Segmentation Granularity in Dependency Representations for Korean

Jungyeul Park
Department of Linguistics
University of Arizona
jungyeul@email.arizona.edu

Abstract

Previous work on Korean language processing has proposed different basic segmentation units. This paper explores different possible dependency representations for Korean using different levels of segmentation granularity — that is, different schemes for morphological segmentation of tokens into syntactic words. We provide a new Universal Dependencies (UD)-like corpus based on different levels of segmentation granularity for Korean. The corpus contains 67K words in 5,000 sentences which are split into training, development and evaluation data sets. We report parsing results using the new dependency corpus for Korean and compare them with the previous Korean UD corpus.

1 Dependency Parsing and the Korean Language

Language processing including morphological analysis for Korean has traditionally been based on the eojeol, which is a basic segmentation unit delimited by a blank in the sentence. Let us consider the sentence in (1), which contains ten eojeols (the corresponding morphological analysis is found in Figure 1). The number of eojeols is entirely based on the blank space character and the tenth eojeol in (1) also includes the punctuation mark. Almost all natural-language processing systems that have been previously developed for Korean have used the eojeol as a fundamental unit of analysis. As Korean is an agglutinative language, joining content and functional morphemes is very productive and they can be combined exponentially. For example, yeoghal (‘role’) is a content morpheme (a common noun) and -eul, a case marker (‘ACC’, accusative), is a functional morpheme. They form together a single eojeol yeoghal-eul (‘role + ACC’). A predicate gangjo-ha-ass-da (‘focused’) also consists of the content morpheme gangjo-ha (‘focus’) and its functional morphemes, -ass (‘PAST’, past tense) and -da (‘IND’, indicative), respectively.

In this paper, we analyze different levels of segmentation granularity in dependency representations for syntactic annotation (§2). We then propose a scheme to build a new Universal Dependencies (UD)-like corpus for Korean based on segmentation granularity (§3). UD has been developed cross-linguistically using a consistent treebank annotation scheme for many languages. We provide 5,000 sentences based on each of the segmentation granularity possibilities described in this paper. We also present its UD parsing results, compare them with previously proposed UD for Korean (§4), and discuss future perspectives of dependency annotation and parsing for Korean (§5).

2 Segmentation Granularity for Korean

We define the following four different levels of segmentation granularity for Korean. These granularity levels have been independently proposed in previous work on Korean language processing as different basic segmentation units.

2.1 Eojeols

Most language processing systems and corpora developed for Korean have used the eojeol as a fundamental unit of analysis (Figure 2). For example, the Sejong corpus, the most widely-used corpus for Korean, uses the eojeol as the basic unit of analysis as presented in (1). Most morphological analysis systems have been developed based on eojeol segmentation,

---

1 For convenience sake, we add the hyphen-minus (-) at the beginning of functional morphemes, such as -eul to distinguish boundaries between content and functional morphemes. The accusative case marker -eul or -leul vary depending on the preceding character.

2 http://universaldependencies.org
2.2 Separating words and punctuation

As eojeols have been used as a basic analysis unit in Korean corpora, the tokenization task is often ignored for Korean. However, there are corpora which use an English-like tokenization (Figure 3). Words in these corpora are already preprocessed: for example, the Penn Korean treebank (Han et al., 2002), in which punctuation marks are separated from words. Note that among existing corpora for Korean, only the Sejong treebank separates quotation marks from the word. Other Sejong corpora including the morphologically analyzed corpus do not separate the quotation marks. While the Korean Penn treebank separates all punctuation marks, quotation marks are the only symbols that are separated from words in the Sejong treebank. Chung and Gildea (2009) used this granularity of separating words and symbols for a baseline tokenization system for a machine translation system. Park et al. (2014) also used this granularity to develop Korean FrameNet lexicon units.

2.3 Separating case markers

The Sejong corpus has been criticized for the scope of the case marker, in which only a final noun (usually the lexical anchor) in the noun phrase is a modifier of the case marker. For example, Emmanuel Ungaro-ga in the Sejong corpus is annotated as (NP (NP Emmanuel) (NP Ungaro-ga)), in which only Ungaro is a modifier of ‘ga’ (‘NOM’). The Korean Penn treebank does not explicitly represent this phenomenon. It just groups a noun phrase together: e.g. (NP Emmanuel Ungaro-ga). Collins’ preprocessing for parsing the Penn treebank adds intermediate NP terminals for the noun phrase (Collins, 1997; Bikel, 2004), and NPs in the Korean Penn treebank will have a similar NP structure in the Sejong corpus (Chung et al., 2010). To fix the problem in the previous treebank annotation scheme, there are other annotation schemes proposed in the corpus and lexicalized parsing grammars for the purpose to correctly express the scope of the case marker (Figure 4). Park (2006) considered case markers (or postpositions) as independent elements in Tree adjoining grammars (Joshi et al., 1975). Therefore, he defined case markers as an auxiliary tree to be ad-
joined to a noun phrase. For example, the single
token jagga-deul-do becomes two tokens, jagga-
deul (‘author’) and -do (‘also’). However, verbal
endings on the inflected forms of predicates are
still in the eojeol and they are represented as ini-
tial trees for Korean TAG grammars. The lemma
of the predicate and its verbal endings are dealt
with as inflected forms instead of separate func-
tional morphemes.

2.4 Separating verbal endings

Government and binding (GB) theory for Ko-
rean often proposed a syntactic analysis, in which
the entire sentence depends on verbal endings.
For example, gangjo-ha-ass-da becomes gangjo-
ha (‘emphasize’), -ass (‘PAST’), and -da (‘IND’) as described in Figure 3.

The Kaist treebank (Choi et al., 1994), the first
treebank created for Korean adapted this type of
analysis (Figure 6). While the Kaist treebank sep-
arates case markers and verbal endings with their
lexical morphemes, punctuation marks are not
separated and they are still a part of the preceding
token. Therefore, strictly speaking, this granular-
ity level is not exactly same as in the Kaist tree-
bank.

2.5 Discussion

The different levels of segmentation granular-
ity described in this section have been proposed
mainly because of different syntactic analysis in several previously proposed Korean treebank
datasets: Kaist [Choi et al., 1994], Sejong [26] and Penn [Han et al., 2002] treebanks. Even for the segmentation granularity which we deal with, syntactic theory is implicitly presented in the corpus for Korean words. Granularity described in §2.1 and §2.2 is based on the Sejong treebank. Granularity described in §2.3 and §2.4 is based on the Korean Penn treebank and the Kaist treebank, respectively.

Many applications for Korean language processing are based on another level of segmentation granularity, in which all morphemes are separated: phrase-structure parsing [Choi et al., 2012] [Park et al., 2016] and statistical machine translation (SMT) [Park et al., 2016], etc. Such morpheme-

---

**Figure 4:** CoNLL-U format by separating case markers, which requires a dependency relationship between the noun phrase and the case marker (case), for example yeoghal (‘role’) and eul (‘ACC’).

**Figure 5:** X-bar schema for gangjo-ha, -ass, and -da in Korean

---

3 **UD for Korean**

Since Universal Dependencies (UD) has been released [Nivre et al., 2016], several studies have been published, both theoretical [Schuster and Manning, 2016] and practical [Zeman et al., 2017]. As for other morphologically rich languages, specific Universal Dependencies for Japanese were introduced relatively recently to meet the requirement of UD’s cross-linguistically consistent treebank annotation [Tanaka et al., 2016]. In the current UD, other morphologically rich languages such as Kazakh [Tyers and Washington, 2015] and Turkish [Sulubacak et al., 2016] are also available. In this section, we describe how to build UDs for Korean based on the different levels of segmentation granularity.

---

https://www.sejong.or.kr
3.1 Universal POS

Using the eoejol and morpheme level mapping tables to Universal POS tags for the Sejong tagset proposed in Petrov et al. (2012) and Park et al. (2016), we can convert the single tags (morphemes) and the sequences of tags (eoejols and tokens) in the Sejong corpus into Universal POS tags. We also use additional mapping rules by using the approach to find Universal POS tags described in Oh et al. (2011) in which they predict phrase tags for the eoejol. In addition, the Sejong tags (morphemes) and the sequence of tags (tokens and eoejols) represented as immediate non-terminal nodes in the eventual parse tree can be used as a language-specific part-of-speech tag in the CoNLL-U format. Figure 6 shows example mapping rules for each segmentation granularity level. Tagsets in the Sejong corpus are mapped to the Universal POS tag sets either individually (NNP → PROPN) or by a sequence of the POS tags (NNP+JKS → PROPN). Figure 7 represents the 1-to-1 mapping from the POS tags in the Sejong corpus to Universal POS tags described in Park et al. (2016). These 1-to-1 mapping rules are used throughout segmentation granularity schemes described [21] to [24] if the eoejol is composed only by a single morpheme.

3.2 Universal features

Park (2006) detailed an approach to extract features from the Sejong treebank. Syntactic tags and morphological analysis allow us to extract syntactic features automatically and to develop universal features. For example, NP-SBJ syntactic tag is changed into NP and a syntactic feature Case=Nom is added. Syntactic tags which end with -sbj (subject), -obj (object) and -CMP (attribute), we extract Case features which describe argument structures in the sentence. Alongside Case features, we also extract Mood and Tense from the morphological analyses in the Sejong treebank. Since however morphological analyses for verbal and adjectival endings in the Sejong treebank are simply divided into ep (non-final endings), ef (final endings) and ec (conjunctive endings), Mode and Tense features cannot be extracted directly. Park (2006) analyzed 7 non-final endings and 77 final endings used in the Sejong treebank to extract automatically Mood and Tense features. In general, ef carries Mood inflections, and ep carries Tense inflections. Conjunctive endings are not concerned with Mood and Tense features and we only extract ec features with their string value. We also add Hor for the honorific feature, which we can extract from lexical information of non-final endings.
such as -sì.

### 3.3 Universal dependency representations

We use basic dependencies (core, non-core, noun dependents) for eojeols for segmentation granularity in \(2.1\) We add punct between word and punctuation marks (\(2.2\)), and case between noun phrase and case markers (\(2.3\)). We also employ fixed for verbal endings (\(2.4\)). Initial dependency labels are based on phrase information in the Sejong treebank such as np-sub, np-obj, etc. We create conversion rules to conform to Universal Dependency relations.

- nsubj (nominal subject) and csubj (clausal subject) can be assigned in which np-sbj occurs and nouns ended with either jks (nominative marker) or jx (topic marker). We distinguish nsubj and csubj as follows:

  - if a subject noun is a derivational noun from the verb or the adjective, which are usually ended with etn+jks or etn+jx (where etn is a derivational morpheme for the noun), then csubj.
  - otherwise, nsubj.

(2) a. **unggaro-ga ... naseo-eoss-da**

  Ungaro-jx ... become-PAST-IND

  ‘Ungaro became ...’

b. **unggaro-ga naseo-gi-ga ... sib-eoss-da**

  Ungaro-NOM become-etn-jx ... easy-PAST-IND

  ‘Ungaro’s becoming ... was easy’

While the previous UD for Korean uses **nsubj:pass** for the passive construction in Korean, we do not use it for the following two reasons: First, passive and causative verbs are often in the same form if they use passive or causative derivational morphemes such as -i, -hi, etc. and they are very ambiguous. Second, intransitive verbs are also allowed in the passive construction unlike in English.

- obj (direct object) can be assigned in which np-obj occurs and nouns ended with jko (accusative marker). There are several cases where nouns can be ended with jx (topic marker). There are also some cases where nouns can be ended with jx (topic marker) for obj. iobj (second core dependent) can be assigned when np-alt (NP adjunct) occurs and nouns ended with jkb (auxiliary marker) such as -ege, -e, -gge (dative markers).

  (3) ... **sagoa-leul unggaro-ege ja-eoss-da**

  ... apple-jko Ungaro-jkb give-PAST-IND

  ‘... gave an apple to Ungaro’

- ccomp (clausal complement) can be assigned when vp-cmp or vnp-cmp occurs. ccomp normally ends with ec and we identify 71 verbal
Figure 8: POS tags in the Sejong corpus and their 1-to-1 mapping to Universal POS tags.

Sejong POS (S) | description | Universal POS (U)
---|---|---
NNG, NNB, NR, XR | Noun related | NOUN
NNP | Proper noun | PROPN
NP | Pronoun | PRON
MAG. | Adverb | ADV
MAJ | Conjunctive adverb | CCONJ
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, JK, JKV, JQ, JC | Postpositions (case markers) | ADP
XPN, XSN, XSA, XSV | Suffixes | PRT
SF, SP, SE, SO, SS | Punctuation marks | PUNCT
SH, SL | Foreign characters | X
SN | Number | NUM
NA, NF, NV | Unknown words | X

 UD for Korean annotates acl:rel instead of acl to specify a relative clause for the verb ended with etm (verbal/adjectival ending for the relative clause) and it modifies a noun, we assign acl.

• if a verb ends with etm (verbal/adjectival ending for the relative clause) and it modifies a noun, we assign acl.

• if a adj ends with etm and it modifies a noun, we assign amod.

UD for Korean annotates acl:rel instead of acl to specify a relative clause for the verb ended with etm (adjunct) or nmod (nominal dependents) can be assigned where np-ajt or np occurs, respectively. det:poss is assigned for noun ended with jkg (genitive marker). Other UD relations such as advmod, det, etc can be assigned as a 1-to-1 mapping table by using Sejong POS labels as described in Table 1.

| Sejong POS | UD relations |
|---|---|
| mag | advmod |
| maj | cc |
| mm | det |
| sn | nummod |

Table 1: Miscellaneous conversion between Sejong POS labels and UD representations

... unggaro-ga ... naseo-eoss-dago malha-eoss-da ... Ungaro-NOM ... become-PAST-ec tell-PAST-IND

‘... told that Ungaro became...’

acl (adnominal clause) and amod (adjectival modifier) for Korean, in which vp-mod occurs, are defined as follows:

• if a verb ends with etm (verbal/adjectival ending for the relative clause) and it modifies a noun, we assign acl.

• if a adj ends with etm and it modifies a noun, we assign amod.

UD for Korean annotates acl:rel instead of acl to specify a relative clause for the verb ended with etm (adjunct) or nmod (nominal dependents) can be assigned where np-ajt or np occurs, respectively. det:poss is assigned for noun ended with jkg (genitive marker). Other UD relations such as advmod, det, etc can be assigned as a 1-to-1 mapping table by using Sejong POS labels as described in Table 1.

4 Experiments and Results

We collected sentences from news articles in one of Korean News websites published during 2016. We select the length of sentences in which there are words (eojeols) between 10 and 20 and the sentence should end with the final verbal ending such as -da (IND) or -gga (INT) and the punctuation mark such as period or question mark (sf).

We perform initial automatic preprocessing tasks using existing tools for Korean such as POS tagging (Hong, 2009), assigning Universal POS labels (Petrov et al., 2012; Park et al., 2016), and MaltParser-based dependency parsing (Park et al., 2016). We manually correct the initial preprocessing tasks especially focused on dependency relation as described in §3.3. First, we build a corpus as described in §2.1 then convert it into other levels of segmentation granularity as described in from §2.2 to §2.4. As a result, we provide a new UD for Korean which contain 5,000 sentences. We split them into 3K-1K-1K sentences for training, development, and evaluation data sets. Table 2 shows the brief statistics of the new UD for Korean. The number of words indicates the number of eojeols as described in §2.1. We train and evaluate four different dependency segmentations.

Table 2 shows the brief statistics of the new UD for Korean. The number of words indicates the number of eojeols as described in §2.1. We train and evaluate four different dependency segmentations.

Similar criteria for selecting sentences are used for the Kaist and the Penn Korean treebank.

Manual verification was done by two linguists in a month.
Table 2: Statistics of the new UD for Korean. A number based on granularity \( \S 2.1 \)

|        | sentences | words  |
|--------|-----------|--------|
| training | 3,000     | 40,648 |
| dev     | 1,000     | 13,492 |
| eval    | 1,000     | 13,623 |
| total   | 5,000     | 67,763 |

Table 3: POS tagging and parsing results using UDPipe trained with four different UDs for Korean.

|     | eojel | symbol | case | verbal |
|-----|-------|--------|------|--------|
| \S 2.1 | \S 2.2 | \S 2.3 | \S 2.4 |
| UPOS  | 93.04 | 94.13  | 97.12 | 98.31  |
| XPOS  | 82.59 | 85.22  | 90.63 | 95.19  |
| UAS   | 62.08 | 65.72  | 76.19 | 79.59  |
| LAS   | 40.51 | 48.44  | 71.29 | 78.07  |

5 Conclusion

The different levels of segmentation granularity described in this paper are mainly due to different representations of syntactic structure in the various Korean treebank datasets. They have used different word segmentation depending on their linguistic and computational requirements. While a certain segmentation granularity may be well suited for some linguistic phenomena or applications, it does not mean that this granularity is a better representation than the other in general. We need to find the most adequate segmentation granularity to adapt to our requirements for Korean language processing. The UDs corpus for Korean based on different levels of segmentation granularity will be publicly available.

Acknowledgement

We thank Francis Morton Tyers, Loïc Dugast, and the anonymous reviewers for their helpful comments and suggestions.

References

[Bengoetxea and Gojenola2010] Kepa Bengoetxea and Koldo Gojenola. 2010. Application of Different Techniques to Dependency Parsing of Basque. In Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages, pages 31–39, Los Angeles, CA, USA. Association for Computational Linguistics.

[Bikel2004] Daniel M. Bikel. 2004. Intricacies of Collins’ Parsing Model. Computational Linguistics, 30(4):479–511.

[Choi et al.1994] Key-Sun Choi, Young S Han, Young G Han, and Oh W Kwon. 1994. KAIST Tree Bank Project for Korean: Present and Future Development. In Proceedings of the International Workshop on Sharable Natural Language Resources, pages 7–14.

[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.

[Collins1997] Michael Collins. 1997. Three Generative, Lexicalised Models for Statistical Parsing. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pages 16–23, Madrid, Spain. Association for Computational Linguistics.

[Han et al.2002] Chung-Hye Han, Na-Rae Han, Eon-Suk Ko, Heejong Yi, and Martha Palmer. 2002. Penn Korean Treebank: Development and Evaluation. In Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation.

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

[Joshi et al.1975] Aravind K. Joshi, Leon S. Levy, and Masako Takahashi. 1975. Tree Adjunct Grammars. Journal of Computer and System Sciences, 10(1):136–163.

[Nivre and Fang2017] Joakim Nivre and Chiao-Ting Fang. 2017. Universal Dependency Evaluation. In Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pages 86–95, Gothenburg, Sweden. Association for Computational Linguistics.

[Nivre et al.2016] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In Luis von Ahn, editor, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. European Language Resources Association (ELRA).

[Nivre et al.2017] Joakim Nivre, Željko Agić, Lars Ahrenberg, et al. 2017. Universal dependencies 2.0 – CoNLL 2017 shared task development and test data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

[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 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.

[Park et al.2011] Jungyeul Park, Jeong-Won Cha, and Seok Woo Jang. 2011. Korean POS Tagging using Noisy Channel Model with Syllable Lattice Based OOV Words Resolution. Information - an international interdisciplinary journal, 14(8):2835–2843.

[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.

[Park et al.2014] Jungyeul Park, Sejin Nam, Youngsik Kim, Younggyun Hahn, Dosam Hwang, and Key-Sun Choi. 2014. Frame-Semantic Web : a Case Study for Korean. In Proceedings of ISWC 2014 : International Semantic Web Conference 2014 (Posters and Demonstrations Track), pages 257–260.

[Park et al.2016] Jungyeul Park, Jeen-Pyo Hong, and Jeong-Won Cha. 2016. Korean Language Resources for Everyone. In Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation (PACLIC 30), pages 49–58, Seoul, Korea.

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

[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).

[Schuster and Manning2016] Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, 5. European Language Resources Association (ELRA).

[Straka et al.2016] Milan Straka, Jan Hajic, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, 5. European Language Resources Association (ELRA).
[Stratos et al.2016] Karl Stratos, Michael Collins, and Daniel Hsu. 2016. Unsupervised Part-Of-Speech Tagging with Anchor Hidden Markov Models. Transactions of the Association for Computational Linguistics, 4:245–257.

[Sulubacak et al.2016] Umut Sulubacak, Memduh Gökürmak, Francis M. Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal dependencies for Turkish. In Proceedings of COLING 2016.

[Tanaka et al.2016] Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shin’ichirō Morita, and Yuji Matsumoto. 2016. Universal Dependencies for Japanese. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, 5. European Language Resources Association (ELRA).

[Tyers and Washington2015] Francis Morton Tyers and Jonathan North Washington. 2015. Towards a Free/Open-source Universal-dependency Treebank for Kazakh. In Proceedings of the 3rd International Conference on Turkic Languages Processing (TurkLang 2015), pages 276–289.

[Zeman et al.2017] Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökürmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajic jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dina Taji, Nizar Habash, Herman Lelong, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Hans Uszkoreit, Vivien Macketzan, Aljoscha Burkhardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez-Alcalde, Jana Strnadova, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sooyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics.