K-UniMorph: Korean Universal Morphology and its Feature Schema

Eunkyul Leah Jo[1*] **Kyuwon Kim**[1,2*] **Xihan Wu**[1*] **KyungTae Lim**[3]
Jungyeul Park[1] **Chulwoo Park**[4]
[1]The University of British Columbia, Canada [2]Seoul National University, South Korea
[3]SeoulTech & Teddysum, South Korea [4]Anyang University, South Korea
{eunkyul,wuxihan}@student.ubc.ca guwon0406@snu.ac.kr ktlim@seoultech.ac.kr
jungyeul@mail.ubc.ca cwpa@anyang.ac.kr

Abstract

We present in this work a new Universal Morphology dataset for Korean. Previously, the Korean language has been underrepresented in the field of morphological paradigms amongst hundreds of diverse world languages. Hence, we propose this Universal Morphological paradigms for the Korean language that preserve its distinct characteristics. For our K-UniMorph dataset, we outline each grammatical criterion in detail for the verbal endings, clarify how to extract inflected forms, and demonstrate how we generate the morphological schemata. This dataset adopts morphological feature schema from Sylak-Glassman et al. (2015) and Sylak-Glassman (2016) for the Korean language as we extract inflected verb forms from the Sejong morphologically analyzed corpus that is one of the largest annotated corpora for Korean. During the data creation, our methodology also includes investigating the correctness of the conversion from the Sejong corpus. Furthermore, we carry out the inflection task using three different Korean word forms: letters, syllables and morphemes. Finally, we discuss and describe future perspectives on Korean morphological paradigms and the dataset.

1 Introduction

The Universal Morphology (UniMorph) project is a collaborative effort providing broad-coverage morphological paradigms for diverse world languages (McCarthy et al., 2020; Kirov et al., 2018). UniMorph consists of a lemma and bundle of morphological features related to a particular inflected word form as follows, for example:

나서다*naseoda* 나섰다*naseossda* V;DECL;PST

where 나서다*naseoda* is the lemma form and 나섰다*naseossda* ('became') is the inflected form with V;DECL;PST (verb, declarative, and past tense) as morphological schema.

It started in 2016 as a SIGMORPHON shared task (Cotterell et al., 2016) for the problem of morphological reinflection, and it introduced morphological datasets for 10 languages. The inflection task, using the given lemma with its part-of-speech to generate a target inflected form, has been continued through the years: CoNLL–SIGMORPHON 2017 Shared Task (Cotterell et al., 2017), CoNLL–SIGMORPHON 2018 Shared Task (Cotterell et al., 2018), SIGMORPHON 2019 Shared Task (McCarthy et al., 2019), SIGMORPHON 2020 Shared Task (Gorman et al., 2020) and SIGMORPHON 2021 Shared Task (Pimentel et al., 2021). However, the Korean language has not been a part of the shared task because of the lack of the dataset.

Nonetheless, although rarely, morphological paradigms for Korean have been explored in the context of computational linguistics. Yongkyoon (1993) defined the inflectional classes for verbs in Korean using word-and-paradigm (WP) (Hockett, 1954) approaches. His fifteen classes of the verb which can be joined with seven different types of verbal endings, are based on inflected forms of the verb. Seokjoon (1999) systematized the list of final endings and their properties, which are also used as conjunctive endings in Korean. Otherwise, properties of verbs such as mood, tense, voice, evidentiality, interrogativity have been extensively studied in Korean linguistics independently: for example, *inter alia*, tense (Byung-sun, 2003), grammatical voice (Chulwoo, 2007), interaction of tense–aspect–mood marking with modality (Jae Mog, 1998), evidentiality (Donghoon, 2008), and interrogativity (Donghoon, 2011).

In continuation of the efforts, this paper proposes a new Universal Morphology dataset for Korean. We adopt morphological feature schema from Sylak-Glassman et al. (2015) and Sylak-Glassman (2016) for the Korean language and extract inflected verb forms from the Sejong morphologi-

[*]Equally contributed authors.

July 9-14, 2023 ©2023 Association for Computational Linguistics

cally analyzed corpus over 0.6M sentences with 9.5M words. We set the criteria in detail by explaining how to extract inflected verbal forms (Section 2), and carry out the inflection task using different Korean word forms such as letter, syllable and morpheme (Section 3). Finally, we discuss future perspectives on a Korean UniMorph dataset (Section 4).

2 UniMorph Features Schema

Verbal endings in the inflected forms of the predicate has been considered as still being in the part of the word as proposed in several grammar formalisms for Korean such as lexicalized tree adjoining grammars (Park, 2006), head driven phrase structure grammars (Ko, 2010), and combinatory categorial grammars (Kang, 2011) in contrast to government and binding (GB) theory (Chomsky, 1981, 1982) for Korean in which the entire sentence depends on separated verbal endings. This idea goes back to Maurice Gross's lexicon grammars (Gross, 1975), and his students who worked on a descriptive analysis of Korean in which the number of predicates in Korean could be fixed by generating possible inflection forms: *e.g.* Pak (1987); Nho (1992); Nam (1994); Shin (1994); Park (1996); Chung (1998); Han (2000). However, we have separated the postposition from the substantive such as noun phrases instead of keeping themselves together. Therefore, with the current Korean dataset, we decide to annotate morphological data for verbs (V).

Table 1 shows the morphological schema for Korean UniMorph where we adopt features from Sylak-Glassman et al. (2015) and Sylak-Glassman (2016) for the Korean language. In addition to the features schema, we consider following these four different types of verbal endings, in which they convey grammatical meanings for the predicate: sentence final ending (ef), non-final ending (ep), conjunctive ending (ec), and modifier ending (etm).

Evidentiality It is a grammatical category that reflects the source of information that a speaker conveys in a proposition. It is often expressed through morphological markers such as sentence final endings (ef) 대*dae*, 내*nae*, and 래*lae* bring in hearsay (HRSY), and non-final endings (ep) 겠*gess* introduce inferred (INFER). Since the suffix for the quotative (QUOT) is denoted with a postposition (jkq) in Korean instead of the verbal ending, it is excluded from the current set of schemata.

Interrogativity It indicates either to express a statement (DECL) or a question (INT). We consider all sentence final ending (ef) ended with 다*da* as declarative DECL, and sentence final ending (ef) included 가*ga* and 까*kka* as interrogative INT.

Mood The grammatical mood of a verb indicates modality on a verb by the morphological marking. Realis (REAL) and irrealis (IRR) are represented by a verbal modifier ending (also known as an adnominal ending) (etm), ㄴ*n* and ㄹ*l*, respectively. The usage of adnominal endings consists of (i) collocation such as 인한*inhan*, 치면*chimyeon*, 대한 *daehan*, (ii) modifiers and (iii) relative clauses. Realis and irrealis are concerned with regardless of modifiers or relative clauses. General purposive (PURP) is decided by 려고*lyeogo* and 하러*haleo*, and obligative (OBLIG) is introduced by 야*ya*. It is worthwhile to note that we do not consider indicative (IND) because we specify declarative DECL.

Tense It refers to the time frame in which a verb's action or state of being occurs. Non-final endings (ep) such as 았*ass* and 었*eoss* and final endings (ef) such as ㄴ다*nda* 는다*neunda* can represent the past (PAST) and the present (PRS) tenses, repectively. Since the future tense (FUT) has been considered as irrealis (IRR) in Korean, we don't annotate it here.

Voice We deduce the passive (PASS) from the verb stem instead of the verbal ending such as *jabhi* ('be caught'). Whereas the verb *jab* ('catch') and the passive suffix *hi* might be segmented, the current criteria of the Sejong corpus combines them together as a single morpheme. 이히리기*i*, *hi*, *li*, *gi* are verbal endings known for both the passive and the causative. If the verb has a verbal ending 게 *ge* such as verb stem+{이*i*|히*hi*|리*li*|기*gi*}+게*ge* {하*ha*|만들*mandeul* ('make')}, then it is causative (CAUS), otherwise passive (PASS).

Other schema For politeness, we introduce only polite (POL) using the non-final ending (ep) 시*si* as the direct encoding of the speaker-addressee relationship (Brown and Levinson, 1987, p.276). Lastly, since we are not able to deduce the valency of the verb from morphemes, we do not include INTR (intransitive), TR (transitive) and DITR (ditransitive). However, we leave them for future work because the valency might still be valid morphological feature schemata for Korean.

Evidentiality	HRSY	hearsay: 일*il* (‘work’)/NNB 이*i* (‘COP’)/VCP + **래*lae* (‘HRSY’)/EF** (‘happen’)
	INFER	inferred: 괜찮*gwaenchanh* (‘fine’)/VA + **겠*gess* (‘INFER’)/EP** + 다*da* (‘DECL’)/EF
Interrogativity	DECL	declarative: 모이*moi* (‘gather’)/VV + **ㄴ다*nda* (‘DECL’)/EF**
	INT	interrogative: 배우*baeu* (‘study’)/VV + **는가*neunga* (‘INT’)/EF**
Mood	REAL	realis: 얻*eod* (‘get’)/VV + **은*eun* (‘REAL’)/ETM**
	IRR	irrealis: 잊*ij* (‘forget’)/VV + **을*eul* (‘IRR’)/ETM**
	PURP	general purposive: 달래*dallae* (‘appease’)/VV + **려고*lyeogo* (‘PURP’)/EC**
	OBLIG	obligative: 이어지*ieoji* (‘connect’)/VV + **어야*eoya* (‘OBLIG’)/EC** (‘should be connected’)
Tense	PRS	present: 들리*deulli* + (‘hear’)/VV + **ㄴ다*nda* (‘PRS,DECL’)/EF**
	PST	past: 나타나*natana* (‘appear’)/VV + **았*ass* (‘PST’)/EP** + 다*da* (‘DECL’)/EF
Voice	CAUS	causative: 보이*boi* (‘show’)/VV + **게*ge* (‘CAUS’)/EC**
	PASS	passive: **잡히*jabhi* (‘be caught’)/VV** + 었*eoss* (‘PAT’)/EP + 다*da* (‘DECL’)/EF

Table 1: Korean UniMorph schema for verbs: vv for verb, va for adjective, vcp for copula, and nnb for bound noun,

3 Experimental Results

3.1 Data creation

We prepare the data by extracting inflected verb forms from the Sejong morphologically analyzed corpus (sjmorph) over 676,951 sentences with 7,835,239 eojeols (word units separated by space) which represent 9,537,029 tokens. We are using the same training/dev/test data split that Park and Tyers (2019) proposed for Korean part of speech (POS) tagging. However, the current sjmorph doesn't contain POS labels for the eojeol (the word). Instead, it contains the sequence of POS labels for morphemes as follows:

나섰다*naseossda* 나서*naseo*/VV+었*eoss*/EP+다*da*/EF

where it contains only each morpheme's POS label: a verb 나서*naseo* (‘become’), a non-final ending 었*eoss* (‘PST’), and a final ending 다*da* (‘DECL’), and it does not show whether the word 나섰다 *naseossda* (‘became’) is a verb. Previous works (Petrov et al., 2012; Park et al., 2016; Park and Tyers, 2019; Kim and Colineau, 2020) propose a partial mapping table between Sejong POS (and the sequence of Sejong POSs) (XPOS) and Universal POS (UPOS) labels where UPOS represents the grammatical category of the word. However, no study has presented the correctness of their conversion rules. Therefore, we utilize UD_Korean-GSD (McDonald et al., 2013) in Universal Dependencies (Nivre et al., 2016, 2020) that provides Sejong POS(s) and Universal POS labels for each word. Nevertheless, we observed several critical POS annotation errors in UD_Korean-GSD. For this reason, we proceeded to revise GSD's Sejong POS(s) and Universal POS to evaluate our criteria of getting verbs (inflected forms and their lemmas) from sjmorph. This approach involved randomly selecting 300 sentences from the GSD and manually revising their POS labels based on the Sejong POSs. For thorough verification, they were examined by our linguist for over 60 hours over 3 weeks. The main places of error that we noticed were how words for proper nouns were labeled as NOUN even with its XPOS of proper nouns (NNP). They were corrected to the UPOS label of PROPN. Another common place of error was how the dataset recognized and labeled words according to their roles as constituent parts of the sentence they are in, instead of the word's own category. For example, the temporal nouns was usually annotated as ADV instead of NOUN. We changed this mislabeling by acknowledging the word itself, separate from the sentence. Again, the Sejong POS labels were revised based on the criteria of the Sejong corpus. After correcting 738 words for Sejong POS labels and 705 words for Universal POS labels from 300 sentences in the development file, we trained the sequence of Sejong POS labels using semi-supervised learning to predict the Universal POS label for each word. Among 3674 predictions, there were only 332 UPOS prediction errors, and an error scarcely occurs for VERB labels, which we attempted to ex-

	train	dev	test
lemma	41,631	7505	7595
inflected	197,774	19,251	27,846

Table 2: Statistics of Korean UniMorph

	Source	Target
letter (L)	ㄴㅏㅅㅓㄷㅏ	ㄴㅏㅅㅓㅆㄷㅏ
syllable (S)	나서다	나섰다
morpheme (M)	나서다	나서었다
surface form	나서다*naseoda*	나섰다*naseossda*

Table 3: Example of the surface form and its different representation using letters, syllables and morphemes.

	L	S	M
baseline	26.88	27.75	31.29
neural	51.97	49.72	54.26

Table 4: Experimental results (accuracy)

	UniMorph 4.0 Korean	K-UniMorph
Evide.	-	**HRS**, **INFER**
Finit.	FIN, NFIN	-
Inter.	DECL, INT, IMP	DECL, INT
Mood	COND, PURP	**REAL**, **IRR**, PURP, **OBLIG**
Tense	PRS, PST, FUT	PRS, PST
Voice	CAUS	CAUS, **PASS**
Polit.	FORM, INFORM, POL ELEV	POL
Per.	1, 2	-
Num.	PL	-

Table 5: Feature schema comparison between UniMorph 4.0 Korean K-UniMorph.

tract from sjmorph. Therefore, we consider this current error rate for the verb to be negligible. Finally, we extract 244,871 inflected verbal forms for 43,959 lemma types from sjmorph. Then, we remove all duplicated items from train+dev datasets compared to the test dataset. In Table 2 is the brief statistics of the current dataset.

3.2 Morphological reinflection

The goal of the morphological reinflection task creates the generative function of morphological schema to produce the inflected form of the given word. For Korean, we use 나서다*naseoda* and V;DECL;PST to predict 나섰다*naseossda* by using the composition of alphabet letters (L), syllables (S) and morphemes (M) of the word as shown in Table 3. The word is decomposed into the sequence of consonants and vowels by Letter, the sequence of units constructed with two or three letters by syllable, and the sequence of morphological units by morpheme. The conversion from the target form of each representation to the surface form and vice versa are straightforward in technical terms.

For our task, we use the baseline system from The CoNLL–SIGMORPHON 2018 Shared Task (Cotterell et al., 2018).[1] The system uses alignment, span merging and rule extraction to predict the set of all inflected forms of a lexical item (Durrett and DeNero, 2013). We also build a basic neural model using fairseq[2] (Ott et al., 2019) and Transformer (Vaswani et al., 2017). Table 4 shows the experimental results for Korean UniMorph using the three different representation forms. It is notable that the morpheme forms outperform the other surface representation forms such as by letters and syllables of the word. This is because morpheme forms imply lemma forms for both source and target data. While the average number of inflected forms per lemma is 8.285, there are 22 verb lemmas that have more than 400 different inflected forms. The average number of inflected forms per lemma and morphological feature pair is also 5.634, and this makes Korean difficult to predict the inflected form.

3.3 Comparison with UniMorph 4.0 Korean

UniMorph 4.0 (Batsuren et al., 2022) includes a Korean dataset, which provides 2686 lemma and 241,323 inflected forms that are automatically extracted from Wiktionary. It is mainly comprised of adjectives and verbs with totals of 52,387 and 188,821, respectively.[3] Thoroughly, we inspected the verbs in UniMorph 4.0 Korean to compare with K-UniMorph: Among the 152,454 inflected forms of verbs in UniMorph 4.0 Korean, there are only 16,489 forms that appear in 9.5M words of the Sejong corpus, and 135,965 forms (89.18%) that never occur. UniMorph 4.0 Korean annotated all verbs (V) as FIN and all participles (V.CPTP) as NFIN. We can consider adding FIN for all verbs endings with ef (final verbal endings) and NFIN for all verbs ending with etm (adnominal endings, which are utilized for relative clauses, modifiers, and a part of collocations). To inspect this, UniMorph 4.0 Korean provides the imperative-jussive modality IMP which consists of 1;PL and 2, but it seems that Number (PL) occurs only with 1 (Person). While K-UniMorph considers only 시*si* (an honorific for the agent) as POL, UniMorph 4.0 Korean uses ELEV

[1] https://github.com/sigmorphon/conll2018

[2] https://github.com/facebookresearch/fairseq

[3] The counts are short of some numbers because the errors, 92 forms without morphological schema, are excluded.

Core case	NOM	nominative which marks the subject of a verb: 병원*byeongwon* (‘hospital’)/NNG + **이*i* (‘NOM’)/JKS**
	ACC	accusative which marks the object of a verb: 원인*wonin* (‘cause’)/NNG + **을*eul* (‘ACC’)/JKO**
Non-core, non-local case	DAT	dative which marks the indirect object: 국민*gugmin* (‘people’)/NNG + 에게***ege* (‘DAT’)/JKB**
	GEN	genitive which marks the possessor: 사회*sahoe* (‘society’)/NNG + **의*ui* (‘GEN’)/JKG**
	INS	instrumental which marks means by which an action occurred: 대리석*daeliseog* (‘marble’)/NNG + **으로*eulo* (‘INS’)/JKB**
	COM	comitative which marks the accompaniment: 망치*mangchi* (‘hammer’)/NNG + **와*wa* (‘COM’)/JC**
	VOC	vocative which indicate the direct form of address: 달*dal* (‘moon’)/NNG + **아*a* (‘VOC’)/JKV**
Local case	ALL	allative which marks a type of locative grammatical case: 길*gil* (‘road’)/NNG + **로*lo* (‘ALL’)/JKB**
	ABL	ablative which expresses motion away from something: 밑*mit* (‘bottom’)/NNG + **에서부터*eseobuteo* (‘ABL’)/JKB**
Comparison	CMPR	comparative: 예상*yesang* (‘expectation’)/NNG + **보다*boda* (‘CMPR’)/JKB**
Information structure	TOP	topic which is what is being talked about: 사람*salam* (‘people’)/NNG + **은*eun* (‘TOP’)/JX**

Table 6: Korean UniMorph schema for nouns.

for 시*si*, and POL comes from verbal endings 요*yo* and 습니다*seubnida* with either FORM or INFM. However, FORM.ELEV is to elevate the referent. Therefore, it should be with IMP;2|3 and instead, FORM.HUMB can be introduced with IMP;1 for 습니다*seubnida*, and INFM.ELEV|INFN.HUMB for 요*yo*. Hence, K-UniMorph provides a richer feature schema based on linguistics analysis. Table 5 summarises the different usage of the feature schema between UniMorph 4.0 Korean K-UniMorph.

4 Discussion and Future Perspectives

We have dealt with UniMorph schema for verbs, and obtained experimental results for the morphological reinflection task using the different representation forms of the word. Nouns in Korean have been considered by separating postposition from the lemma of the noun instead of keeping themselves together (*e.g.* 프랑스*peulangseu* (‘France’) and 의*ui* (‘GEN’) instead of 프랑스의*peulangseuui*) in several grammar formalisms for Korean. However, in addition to exogenously given interests such as *inflection in context*,[4] recent studies insist the functional morphemes including both verbal endings and postpositions in Korean should be treated as part of a word, with the result that their categories do not require to be assigned individually in a syntactic level (Park and Kim, 2023). Accordingly, it would be more efficient to assign the syntactic categories on the fully inflected lexical word derived by the lexical rule of the morphological processes in the lexicon. Therefore, we will investigate how we adopt features for nouns such as cases including non-core and local cases such as NOM (nominative), ACC (accusative), comparison (CMPR), and information structure TOP (topic) (Table 6). It will also include a typology of jkb (adverbial marker), which raises ambiguities. An adverbial marker can represent ‘dative’ which marks the indirect object, ‘instrumental’ which marks means by which an action occurred, ‘allative’ which marks a type of locative grammatical case, ‘ablative’ which expresses motion away from something, or ‘comparative’ (CMPR, 예상*yesang*. We leave a detailed study on nouns and other grammatical categories for future work. All datasets of K-UniMorph are available at https://github.com/jungyeul/K-UniMorph to reproduce the results.

[4] https://sigmorphon.github.io/sharedtasks/2018/task2/

Acknowledgement

We would like to thank Ekaterina Vylomova and Omer Goldman at the UniMorph project for their help and support. We also wish to thank three anonymous reviewers for providing us with helpful feedback. This research was based upon work partially supported by the *Students as Partners Course Design Grants* through the Office of the Provost & Vice-President Academic at the University of British Columbia to Eunkyul Leah Jo, and by *Basic Science Research Program* through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1F1A1063474) to KyungTae Lim. This research was also supported in part through computational resources and services provided by Advanced Research Computing at the University of British Columbia.

References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahó\vga, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Hwang Byung-sun. 2003. A Study on Interpretation of the Korean Tense. *The Korean Language and Literature*, 79(1):309–346.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Studies in Generative Grammar. Foris Publications, Dordrecht, The Netherlands.

Noam Chomsky. 1982. *Some Concepts and Consequences of the Theory of Government and Binding*. Linguistic Inquiry Monograph 6. The MIT Press, Cambridge, MA.

Park Chulwoo. 2007. The Grammatical Voice in Korean: an Interface Phenomenon between Syntax and Semantics. *Korean Linguistics*, 37(1):207–228.

Min-Chung Chung. 1998. *Les nominalisations d'adjectifs en coréen : constructions nominales à support issda (il y avoir)*. Ph.D. thesis, Université Paris 7 - Denis Diderot, Paris, France.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 Shared Task—Morphological Reinflection. In *Proceedings of the 2016 Meeting of SIGMORPHON*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Lim Donghoon. 2008. The Mood and Modal systems in Korean. *Korean Semantics*, 26(2):211–248.

Lim Donghoon. 2011. Sentence types in Korean. *Journal of Korean Linguistics*, 60(1):323–359.

Greg Durrett and John DeNero. 2013. Supervised Learning of Complete Morphological Paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195, Atlanta, Georgia. Association for Computational Linguistics.

Kyle Gorman, Lucas F. E. Ashby, Aaron Goyzueta, Arya McCarthy, Shijie Wu, and Daniel You. 2020. The SIGMORPHON 2020 Shared Task on Multilingual Grapheme-to-Phoneme Conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 40–50, Online. Association for Computational Linguistics.

Maurice Gross. 1975. *Méthodes en syntaxe*. Hermann.

Sunhae Han. 2000. *Les predicats nominaux en coreen : Constructions a verbe support hata*. Ph.D. thesis, Université Paris 7 - Denis Diderot, Paris, France.

Charles F. Hockett. 1954. Two Models of Grammatical Description. *WORD*, 10(2-3):210–234.

Song Jae Mog. 1998. Semantic functions of the non - terminal suffix - te - in Korean : from a typological perspective. *Journal of Korean Linguistics*, 32(1):135–169.

Juyeon Kang. 2011. *Problèmes morpho-syntaxiques analysés dans un modèle catégoriel étendu : application au coréen et au français avec une réalisation informatique*. Ph.D. thesis, Université Paris IV - Paris-Sorbonne, Paris, France.

Myung Hee Kim and Nathalie Colineau. 2020. An Enhanced Mapping Scheme of the Universal Part-Of-Speech for Korean. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3826–3833, Marseille, France. European Language Resources Association.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1868–1873, Miyazaki, Japan. European Language Resources Association (ELRA).

Kil Soo Ko. 2010. *La syntaxe du syntagme nominal et l'extraction du complément du nom en coréen : description, analyse et comparaison avec le français*. Ph.D. thesis, Université Paris 7 - Denis Diderot, Paris, France.

Arya D McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 Shared Task: Morphological Analysis in Context and Cross-Lingual Transfer for Inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Jee-Sun Nam. 1994. *Classification syntaxique des constructions adjectivales en coréen*. Ph.D. thesis, Université Paris 7 - Denis Diderot, Paris, France.

Yun-Chae Nho. 1992. *Les constructions converses du coréen : études des prédicats nominaux*. Ph.D. thesis, Université Paris 7 - Denis Diderot, Paris, France.

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 *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, page 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics (Demonstrations), pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Hyong-Ik Pak. 1987. *Lexique-grammaire du coréen : construction à verbes datifs*. Ph.D. thesis, Université Paris 7 - Denis Diderot, Paris, France.

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, Paris, France.

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: Oral Papers (PACLIC 30)*, pages 49–58, Seoul, Korea. Pacific Asia Conference on Language, Information and Computation.

Jungyeul Park and Mija Kim. 2023. A role of functional morphemes in Korean categorial grammars. *Korean Linguistics*, 19(1):1–30.

Jungyeul Park and Francis Tyers. 2019. A New Annotation Scheme for the Sejong Part-of-speech Tagged Corpus. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 195–202, Florence, Italy. Association for Computational Linguistics.

Sounnam Park. 1996. *La construction des verbes neutres en coreen*. Ph.D. thesis, Université Paris 7 - Denis Diderot, Paris, France.

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).

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 Shared Task on Morphological Reinflection: Generalization Across Languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Park Seokjoon. 1999. A few notes on the systematization of final endings in modern Korean: Focusing on '-geodeun' and '-lyeogo'. *Journal of Educational Research*, 7(1):225–244.

Kwang-Soon Shin. 1994. *Le verbe support hata en coréen contemporain : morpho-syntaxe et comparaison*. Ph.D. thesis, Université Paris 7 - Denis Diderot, Paris, France.

John Sylak-Glassman. 2016. The Composition and Use of the Universal Morphological Feature Schema (UniMorph Schema). Technical report, Johns Hopkins University, Baltimore, MD, USA.

John Sylak-Glassman, Christo Kirov, Matt Post, Roger Que, and David Yarowsky. 2015. A Universal Feature Schema for Rich Morphological Annotation and Fine-Grained Cross-Lingual Part-of-Speech Tagging. In *Systems and Frameworks for Computational Morphology*, pages 72–93, Cham. Springer International Publishing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

No Yongkyoon. 1993. Deciding on Inflectional Classes in a Word-and-Paradigm Morphology. In *Proceedings of the 5th Annual Conference on Human and Cognitive Language Technology*, pages 405–411, Daejeon, South Korea. Special Interest Group of Human and Cognitive Language Technology.

A Neural Experiment Description

We use the default setting of `fairseq` for the neural experiment for the Table 4 in §3.2 as described in Table 7.

fairseq `fairseq-preprocess`, `fairseq-train` and `fairseq-interactive`.

GPU around 1 hour of GPU has been consumed for the training step for each experiment.

Total runtime It takes about 2 to 3 hours for completing one experiment including all steps (preprocessing, training and evaluation).

Results A single run with a seed number

task	translation
arch	transformer
dropout	0.3
learning rate	0.0001
lr-scheduler	inverse_sqrt
attention-dropout	0.3
activation-dropout	0.3
activation-fn	relu
encoder-embed-dim	256
encoder-ffn-embed-dim	1024
encoder-layers	4
encoder-attention-heads	4
decoder-embed-dim	256
decoder-ffn-embed-dim	1024
decoder-layers	4
decoder-attention-heads	4
optimizer	adam
adam-betas	(0.9, 0.98)
clip-norm	1.0
warmup-updates	4000
label-smoothing	0.1
batch-size	400
max-update	20000

Table 7: Hyperparameter

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
5

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

□ A4. Have you used AI writing assistants when working on this paper?
Not applicable. Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3.2

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We will follow UniMorph's policy for data distribution

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
1

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
3

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3, table 2

C ☑ Did you run computational experiments?

3, Appendix A

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix A

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
3

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Annotator is one of authors

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Because it is CC BY-NC-SA

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.