonterminal nodes and finding dependencies in the next step. Lexical anchor rules give priorities to the rightmost child node, which inherits mostly the same phrase tag. Exceptionally, in case of “VP and VP" (or “S and S"), the leftmost child node is assigned as an anchor. Then, we can find dependency relations between terminal nodes using the anchor information as follows:

1. The head is the anchor of the parent of the parent node of the current node.
2. If the anchor is the current node and
   (a) if the parent of the parent node does not have another right sibling, the head is itself.
   (b) if the parent of the parent node have another right sibling, the head if the anchor of the right sibling.

Results from the conversion can allow to train existing dependency parsers. Figure 5 presents an example of the original Sejong Treebank (above) and its automatically-converted dependency representation.[3] The address of terminal nodes (underneath) and the anchor of nonterminal node (on its right) are arbitrarily assigned for dependency conversion algorithm using lexical head rules. The head of the terminal node 1 is the node 4, which is the anchor of the parent of the parent node (NP:4). The head of the terminal node 4 is the node 6 where the anchor of its ancestor node is changed from itself (NP-SBJ:6). The head of the terminal node 11 is itself where the anchor of the root node and itself are same (S:11).

The parsing model of MaltParser (Nivre et al., 2006) is provided for dependency parsing for Korean.[4] As preprocessing tools, we provide `MakeMaltTestIn`. It converts ESPRESSO’s output into the MaltParser’s input by generating required features for MaltParser. Figure 6 shows example of the input and the output of MaltParser. We use the data format of CoNLL-X dependency parsing, described in Figure 7 (partially presented). See `http://ilk.uvt.nl/conll` for other information about the data format of CoNLL-X that MaltParser requires. From word and POS information, we convert them into features that MaltParser requires for Korean dependency parsing.

[3] The figure originally appeared in Park et al. (2013) with minor errors, and we corrected them.

[4] `http://www.maltparser.org`

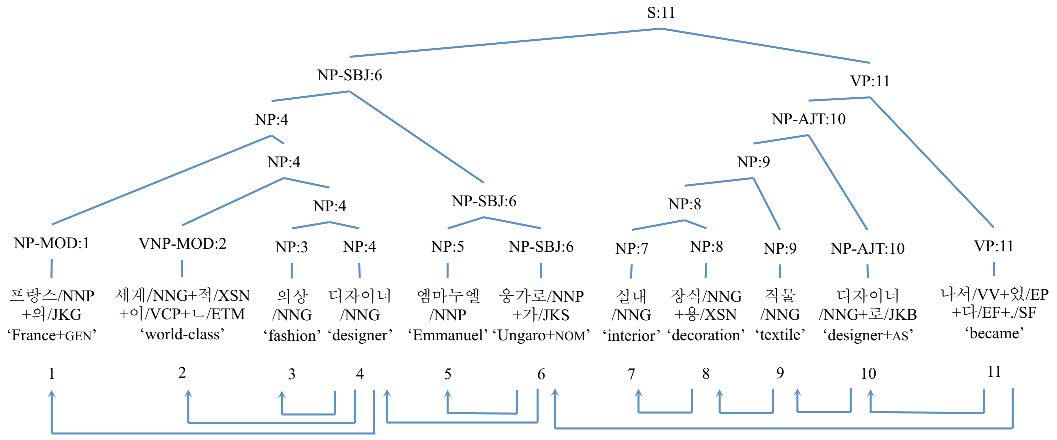

Figure 5: Example of the original Sejong Treebank (above) and its automatically-converted dependency representation (below)

4.3 Discussion on parsing for Korean

In previous work on parsing for Korean, either phrase structure or dependency parsing, while Park et al. (2013) proposed the 80-10-10 corpus split for training, development and evaluation, others often used cross validation (Oh and Cha, 2010; Choi et al., 2012; Oh and Cha, 2013).

For phrase structure parsing, Choi et al. (2012) obtained up to 78.74% F_1 score. For dependency parsing, Oh and Cha (2013) obtained 87.03% (10-fold cross validation) and Park et al. (2013) up to 86.43% (corpus split) by using external case frame information.

Currently, we distribute only parsing models instead of parsers and training data themselves because of following reasons. First, the Sejong Treebank that we use to train and evaluate is not allowed to be distributed by third parties. Corpus users should ask directly to National Institute of the Korean Language[5] for their own usage. Therefore, it would be easy that we only make current parsing models publicly available instead of actual training data. Second, multilingualism becomes more and more important. Many natural language processing (NLP)-related works rely on a single system to deal with multiple languages homogeneously. Berkeley parser and MaltParser in which we provide parsing models have been developed for many other languages and users can easily obtain their up-to-dated parsing systems and models for several other languages.

We provide parsing models trained only on the training data, which can be subject to the baseline parsing system for Korean to be compared in future work. Table 1 presents the current baseline parsing results using phrase structure grammars by the Berkeley parser. We performed 5-fold and 10-fold cross-validation as well as corpus split evaluation for comparison purpose. We also tested both cases in which Berkeley parser selects POS tags by itself during the parsing task (parser) and we provided gold POS tags before parsing (gold). Reported results are improved compared to Choi et al. (2012) because we have corrected syntactic and POS tagging errors in the Sejong Treebank for the current work. Since results between different evaluation methods are not statistically significant, we propose to use 80-10-10 corpus split evaluation using the current distributed parsing model. For the current baseline parsing results using dependency grammars trained using corpus split, Park et al. (2013) reported that MaltParser on the Sejong Treebank can obtain 85.41% for the unlabeled attachment score (UAS). We provide the development data (10% of the corpus) and the evaluation data set (last 10%) as well as the parsing model (trained on first 80% of the corpus) for phrase-structure and dependency parsing.

[5] http://www.korean.go.kr

Input:

```
1   프랑스의     프랑스      NNP            NNP+JKG              JKG
2   세계적인     세계적이    NNG+XSN+VCP    NNG+XSN+VCP+ETM      ETM
3   의상         의상        NNG            NNG                  _
4   디자이너     디자이너    NNG            NNG                  _
5   엠마누엘     엠마누엘    NNP            NNP                  _
6   웅가로가     웅가로      NNG            NNG+JKS              JKS
7   실내         실내        NNG NNG        _
8   장식용       장식용      NNG            NNG                  _
9   직물         직물        NNG            NNG                  _
10  디자이너로   디자이너    NNG            NNG+JKB JKB
11  나섰다.      나서        VV             VV+EP+EF+SF          EP|EF|SF
```

Output :

```
1   프랑스의     프랑스      NNP            NNP+JKG              JKG        4    NP-MOD     _   _
2   세계적인     세계적이    NNG+XSN+VCP    NNG+XSN+VCP+ETM      ETM        4    VNP-MOD    _   _
3   의상         의상        NNG            NNG                  _          4    NP         _   _
4   디자이너     디자이너    NNG            NNG                  _          6    NP         _   _
5   엠마누엘     엠마누엘    NNP            NNP                  _          6    NP         _   _
6   웅가로가     웅가로      NNG            NNG+JKS              JKS        11   NP-SBJ     _   _
7   실내         실내        NNG            NNG                  _          8    NP         _   _
8   장식용       장식용      NNG            NNG                  _          9    NP         _   _
9   직물         직물        NNG            NNG                  _          10   NP         _   _
10  디자이너로   디자이너    NNG            NNG+JKB              JKB        11   NP-AJT     _   _
11  나섰다.      나서        VV             VV+EP+EF+SF          EP|EF|SF   0    ROOT       _   _
```

Figure 6: Input and output examples for Korean dependency parsing

column	name	description
3	LEMMA	Lexical morphemes, where functional morphemes are excluded from FORM.
4	CPOSTAG	POS tags for lexical morphemes.
5	POSTAG	Fine-grained part-of-speech tag.
6	FEATS	POS tags for functional morphemes.

Figure 7: Data format of CoNLL-X dependency parsing for Korean

	5-fold	10-fold	80-10-10
parser	84.90	84.83	84.34
gold	85.88	85.75	85.12

Table 1: Baseline phrase structure parsing results

5 Statistical Machine Translation

Actually, statistical machine translation (SMT) for Korean has not been frequently investigated. Previous work on SMT involving Korean often suffers from the lack of openly available bilingual language resources. Lee et al. (2006) used a Korean-English bilingual sentence-aligned corpus which contains 41,566 sentences and 190,418 eojeols. It was manually collected from travel guide books. Xu et al. (2009) used an in-house collection of Korean-English parallel documents. Unfortunately, they did not present the size, or the domain of the corpus. Hong et al. (2009) used about 300K sentences which were collected from the major bilingual news broadcasting sites and randomly selected 5,000 sentence pairs from the Sejong parallel corpus for tuning, development and evaluation. Chung and Gildea (2009) collected the Korean-English parallel data from news websites and used subsets of the parallel corpus consisting of about 2M words and 60K sentences on the English side. Tu et al. (2010) carried out an experiment on Korean-Chinese translation. The training corpus contains about 8.2M Korean words and 7.3M Chinese words. Most of the datasets in previous work are independently collected from various sources and more than anything else they are not currently publicly accessible.

5.1 Tokenization for Korean SMT

For Korean SMT, we tokenize Korean words based on morphological analysis instead of directly using words themselves by which we empirically found that we are able to get the best results for Korean SMT rather than other unsupervised syllable-based tokenization method described in Chung and Gildea (2009). In addition, by tokenizing Korean sentences based on morphological analysis, we can deal with compound words, in which they appear frequently in Korean. Such compounds may be written with or without a blank and they easily lead to the lexicon sparsity problem in SMT. In many cases, compound words become out-of-vocabulary words (OOV) if they do not appear in training data.

5.2 Korean-English parallel data

There are several existing Korean-English parallel data. Sejong parallel data are available directly from National Institute of the Korean Language and News Commentary data are available from the Korean parallel data site[6]. Sejong parallel data are from various sources including novels, government document, and transcribed speech documents. News Commentary data had been crawled from Yahoo! Korea[7] and Joins CNN[8] during 2010-2011. There are also several Korean-English parallel data from OPUS (Tiedemann, 2012)[9,10]. OPUS parallel corpora consist of movie subtitles (OpenSubtitles 2012, 2013, and 2016), technical documents (GNOME, KDE4, and Ubuntu) and religious texts (Tanzil). Since there are alignment errors, we use only some of parallel data from OPUS, in which we judged them to be proper enough to use. For example, PHP data from OPUS, in which the language identification task fails in the corpus, are not utilized. We summarize the brief statistics of currently available parallel corpora in Table 2. Note that the size indicates the number of words of the target language (English).

	size	description
Sejong parallel	0.8M	various
News commentary	2.3M	newswire
OpenSubtitles (OPUS)	3.5M	subtitles
Technical (OPUS)	0.4M	technical
Tanzil (OPUS)	2.8M	religious

Table 2: Previous Korean-English parallel data. These are publicly available.

Actually, there is no standard evaluation data for Korean-English machine translation. Previously existing parallel corpora are mostly automatically created without human intervention and judgment, and there exists inevitable sentence alignment errors. These errors make existing parallel corpora for Korean be difficult to use as standard evaluation data. Moreover, they are not written for translation studies and they might contain translation gaps between source and target languages, which still make them use as proper evaluation data for machine translation. Therefore, we decide to create new evaluation data for Korean-English machine translation (MT). Junior High English evaluation data for Korean-English machine translation (JHE) are the Korean-English parallel corpus which contains sentences from English reading comprehension exercises for Junior high students. We extracted Korean-English sentences from English reference materials and we manually aligned them to build a parallel data. We manage to produce a set of parallel sentences with high precision alignment, for the sake of future evaluation tasks. The average number of words in the sentence is 12 words in Korean, and it contains various topic including news articles, short stories, letters and advertisements. Table 3 describes the statistics of the newly created evaluation data. They are originally written in English (about 60%) and Korean (40%), and they are translated into counterpart languages. Since they are from educational materials, they keep well formal equivalence between source languages and their translation. We believe that JHE data should be suitable to evaluate the correctness and the robustness of MT systems for Korean regardless of their domain.

5.3 Baseline system for SMT

Table 4 shows results on machine translation using existing parallel corpora (Korean into English).

[6] http://site.google.com/koreanparalleldata

[7] https://www.yahoo.co.kr

[8] http://www.joins.com

[9] http://opus.lingfil.uu.se

[10] Accessed on 22 April 2016.

	sentences		words
dev	720	7,608	8,702
eval	720	7,491	8,529
		(Korean)	(English)

Table 3: Junior High English evaluation data for Korean-English machine translation

	internal	JHE
Sejong parallel	1.34	4.48
News commentary	9.12	**7.92**
OpenSubtitles (OPUS)	7.67	6.60
Technical (OPUS)	10.45	0.92
Tanzil (OPUS)	14.95	0.96
News + OpenSubtitles	8.85	8.18

Table 4: Extrinsic evaluation results for the quality of the existing parallel corpora

Internal results presents BLEU scores (an automatic metric for evaluating the quality of machine-translated text) using held-out data from their own corpus (each 1,000 sentences for development and evaluation datasets, respectively). JHE results presents BLEU scores using JHE evaluation data. Bad internal results on the Sejong parallel corpus are understandable because they consist of various sources and held-out data can be a quite different domain from training data. While parallel data of specific domains such as technical and religious can obtain good internal results, it is very difficult to expect to equivalent results on texts of the general domain. We tested all possible combinations with Sejong, News, and OpenSubtitles and only News + OpenSubtitles improves the result. We provide the baseline SMT system using Korean-English News commentary and OpenSubtitles data for future comparison purpose.

6 Summary

In this paper, we present following data, models, tools, and systems for Korean:

- ESPRESSO for sentence segmentation, morphological analysis and POS tagging.

- Berkeley parser models for phrase structure syntactic parsing.

- A pipeline script from ESPRESSO to the Berkeley parser: `MakeBerkeleyTestIn` and `MakeBerkeleyTestWithPOSIn`.

- MaltParser models for dependency analysis.

- A pipeline script from ESPRESSO to Malt-Parser: `MakeMaltTestIn`.

- Baseline Korean-English SMT system using News commentary data and OpenSubtitles.

- Junior High English evaluation data for Korean-English machine translation.

Everyone's Korean language resources described in this paper is available at `https://air.changwon.ac.kr/software/everyone`.

7 Conclusion and Future Perspectives

In this paper, we provided the entire working pipeline for Korean from POS tagging to syntactic analysis. We also described the standard evaluation data and the baseline system for Korean-English statistical machine translation. We hope that these language resources for Korean will pave the way for further development regarding Korean language processing for everybody. For future work, we are planning to distribute other NLP-related systems and models for Korean such as named entity recognition (NER) and semantic role labeling (SRL).

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments.

References

[Bikel2004] Daniel M. Bikel. 2004. *On the Parameter Space of Generative Lexicalized Statistical Parsing Models*. Ph.D. thesis, University of Pennsylvania.

[Cha et al.1998] Jeong-Won Cha, Geunbae Lee, and Jong-Hyeok Lee. 1998. Generalized Unknown Morpheme Guessing for Hybrid POS Tagging of Korean. In *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, Quebec, Canada.

[Choi and Palmer2011] Jinho D. Choi and Martha Palmer. 2011. Statistical Dependency Parsing in Korean: From Corpus Generation To Automatic Parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics.

[Choi et al.2012] DongHyun Choi, Jungyeul Park, and Key-Sun Choi. 2012. Korean Treebank Transformation for Parser Training. In *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, pages 78–88, Jeju, Republic of Korea. Association for Computational Linguistics.

[Chung and Gildea2009] Tagyoung Chung and Daniel Gildea. 2009. Unsupervised Tokenization for Machine Translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 718–726, Singapore. Association for Computational Linguistics.

[Chung et al.2010] Tagyoung Chung, Matt Post, and Daniel Gildea. 2010. Factors Affecting the Accuracy of Korean Parsing. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 49–57, Los Angeles, CA, USA. Association for Computational Linguistics.

[Chung2004] Hoojung Chung. 2004. *Statistical Korean Dependency Parsing Model based on the Surface Contextual Information*. Ph.D. thesis, Korea University.

[Crystal2008] David Crystal. 2008. *Dictionary of Linguistics and Phonetics*. Wiley-Blackwell, the langua edition.

[Hong et al.2009] Gumwon Hong, Seung-Wook Lee, and Hae-Chang Rim. 2009. Bridging Morpho-Syntactic Gap between Source and Target Sentences for English-Korean Statistical Machine Translation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 233–236, Suntec, Singapore. Association for Computational Linguistics.

[Hong2009] Jeen-Pyo Hong. 2009. *Korean Part-Of-Speech Tagger using Eojeol Patterns*. Master's thesis. Changwon National University.

[Kang et al.2007] Mi-Young Kang, Sung-Won Jung, Kyung-Soon Park, and Hyuk-Chul Kwon. 2007. Part-of-Speech Tagging Using Word Probability Based on Category Patterns. *Computational Linguistics and Intelligent Text Processing (Lecture Notes in Computer Science)*, 4394:119–130.

[Klein and Manning2003] Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan. Association for Computational Linguistics.

[Lee and Rim2004] Do-Gil Lee and Hae-Chang Rim. 2004. Part-of-Speech Tagging Considering Surface Form for an Agglutinative Language. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 130–133, Barcelona, Spain. Association for Computational Linguistics.

[Lee et al.2006] Jonghoon Lee, Donghyeon Lee, and Gary Geunbae Lee. 2006. Improving phrase-based Korean-English statistical machine translation. In *Proceeding of: INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing*, Pittsburgh, PA, USA.

[Lee2011] Jae Sung Lee. 2011. Three-Step Probabilistic Model for Korean Morphological Analysis. *Journal of KIISE:Software and Applications*, 38(5):257–268.

[Nivre et al.2006] Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

[Oh and Cha2010] Jin-Young Oh and Jeong-Won Cha. 2010. High Speed Korean Dependency Analysis Using Cascaded Chunking. *Korean Simulation Journal*, 19(1):103–111.

[Oh and Cha2013] Jin-Young Oh and Jeong-Won Cha. 2013. Korean Dependency Parsing using Key Eojoel. *Journal of KIISE:Software and Applications*, 40(10):600–6008.

[Oh et al.2011] Jin Young Oh, Yo-Sub Han, Jungyeul Park, and Jeong-Won Cha. 2011. Predicting Phrase-Level Tags Using Entropy Inspired Discriminative Models. In *International Conference on Information Science and Applications (ICISA) 2011*, pages 1–5.

[Pantel and Pennacchiotti2006] Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 113–120, Sydney, Australia. Association for Computational Linguistics.

[Park et al.2013] Jungyeul Park, Daisuke Kawahara, Sadao Kurohashi, and Key-Sun Choi. 2013. Towards Fully Lexicalized Dependency Parsing for Korean. In *Proceedings of The 13th International Conference on Parsing Technologies (IWPT 2013)*, Nara, Japan.

[Park2006] Jungyeul Park. 2006. *Extraction automatique d'une grammaire d'arbres adjoints à partir d'un corpus arboré pour le coréen*. Ph.D. thesis, Université Paris 7 - Denis Diderot.

[Petrov and Klein2007] Slav Petrov and Dan Klein. 2007. Improved Inference for Unlexicalized Parsing. In *Human Language Technologies 2007: The Conference of* *the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York. Association for Computational Linguistics.

[Petrov et al.2012] Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

[Sarkar and Han2002] Anoop Sarkar and Chung-Hye Han. 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 (TAG+ 6)*, pages 48–56, Venice, Italy.

[Tiedemann2012] Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

[Tu et al.2010] Zhaopeng Tu, Yang Liu, Young-Sook Hwang, Qun Liu, and Shouxun Lin. 2010. Dependency Forest for Statistical Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1092–1100, Beijing, China. Coling 2010 Organizing Committee.

[Xu et al.2009] Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. 2009. Using a Dependency Parser to Improve SMT for Subject-Object-Verb Languages. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 245–253, Boulder, Colorado. Association for Computational Linguistics.