Automatic analysis of caregiver input and child production

Insight into corpus-based research on child language development in Korean

Gyu-Ho Shin
Palacký University Olomouc, Czech Republic

The present study explores the applicability of Natural Language Processing (NLP) techniques to investigate child corpora in Korean. We employ caregiver input and child production data in the CHILDES database, currently the largest and open-access Korean child corpus data, and apply NLP techniques to the data in two ways: automatic Part-of-Speech tagging by adapting a machine learning algorithm, and (semi-)automatic extraction of constructional patterns expressing a transitive event (active transitive and suffixal passive). As the first empirical report on NLP-assisted analysis of Korean child corpora, this study is expected to reveal its advantages and drawbacks, thereby opening the window to furthering corpus-mediated research on child language development in Korean. Implications of this study's findings will also contribute to research practice regarding developmental studies on Korean through child corpora, ensuring the reproducibility of procedures and results, which is often lacking in previous corpus-based research on child language development in Korean.

Keywords: Natural Language Processing, caregiver input, child production

1. Introduction

A usage-based constructionist approach highlights input, together with domain-general learning capacities, as a nucleus for language development. This approach favours the idea that speakers' actual experience with language affects their cognitive representations of language to a great extent (e.g., Behrens 2009, Tomasello 2003), and that humans' built-in sensitivity to frequency information modulates the degree to which (non-)linguistic resources engage in language development for a person's entire life (e.g., Ambridge et al. 2015, Ellis 2002). Indeed, there is

https://doi.org/10.1075/kl.20002.shi
Korean Linguistics 18:2 (2022), pp. 125–158. ISSN 0257-3784 | E-ISSN 2212-9731

ample evidence for the major contribution of input in shaping the course of language development (e.g., Ellis & Ferreira-Junior 2009, Goldberg et al. 2004, Wonnacott et al. 2012). In this respect, language – a structured inventory of linguistic repertoires through speakers' perceptual experience – emerges and grows by virtue of concrete language use.

The use of corpora to study frequency effects and distributional properties as a proxy for the input that children receive is now common in child language development literature (e.g., Behrens 2006, Cameron-Faulkner et al. 2003, Stoll et al. 2009). To illustrate, Cameron-Faulkner et al. (2003) show that half of the English-speaking caregivers' utterances from the Child Language Data Exchange System (CHILDES) database (MacWhinney 2000) consist of simple, item-based phrases mostly with two words, and that child utterances tend to mimic these phrases in proportion to the caregivers' use of the target phrases. By comparing English (restrictive word order and little morphology) to Russian (flexible word order and rich morphology) and German (in between English and Russian for its morpho-syntactic properties), Stoll et al. (2009) add to the cross-linguistic evidence for the relation between the way that maternal input is structured and the types of child production in the beginning stage of language development.

A few studies on Korean join the literature by showing a close relation between caregiver input and child production (e.g., Cho 1982, Chung 1994) and developmental aspects (e.g., Choi 1999, Lee & Cho 2009). However, implications from the literature seem to be diluted because of a lack of clarity on the quantity of the data analysed and the accessibility of the data. Moreover, data analysis has mostly been done by hand, which makes it demanding to deal with large-scale child corpora in Korean. One promising remedy for these issues is to apply Natural Language Processing (NLP) to corpus analysis: the recent advancement of NLP techniques allows us to handle big data with much less effort and more compatibility with language-specific challenges.

The present study aims to employ currently available NLP techniques to analyse caregiver input and child production data in the CHILDES database, which is the largest and open-access child language dataset in Korean. As the first NLP-assisted investigation of Korean child corpora, instead of giving an in-depth investigation regarding properties of caregiver input and child production in the data, this study focusses primarily on providing a methodological report on challenges and prospects for automatic processing of the data in two ways. One involves the development of a Part-Of-Speech (POS) tagger for the data by adapting a machine learning algorithm. The other involves (semi-)automatic extraction of argument structure constructions expressing a transitive event – active transitives and suffixal passives – from child corpora with revised POS-tagging. Findings of this study will open the window to NLP-assisted corpus research on child language development in Korean, with particular emphasis on possible directions for exploring input-output relations in child language development in Korean. Moreover, descriptions about methodological details will enhance reproducibility of procedures and results, which has been often unstressed in previous corpus-based developmental research on Korean.

2. Research on child corpora in Korean

Korean is a Subject-Object-Verb language with overt case-marking by dedicated markers (1a). These structural cues allow scrambling of pre-verbal arguments if that reordering preserves the original meaning with no ambiguity (1b).

(1) a. Active transitive: canonical
Minsu-ka Yengci-lul an-ass-ta.
Minsu-NOM Yengci-ACC hug-PST-SE
'Minsu hugged Yengci.'
b. Active transitive: scrambled
Yengci-lul Minsu-ka an-ass-ta.
Yengci-ACC Minsu-NOM hug-PST-SE
'Minsu hugged Yengci.'

One core characteristic of Korean is to omit elements in a sentence as long as the omitted information can be inferred properly from the context. The omission of case-marking is less restricted than that of an argument. The optionality of certain case markers such as a nominative case marker *-i/ka*, an accusative case marker *-(l)ul*, and a dative marker *-eykey/hanthey* is observed particularly in colloquial speech (e.g., Sohn 1999).

Corpus-mediated research on Korean child corpora goes back to the 1980s. Cho (1982) offers the first official report on this topic by exploring developmental aspects pertaining to word order and case-marking in Korean. The analysis of spontaneous speech of three children and their mothers that she collected showed a correlation between the mothers' and children's utterances in word order such that SV and OV were the dominant patterns that the two interlocutors employed. She also found an asymmetry involving case-marking: whereas use of the nominative case marker was more than omission of the nominative marker, use of the accusative case marker was less than omission of the accusative marker. The children followed these characteristics such that they generally acquired the nominative case marker earlier than the accusative one. A similar topic was investigated by Chung (1994), who focussed more on erroneous patterns of case-marking by collecting audio-tapes and diary notes from four children and their parents. She reported discrete stages of how the children acquired individual case markers and word order facts, claiming that children in this age group prefer word order over case-marking for the indication of grammatical functions of nominals in a sentence.

A seminal study by Choi (1999) addressed the issue of acquisition of verb-argument constructions for young Korean-speaking children through corpus analysis. She collected data from two children and their mothers through written reports and video recordings of spontaneous interaction. Analysis of the data revealed that the children initially acquired argument structures tied to specific verbs, supporting the verb-island hypothesis (Tomasello 1992). It was also found that, after a short period of this lexically specific stage, the children manifested verb-argument constructions systematically and consistently from around the age of two (e.g., transitive verbs with objects; intransitive verbs with subjects). More crucially, the study showed that characteristics that the children manifested were anchored by the nature of the caregiver input, which highlights the role of child-directed speech that encodes the preferred association between a particular verb and a particular argument structure construction that caregivers favour.

A few more studies further report various aspects of child language development through corpus analysis. For example, Lee (2004) collected data from two children and their mothers and explored how the children employed grammatical morphemes to indicate a subject/topic. Her analysis showed a notable production rate of the nominative case marker and the topic marker when 2-year-old children indicated the subject/topic, with varying degrees of individual differences in the course of acquisition, and suggested an influence of the mothers' utterances on the children's use of the topic marker as a contrast function. Lee and Cho (2009) focussed more on children's production of the subject/topic markers over time. They analysed the pre-existing child corpora from various researchers and showed developmental stages before the age of four as to how these markers emerged based on their functions.

Despite the importance of the previous research, there remain two major concerns regarding research practice. One is that the size of corpora that the researchers investigated was never reported. As Table 1 illustrates, no study indicates the number of analysed utterances. The information about the duration and frequency of data collection was stated, but this does not ensure the representativeness and generalisability of these studies' findings. Calculating the totals for all the subtypes of sentences reported in a study does not help to bypass this issue because the totals do not represent the entire amount of data collected or analysed in the study. We still do not know whether findings of the study are drawn from the majority of the entire data or from only a small portion of the data. This not only weakens the credibility of the study's findings but also makes it difficult to apply informative corpus-internal measurement – which is utilised typically in corpus linguistics – to the reported Korean data.

Table 1. Information about corpora used in previous studies on Korean

	Caregiver	Child / age range	Duration / frequency	Size
Cho (1982)	M	Alicia / 2;2–2;9	1-hour recording / biweekly	–
	M	Paul / 2;7–3;2		
	M	Anne / 2;10–3;5		
Chung (1994)	M & F	Hyuck / 1;0–3;0	occasional video recording until 1;6 biweekly; 0.5-to-0.75-hour audio recording until 2;5 monthly; 0.75-to-1-hour audio recording from 2;6	–
	M	MJ / 1;10–2;9	biweekly; 0.75-hour audio recording	
	NA	SK / 1;11–2;4	diary notes only	
	NA	CK / 1;0–2;4	diary notes only	
Choi (1999)	M M	JS & TN / 1:1–2:5	every three to four weeks; 0.5-hour recording until 1;6 & 1-hour recording from 1;7	–
Lee (2004)	M M	JK & JW / 2;0–2;10	1-hour recording / biweekly	–
Lee & Cho (2009)	M	AL / 2:2–2:9	bi-weekly	–
	MNS	AN / 2:10–3:5	bi-weekly	
	NS	C / 2:0–2:2	weekly	
	M, GM, & N	HS / 1:8–2:11	weekly	
	NA	JK / 0:1–3:0	weekly and bi-weekly	
	NS	CK / 1:3–3:11	every day in principle	
	M	JW / 2:0–3:3	bi-weekly	
	NS	PL / 2:7–3:2	bi-weekly	
		Y / 1:3–3:11	every day in principle	
			Hour of recording not reported in all cases	

Note. F = father; GM = grandmother; M = mother; N = nanny; NA = not applicable.

The other concern is about the nature of the data: all the corpora used in these studies are privately possessed and thus not easily available to other scholars. Researchers mostly use their own collection for their work, or they request corpora from other researchers who already hold private ones. This characteristic circumscribes the reproducibility of the previous findings to a great extent.

With these concerns in mind, the following sections provide detailed methodological reports on how NLP-assisted analysis of child corpora was conducted in this study, by utilising the Korean dataset from the CHILDES database.[1] This dataset is currently the largest open-access child corpus data in Korean, which consists of 81,593 sentences (320,068 eojeols)[2] from nine caregivers and 38,388 sentences (70,928 eojeols) from four children whose ages range from 1;3 to 3;10 (Table 2).

Table 2. Information about Korean child corpora in the CHILDES database

Name of corpus	Caregiver	Child / age range	Time of collection (year)	Quantity (lines)	
				Caregiver	Child
Jiwon	M & F	Jiwon / 2;0–2;3	1992	10,602	6,443
Ryu	GM, GF, & M	Jong / 1;3–3;5	2009–2011	28,657	13,698
	GM, M, & F	Joo / 1;9–3;10	2010–2011	27,071	11,730
	M	Yun / 2;3–3;9	2009–2010	15,263	6,517

Note. F = father; GM = grandmother; GF: grandfather; M = mother.

The currently available open-to-public pipelines[3] from tokenisation to dependency parsing (e.g., *UDPipe*: Straka & Straková 2017; *StanfordNLP*: Qi et al. 2018) are not promising for this study since they are based mostly on written genres. Taking dependency relations into account seems to be unrealistic given the characteristics of the corpora under investigation (e.g., partial/incomplete utterances, repetition of onomatopoeia and mimetic words). Furthermore, CLAN, a default program provided by CHILDES for data analysis and editing, is not supported for Korean. Our analysis thus started from the creation of a Part-of-Speech (POS) tagger covering XPOS (a language-specific POS tag set; the Sejong tag[4] proposed by Kim et al. 2007) and UPOS (the universal POS tag set; Petrov et al. 2012) suit-

---

[1] https://childes.talkbank.org/browser/index.php?url=EastAsian/Korean/

[2] An eojeol is defined as a unit with white space on both sides that serves as the minimal unit of sentential components (Lee 2011). It therefore corresponds roughly to what we call a (tokenised) word in English.

[3] A pipeline (in Natural Language Processing) is defined as a series of steps where the output of one step feeds to the input of the next step. Normally, the pipeline is composed of a tokeniser, a tagger, a parser, and other specific functions required for data processing.

[4] This tag set is particularly influential in Korean. The system has 45 labels under seven categories, and employs relatively detailed classification for the postpositions and dependency-related items by function, which reflects linguistic characteristics of Korean. The basic unit of POS tagging in this system is a morpheme within an eojeol.

able for Korean child corpora, and proceeded to pattern-finding through a series of Python programming. Processing the caregiver input was the primary interest because this comprised the majority of the entire dataset.

3. Towards automatic processing of child corpora: POS tagging

3.1 Issues with POS tagging in Korean

There are three major issues regarding the tagging of XPOS and UPOS for Korean corpora. First, white-space tokenisation is not always effective in detecting an appropriate range for tagging. For example, in English, most words separated by a space provide a good estimate to assign proper tags (2).

(2) I love you. → I/PRP love/VBP you/PRP
Note. PRP = proper noun; VBP = verb, general (from the Penn Treebank tag set)

In Korean, however, an eojeol is decomposable into a sequence of morphemes. This property requires an additional breakdown of a syllable involving the physical detachment of a letter from the syllable (3) or resyllabification (4).

(3) Physical detachment: -*n* 'present tense'
Mia-ka Kenwu-lul cha-**n**-ta.
Mia-NOM Kenwu-ACC kick-PRS-SE
'Mia kicks Kenwu.' → Mia/NNP-ka/JKS Kenwu/NNP-lul/JKO cha/VV-**n**/EP-ta/EF[5]

(4) Resyllabification: *mwu-* ~ *mwul-* 'to bite'
tali-lul **mwu**-n-ta.
leg-ACC bite-PRS-SE
'(I) bite the leg.' → tali/NNG-lul/JKO **mwul**/VV-n/EP-ta/EF

More problematic is a seemingly identical form can be segmented differently: *cal* is either a single morpheme (5a) or a combination of *ca* and *l* (5b).

(5) a. *cal* as a single morpheme
na-nun cal mek-nun-ta.
I-TOP well eat-PRS-SE
'I eat well.'

---

[5] See Appendix for the full list of XPOS/UPOS tag sets.

b. *cal* as a combination of morphemes
na-nun ca-l saynkgak-i-ta.
I-TOP sleep-REL thought-be-SE
'I am thinking of sleeping.'

Second, homonymy involving a morpheme complicates the tagging process. To illustrate, the same morpheme *un* can be a lexical morpheme (6a) or attached to a noun as a topic marker (6b), requiring different tags for each instance.

(6) a. *un* as a lexical morpheme
uncangsik
silver.decoration
'A silver decoration'
b. *un* as a topic marker
kaul-un nalssi-ka coh-ta.
autumn-TOP weather-NOM good-SE.
'As for Autumn, the weather is good.

Third, the relation between XPOS and UPOS is not always straightforward in Korean. One good example of this case is found in the same sequence of XPOS tags which must bear different UPOS tags (7a–b).

(7) a. noun-verb as a verb
kongpu-ha-ta
study-do-SE
'to study'
→ kongpu/NNG-ha/VV-ta/EF → VERB
b. noun-verb as an adjective
sinsen-ha-ta
freshness-do-SE
'to be fresh'
→ sinsen/NNG-ha/VV-ta/EF → ADJ

Park, Hong and Cha (2016) suggested one-to-one conversion relationships between the Sejong tags and the Universal tags. However, as the authors admit, its application is limited to the case where an eojeol consists of a single morpheme. For a multi-morphemic eojeol, it is difficult to assume this way of one-to-one relations between XPOS and UPOS.

One way to handle these issues is to incorporate the combinatorial properties of an eojeol into the tagging process. This can be achieved by attaching indices that represent relative positions of morphemes within one eojeol to these morphemes. For example, the decision of the proper XPOS tags may be improved by differentiating *cal* in (5a) from *ca-l* in (5b) with indexing information within the same eojeol, which yields *cal*_0 and *ca*_0+*l*_1, respectively. The distinction of *un* in (6a–b) may also benefit from the same approach. *un* in (6a) can appear either in the eojeol-initial, eojeol-medial, or eojeol-final position, but the same morpheme in (6b) can occur only in the eojeol-final position. If the tagger detects *un*_0 at the eojeol-initial position, it is impossible for the morpheme to be assigned to a topic marker (JX) tag. In contrast, if the tagger detects *un*_2 at the eojeol-final position, it is highly probable that this morpheme is a topic marker, thus assigning it to JX. In other words, information about the relative positions of morphemes within one eojeol, represented as numeric values, may promote better discrimination of each XPOS tag. The issue with the determination of UPOS tags may also be alleviated with such treatment, by enhancing feature sets through information about morphemes and their corresponding XPOS tags with indices altogether. Our testing of the possibilities raised here is discussed in the following sections.

3.2 Developing a POS tagger for Korean child corpora[6]

3.2.1 *Pre-processing*

There is no previous work on automatic processing of Korean caregiver input, so the first task for developing a POS tagger was to create a dataset for training and evaluating the tagger. For this purpose, the raw child-directed speech data (with typos and spacing errors corrected) were entered into the existing Pythonic pipeline for general-purpose corpora – *UDPipe* (Straka & Straková 2017) – from tokenisation up to XPOS/UPOS tagging. After exploring the processed data, grave problems were found such as improper tokenisation (8a), mis-tagging (8b), a nonsensical relation between XPOS and UPOS (8c), and inconsistency in tagging (8d).

(8) a. Improper tokenisation
있었어 → 있었+어 VERB VV+EF
issesse → issess+e
(*issesse* should be *iss+ess+e* ‘exist+PST+SE’)
b. Mis-tagging
아빠네 → 아빠네 NUM MM
appaney → appaney
(*appaney* should be *appa+ney* ‘father+SE’, which should also obtain VERB (UPOS) and NNG+EF (XPOS))

---

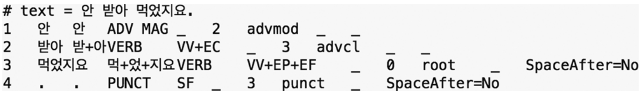
[6] See the github page for the Python code used for the POS tagger.

c. Nonsensical relation between XPOS and UPOS
배운단다 → 배운단다 ADJ VV+EF
paywuntanta → paywu-n-tanta ('learn-PRS-SE')
(Apart from the problem of tokenisation, the combination of VV+EF should be VERB, not ADJ)

d. Inconsistency in tagging
기침 → 기침 NOUN NNG
VERB NNG+NNG
kichim → kichim ('cough')
(The same word returned the two different XPOS-UPOS pairs)

Because the performance of the existing pipeline was not satisfactory for the POS tagging, all of the tagged data were revised manually in order to ensure that each morpheme and word was assigned to an appropriate tag. During this revision, we particularly focussed on correcting tokenisation and tag information about case-marking and verbal morphology, which are often mis-analysed in the currently available pipelines for Korean. Utterances whose length was less than 5 strings (e.g., 까꿍 까꿍) were excluded at this stage, and this resulted in 69,498 sentences (285,350 eojeols) for the actual analysis, which occupied 85.18% of the entire dataset.

As the format of the revised data did not fit the intention of this study, additional formatting adjustments were made in the following ways. First, in a pair of the raw sentence (starting from '*# text =* ') and its corresponding tagged sentence (Figure 1), the raw sentence line was excluded as it was not informative for the purpose of developing the tagger.



```
# text = 안 받아 먹었지요.
1   안   안   ADV MAG _   2   advmod  _   _
2   받아 받+아VERB    VV+EC   _   3   advcl   _   _
3   먹었지요   먹+었+지요VERB   VV+EP+EF    _   0   root    _   SpaceAfter=No
4   .   .   PUNCT   SF  _   3   punct   _   SpaceAfter=No
```

Figure 1. Data format (before adjustment)

Next, individual morphemes and the corresponding tags carried their relative locations within one eojeol with indices attached sequentially from the leftmost morpheme. This treatment aimed to handle the varying structures of an eojeol contingent upon morpheme combinations and to manage the issue of homonymy involving an individual morpheme (see Section 3.1). The final data structure for training the tagger is schematised in (9).

(9) Data structure for training POS tagger
[[(mm xx u), …] …]
→ [[(m_0, x), (m_1, x), …] …] (for XPOS tagging)

→ [[(m_0+m_1/x_0+x_1, u), …] …] (for UPOS tagging)
Note. m = morpheme; u = UPOS tag; x = XPOS tag; _# = index

3.2.2 *Machine learning algorithm for POS tagging: Perceptron*

As a machine learning algorithm for the tagger, this study employed the perceptron. This is one type of hypothetical model in the brain, operating probabilistically in storing and organising information: exposure to a large number of stimuli creates biases for or against a certain response, which modulates the strength of connections between input and output currently in progress (Rosenblatt 1958). The perceptron learning algorithm continuously updates the degree of biases (i.e., weight) through facilitatory or inhibitory forces from the prior predictions (Ghosh et al. 2008). During the learning phase, instead of referring to the entire data, the algorithm looks at only one instance at a time such that updating the weight for features occurs online in a sequential manner (e.g., Daumé III 2015).

Of the various types of perceptron learning algorithms (e.g., Collins & Duffy 2002, Freund & Schapire 1999), the current study employed the averaged perceptron for the tagging. This algorithm utilises an averaged value of a collection of weights attested previously in order to make the current weight more informative for future judgment of features (e.g., Daumé III 2015), not letting it dissipate quickly. This modification ensures a longer life expectancy of individual weights for features, which yields better performance than the strict version of the perceptron learning algorithm. Evidence supports the effectiveness of the averaged perceptron model on various NLP works (e.g., Honnibal et al. 2013, Honnibal & Johnson 2015).

We utilised an open-source averaged perceptron tagger from Honnibal (2013)[7] for the POS tagging of Korean child corpora. Whilst the tagger in Honnibal (2013) was originally designed for English, we explored whether and how this classic version of a neural network algorithm extends to tagging non-English language. The entire set of code was optimised in two ways.

First, the range of reference to the existing tags were expanded up to two items forwards (by adding *post* and *post2* values). When the tagger accrues information for the determination of the current POS tag, the original algorithm allowed the tagger only to consider the previous one or two tags (through *prev* and *prev2* values). The same approach may not be effective in Korean. As exemplified in (4a–b) and (5a–b), the previous tags of *ca(l)* and *un* may not provide crucial information about the decision of the current tag for the morpheme of interest. By allowing the tagger to refer to a wider linguistic environment for the current morpheme both

---

[7] https://explosion.ai/blog/part-of-speech-pos-tagger-in-python

backwards and forwards, the performance of the tagger was enhanced for the precise identification of the POS tags.

Second, on top of an individual morpheme, the combination of that morpheme and the index of the morpheme (for the XPOS tagging) or that of the XPOS tag (for the UPOS tagging) was incorporated into the process of obtaining features for the determination of POS tags. Including information about positional specifications were intended to alleviate the aforementioned language-specific challenges. In this way, the performance of the tagger could be ameliorated such that its operation accommodates the linguistic nature of Korean.

3.2.3 *Model performance*

The final data was split into the training (90% of the entire data) and the test (10% of the entire data) sets. After the tagger was trained, it tagged the untagged test set, and the newly tagged test set was compared with the original test set in view of POS tagging. The performance of the tagger was calculated as an F1 score, the harmonic mean of precision (i.e., the ratio of relevant instances amongst the retrieved instances) and recall (i.e., the ratio of relevant instances retrieved over the total number of relevant instances).

3.3 Results and discussion

Table 3 presents the performance of the POS tagger in terms of the accuracy levels of the XPOS tagging. We obtained an accuracy of 0.95 for the test set. Given the characteristics of child corpora (e.g., partial/incomplete utterances, repetition of onomatopoeia and mimetic words), the performance of our model was comparable to that reported in previous studies on POS tagging for general-purpose Korean corpora (e.g., 0.95 for Park 2017; 0.96 for Hong 2009). The low accuracy in JKC (complementative marker, F1 score: 0.29) was ascribable to its formal similarity to JKS (nominative case marker), which makes it difficult to distinguish between them on the basis of morphology. The accuracy in NNP (proper noun, F1 score: 0.69) may be due to a large degree of productivity, thus being less predictable than the other tags. Other than that, the rates of accuracy below 0.9 (JKV and SN) were ignorable since these tags were periphery for the purpose of the current study.

Table 4 shows the performance of the POS tagger in terms of the accuracy levels of the UPOS tagging. We obtained an accuracy of 0.99 for the test set. This accuracy rate outnumbered previous reports on the model performance for this task (e.g., 0.94 for Straka & Straková 2017; 0.96 for Qi et al. 2018). Of the individual UPOS tags' performance, ADP (adposition) showed a relatively lower accuracy rate (F1 score: 0.7). There is no clear reason for this at present; however, there were only 13 instances involving ADP, which is ignorable.

Table 3. Accuracy of XPOS tagging

Tag	Precision	Recall	F1 score	Tag	Precision	Recall	F1 score
NNG	0.89	0.97	0.93	EC	0.95	0.93	0.94
NNP	0.94	0.55	0.69	ETN	0.94	0.96	0.95
NNB	0.93	0.91	0.92	ETM	0.98	0.97	0.97
NR	0.96	0.84	0.9	XPN	0	0	0
NP	0.95	0.96	0.96	XSN	0.98	0.97	0.97
JKS	0.98	1	0.99	XSV	0.95	0.96	0.96
JKC	0.75	0.18	0.29	XSA	0.98	0.98	0.98
JKG	1	0.99	0.99	XR	0	0	0
JKO	0.99	0.99	0.99	MAG	0.96	0.92	0.94
JKB	0.98	0.97	0.98	MAJ	0.97	0.93	0.95
JKV	0.99	0.81	0.89	IC	0.96	0.97	0.97
JKQ	0.9	0.89	0.9	SF	0	0	0
JC	0.9	0.9	0.9	SE	0	0	0
JX	0.96	0.98	0.97	SS	0	0	0
VV	0.95	0.96	0.95	SP	0	0	0
VX	0.92	0.89	0.9	SO	0	0	0
VCP	0.99	0.94	0.96	SW	0	0	0
VCN	0.93	0.96	0.94	SH	0	0	0
VA	0.97	0.93	0.95	SL	0	0	0
MM	0.97	0.97	0.97	SN	0.83	0.91	0.87
EP	0.99	0.98	0.99	NF	0	0	0
EF	0.95	0.96	0.96	NV	0	0	0

Note. ‘0’ refers to no case for calculation, not a zero rate of accuracy.

Table 4. Accuracy of UPOS tagging

	Precision	Recall	F1 score
NOUN	0.99	0.99	0.99
PRON	0.99	0.98	0.98
VERB	0.99	0.99	0.99
ADJ	0.97	0.98	0.98
ADV	1	0.98	0.99
DET	0.99	0.99	0.99
NUM	0.99	0.98	0.99
CCONJ	0.98	0.99	0.99
ADP	1	0.54	0.7

In summation, despite the relatively low accuracy of tagging performance for several XPOS and UPOS tags, the model in the present study excelled in the POS tagging for Korean child corpora. The success of this tagger suggests that the perceptron tagger, with enhanced input through indexing information, works effectively for the tag-decision process even in Korean child corpora. The results lend support to the effectiveness of NLP techniques on analysis of large-scale child corpora in Korean, alleviating language-specific issues for this task as well. The fact that Perceptron, despite its being an early neural-network-based machine learning algorithm, worked for this task at a satisfactory level further leads to considering cutting-edge machine learning algorithms such as *tensorflow* (https://www.tensorflow.org/) in developing automatic tools for Korean child corpora.

4. Towards automatic processing of child corpora: Construction identification[8]

4.1 Challenges in automatic processing of active transitives and suffixal passives in Korean

This study, particularly in the task of pattern-finding, narrows the scope of investigation into two contrastive types of construction for expressing a transitive event (active transitives and suffixal passives).[9] A canonical active transitive construction (1a), re-stated in (10a), typically occurs with the NOM-marked actor,[10] followed by the ACC-marked undergoer. The verb carries no dedicated active morphology *per se*. A canonical suffixal passive construction (11a) occurs with the NOM-marked undergoer, followed by the DAT[11]-marked actor. The verb carries

---

[8] This section was adapted from Shin (2020). See Shin (2020) for the detailed report on the construction-identification results and their implications on first language development, along with comprehension performance measured through behavioural experiments.

[9] Of the three types of passive constructions in Korean – suffixal, lexical, and periphrastic (Sohn 1999, Song & Choe 2007, but Yeon 2015), the suffixal passive is found to be the most frequent and thus representative type that Korean-speaking children are most likely to encounter (Shin 2020). See Shin (2020) for the first empirical report on the frequency of the individual passive types in caregiver input.

[10] Actor ≈ Agent; Undergoer ≈ Theme. The notation convention for thematic roles follows the tradition of the usage-based constructionist approach.

[11] This marker has variants in their forms and the environment where it occurs: *-eykey* in written and formal contexts, *-hanthey* in spoken and casual contexts, *-kkey* for an honorific recipient, and *-tele/poko* only for 'telling' verbs in colloquial settings (Choo & Kwak 2008).

dedicated passive morphology as one of the four passive suffixes: *-i-*, *-hi-*, *-li-*, and *-ki-*. The two patterns can be scrambled as in (10b) and (11b).

(10) a. Active transitive: canonical
Minsu-ka Yengci-lul an-ass-ta.
Minsu-NOM Yengci-ACC hug-PST-SE
'Minsu hugged Yengci.'
b. Active transitive: scrambled
Yengci-lul Minsu-ka an-ass-ta.
Yengci-ACC Minsu-NOM hug-PST-SE
'Minsu hugged Yengci.'

(11) a. Suffixal passive: canonical
Yengci-ka Minsu-eykey an-ki-ess-ta.
Yengci-NOM Minsu-DAT hug-PSV-PST-SE
'Yengci was hugged by Minsu.'
b. Suffixal passive: scrambled
Minsu-eykey Yengci-ka an-ki-ess-ta.
Minsu-DAT Yengci-NOM hug-PSV-PST-SE
'Yengci was hugged by Minsu.'

Previous reports on Korean-speaking children's acquisition of the two constructions are summarised in several trends. One is that children acquire and use active transitives earlier and more reliably than suffixal passives (e.g., Kim et al. 2017, Shin 2020). This converges upon the attested challenge that the passive poses across languages (e.g., Abbot-Smith et al. 2017, Huang et al. 2013). Another trend is that, within active transitives, the canonical pattern is interpreted more reliably than its scrambled counterpart (e.g., Jin et al. 2015, Kim et al. 2017, Shin 2020). This is due to children's predisposition that interprets the initial argument as the actor, regardless of its canonicity (e.g., Kim et al. 1995, No 2009), which aligns with the oft-mentioned agent-first strategy (cf. Abbot-smith et al. 2017, Sinclair & Bronckart 1972, Slobin & Bever 1982). The other trend is that children, although imperfect, start to reliably apply knowledge about passive morphology to their comprehension of suffixal passives from the age of five or six (Kim 2010, Kim et al. 2017, Shin 2020).

There are three major challenges in the automatic processing of Korean corpora with respect to active transitives and suffixal passives. First, identification of these constructions is tricky since core elements for the constructions such as case-marking and verbal morphology are sometimes mis-tagged and/or ignored in the currently available Korean corpora. To illustrate, open-to-public pipelines do not distinguish clearly between the nominative case marker *-i* and a suffix *-i* which appears after a consonant (e.g., *Caykyeng-i* is often analysed as a combination of a proper noun and the nominative case marker, but *-i* in this case is not the case marker but the suffix only for phonological considerations). They are also poor at recognising verbal morphology, largely due to imperfect tokenisation from the outset (e.g., *ssuye* 'to be used' is tokenised as *ssui-e*, not *ssu-i-e*, and this results in tagging the verb *ssu-* and the passive morphology *-i* altogether as one single verb, ignoring information about passive morphology). Indeed, similar pitfalls are observed in the Sejong corpus, which is a popular open-access dataset in Korean and is widely used as a mother corpus for the development of Korean NLP tools. To overcome this, we enhance the child corpora used in this study with regard to proper tokenisation and tagging of case-marking and verbal morphology to better capture the constructional patterns of interest.

The next challenge pertains to the determination of canonicity involving these constructions. One way to meet this challenge is to utilise information about relative positions of individual markers within a sentence. Given the assumption that sentence composition in child corpora is mostly simple (i.e., mono-clause), we may determine the canonicity of a sentence by way of comparing the numeric location of an initial marker to that of a non-initial one. In a Python environment, a text is treated as a sequence of characters (i.e., strings) numbered sequentially from the left end. As an illustration, the text *hello* consists of five strings in the Python environment, 0 being assigned to *h* and 4 to *o*. Strings can then be searched and compared on the basis of these reference numbers. This characteristic allows us to determine the canonicity of a sentence by extracting information about the relative locations of each marker (expressed as the strings' reference numbers) as long as the sentence has dedicated markers at the designated place. For instance, in the pattern *noun-DAT noun-NOM verb-psv*, the DAT has smaller reference numbers than the NOM, which indicates that the DAT occurs earlier than the NOM. The pattern finder thus classifies this pattern as the scrambled suffixal passive. If one of the markers is omitted, we can still use information about the relative positions of the other marker and the case-less noun. Take the pattern *noun-~~NOM~~ noun-ACC verb* as an example: the ACC occurs after any noun, and this characteristic allows the ACC to have larger numeric values than any noun has, which allows this pattern to be classified as the canonical active transitive. There are very few cases in naturalistic conversation where two markers are dropped in the two constructions (e.g., Chung 1994), so we do not consider this possibility as of yet.

A further challenge, omission of sentential components (e.g., argument, case-marking), is a major difficulty in automatic processing of Korean corpora in general. Several methodological proposals have been made such as consideration of dependency relations (e.g., Choi & Palmer 2011), application of case frames (e.g., Kim & Ock 2015), and development of a dictionary with information about the argument structure of particular verbs (e.g., Lee & Choi 2013). However, these studies – all of which targeted general-purpose corpora – have varying accuracy rates (from around 0.7 to 0.95), and most importantly, there is no empirical report on the application of these proposals to child corpora in Korean. In the present study, rather than developing a new system dedicated only to this task, we find target patterns involving omission of constructional components in a semi-automatic way, first sorting out possible candidates automatically and then extracting the precise instances of these patterns manually.

4.2 Construction identification: Caregiver input[12]

By employing the same dataset used in the development of the POS tagger (69,498 sentences; 285,350 eojeols), the pattern-finding task was conducted with a focus on the four construction types: active transitives (10a–b) and suffixal passives (11a–b) with canonical and non-canonical word order. We also investigated cases involving omission of required arguments and/or markers for each pattern. In addition, we examined the use of individual markers (NOM, ACC, and DAT) with respect to active transitives and suffixal passives.

The tagged data were subjected to a pattern-finding process. All the information about individual morphemes and their corresponding tags in one sentence was transformed into a sequence of strings in an eojeol-by-eojeol basis as in (12).

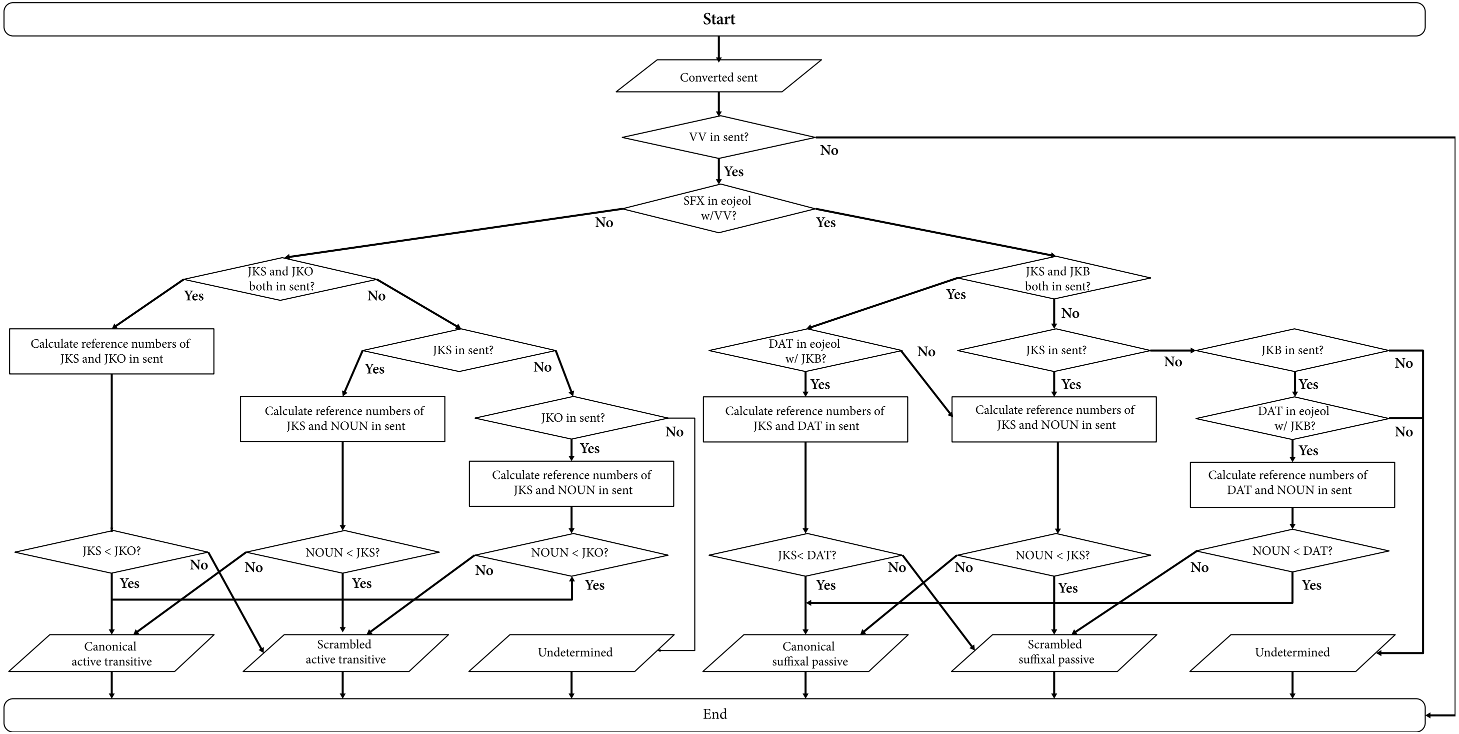
(12) Example of a sentence for pattern-finding
안/안/MAG/ADV 받아/받+아/VV+EC/VERB 먹었지요/먹+었+지요/
VV+EP+EF ./././SF/PUNCT
Note. One eojeol string consists of an eojeol, a sequence of morphemes, XPOS tags corresponding to each morpheme, and a UPOS tag corresponding to the entire eojeol.

The transformed sentences were inputted to an automatic search process whereby the two construction types by canonicity and patterns relating to these constructions were extracted, as schematised in Figure 2.

To illustrate, the canonical active transitive was searched through the following steps: sorting out sentences with a verb; extracting sentences with both JKS (for the NOM) and JKO (for the ACC); and outputting sentences where JKS precedes JKO. Every list of sentences for each extraction was checked manually to ensure the accuracy of the results. Patterns in which main verbs appeared sentence-initially or sentence-medially were excluded at this stage.

---

[12] See the github page for the entire Python code used for the construction identificaation process.


Start
Converted sent
VV in sent?
No
Yes
SFX in eojeol w/VV?
No
Yes
JKS and JKO both in sent?
Yes
No
Calculate reference numbers of JKS and JKO in sent
JKS in sent?
Yes
No
Calculate reference numbers of JKS and NOUN in sent
JKO in sent?
No
Yes
Calculate reference numbers of JKS and NOUN in sent
JKS and JKB both in sent?
Yes
No
DAT in eojeol w/ JKB?
No
Yes
Calculate reference numbers of JKS and DAT in sent
JKS in sent?
Yes
No
Calculate reference numbers of JKS and NOUN in sent
JKB in sent?
No
Yes
DAT in eojeol w/ JKB?
No
Yes
Calculate reference numbers of DAT and NOUN in sent
JKS < JKO?
No
Yes
NOUN < JKS?
No
Yes
NOUN < JKO?
No
Yes
JKS< DAT?
No
Yes
NOUN < JKS?
No
Yes
NOUN < DAT?
No
Yes
Canonical active transitive
Scrambled active transitive
Undetermined
Canonical suffixal passive
Scrambled suffixal passive
Undetermined
End

Note. ‘sent’ stands for a sentence. DAT and SFX are not search terms but cover terms (used only in this flow chart) representing key morphemes used in suffixal passives (DAT = *-eykey*, *-hanthey*, *-ey*; SFX = *-i-*/*-hi-*/*-li-*/*-ki-*).

Figure 2. Flow chart: Construction identification

In addition to raw frequency information about each pattern, we calculated ΔP, a unidirectional statistics for association strength that estimates the degree to which a cue co-occurs with an outcome (e.g., Allan 1980, Desagulier 2016). A ΔP score (which ranges from −1 to 1) is computed through a contingency table (Table 5), following the mathematical formula (13), where the probability of the outcome is conditioned upon that of the cue.

Table 5. Association strength: ΔP

	Outcome	**¬ Outcome**
Cue	a	b
¬ Cue	c	d

Note. ¬ stands for 'not'.

(13) $\Delta P_{(\text{outcome}|\text{cue})} = p(\text{Outcome}|\text{Cue}) - p(\text{Outcome}|\neg\ \text{Cue}) = \text{a}/(\text{a}+\text{b}) - \text{c}/(\text{c}+\text{d})$

For the interpretation of individual ΔP scores, the closer $\Delta P_{(\text{outcome}|\text{cue})}$ is to 1, the more likely the cue co-occurs with the outcome; the closer $\Delta P_{(\text{outcome}|\text{cue})}$ is to −1, the more unlikely the cue co-occurs with the outcome. We applied this technique to the two construction types in order to better ascertain the status of these constructions and case-marking in expressing a transitive event.

4.3 Construction identification: Child production

Although the caregiver input was the main data for analysis, the child production data were also analysed in order to further show how to use child corpora for studies on child language development, following the same method by which the caregiver input was processed. Since there is no optimised tokeniser dedicated to child corpora, we borrowed the tokenisation function from *UDPipe* (Straka & Straková 2017). To sustain the performance of the tokeniser, utterances were excluded whose length was less than 16 strings and which consisted of meaningless/incomprehensible repetition of onomatopoeia and/or mimetic words (e.g., 총총 청청 여기 히베 바으 에 해태 히투바 페테 운으아) before the tokenisation process. This treatment yielded 1,985 sentences (25,047 eojeols) for the actual analysis, which occupied 35.31% of the entire dataset. The tokenised data were then inputted to the same tagger and pattern-finder that was developed for the caregiver input. Every automatic process was complemented by a manual checkup to ensure its accuracy.

4.4 Model performance

The performance of the pattern-finder was measured in the same way as the tagger, by calculating an F1 score (the harmonic mean of precision and recall). Creating a golden set manually using the entire dataset, which consisted of around 70,000 sentences, requires huge amounts of time and human resources. We thus opted to use a small set of data with 100 sentences obtained randomly from the original dataset.

4.5 Results and discussion

4.5.1 *Accuracy of pattern-finder*

In the golden set, there were three constructional patterns of interest in this study: canonical active transitive with no omission of arguments and case-marking (N-NOM N-ACC V), canonical active transitive with no ACC (N-NOM N-~~ACC~~ V), and suffixal passive with undergoer-NOM only (N-NOM V-psv). As Table 6 shows, there were no false negatives (i.e., items that should be included in the target category but are excluded actually) but only false positives (i.e., items that should not be included in the target category but are included actually) in the extraction results.

Table 6. By-pattern F1 score: pattern-finder

	True positive	**False negative**	**False positive**	**F1 score**
Canonical active transitive, no omission	32	0	3	0.955
Canonical active transitive, no ACC	5	0	4	0.714
Suffixal passive, undergoer-NOM only	7	0	3	0.824

Note. Positive: an instance included after a process; Negative: an instance not included after a process; True: an instance that should be included after a process; False: an instance that should not be included after a process

This indicates that the pattern-finder extracted more potential instances of these patterns than the exact number of instances from the manual extraction. Meanwhile, the fact that no false negative was included in the fully automatic extraction result allowed for automatic extraction of constructional patterns supplemented by manual checking.

The performance of the pattern-finder varied by specific constructional patterns. The accuracy of the canonical active transitive with no omission was reasonably high. In contrast, the accuracy levels of the other two patterns were lower than the accuracy of the canonical active transitive with no omission. The false negative instances of the canonical active transitive with no ACC and the suffixal passive with only the undergoer-NOM pairing included noun sequences (e.g., Jio-NOM grandma house-LOC go) and morphological causatives (e.g., Jio-NOM eat-CST), respectively. Inclusion of these instances again necessitates manual checking of automatic extraction results.

Based on the pattern-finder, the following sections provide a preliminary report on the use of constructional patterns in expressing a transitive event attested in child corpora. However, although this study employed the largest dataset of child corpora currently available, a substantial portion of the data was left untouched due to the focus of investigation (i.e., constructions involving a transitive event).

4.5.2 *Use of active transitives and suffixal passives: caregiver input*

4.5.2.1 *By-construction use*

Table 7 presents the frequency of occurrence involving active transitives and suffixal passives by canonicity with no omission of arguments and case-marking in the caregiver input. There was a substantial difference in the frequency of occurrence in active transitives by canonicity: the canonical pattern (1,757 instances) occurred far more frequently than the scrambled pattern (51 instances). Suffixal passives were extremely rare in their use, occurring two instances in the canonical pattern and one instance in the scrambled pattern. These asymmetries across these constructions and those within active transitives parallel previous findings from the general-purpose corpora (e.g., Shin 2006).

Table 7. Frequency of active transitives and suffixal passives in the caregiver input (no omission of arguments or case-marking)

	Active transitive		**Suffixal passive**	
	#	**%** [1]	**#**	**%** [1]
Canonical	1,757	97.02	2	0.11
Scrambled	51	2.82	1	0.06

Note. (1) was calculated out of the four constructional patterns (1,811 instances).

ΔP scores of the two construction types (Table 8) reveal varying degrees of association that a transitive event and each construction type manifest in the caregiver input. As calculated in $\Delta P_{(B|A)}$, the four patterns served as equally strong cues to introduce a transitive event, showing more than a score of 0.97 across the board. However, the reversed direction $\Delta P_{(A|B)}$ showed that a transitive event was most likely by far to be expressed as the canonical active transitive and least likely to be encoded as the passive.

Table 8. *ΔP* scores: Active transitives and suffixal passives for a transitive event in the caregiver input (no omission of argument and case-marking)

	Canonical active transitive	Scrambled active transitive	Canonical suffixal passive	Scrambled suffixal passive
$\Delta P_{(B\vert A)}$	0.999	0.975	0.974	0.974
$\Delta P_{(A\vert B)}$	0.979	0.028	0.001	0.000

Note. A = individual construction; B = transitive event.

The strong bi-directionality between the canonical active transitive and a transitive event suggests that, of the four candidates, the canonical active transitive is the default construction for expressing this type of event in caregiver input. In contrast, the asymmetric strength of association that the other three patterns demonstrated with respect to a transitive event indicate that, although they could be used to express a transitive event, their use is not preferred over that of the canonical active transitive in caregiver input.

Table 9 presents frequency information about all the patterns, with varying degrees of omission of sentential components, for a transitive event in the caregiver input. As for the active patterns, whereas the ACC tended to be omitted more often than the NOM within the patterns with two overt arguments (268 + 6 instances vs. 19 instances), the undergoer-ACC pairing appeared more frequently than the actor-NOM pairing when the patterns retained only one overt argument (935 instances vs. 1,938 instances). When two arguments were attested in active transitives, the NOM-marked argument occurred initially (1,757 + 268 = 2,025 instances) more than non-initially (51 + 6 = 57 cases). In contrast, the ACC-marked argument showed the reverse tendency, appearing non-initially (1,757 + 19 = 1,776 cases) more than initially (51 cases). The passive patterns were rare in the input compared to the active ones (4,974 instances vs. 423 instances) in general. However, the number of passive patterns with only one case-marked argument was relatively large (407 + 13 instances).

Table 10 presents frequency of case-less patterns expressing a transitive event in the caregiver input. Note that these patterns involve no overt case-marking attached to argument(s) and so interpretation of thematic roles of argument(s) can be ambiguous, thus necessitating human judgment through manual inspection. Regarding the one-argument active pattern without case-marking, the number of instances where the argument expresses the undergoer (i.e., the ACC is omitted) outnumbered the number of instances where that argument expresses the actor (i.e., the NOM is omitted). As for the corresponding passive pattern, all the instances fell into a case in which the argument expresses the undergoer (i.e., the NOM is omitted). There were only three instances that consist of two

Table 9. Frequency of patterns for a transitive event in the caregiver input

Type	Example	Frequency (#)
Canonical active transitive	police-NOM thief-ACC catch	1,757
Scrambled active transitive	thief-ACC police-NOM catch	51
Canonical suffixal passive	thief-NOM police-DAT catch-psv	2
Scrambled suffixal passive	police-DAT thief-NOM catch-psv	1
Canonical active transitive, no ACC	police-NOM thief-~~ACC~~ catch	268
Canonical active transitive, no NOM	police-~~NOM~~ thief-ACC catch	19
Scrambled active transitive, no ACC	thief-~~ACC~~ police-NOM catch	6
Scrambled active transitive, no NOM	thief-ACC police-~~NOM~~ catch	0
Canonical suffixal passive, no DAT	thief-NOM police-~~DAT~~ catch-psv	0
Canonical suffixal passive, no NOM	thief-~~NOM~~ police-DAT catch-psv	0
Scrambled suffixal passive, no DAT	police-~~DAT~~ thief-NOM catch-psv	0
Scrambled suffixal passive, no NOM	police-DAT thief-~~NOM~~ catch-psv	0
Active transitive, actor-NOM only	police-NOM catch	935
Active transitive, undergoer-ACC only	thief-ACC catch	1,938
Suffixal passive, undergoer-NOM only	thief-NOM catch-psv	407
Suffixal passive, actor-DAT only	police-DAT catch-psv	13
SUM		5,397

overt arguments without case-marking altogether, all of which fell into the actor-undergoer ordering.

Table 10. Frequency of case-less patterns for a transitive event in the caregiver input

Pattern	Thematic role	Frequency
$\mathrm{N}_{\cancel{\mathrm{CASE}}}\mathrm{V}_{\mathrm{act}}$	Actor	53
	Undergoer	1,155
	Undetermined	40
$\mathrm{N}_{\cancel{\mathrm{CASE}}}\mathrm{V}_{\mathrm{psv}}$	Actor	0
	Undergoer	20
	Undetermined	0
$\mathrm{N}_{\cancel{\mathrm{CASE}}}\mathrm{N}_{\cancel{\mathrm{CASE}}}\mathrm{V}_{\mathrm{act}}$	Actor-undergoer	3
	Undergoer-actor	0
	Undetermined	0
SUM		1,268

4.5.2.2 *By-marker use*

Based on the construction-wise results, additional analyses were conducted in light of case-marking use in the caregiver input. Table 11 presents frequency information about the NOM based on the thematic role associated with it and whether/where the case-marked argument appears in the patterns extracted from the caregiver input. The NOM was used as an indication of the actor (935 + 2,025 + 57 = 3,017 instances) more than an indication of the undergoer (407 + 2 + 1 = 410 instances). This marker was also used overtly (935 + 2,025 + 57 + 407 + 2 + 1 = 3,427 instances) more than it was omitted (53 + 22 + 20 = 95 instances). Within the one-argument patterns, the marker was present (935 instances for the actor; 407 instances for the undergoer) considerably more than it was absent (53 instances for the actor; 20 instances for the undergoer). In the two-argument active transitive patterns, the marker was used initially (2,025 instances) more than non-initially (57 instances).

Table 11. Frequency of NOM in the caregiver input

Thematic role	Appeared?	Where?	Pattern type	Frequency (#)
Actor	Yes	Initially	One-argument	935
			Two-argument, canonical	2,025
		Non-initially	Two-argument, scrambled	57
	No	Initially	One-argument	53
			Two-argument, canonical	22
		Non-initially	Two-argument, scrambled	0
Undergoer	Yes	Initially	One-argument	407
			Two-argument, canonical	2
		Non-initially	Two-argument, scrambled	1
	No	Initially	One-argument	20
			Two-argument, canonical	0
		Non-initially	Two-argument, scrambled	0

The ΔP scores substantiate the strong bi-directional association between the NOM and the actor in the context of a transitive event. The NOM was an extremely reliable cue for the actor role ($\Delta P_{(\text{ACTOR}|\text{NOM})} = 0.853$) and vice versa ($\Delta P_{(\text{NOM}|\text{ACTOR})} = 0.856$). In contrast, the NOM was very unlikely to introduce the undergoer ($\Delta P_{(\text{UNDERGOER}|\text{NOM})} = -0.868$) and vice versa ($\Delta P_{(\text{NOM}|\text{UNDERGOER})} = -0.905$). This reveals the strong reliability of the NOM for the actor and vice versa in child-directed speech.

Table 12 presents frequency information about the ACC based on whether and where the case-marked argument appears in the patterns extracted from the caregiver input. The ACC was used overtly (1,938 + 51 + 1,776 = 3,765 instances) more than it was omitted (1,155 + 6 + 271 = 1,432 instances). Within one-argument patterns, this marker was present (1,938 instances) more than it was omitted (1,155 instances). However, its omission in one-argument patterns occurred proportionally more than that of the NOM. Thus, the rate at which the ACC was dropped (0.373) was much higher than the rate at which the NOM (indicating the actor) was dropped (0.054). In the two-argument active transitive patterns, the ACC was used non-initially (1,776 instances) more than initially (51 instances).

Table 12. Frequency of ACC in the caregiver input

Thematic role	Appeared?	Where to appear?	Pattern type	Frequency (#)
Undergoer	Yes	Initially	One-argument	1,938
			Two-argument, scrambled	51
		Non-initially	Two-argument, canonical	1,776
	No	Initially	One-argument	1,155
			Two-argument, scrambled	6
		Non-initially	Two-argument, canonical	271

Note. Because the focus of analysis was patterns involving a transitive event, any ditransitive pattern was excluded.

The ΔP scores show that the association between the ACC and the undergoer within a transitive event was moderately reliable: the ACC was a good cue for the undergoer ($\Delta P_{(\text{UNDERGOER}|\text{ACC})} = 0.350$) and vice versa ($\Delta P_{(\text{ACC}|\text{UNDERGOER})} = 0.670$) but not extremely strong as in the case of the NOM and the actor. This is due to the high omission rate for the ACC compared to the case of the NOM, by increasing the impact of '¬ cue' on the calculation of ΔP.

Regarding the DAT, whereas there were 269 instances in which the DAT indicates a recipient (in ditransitives), there were only 16 instances in which the DAT marked an actor (in the passive). ΔP scores for the DAT showed that the marker was not likely to be associated with the actor ($\Delta P_{(\text{ACTOR}|\text{DAT})} = -0.410$) or vice versa ($\Delta P_{(\text{DAT}|\text{ACTOR})} = -0.066$). Although the active patterns involving the DAT are ditransitives (and therefore do not count as relevant patterns expressing a simple transitive event), these patterns were considered only here because the DAT is often used as an indicator of a recipient in the active and thus a potential competitor of the actor-DAT pairing in the passive.

4.5.2.3 *Summary of findings: Caregiver input*

Three major findings were noted. First, of the core constructional patterns with no omission of arguments and case-marking, the canonical active transitive occurred far more frequently than its scrambled counterpart, and the passives were extremely rare, regardless of canonicity. Second, in the two-argument active transitive patterns, the NOM-marked and ACC-marked arguments tended to appear initially and non-initially, respectively. Third, the degree of association between individual markers and thematic roles was asymmetric: the NOM was a very strong cue for the actor (and vice versa), the ACC was a moderately good cue for the undergoer (and vice versa), and the DAT was not likely to occur with the actor (and vice versa).

4.5.3 *Use of active transitives and suffixal passives: Child production*

Table 13 presents frequency information about all the patterns, with varying degrees of omission of sentential components, for a transitive event in the child production. In expressing a transitive event, the children used only a few patterns intensively such as the canonical active transitive with no omission (37 instances) and the one-argument active patterns either with the undergoer-ACC pairing (25 instances) or with the actor-NOM pairing (21 instances). There were only 9 instances of the one-argument passive pattern with the undergoer-NOM pairing.

Table 13. Frequency of patterns for a transitive event in child production

Type	**Example**	**Frequency (#)**
Canonical active transitive	police-NOM thief-ACC catch	37
Scrambled active transitive	thief-ACC police-NOM catch	0
Canonical suffixal passive	thief-NOM police-DAT catch-psv	0
Scrambled suffixal passive	police-DAT thief-NOM catch-psv	0
Canonical active transitive, no ACC	police-NOM thief-~~ACC~~ catch	14
Canonical active transitive, no NOM	police-~~NOM~~ thief-ACC catch	0
Scrambled active transitive, no ACC	thief-~~ACC~~ police-NOM catch	0
Scrambled active transitive, no NOM	thief-ACC police-~~NOM~~ catch	0
Canonical suffixal passive, no DAT	thief-NOM police-~~DAT~~ catch-psv	0
Canonical suffixal passive, no NOM	thief-~~NOM~~ police-DAT catch-psv	0
Scrambled suffixal passive, no DAT	police-~~DAT~~ thief-NOM catch-psv	0
Scrambled suffixal passive, no NOM	police-DAT thief-~~NOM~~ catch-psv	0
Active transitive, actor-NOM only	police-NOM catch	21
Active transitive, undergoer-ACC only	thief-ACC catch	25

Table 13. *(continued)*

Type	Example	Frequency (#)
Suffixal passive, undergoer-NOM only	thief-NOM catch-psv	9
Suffixal passive, actor-DAT only	police-DAT catch-psv	0
SUM		106

Although the analysis of the child production data in this study was preliminary, this disproportionate use of the individual constructional patterns observed in the data largely mirrored the characteristics of the caregiver input (see Table 9), which prioritised some patterns over others with respect to a transitive event.

5. Conclusion: Implications on automatic processing of Korean child corpora for developmental research on Korean

In response to the lack of research on automatic processing of Korean child corpora, the present study conducted a series of NLP-assisted analyses of caregiver input and child production in the CHILDES database. We reported how corpus-based research was done previously, what kind of language-specific properties pose challenges to automatic processing of Korean child corpora, and how these obstacles can be alleviated in this task. Two empirical works were then conducted for this task. One was to develop a POS tagger through a machine learning algorithm, together with enhanced input and feature sets. The tagger demonstrated the state-of-the-art accuracy rates in performance of XPOS and UPOS tagging for Korean child corpora. The other was to extract two construction types in expressing a transitive event (active transitives and suffixal passives), with scrambling and omission of constructional components manifested, from the child corpora. This pattern-finding work revealed a considerable overlap in caregiver input and child production in the employment of these constructions when it comes to a transitive event. Together, findings of this study suggest the applicability of NLP techniques to research on Korean child corpora, which has not been considered in the field of developmental research on Korean.

Despite its (methodological) success, this study still has limitations, which await further investigation. First of all, we did not demonstrate full-fledged automatic processing for the analysis of caregiver input in Korean. The currently available pipelines for handling corpus data are mostly based on general-purpose corpora, which reduces the applicability of the open-access pipelines to the analysis of child corpora. Characteristics of child corpora such as onomatopoeia and mimetic words also complicate the analysis. In addition, language-specific properties such as scrambling and omission of sentential components remain a significant challenge. To bypass these thorny issues, we took a semi-automatic approach to pattern-finding, but we acknowledge that this is not the ultimate solution. We are optimistic that analyses of child corpora will benefit from techniques that employ probabilistic dependency relations, which several studies (targeting general-purpose written corpora) have suggested in different language-use settings (e.g., Park et al. 2016). Future research should be directed towards measuring the extent to which cutting-edge methods for general-purpose corpora can overcome the challenges associated with automatic processing of Korean child corpora.

Next, in a broader context, implications of findings from corpus analysis should be complemented and re-assessed by way of denser corpus data and/or other methods of investigation such as behavioural experiments and computational modelling. The frequency and time period in which the CHILDES database was collected for Korean child corpora were not dense enough to ensure the representativeness and generalisability of findings from this dataset; that is, we may have missed something in between the collections. In addition, various factors are known to modulate frequency effects in learning: consistency of mapping between form and function (e.g., Cameron-Faulkner et al. 2007); linguistic environments where target language items occur (e.g., Dąbrowska 2008); informativeness of the current stimulus against the prior experience (e.g., Dittmar et al. 2008); and domain-general factors (e.g., Stefanowitsch 2011, Theakston 2004). We supplemented frequency of occurrence involving the target constructional patterns and case-marking with ΔP scores, but we still need further verification as to what we found. Future studies on corpus analysis should involve much denser corpora for child language (cf. Thomas corpus),[13] assisted by experimentation (e.g., Shin 2020) and simulation work (e.g. Alishahi & Stevenson 2008), for a more precise understanding of child language development.

This study focussed on providing a methodological report on how NLP techniques can be applied to analyse of child corpora in Korean, and therefore, the current study falls short of revealing a complete picture of developmental aspects in child language development in Korean (which is not a focal point in this study). As one reviewer pointed out, frequency counting and association measurement are not enough to address the whole process of language acquisition. The POS tagger and pattern-finder proposed in this study utilise morpho-syntactic features as their unit of analysis, which renders our understanding of child corpora unable to deal with other important aspects of language such as semantics and pragmatics/discourse. Nevertheless, we believe that our work in this study serves to initiate alternative ways of analysing Korean child corpora, and future research should

---

[13] https://childes.talkbank.org/access/Eng-UK/Thomas.html

answer questions about 'invisible' factors at a more detailed level. One promising area for subsequent research in this respect includes analysis of local sequences of mother-child conversational turns to better ascertain learning mechanisms.

Automatic processing of child corpora in Korean and its application to developmental research on Korean are still in their infancy. As the first methodological report on these topics, empirical findings of this study shed light on promising ways of corpus-based research on child language development in Korean. This also contributes to research practice, ensuring the reproducibility of procedures and results, with respect to studies on Korean child corpora (and perhaps beyond).

Abbreviations

ACC	accusative case marker
CST	causative marker
DAT	dative marker
LOC	locative marker
NOM	nominative case marker
PRS	present tense marker
PSV	passive marker
PST	past tense marker
REL	relativiser
SE	sentence ender
TOP	topic marker

References

Abbot-Smith, Kirsten, Franklin Chang, Caroline Rowland, Heather Ferguson & Julian Pine. 2017. Do two and three year old children use an incremental first-NP-as-agent bias to process active transitive and passive sentences?: A permutation analysis. *PloS one* 12.10. e0186129. https://doi.org/10.1371/journal.pone.0186129

Alishahi, Afra & Suzanne Stevenson. 2008. A computational model of early argument structure acquisition. *Cognitive Science* 32.5. 789–834. https://doi.org/10.1080/03640210801929287

Allan, Lorraine G. 1980. A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society* 15.3. 147–149. https://doi.org/10.3758/BF03334492

Ambridge, Ben, Evan Kidd, Caroline F. Rowland & Anna L. Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language* 42.2. 239–273. https://doi.org/10.1017/S030500091400049X

Behrens, Heike. 2006. The input-output relationship in first language acquisition. *Language and Cognitive Processes* 21.1–3. 2–24. https://doi.org/10.1080/01690960400001721

Behrens, Heike. 2009. Usage-based and emergentist approaches to language acquisition. *Linguistics* 47.2. 383–411. https://doi.org/10.1515/LING.2009.014

Cameron-Faulkner, Thea, Elena Lieven & Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive Science* 27.6. 843–873. https://doi.org/10.1207/s15516709cog2706_2

Cameron-Faulkner, Thea, Elena Lieven & Anna Theakston. 2007. What part of no do children not understand? A usage-based account of multiword negation. *Journal of Child Language* 34.2. 251–282. https://doi.org/10.1017/S0305000906007884

Cho, Sook Whan. 1982. The acquisition of word order in Korean. MA thesis, University of Calgary.

Choi, Soonja. 1999. Early development of verb structures and caregiver input in Korean: Two case studies. *International Journal of Bilingualism* 3.2–3. 241–265. https://doi.org/10.1177/13670069990030020701

Choi, Jinho D. & 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*, 1–11.

Choo, Miho & Kwak, Hye-Young. 2008. *Using Korean*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139168496

Chung, Gyeonghee No. 1994. Case and its acquisition in Korean. Ph.D. dissertation, University of Texas at Austin.

Collins, Michael & Nigel Duffy. 2002. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 263–270.

Dąbrowska, Ewa. 2008. The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language* 58.4. 931–951. https://doi.org/10.1016/j.jml.2007.11.005

Daumé III, Hal. 2015. A Course in machine learning (Ch3. The perceptron). http://ciml.info/

Desagulier, Guillaume. 2016. A lesson from associative learning: asymmetry and productivity in multiple-slot constructions. *Corpus Linguistics and Linguistic Theory* 12.2. 173–219. https://doi.org/10.1515/cllt-2015-0012

Dittmar, Miriam, Kirsten Abbot-Smith, Elena Lieven & Michael Tomasello. 2008. German children's comprehension of word order and case marking in causative sentences. *Child Development* 79.4. 1152–1167. https://doi.org/10.1111/j.1467-8624.2008.01181.x

Ellis, Nick. C. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24. 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, Nick C. & Fernando Ferreira-Junior. 2009. Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal* 93.3. 370–385. https://doi.org/10.1111/j.1540-4781.2009.00896.x

Freund, Yoav & Robert E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning* 37.3. 277–296. https://doi.org/10.1023/A:1007662407062

Ghosh, Devyani, John B. Carter & Hal Daumé III. 2008. Perceptron-based Coherence Predictors. In *Proceedings of the 2nd Workshop on chip multiprocessor memory systems and interconnects.*

Goldberg, Adele E., Devin M. Casenhiser & Nitya Sethuraman. 2004. Learning argument structure generalizations. *Cognitive Linguistics* 15.3. 289–316. https://doi.org/10.1515/cogl.2004.011

Honnibal, Matthew. 2013. A good part-of-speech tagger in about 200 lines of Python. https://explosion.ai/blog/part-of-speech-pos-tagger-in-python

Honnibal, Matthew, Yoav Goldberg & Mark Johnson. 2013. A non-monotonic arc-eager transition system for dependency parsing. In *Proceedings of the 7th Conference on Computational Natural Language Learning*, 163–172.

Honnibal, Matthew & Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1373–1378. https://doi.org/10.18653/v1/D15-1162

Huang, Yi Ting, Xiaobei Zheng, Xiangzhi Meng & Jesse Snedeker. 2013. Children's assignment of grammatical roles in the online processing of Mandarin passive sentences. *Journal of Memory and Language* 69.4. 589–606. https://doi.org/10.1016/j.jml.2013.08.002

Jin, Kyong-sun, Min Ju Kim & Hyun-joo Song. 2015. The development of Korean preschooler' ability to understand transitive sentences using case-markers. *The Korean Journal of Cognitive and Biological Psychology* 28.3. 75–90.

Kim, Hung-gyu, Beom-mo Kang & Jungha Hong. 2007. 21seyki seycongkyeyhoyk hyentaykwuke kichomalmwungchi sengkwawa cenmang [21st century Sejong modern Korean corpora: Results and expectations]. In *Proceedings of annual conference on human and language technology 31*, 311–316.

Kim, Meesook. 2010. Syntactic priming in children's production of passives. *Korean Journal of Applied Linguistics* 26.2. 271–290.

Kim, Seongchan, William O'Grady & Sookeun Cho. 1995. The acquisition of case and word order in Korean: A note on the role of context. *Language Research* 31.4. 687–695.

Kim, Shin-Young, Jee Eun Sung & Dongsun Yim. 2017. Sentence comprehension ability and working memory capacity as a function of syntactic structure and canonicity in 5-and 6-year-old children. *Communication Sciences & Disorders* 22.4. 643–656. https://doi.org/10.12963/csd.17420

Kim, Wansu & Cheol Young Ock. 2015. hankwuke kyekthul sacenkwa uymiyek pinto cengpolul sayonghan hankwuke uymiyek kyelceng [Korean semantic role labeling using case frame and frequency]. *Journal of Korean Institute of Information Technology* 11.2. 161–167.

Lee, Chungmin & Sook Whan Cho. 2009. Acquisition of the subject and topic nominals and markers in the spontaneous speech of young children in Korean. In *The Handbook of East Asian Psycholinguistics 3* ed by Chungmin Lee, Greg Simpson and Youngjin Kim, 23–33. New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511596865.003

Lee, Hee Ran. 2004. 2sey hankwuk atonguy cwue paltal thukseng [A study of early subject acquisition in Korean]. *Communication Sciences and Disorders* 9.2. 19–32.

Lee, Ikseop. 2011. *kwukehakkaysel* [Introduction to Korean linguistics]. Seoul: Hakyensa.

Lee, Sun-Ar & Jin-Tak Choi. 2013. hankwuke Verb_OntoNetuy selkyeywa kwuchwuk [Design and implementation of Korean Verb_OntoNet]. *Journal of Korean Institute of Information Technology* 11.2. 161–167.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for analyzing talk. Third Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

No, Gyeong Hee. 2009. Acquisition of case markers and grammatical functions. In *The Handbook of East Asian Psycholinguistics 3* ed by Chungmin Lee, Greg Simpson and Youngjin Kim, 23–33. New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511596865.005

Park, Jungyeul, Jeen-Pyo Hong & Jeong-Won Cha. 2016. Korean language resources for everyone. In *JProceedings of the 30th Pacific Asia conference on language, information and computation: Oral Papers*, 49–58.

Petrov, Slav, Dipanjan Das & Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, 2089–2096.

Qi, Peng, Timothy Dozat, Yuhao Zhang & Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 160–170. https://doi.org/10.18653/v1/K18-2016

Rosenblatt, Frank. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65.6. 386–408. https://doi.org/10.1037/h0042519

Shin, Gyu-Ho. 2020. Connecting input to comprehension: First language acquisition of active transitives and suffixal passives by Korean-speaking preschool children. Ph.D. dissertation, University of Hawaiʻi at Mānoa.

Shin, Seo-in. 2006. kwumwun pwunsek malmwungchilul iyonghan hankwuke mwunhyeng yenkwu [A study on Korean sentence patterns using a parsed corpus]. Ph.D. dissertation, Seoul National University.

Sinclair, Hermina & Jean-Paul Bronckart. 1972. S.V.O A linguistic universal? A study in developmental psycholinguistics. *Journal of Experimental Child Psychology* 14. 329–348. https://doi.org/10.1016/0022-0965(72)90055-0

Slobin, Dan I. & Thomas G. Bever. 1982. Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition* 12.3. 229–265. https://doi.org/10.1016/0010-0277(82)90033-6

Sohn, Ho Min. 1999. *The Korean language*. Cambridge University Press.

Song, Sanghoun & Jae-Woong Choe. 2007. Type hierarchies for passive forms in Korean. In *Proceedings of the 14th international conference on Head-Driven Phrase Structure Grammar, Stanford Department of Linguistics and CSLI's LinGO Lab*, 250–270. https://doi.org/10.21248/hpsg.2007.15

Stefanowitsch, Anatol. 2011. Constructional preemption by contextual mismatch: A corpus-linguistic investigation. *Cognitive Linguistics* 22.1. 107–129. https://doi.org/10.1515/cogl.2011.005

Stoll, Sabine, Kirsten Abbot-Smith & Elena Lieven. 2009. Lexically restricted utterances in Russian, German, and English child-directed speech. *Cognitive Science* 33.1. 75–103. https://doi.org/10.1111/j.1551-6709.2008.01004.x

Straka, Milan & Jana Straková. 2017. Tokenizing, POS Tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. https://doi.org/10.18653/v1/K17-3009

Theakston, Anna L. 2004. The role of entrenchment in children's and adults' performance on grammaticality judgment tasks. *Cognitive Development* 19.1. 15–34. https://doi.org/10.1016/j.cogdev.2003.08.001

Tomasello, Michael. 1992. *First verbs: A case study of early grammatical development*. New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511527678

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Wonnacott, Elizabeth, Jeremy K. Boyd, Jennifer Thomson & Adele E. Goldberg. 2012. Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language* 66.3. 458–478. https://doi.org/10.1016/j.jml.2011.11.004

Yeon, Jaehoon. 2015. Passives. In *The handbook of Korean linguistics* ed by Lucien Brown & Jaehoon Yeon, 116–136. Oxford: John Wiley & Sons. https://doi.org/10.1002/9781118371008.ch7

Appendix. Tag set: XPOS (Sejong tag set) & UPOS (Universal tag set)

XPOS: Sejong tags

Tag	Meaning	Tag	Meaning
NNG	Noun, general	ETN	Ending, transformative (noun)
NNP	Noun, proper	ETM	Ending, transformative (determiner)
NNB	Noun, dependent	XPN	Prefix (for noun)
NR	Noun, number	XSN	Suffix (for noun)
NP	Pronoun	XSV	Suffix (for verb)
JKS	Postposition, nominative	XSA	Suffix (for adjective)
JKC	Postposition, complement	XR	Root
JKG	Postposition, possession	MAG	Adverb, general
JKO	Postposition, accusative	MAJ	Adverb, connective
JKB	Postposition, adverbial	IC	Interjection
JKV	Postposition, vocative	SF	Symbol (period, question mark, exclamation mark)
JKQ	Postposition, quotative	SE	Symbol (ellipsis)
JC	Postposition, conjunctive	SS	Symbol (quotation mark, parenthesis, dash)
JX	Postposition, auxiliary	SP	Symbol (comma, interpunct, colon, slash)
VV	Verb, general	SO	Symbol (hyphen)
VX	Verb, auxiliary	SW	Symbol (others)
VCP	Verb, copular (positive)	SH	Symbol, Chinese character
VCN	Verb, copular (negative)	SL	Symbol, Foreign language character
VA	Adjective	SN	Symbol, number
MM	determiner	NF	Not clear (estimated to be noun)
EP	Ending, pre-final	NV	Not clear (estimated to be predicate)
EF	Ending, final	NA	Not clear (estimation impossible)
EC	Ending, connective		

UPOS: Universal Dependencies POS tags (https://universaldependencies.org/u/pos/index.html)

Tag	Meaning
ADJ	Adjective
ADP	Adposition
ADV	Adverb
AUX	Auxiliary
CCONJ	Coordinating conjunction
DET	Determiner
INTJ	Interjection
NOUN	Noun
NUM	Numeral
PART	Particle
PRON	Pronoun
PROPN	Proper noun
PUNCT	Punctuation
SCONJ	Subordinating conjunction
SYM	Symbol
VERB	Verb
X	Other

Address for correspondence

Gyu-Ho Shin
tř. Svobody 26
779 00 Olomouc
Czech Republic
gyuho.shin@upol.cz

Publication history

Date received: 24 February 2020
Date accepted: 11 October 2020