Automatic analysis of passive constructions in Korean

Written production by Mandarin-speaking learners of Korean

Gyu-Ho Shin and Boo Kyung Jung
Palacký University Olomouc | University of Pittsburgh

The present study aims to explore the applicability of automatic analysis to L2-Korean learner corpora, with a special focus on learners' use of a clause-level construction. For this purpose, we investigate L1-Mandarin L2-Korean learners' written production of two passive construction types in Korean – suffixal and periphrastic – by devising a pattern-extraction process through NLP techniques. We focus on reporting how the passive constructions are identified and extracted from learner writing automatically, given language-specific features involving the passive. A total of 72 essays were analysed by adapting an existing pipeline (developed by Shin, forthcoming), with enhanced tokenisation and annotation through manual revision of the data. Results showed that our automatic pattern-finder identified more instances than manual extraction for the suffixal passive and yielded a perfect match with manual extraction for the periphrastic passive. Implications of the findings are discussed in regard to strengths and drawbacks of the automatic analysis of learner writing, with suggestions for improving currently available tools for learner corpus research in Korean.

Keywords: natural language processing, learner corpus, passive construction, Korean

1. Introduction

The use of learner corpora has played a crucial role for the understanding of developmental aspects that second language (L2) learners manifest (e.g., Biber, Conrad, & Cortes, 2004; Biber, Gray, & Poonpon, 2011; Crossley, Kyle, & McNamara, 2016; Ellis & Ferreira-Junior, 2009; Gablasova, Brezina, & McEnery, 2017; Gilquin, 2008). Recent advancement of computational approaches to data analysis promotes

https://doi.org/10.1075/ijlcr.20002.shi
International Journal of Learner Corpus Research 7:1 (2021), pp. 53–82. issn 2215-1478 | e-issn 2215-1486

three specific areas in relation to learner corpus research: annotation of learner production (e.g., de Haan, 2000; de Mönnink, 2000), examination of specific features for learner language in corpora (e.g., Bestgen & Granger, 2014; de Felice & Pulman, 2009; Kyle & Crossley, 2017), and development of automatic tools to analyse learner corpora (e.g., Kyle, Crossley, & Berger, 2018; Lu, 2010) (see also Meurers, 2015 for a comprehensive review of each area). There are challenges for the pursuit of these interdisciplinary areas due to characteristics of learner language such as variability and non-standard use of target language systems (e.g., Meurers & Dickenson, 2017). However, it is clear that automatic processing of learner corpora is gaining momentum for a better understanding of properties of learner language, which also ensures reproducibility of procedures and results across various learner corpora.

In the present study, we aim at applying Natural Language Processing (NLP) techniques to learner corpora produced by Mandarin-speaking learners of Korean, with a particular focus on argument structure constructions in written production. An argument structure construction is defined as a clause-level form-function pairing which provides a means for delivering a basic proposition in language (e.g., Goldberg, 1995, 2006). Previous research has revealed L2 learners' increasing ability to employ complex constructions such as causatives and resultatives, along with less typical associations between constructions and verbs, in proportion to learner proficiency (e.g., Ellis & Ferreira-Junior, 2009; Kim & Rah, 2016; Kim, Shin, & Hwang, 2020; Kyle & Crossley, 2017; Sung & Kim, 2020). Crucially, however, the majority of this line of research has been skewed towards frames with concrete lexical items such as verb and preposition as a pivot (e.g., Ellis & Ferreira-Junior, 2009; Kyle, 2016; Kyle & Crossley, 2017; Römer, Roberson, O'Donnell, & Ellis, 2014). It is thus uncertain whether their findings suffice to clarify the relation between L2 learners' use of these constructions and trajectories of L2 development. This calls for a closer look at the direct contribution of argument structure constructions themselves to L2 development.

Of the various types of argument structure constructions, we focus on passive constructions. It is often argued that the acquisition of passives is challenging for both first language (L1) (e.g., Abbot-Smith, Chang, Rowland, Ferguson, & Pine, 2017; Brooks & Tomasello, 1999; Huang, Zheng, Meng, & Snedeker, 2013; Shin, 2020) and L2 (e.g., Dąbrowska & Street, 2006; Jeong, 2014; Xiao, 2007) learners. Most of the L2 studies on the passive heavily emphasise instructional effects on the acquisition of L2-English passives (e.g., Birjandi, Maftoon, & Rahemi, 2011; Hinkel, 2004; Izumi & Lakshmanan, 1998; Ju, 2000; Lee, 2007). This obscures linguistic considerations about challenges involving the passive in non-English L2 acquisition. Moreover, literature on NLP-assisted investigation of the passive remains thin (cf. Shin, forthcoming), with myriad unanswered questions regarding the feasibility of automatic analysis of the passive to learner writing (see Meurers

& Dickenson, 2017 for broader issues on the adaptability of NLP techniques for learner language data).

Against this background, the present study probes into L2 learners' production of passive constructions in writing by way of NLP techniques. We demonstrate automatic extraction of passives from learner writing produced by Mandarin-speaking learners of Korean, and provide a preliminary analysis of how learners use the target construction type in written production. As the first report on automatic processing of L2-Korean learner corpora with respect to Korean passives, this study will shed light on strengths and weaknesses of using NLP techniques for learner writing analysis, given language-specific features involving passives. Findings of this study are also expected to contribute to extending our current understanding of automatic analysis of learner corpora for investigating L2 development of Korean (and beyond).

2. Literature review

2.1 Trend of learner corpus research on L2 Korean

Recent learner corpus research on L2 Korean has been conducted in three main strands: error analysis, use of linguistic items (e.g., content word, case-marking, auxiliary verb), and complexity-accuracy-fluency (CAF) (see Appendix A for the summary of recent learner corpus research on L2 Korean).

The automatic analysis of L2-Korean corpora is a very recent research trend. Many studies relied on pre-made L1-Korean POS-tagging programmes[1] (Kwak, 2016; J. Lee, 2017; S. Lee, 2017; Nam & Hong, 2014; Park & Lee, 2017; Ryu, 2017), often supplemented with manual annotation and general-purpose programmes for text editing. To illustrate, Nam and Hong (2014) created their own corpus data by collecting L2 spoken data from story re-telling, communicative tasks, and naturalistic conversation, in the following steps: converting the pre-processed transcription

---

[1] There are various morphological analysers in Korean, all of which were developed for the purpose of general-purpose L1-Korean corpora: Hannanum (http://semanticweb.kaist.ac.kr/home/index.php/HanNanum); Kkma (http://kkma.snu.ac.kr/); Komoran (https://github.com/shineware/KOMORAN); MeCab-ko (https://bitbucket.org/eunjeon/mecab-ko-dic/src/master/); Open Korean Text (https://github.com/open-korean-text/open-korean-text), to name a few. They are all Java-based (cf. Python wrapper: KoNLPy [Park & Cho, 2014]) and employ different POS tag-sets. Recently, the Electronics and Telecommunications Research Institute released another open-source morphological analyser (http://aiopen.etri.re.kr/guide_wiseNLU.php), which is compatible with various computer languages but limits daily use (5,000 trials; 10,000 characters per trial).

into text (*.txt*) files, analysing the text files through the Sejong POS tag set,[2] and converting the POS-tagged data into a new database with their own coding rubrics using Microsoft Excel. Based on the new dataset, they conducted a sample analysis of how L2 learners (two beginners and two intermediates) produced case-marking across proficiency levels of Korean. In this regard, the use of the automatic processing tool serves mostly as another pre-processing stage for the actual analysis.

Some studies aimed at developing automatic processing programmes for their own purposes, in combination with pre-existing NLP tools (Cho & Park, 2018; Kim, Park, Kim, Kim, Choi, Suh, & Kwak, 2016; Lee, Dickenson, & Israel, 2016; Park, Kim, Lee, & Lee, 2017). Lee et al. (2016),[3] for example, created learner corpora including essays of 100 learners of Korean by crossing two proficiency levels (beginners and intermediates) with two learning contexts (foreign and heritage language), and they explained annotation processes involving the corpora in detail. They put emphasis on describing challenges for annotation of learner writing, particularly in dealing with learner errors. Kim et al. (2016) [4] suggested an advanced system that allows pattern identification in learner corpora with error-fixed essays of around 500 learners of Korean. They focused on describing types of errors and rules for searching patterns on the basis of POS-tagging produced by a pre-made tool. Cho and Park (2018) conducted text quality research with error-fixed learner writing from 16 English-speaking learners of Korean. They employed several morphological analysers provided by *KoNLPy* (Park & Cho, 2014) and calculated similarity scores by employing term frequency-inverse document frequency through *scikit-learn* (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, & Vanderplas, 2011).

Despite the increasing interest in automatic processing of learner corpora in Korean, there are two obvious limitations in the previous reports. One is the lack of 'how-to' descriptions. Except for three studies (Kim et al., 2016; Lee et al., 2016; Nam & Hong, 2014), most did not provide details about how they conducted their analyses. It is thus impossible to determine what difficulties arose when they dealt with learner writing in Korean, and also how they alleviated or bypassed the difficulties in the actual procedures on (automatic) processing of the data. These

---

[2] The Sejong POS tag set (Kim, Kang, & Hong, 2007) is particularly influential in the Korean context. The system has 45 labels under seven categories and employs relatively detailed classification for the postpositions and dependency-related items by function, reflecting linguistic characteristics of Korean. The basic unit of POS tagging in this system is a morpheme within an eojeol: a white-space-based unit serving as the minimal unit of sentential components (Lee, 2011).

[3] http://cl.indiana.edu/~kolla/.

[4] http://www.yskli.com:8080/lcms/login/userLogin.do ('currently not available').

limited descriptions do not ensure reproducibility of the procedures and results that they reported. The other limitation concerns the scope of investigation. A large portion of previous research focused on the learners' developmental aspects involving individual lexical items, and this practice does not reveal how learners of Korean acquire knowledge about argument structure constructions, which deliver complete propositions in language (Goldberg, 1995, 2006). No research on Korean as an L2 touches upon possible connections between learners' production of these clause-level constructions and learner characteristics such as proficiency, nor the application of NLP techniques to the automatic identification/extraction of these constructional patterns.

2.2 Passive constructions in Korean and Mandarin

Korean is a Subject-Object-Verb language with overt case marking by dedicated markers. These structural cues allow scrambling of pre-verbal arguments as long as that reordering unambiguously preserves the original intention as in (1a) and (1b).

(1) a. Canonical active transitive
kyengchal-i totwuk-ul cap-ass-ta.
police-NOM thief-ACC catch-PST-SE
'The police caught the thief.'
b. Scrambled active transitive
totwuk-ul kyengchal-i cap-ass-ta.
thief-ACC police-NOM catch-PST-SE
'The police caught the thief.'

Korean also permits the omission of almost all elements in a sentence if the omitted information can be inferred from the context. This omission applies to a marker (2a), an argument coupled with a marker (2b), and even a predicate (2c).

(2) a. Omission: marker
kyengchal-i totwuk-~~ul~~ cap-ass-ta.
police-NOM thief-~~ACC~~ catch-PST-SE
'The police caught the thief.'
b. Omission: argument + marker
~~kyengchal-i~~ totwuk-ul cap-ass-ta.
~~police-NOM~~ thief-ACC catch-PST-SE
'(The police) caught the thief.'

c. Omission: predicate
kyengchal-i totwuk-ul ~~cap-ass-ta~~.
police-NOM thief-ACC ~~catch-PST-SE~~
'The police (caught) the thief.'

Of main interest in this study are two types of passive constructions in Korean: suffixal and periphrastic (Yeon, 2015).[5] A suffixal passive is formed by attaching one of the passive markers (*-i-, -hi-, -li-, -ki-*) to a verb stem with a nominative-marked subject indicating an undergoer and a dative-marked oblique indicating an actor (3). In a periphrastic passive (4), the undergoer is expressed by the nominative case marker but the actor is expressed mostly by *-ey uyhay*, not by the dative marker. This type of passive has a combination of a suffix *-e/a* and an inchoative verb *ci-* 'to become' after the verb stem. The canonical word order of these passive construction types follows an undergoer-before-actor order, and they can be scrambled (i.e., actor-before-undergoer) with varying degrees of omission of sentential components.

(3) Suffixal passive
totwuk-i kyengchal-hanthey cap-hi-ess-ta.
thief-NOM police-DAT catch-PSV-PST-SE
'The thief was caught by the police.'

(4) Periphrastic passive
chayk-i Chelswu-ey uyhay ccic-eci-ess-ta.
book-NOM Chelswu-by tear-become.PSV-PST-SE
'The book was torn by Chelswu.'

All these types of passives are rare in language use. Particularly in L2 learning-teaching contexts, presentation of the passive lacks systematicity, and instructors tend to focus on rote memorisation of limited ranges of chunks (e.g., Kim, 2019).

---

[5] Some researchers claim a lexical passive (a) as another passive construction type (e.g., Sohn, 1999; Song & Choe, 2007). This passive construction type involves no passive morphology on the verb, but the meaning of the verb (e.g., *mac-* 'to be hit') is one of affectedness. Use of markers are also the same as the suffixal passive, supporting their status as a passive.

(a) Lexical passive
Chelswu-ka Minho-eykey mac-acc-ta.
Chelswu-NOM Minho-DAT get.hit-PST-SE
'Chelswu was/got hit by Minho.'

However, there is a debate on whether the lexical passive is a genuine type of passive in Korean (e.g., Yeon, 2015), based on the idea that a passive construction should involve passive morphology, or at least constant marking designated for a passive voice, on a verb (Haspelmath, 1990; Siewierska, 2013). We thus exclude this passive construction type from our investigation.

As we focus on L1-Mandarin L2-Korean learners' use of Korean passive constructions, a brief look at Mandarin passives is noteworthy. The basic word order of Mandarin is Subject-Verb-Object (Sun & Givón, 1985), and the word order is restructured when a sentence is passivised such that the undergoer moves to the front of a sentence and the actor is placed between the undergoer (as the subject) and the verb (Li & Thompson, 1981) as in (5).

(5) *Zhangsan bei Lisi da-le.*
Zhangsan PSV Lisi hit-PRF
'Zhangsan was hit by Lisi.' (Example from Liu, 2016: 858)

Corpus findings revealed that the passive in Mandarin is far less frequent than active constructions (e.g., Xiao, McEnery, & Qian, 2006), which is consistent with the rare use of the passive voice in Korean.

L1-Mandarin L2-Korean learners encounter two central challenges when acquiring Korean passives. One comes from verbal morphology. Mandarin has a marker, *bei*, which is exclusively used to signal a passive voice (Huang et al., 2013; Li & Thompson, 1981; Liu, 2016). This marker is functionally similar to the passive suffix in Korean insofar as it indicates the passive voice in a sentence. However, the marker in Mandarin has a regular, fixed form and is not inserted into a verb, which shows a contrast to passive morphology in Korean. Particularly for the Korean suffixal passive, the passive suffixes are also used for morphological causatives (e.g., Sohn, 1999; Song, 2015), so an overlap arises. Considering the fact that the sensitivity to verbal morphology is crucial for Korean passives (e.g., Shin, 2020; Yeon, 2015), learners may have difficulty in employing the passive due to the properties of passive morphology.

The other challenge that Mandarin-speaking learners of Korean must overcome with respect to the Korean passives is case marking. Whereas Mandarin does not have case marking dedicated to the passive, Korean has two overt markers for the passive to indicate the thematic roles of each argument. Moreover, the form-function pairings of each marker involving the Korean passive are infrequent and thus atypical. To illustrate, the nominative case marker usually indicates the actor in the active transitive as in (1), but the same marker indicates the undergoer in the passive as in (3) and (4). The actor in the passive is indicated by the dative marker for the suffixal passive as in (3), which is often used to indicate a recipient in an active sentence, or is indicated by a special marker *-ey uyhay* for the periphrastic passive as in (4). Therefore, learners must discern these case-marking facts simultaneously: an argument indicated by the nominative case marker is not the actor but the undergoer in the passive, and there is a new association of the markers and the actor dedicated only to the passive.

2.3 Issues of Korean passive constructions in learner corpus research

Korean passives pose a variety of major challenges in automatic analysis of learner corpora. First, identification of the passive is tricky because core elements of the constructions such as case marking and verbal morphology are often mis-tagged or ignored in the current, open-to-public pipelines[6] in Korean. For instance, they do not distinguish clearly between the nominative case marker *-i* and the suffix *-i* (appearing after a consonant for phonological considerations), particularly when it occurs with a proper noun. Take (6) as an example.

(6) Ambiguity involving *-i*
Swukyeng-i mwe ha-y?
Swukyeng-SFX what do-SE
'What is Swukyeng doing?'

The performance of the publicly available pipelines is not stable, such that the first eojeol is often analysed as a combination of a human name 'Swukyeng' and the nominative case marker, despite the fact that *-i* in this example is not the case marker but the suffix.[7]

Korean passive constructions are also unsatisfactory with respect to recognising verbal morphology, largely due to imperfect tokenisation from the outset. For example, *ssuye* 'to be used' is often tokenised as *ssui-e* (passive morphology attached to verb stem), not *ssu-i-e* (passive morphology detached from verb stem). This results in tagging the verb *ssu-* and the passive morphology *-i* as a single verb, ignoring information about the passive morphology itself. Moreover, if the passive suffixes are detected properly by any chance, they are marked by the same language-specific POS tag (XPOS) as the suffixes for the morphological causative construction. This further renders the distinction between the two construction types unclear.[8] The suffixal passive and the morphological causative differ occasionally in terms of the number of arguments, use of case marking, and animacy involving arguments as in (7a–b). However, omission of arguments and/or markers as in (8a–b) creates confusion in distinguishing one from the other in the

---

[6] A pipeline in NLP 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.

[7] *-i* in this case could be interpreted as the nominative case marker, but this usage is limited heavily to written registers (e.g., description of action in a novel or a play script), and is thus uncommon.

[8] Indeed, similar pitfalls are observed in the Sejong corpus, which is a popular open-access dataset in Korean and is also widely used as a base corpus for the development of NLP tools for studies on Korean.

pattern-wise automatic search process with no consideration of contextual/discourse information.

(7) Structural distinction between suffixal passive (a) and morphological causative (b)
  a. *umsik-i chinkwu-eykey mek-hi-ess-ta.*
    food-NOM friend-DAT eat-PSV-PST-SE
    'The food was eaten by a friend.'
  b. *nay-ka chinkwu-eykey umsik-ul mek-i-ess-ta.*
    I-NOM friend-DAT food-ACC eat-CST-PST-SE
    'I made (my) friend eat food.'

(8) Ambiguity between suffixal passive (a) and morphological causative (b)
  a. *chinkwu-eykey mek-hi-ess-ta.*
    friend-DAT eat-PSV-PST-SE
    '(The food) was eaten by a friend.'
  b. *chinkwu-eykey mek-i-ess-ta.*
    friend-DAT eat-CST-PST-SE
    '(I) made (my) friend eat (food).'

The same manner of imperfect tokenisation also occurs in the case of the periphrastic passive, but verbal morphology *-e/a ci-* is exclusive to this type of passive, serving as a reliable cue for a morphology-based search. These limitations found in the existing pipelines[9] prevent us from relying entirely on their tokenisation and POS tagging functions, and lead us to complement the results with manual correction to some extent.

Another challenge pertains to the determination of canonicity involving the passive. One way to meet this challenge is to utilise information about relative positions of individual markers in a sentence. In other words, we can determine the canonicity of a sentence by 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,

---

[9] Despite broad descriptions about the performance of the currently available morphological analysers in Korean (e.g., Chun, Han, Hwang, & Choi, 2018; Han & Palmer, 2005; Park, Hong, & Cha, 2016; Qi, Dozat, Zhang, & Manning, 2018; Straka & Straková, 2017; Zeman, Hajič, Popel, Potthast, Straka, Ginter, Nivre, & Petrov, 2018), there is no report on their accuracy precisely touching upon these specific features. We discovered the shortcomings through informal testing with a small set of sentences engaging in the particular linguistic features that we focus on. Measuring the accuracy of automatic processing tools with respect to core linguistic features involving clause-level constructions is out of the scope of the present study, but we hope to pursue this line of research in the near future.

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 reference numbers of the strings) as long as the sentence has dedicated markers at the designated place. For instance, in the pattern *noun-DAT noun-NOM verb*-PSV, the dative marker has smaller reference numbers than the nominative case marker, which indicates that the dative marker occurs prior to the nominative case marker. We thus classify 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-DAT verb*-PSV as an example: the dative marker occurs after any noun, and this characteristic results in larger numeric values for the dative marker than any noun has, which allows this pattern to be classified as the canonical suffixal passive. There are very few instances where two case markers are dropped altogether for the two construction types (e.g., Chung, 1994), so we do not consider this possibility for now. However, one caveat to this approach is that its application is less promising in multi-clause sentences, and this necessitates manually examining the automatic processing results.

The last challenge, omission of sentential components, is understood as a major hardship 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 verb dictionary with information about semantic features of obligatory arguments tied to particular predicates (e.g., Lee & Choi, 2013). The rates of accuracy reported from these studies, all of which targeted general-purpose L1-Korean corpora, varied from around 70 up to 95 percent. However, there is no empirical report on the application of these proposals to learner corpora in Korean. We may set aside this particular challenge for now.

With these challenges in mind, we conduct an automatic analysis of learner writing by adapting NLP techniques. Our particular focus lies in reporting *how* we pursue this task, by asking two specific questions about the automatic processing: what practical issues arise in conducting this work, and what we can manage under the particular treatments that we conduct.

3. Methods

3.1 Learner corpus creation

To create learner corpora, we collected essays from 36 Mandarin-speaking learners of Korean who attended a university in Korea (mean age = 24.2, *SD* = 3.1). The duration of the learners' experience learning Korean varied from two months to more than seven years (mean year = 3.1, *SD* = 2.7). We also measured learner proficiency separately by using the Korean C-test (Excerpts 1 to 4; Lee-Ellis, 2009). In addition to learner writing, we collected essays from 10 native Korean speakers (mean age = 27.5, *SD* = 2.9) as a reference corpus.

Participants were asked to write argumentative essays about two topics, adapted from Test of Proficiency in Korean tests (Topic 1: 'Which do you think is the most important, preservation vs. exploitation of the nature?'; Topic 2: 'What affects success the most, competition or cooperation?'), on a separate sheet of paper, for 20 minutes per essay. The prompts were presented both in Korean and Chinese for the participants' clear understanding of these topics. The two trials of essay writing were interspersed with the four proficiency test excerpts (Excerpts 1 & 2 → writing → Excerpts 3 & 4 → writing); no mobile device was allowed during the entire session. The entire participation took 1.5 hours, and every participant received monetary compensation (approximately $ 10) for their participation.

All the essays were converted into an electronic format (*.txt* file), with typos and errors uncorrected. There was no direct use of the prompts in the essays, so we decided not to delete prompt-like sentences from the essays (if any). This conversion was done by two native Korean speakers, and we verified that their conversions were identical. We then sorted out the two types of passives from the data manually, as a reference for comparison with the automatic extraction results. The legitimate passive instances that we extracted from the data included the case-marking and verbal morphology that we explained in Section 2, with or without the omission of sentential components. The hand-extracted results were also cross-validated by two native speakers of Korean; they reached complete agreement on the manual results. Table 1 provides a summary of the learner corpora that we created.

Table 1. Information about learner corpora

Topic	Size (eojeol)		
	Mean (SD)	Minimum	Maximum
1	107 (36.36)	62	201
2	113 (38.48)	57	203

We note that, although there exists a sizable learner corpus made by the National Institute of Korean Language (NIKL) which was publicised in 2018 (2,021,991 eojeols from 15,983 written essays; 579,391 eojeols from 1,251 oral interviews and presentations), some grave issues involving the dataset inhibited us from relying on this pre-made corpus for our investigation. First of all, there is no way for researchers to ensure whether learners were given the original prompts or if the instructors/collectors modified them. NIKL distributed a separate Excel file ('Learner Corpus Sampling Information') containing information about each essay. In that file, the essays are classified into groups which share the same topics such as (9a) but not the actual prompts that are recommended in the guideline such as (9b).

(9) a. Example of topic (from 'Learner Corpus Sampling Information' by NIKL; translated in English)
'Future plans', '10 years later'
b. Example of prompt (from '2017 Korean Language Learner Corpus Establishment Guidelines' by NIKL; translated in English)
What will your life be like 10 years from now? Why do you think so? Write an essay about this, with the title as 'my plan in 10 years', including the following points: 'How do you see yourself 10 years from now?', 'Why do you think so?', 'What should you prepare?'.

No explanation was provided about whether these topics are shorthand for each prompt, and this renders the impacts of prompts on learner writing uncontrollable in an actual analysis. A possibility thus arises in which the instructors/collectors extracted topics from the standard prompts, presented the topics (not the prompts) to learners, and labelled learner essays with the same topics. This suspicion is justified by the observation that essays under the same topic do not always fit into the specifications of the prompt. Considering that prompts serve as one good source for understanding characteristics of learner writing (e.g., Cho, 2019; Miller, Mitchell, & Pessoa, 2016), it is unfortunate that the NIKL corpus fails to satisfy the rigour of the prompt issue.

Another important issue with using this NIKL learner corpus lies in how the data were produced and collected. Given the current descriptions provided by NIKL, it is unclear whether the learner essays in the corpus were collected on the spot without any revision and support from other materials. There is a possibility that the learners either revised their initial writing (possibly reflecting comments and suggestions from the instructors/collectors) or wrote the essays with the help of resources that are normally unavailable for on-site essay writing. Indeed, we found some examples of essays including statistical data or detailed chronological information, which often require technical references. NIKL provides no explanation on this point, and this does not guarantee that learner writing in the NIKL data fits with our intention in this study.

Apart from the fact that manual coding of this large dataset for the gold annotation set requires a considerable amount of human resources, which is not viable for now, the aforementioned issues were not controlled properly – or at least, NIKL does not fully report how they controlled the issues. This aspect prohibited us from using their corpus data, which led us to collect and utilise our own learner writing data.

3.2 Pattern extraction: Passive constructions in leaner writing

Because we are not aware of any previous work on an automatic tool for analysing L2-Korean learner corpora, we employed a recently proposed programme by Shin (forthcoming) based on Korean child corpora from tokenisation up to tagging of XPOS (the Sejong tag in this study) and UPOS (the universal POS tag set; Petrov, Das, & McDonald, 2012).[10] Because of the challenges that we mentioned in Section 2.3 (also reported in Shin, forthcoming), the performance of the existing open-to-public pipeline was not satisfactory for the pattern extraction task. We thus revised the initially tagged data (through *UDPipe*; Straka & Straková, 2017) manually to ensure that each morpheme and word was assigned to an appropriate tag. During this revision, we focused on correcting tokenisation and tag information about case-marking and verbal morphology, which are crucial for the extraction of the passive but often mis-analysed in the currently available pipelines for Korean.

The tagged data were then inputted to a pattern extraction process. All the information about individual morphemes and their corresponding tags in one sentence was transformed into a single string on an eojeol-by-eojeol basis as in (10).

---

[10] Shin (forthcoming) developed a Python-based pattern-finder to investigate the use of various clause-level constructions in caregiver input and child production from CHILDES (with around 70,000 sentences after pre-processing). The overall performance of this programme was decent, with F1 scores ranging from 0.714 to 0.955 depending on construction types, given the characteristics of child corpora such as partial/incomplete utterances and repetition of onomatopoeia and mimetic words. Despite its satisfactory level of accuracy, we acknowledge that more L2 data (with a manually-coded golden set) are needed to further verify its applicability to learner corpora. See Shin (forthcoming) for detailed descriptions on how this pattern-finder applied to Korean child corpora and what challenges occurred in the development/application of this programme.

(10) Example of a sentence for pattern extraction

a. Original sentence
자연보존이 더 중요하다고 생각한다.
cayenpoconi te cwungyohatako sayngkakhanta.
cayen.pocon-i te cwungyo.ha-ta-ko sayngkak.ha-n-ta.
nature.preservation-NOM more importance.do-SE-CON thought.do-PRS-SE
'(I) think nature preservation is more important.'

b. Converted sentence
자연보존이/자연+보존+이/NNG+NNG+JKS/NOUN 더/더/MAG/ADV
cayenpoconi/cayen+pocon+i te/te
중요하다고/중요+하+다+고/NNG+VV+EF+EC/ADJ
cwungyohatako/cwungyo+ha+ta+ko
생각한다/생각+하+ㄴ다/ NNG+VV+EF/VERB./././SF/PUNCT
sayngkakhanta/sayngkak+ha+nta
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 entered into an automatic search process whereby the two construction types were extracted in the following steps, as schematised in Figure 1. First, we sorted out sentences with verbs (VV) that included key morphology and corresponding tagging information involving these constructions: a passive suffix (*-i-/-hi-/-li-/-ki-*) for the suffixal passive, and *-e/a ci-* for the periphrastic passive. We then verified that the markers dedicated to the passive appeared in the construction and calculated the numeric locations of these markers within each sentence (see Section 2.3). Finally, we classified these extracted instances into three categories per construction type based on information about these reference numbers: canonical (the nominative case marker preceded the other markers for each passive construction type and/or another noun under the UPOS tagging [NOUN]); scrambled (the nominative case marker followed the other markers for each passive construction type and/or another noun under the UPOS tagging [NOUN]); and undetermined (all the remaining instances which did not fall into the two categories due to omission of arguments and/or markers). Lists of instances were outputted into *.txt* files, and every list for each extraction was checked manually to ensure the accuracy of the results.

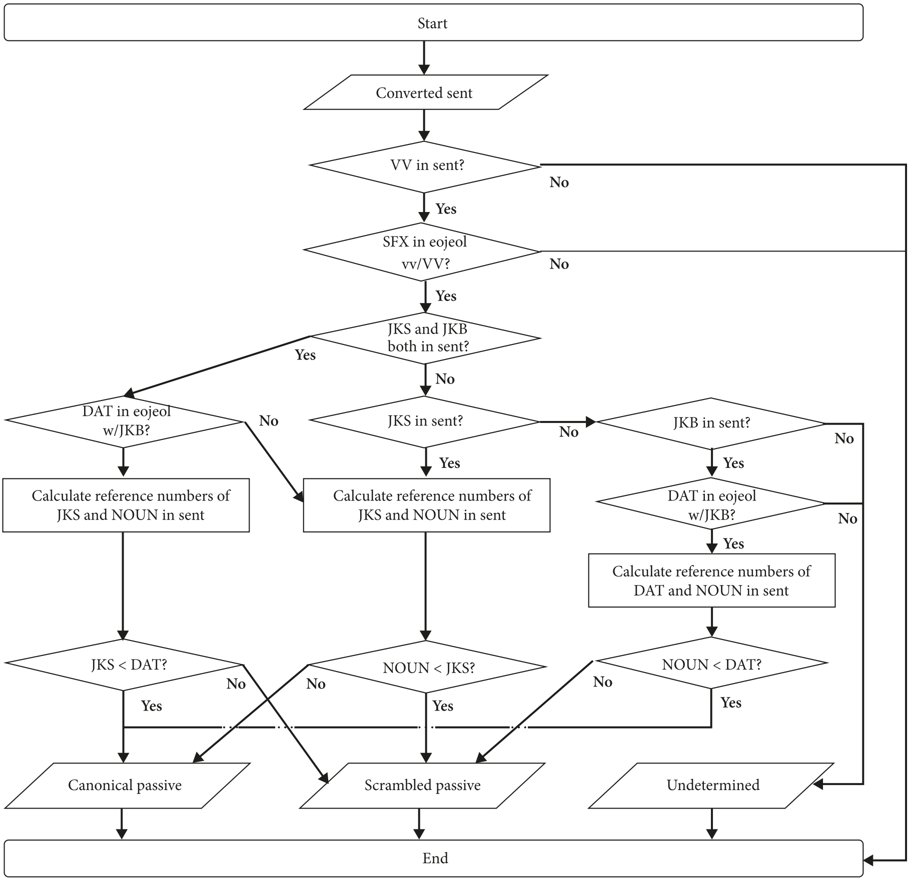

Figure 1. Flow of pattern extraction: Suffixal and periphrastic passive constructions
Note. 'sent' stands for a sentence. DAT and SFX are not search terms but cover terms (used only in this flow chart) representing key morphemes used in the passive (DAT = *-eykey*, *-hanthey*, *-ey (uyhay)*; SFX = *-i-/-hi-/-li-/-ki-* for the suffixal passive, *-e/a ci-* for the periphrastic passive). JKB = XPOS for adverbial marker in general (DAT is one type of JKB); JKS = XPOS for nominative case marker; NOUN = UPOS for noun; VV = XPOS for verb. The horizontal line from the 'NOUN < DAT?' diamond goes straight to the 'canonical passive' parallelogram and is not related to any line towards the 'scrambled passive' parallelogram.

Suppose that we have (3), re-stated as (11a), that we input to the search process. This sentence is converted into a single string with relevant information, as shown in (11b).

(11) Example sentence: suffixal passive
a. *totwuk-i kyengchal-hanthey cap-hi-ess-ta.*
thief-NOM police-DAT catch-PSV-PST-SE
'The thief was caught by the police.'
b. 도둑이/도둑+이/NNG+**JKS**/NOUN
totwuki/totwuk+i
경찰한테/경찰+**한테**/NNG+**JKB**/ADV
kyengchalhanthey/kyengchal+hanthey
잡혔다/잡+히+었+다/**VV**+**XSV**+EP+EF/VERB./../SF/PUNCT
caphyessta/cap+hi+ess+ta

The sentence has the VV tag and an eojeol with VV has the marker *-hi-*, so the programme sequentially looks into whether the sentence involves JKS and JKB and whether JKB is one of the dative markers for the passive. After the sentence passes the two steps, the programme then calculates and compares the numeric reference values of the nominative case marker and the dative marker. Because the value of the nominative case marker is smaller than that of the dative marker, the sentence is classified as the canonical suffixal passive.

We included in our analysis any instance that engages in passive morphology, based on the typological observation that a passive construction has verbal morphology dedicated to a passive voice (Haspelmath, 1990; Siewierska, 2013). Manual inspection revealed that all legitimate instances of passive constructions were grammatical.

The reference corpus – the essays from the 10 native speakers of Korean – was processed according to the same procedures as the learner writing data. It consisted of 3,243 eojeols, with a mean length of 162.15 eojeols per essay.

3.3 Evaluation

We measured how accurately the two passive construction types were extracted under the current automatic pattern extraction system by calculating 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). In addition to the F1 score, we compared the performance of automatic extraction with manual extraction to see how reliable the pattern-finder operated.

4. Results

Table 2 presents the accuracy of the automatic extraction of the two passive constructions from learner writing.

Table 2. Results: Manual extraction vs. automatic extraction (learner writing)

	Suffixal passive		Periphrastic passive	
	Manual	**Automatic**	**Manual**	**Automatic**
Instances (#)	8	12	11	11
Precision	0.667		1.000	
Recall	1.000		1.000	
F1 score	0.800		1.000	

Note. Passive constructions were attested mostly in Topic 1 (seven instances for the suffixal passive; eight instances for the periphrastic passive).

We obtained more potential instances of the suffixal passive through the automatic pattern extraction process than the exact number of passive instances from the manual extraction. The result seems to be understandable in that the verbal morphology of the suffixal passive (*-i-/-hi-/-li-/-ki-*) overlaps with that of the morphological causative construction in Korean. Indeed, there were four morphological causative instances that were classified into the suffixal passive. This overlap is something that cannot be resolved under our pattern identification scheme that considers information about individual eojeols and their corresponding POS tags.

In contrast, we found a perfect match between automatic extraction and manual extraction in the case of the periphrastic passive. This was possible because verbal morphology *-e/a ci-* is exclusively used for this construction type, thus serving as a clear criterion for extraction. The fact that we circumscribed the type of predicates for pattern extraction into a verb (with the VV tag) also contributed to this success – this morphology can also be attached to an adjective, which changes it to an intransitive verb.

Across the two construction types, we found no false negatives (i.e., items that should be included in the target category but are actually excluded), but false positives (i.e., items that should not be included in the target category but are actually included). This suggests that our pattern-finder needs improvement with respect to how to exclude non-target items automatically amongst what it extracts.

Unfortunately, however, we failed to determine the canonicity of each instance through the current automatic pattern identification scheme. No instance fell into the intended categories (canonical and scrambled) accurately, such as (12), and so we had to determine whether or not the instances followed the canonical word order manually. This failure is attributed to sentence-level complexity (e.g., multi-clause sentences and omission of sentential components) and word-level complexity (e.g., multiple form-function mapping in postpositions; cf. Choo & Kwak, 2008), which indicates that additional consideration is necessary in the automatic detection of canonicity.

(12) Example of unsuccessful classification of sentence by canonicity (clipped from a multi-clause sentence)
어느 섬이 관광객에게 열리지 않는다고
enu sem-i kwankwangkayk-eykey yel-li-ci anh-nun-ta-ko
an island-NOM tourist-DAT open-PSV-NML not-PRS-SE-CON
들었다
tul-ess-ta.
hear-PST-SE
'(I) heard that an island is not opened to tourists.'
– correct classification: suffixal passive, undetermined
– actual classification: suffixal passive, canonical

Similar results were found in the native speakers' writing (Table 3): our pattern-finder extracted more instances of the suffixal passive than necessary, and the number of periphrastic passive instances that the pattern-finder extracted was the same as for the manual annotation. The accuracy for the suffixal passive was lower than the accuracy found in learner writing above (0.800). This was because the native speaker participants produced morphological causatives (seven instances) more often than the L2 learner participants did.

Table 3. Results: Manual extraction vs. automatic extraction (native speaker writing)

	Suffixal passive		**Periphrastic passive**	
	Manual	**Automatic**	**Manual**	**Automatic**
Instances (#)	8	15	13	13
Precision	0.533		1.000	
Recall	1.000		1.000	
F1 score	0.696		1.000	

Note. Whereas the suffixal passive was attested mostly in Topic 1 (six instances), the periphrastic passive was attested mostly in Topic 2 (10 instances).

Although we acknowledge that it is difficult to generalise our findings due to the limited size of the learner corpus data and the limited number of passive instances, we conducted a by-proficiency analysis to see any developmental aspects in relation to passive constructions. Table 4 presents the number of instances that each group produced for the two types of passive constructions. We set the advanced learner group and the novice learner group arbitrarily, by aggregating the 10 highest-proficiency learners (for the advanced learner group) and the 10 lowest-proficiency learners (for the novice learner group) based on the proficiency scores. The production of the passive was proportionate to learner proficiency in general. This is consistent with previous findings from behavioural experiments (e.g., Jeong, 2014) that suggest that L2 learners' command of the passive is contingent on their proficiency of the target language. Our finding is also indicative of a possibility for the passive to function as a predictor of L2 development, which requires further investigation with more instances of the passive from larger learner corpora.

Table 4. Production of passive constructions by proficiency

Group	Suffixal passive		Periphrastic passive	
	Raw	Normed	Raw	Normed
Native	8	2,561	13	4,161
Advanced	4	1,468	3	1,101
Novice	1	617	1	617

Note. The raw frequency values were normed per one million words.

As for the passive construction types, our learners demonstrated numerically more frequent use of the periphrastic passive than the suffixal passive. Interestingly, whereas verb use in the periphrastic passive was skewed towards *eps*- 'to not exist' (eight of the 11 instances), the suffixal passive exhibited diverse types of verbs in the learner productions, as Table 5 shows. It is unclear at this stage whether this (non-)skewedness of verb use in each passive construction type was due to topic effects or learner proficiency. We presume that the reason may be interactions between linguistic properties involving the two passive construction types (i.e., use of the suffixal passive is limited to a set of verbs; Sohn, 1999) and input to which the learners are exposed (cf. Shin, 2020); but again, we admit that this way of reasoning is currently speculative.

Table 5. By-construction use of verb in learner writing

Suffixal passive		Periphrastic passive	
Verb	#	Verb	#
po- 'to see'	3	*eps*- 'to not exist'	8
yel- 'to open'	2	*ilwu*- 'to achieve'	1
camku- 'to sink'	1	*mangha*- 'to ruin'	1
tal- 'to hang'	1	*yeki*- 'to regard'	1
tul- 'to lift'	1		

5. Discussion and conclusion

5.1 Implications of findings

Motivated by the lack of research on automatic processing of clause-level constructions such as passive constructions in Korean learner corpora, the present study conducted an NLP-assisted analysis of L1-Mandarin L2-Korean learners' written production by focusing on their use of the two passive construction types in Korean. We reported pevious corpus-based research on L2 Korean, language-specific properties that pose challenges in the automatic processing of learner writing with respect to the passive voice in Korean, and possible ways to deal with these challenges in (semi-)automatic processing of L2-Korean learner corpora. We then reported an automatic pattern extraction process, constructing our own small-scale corpus of learner writing with revised tokenisation and tagging information.

We obtained a good level of accuracy in extracting the two passive construction types relative to the manual extraction result. Despite the overlap in verbal morphology between the suffixal passive and the morphological causative, the automatic extraction of this construction type was relatively satisfactory. The accuracy level of our pattern extraction process for the suffixal passive was even more accurate for the L2 data than for the L1 data, although their production of this passive construction type was numerically not so different from each other. In contrast, the automatic extraction of the periphrastic passive demonstrated a perfect match with the manual extraction, regardless of L1 or L2 data. This was possible due to the particular case-marking and verbal morphology which are dedicated to this passive construction and do not overlap with components of other constructions.

These findings suggest that the performance of our construction identification scheme may have benefitted from the tokenisation and (X)POS tagging information that we modified in the pre-processing stage with special emphasis on case-marking and verbal morphology (the two core elements of Korean passives). This stands as an indication that, as long as learner corpora are properly tokenised and tagged, one can successfully identify the two Korean passive constructions from the corpora under the current pattern extraction process that we proposed.

We also conducted a preliminary analysis of learner writing by proficiency. Despite the small size of the learner writing data, we found a tendency of the advanced learners to use the passive proportionally more than the novice learners. A passive construction is one of the complex argument structure constructions acquired later in language development in general (e.g., Shin, 2020). Moreover, cross-linguistic differences between learners' L1 (Mandarin) and L2 (Korean) were evident in the passive (see Section 2.2). Considering these acquisitional challenges and typological differences involving the passive, our findings lend indirect support to the role of passive constructions in explaining L2-Korean proficiency (albeit less generalisable due to the small corpus size).

5.2 Limitations and future directions

Although we found benefits and potential applications for automatic processing of learner corpora, there are still drawbacks of the NLP-assisted analysis of learner corpora in Korean, which await further investigation.

First of all, we could not demonstrate full-fledged automatic processing of learner writing. The currently available pipelines for data analysis are mostly based on general-purpose L1 corpora, and so they may not be ideal for analysing learner corpora (cf. Meurers & Dickenson, 2017). It is widely known that learner language is qualitatively different from how the target language is used natively (e.g., Meurers & Dickenson, 2017). Hence, researchers should be aware of the possibility that premade NLP tools (developed mostly using L1 corpora) do not comply with features of learner language such as spelling/spacing errors and novel combinations of words and chunks, which possibly aggravates the performance of the tools in exploring linguistic features of interest in learner corpora. The current study did not identify particular issues with the aforementioned characteristics of learner language for the automatic pattern extraction process, but we believe that it was because we focused only on specific construction types. Given the need for the application of NLP techniques to learner corpora to accommodate characteristics of learner language itself, more applications of our scheme to various (learner) corpora in Korean are required to verify the effectiveness of our approach to this kind of task, which we plan to do next.

These open-access pipelines often fall short of their performance in regard to tokenisation and POS tagging of sentential components that are crucial for pattern-wise analysis of corpus data in Korean (see Sections 2.3 and 3.2). This suggests that, unless the performance of the currently available L1-based automatic tools are improved with respect to the tokenisation and tagging issues, learner corpus researchers would still have difficulty in coping with these issues. Together, these shortcomings of the open-to-public pipelines (which are oriented heavily to properties of Korean) render their full applicability to the analysis of written corpora (whether they be L1 data or L2 data) less promising. This in turn necessitates manual inspection to some extent, just as we did. One meaningful investigation in this respect would be to compare the performance of these pipelines synchronically and to see how different pipelines can affect the pattern extraction procedure, which provides an important venue for future research.

Particularly for the passive constructions in Korean, language-specific challenges involving the passive add difficulty to this automatic extraction of the target patterns. To illustrate, the omission of sentential components such as arguments and case marking – which often occurs in Korean – makes it difficult to determine the canonicity of instances. In addition, there is a morphological overlap between passive suffixes in the suffixal passive and causative suffixes in the morphological causative, which is not detectable under the current scheme. Clausal complexity in learner production also adds to the difficulty in general. Chances are that learner corpus analysis may utilise probabilistic dependency relations, as several studies on L1 Korean report decent performance using dependency information (e.g., Park, Hong, & Cha, 2016), or an additional procedure that converts multi-clause sentences into mono-clause ones by using morphemes indicating a clausal boundary. Unfortunately, there is no verified L2-Korean analysis tool that deals with these language-specific issues at a satisfactory level.

To bypass these issues, we took a semi-automatic approach to pattern extraction, but we acknowledge that ours is not the ultimate solution. Future research will benefit from measuring the degree to which cutting-edge methods for general-purpose corpora alleviate these challenges pertaining to automatic processing of Korean learner corpora. Subsequent studies, both on L1 and L2 Korean, need to pursue these lines of inquiry for a better NLP-assisted tool that effectively copes with the aforementioned challenges in the automatic processing of (learner) corpora in Korean.

Lastly, the present study did not address proficiency-related issues pertaining to automatic processing of learner writing in a detailed manner. This is because too few instances of passive constructions prevents us from inferring clear/strong implications on automatic processing of learner writing and proficiency. Given the findings of this study, we may speculate on the relationship between the performance of our automatic pattern identification scheme and by-proficiency use of pssive constructions. For example, considering that native speaker participants produced more morphological causative instances than L2 participants produced, we may claim that the performance of our pattern-finder decreases as proficiency increases. This is something that we cannot verify for now, again due to the rare use of the passive, and will require further research in order to better estimate production of the Korean passive across learner proficiency using larger corpora with various genres.

The automatic processing of learner corpora in Korean and its application to L2 research on Korean are still in their infancy. We believe that, as the first empirical report on these topics, the findings of this study open the door to potential ways of (and directions towards) NLP-assisted learner corpus research on Korean.

Acknowledgments

This study was supported by the European Regional Development Fund through the 'Sinophone Borderlands – Interaction at the Edges' project (CZ.02.1.01/0.0/0.0/16_019/0000791).

Abbreviations

ACC	accusative case maker	PRS	present tense marker
CON	connector	PRF	perfective marker
CST	causative suffix	PST	past tense marker
DAT	dative marker	PSV	passive suffix
NOM	nominative case marker	SE	sentence ender
NML	nominaliser suffix	SFX	suffix

References

Abbot-Smith, K., Chang, F., Rowland, C., Ferguson, H., & Pine, J. (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

Bang, D.-S. (2014). hankwuke kokup haksupcauy ssukiey nathananun hancae olyu pwunsek [A study of sino-Korean errors found in advanced Korean learners]. *kwukhakyenkwulonchong*, 14, 1–21.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26, 28–41. https://doi.org/10.1016/j.jslw.2014.09.004

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?. *TESOL Quarterly*, 45(1), 5–35. https://doi.org/10.1002/(ISSN)1545-7249

Birjandi, P., Maftoon, P., & Rahemi, J. (2011). VanPatten's processing instruction: Links to the acquisition of English passive structure by Iranian EFL learners. *European Journal of Science Research*, 64(4), 598–609.

Brooks, P. J., & Tomasello, M. (1999). Young children learn to produce passives with nonce verbs. *Developmental Psychology*, 35, 29–44. https://doi.org/10.1037/0012-1649.35.1.29

Cho, S., & Park, Y. (2018). Sheffield tayhakkyo hankwuke haksupcauy cakmwun thukseng pwunsek [Characteristics of Korean language writing by students at the University of Sheffield]. *cakmwunyenkwu*, 38, 149–172. https://doi.org/10.31565/korrow.2018.38..006

Cho, Y. (2019). The effects of writing prompt types on L2 learners' writing strategy use and performance. *Studies in English Language & Literature*, 45(3), 295–314. https://doi.org/10.21559/aellk.2019.45.3.013

Choi, B.-S. (2018). oykwukin yuhaksaynguy kwanhyengcel silhyen yangsang yenkwu – cakmwun calyolul cwungsimulo [A study on the aspects of the Korean adnominal clause of overseas students – focused on using writing]. *hanmincokemwunhak*, 79, 61–95. https://doi.org/10.31821/HEM.79.3

Choi, J. D., & Palmer, M. (2011, October). Statistical dependency parsing in Korean: From corpus generation to automatic parsing. In D. Seddah, R. Tsarfaty, & J. Foster (Eds.), *Proceedings of the 2nd Workshop on Statistical Parsing of Morphologically-Rich Languages* (pp. 1–11). Stroudsburg: Association for Computational Linguistics.

Choo, M., & Kwak, H.-Y. (2008). *Using Korean*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139168496

Chun, J., Han, N.-R., Hwang, J. D., & Choi, J. D. (2018). Building Universal Dependency Treebanks in Korean. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32, 1–16. https://doi.org/10.1016/j.jslw.2016.01.003

Cui, Y., & Wang, J. (2018). cwungkwukin haksupcatuluy pocoyongen sayongyangsang kochal – cakmwun pwunsekul thonghan sayongpinto cosa mich olyu pwunsek [Exploring the use of Korean auxiliary verbs among Chinese learners]. *Teaching Korean as a Foreign Language*, 51, 175–202. https://doi.org/10.21716/TKFL.51.175

Dąbrowska, E., & Street, J. (2006). Individual differences in language attainment: Comprehension of passive sentences by native and non-native English speakers. *Language Sciences*, 28(6), 604–615. https://doi.org/10.1016/j.langsci.2005.11.014

de Felice, R., & Pulman, S. (2009). Automatic detection of preposition errors in learner writing. *Calico Journal*, 26(3), 512–528. https://doi.org/10.1558/cj.v26i3.512-528

de Haan, P. (2000). Tagging non-native English with the TOSCA–ICLE tagger. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory* (pp. 69–79). Amsterdam: Rodopi.

de Mönnink, I. (2000). Parsing a learner corpus. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory* (pp. 81–90). Amsterdam: Rodopi.

Ellis, N. C., & Ferreira-Junior, F. (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

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: identifying, comparing, and interpreting the evidence. *Language Learning*, 67(S1), 155–179. https://doi.org/10.1111/lang.12225

Gilquin, G. (2008). Hesitation markers among EFL learners: Pragmatic deficiency or difference. In J. Romero-Trillo (Ed.), *Pragmatics and corpus linguistics: A mutualistic entente* (pp. 119–149). Berlin: Mouton de Gruyter.

Goldberg, A. E. (1995). *Constructions: a construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Han, C.H., & Palmer, M. (2005). A morphological tagger for Korean: Statistical tagging combined with corpus-based morphological rule application. *Machine Translation*, 18(4), 275–297. https://doi.org/10.1007/s10590-004-7693-4

Haspelmath, M. (1990). The grammaticization of passive morphology. *Studies in Language*, 14(1), 25–72. https://doi.org/10.1075/sl.14.1.03has

Hinkel, E. (2004). Tense, aspect and the passive voice in L1 and L2 academic texts. *Language Teaching Research*, 8(1), 5–29. https://doi.org/10.1191/1362168804lr132oa

Huang, Y.T., Zheng, X., Meng, X., & Snedeker, J. (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

Huh, C.-G. (2018). cwungkwukin haksupcauy kulssukiey nathanan hankwuke eswunuy haksup yangsang [The aspects of learning word order of Korean language of Chinese learners]. *tonamemwunhak*, 34, 255–290. https://doi.org/10.17056/donam.2018.34..255

Izumi, S., & Lakshmanan, U. (1998). Learnability, negative evidence and the L2 acquisition of the English passive. *Second Language Research*, 14(1), 62–101. https://doi.org/10.1191/026765898675700455

Jeong, H. (2014). Processing and acquisition of Korean passive voice by Chinese L2 learners. hankwuke kyoyuk [Korean Education], 25(2), 165–186.

Ju, M.K. (2000). Overpassivization errors by second language learners: The effect of conceptualizable agents in discourse. *Studies in Second Language Acquisition*, 22(1), 85–111. https://doi.org/10.1017/S0272263100001042

Kim, H., & Rah, Y. (2016). Effects of verb semantics and proficiency in second language use of constructional knowledge. *The Modern Language Journal*, 100(3), 716–731. https://doi.org/10.1111/modl.12345

Kim, H., Shin, G.-H., & Hwang, H. (2020). Integration of verbal and constructional information in the Second Language processing of English dative constructions. *Studies in Second Language Acquisition*, 42(4), 825–847. https://doi.org/10.1017/S0272263119000743

Kim, H.-G., Kang, B.-M., & Hong, J. (2007). 21seyki seycongkyeyhoyk hyentaykwuke kichomalmwungchi sengkwawa cenmang [21st century Sejong modern Korean corpora: Results and expectations]. In Korean Institute of Information Scientists and Engineers (Ed.), *Proceedings of Annual Conference on Human and Language Technology 31* (pp. 311–316). Korean Institute of Information Scientists and Engineers.

Kim, J.Y., Park, Y.H., Kim, M.J., Kim, H.N., Choi, S.K., Suh, J.H., & Kwak, Y.J. (2016). hankwuke haksupcauy cakmwun malmwungchilul hwalyonghan mwunhyeng yonglyey kemsaykki kaypal yenkwu [A study of developing usage searcher of grammar pattern in the Korean learner's writing corpus]. *Teaching Korean as a Foreign Language*, 44, 131–155. https://doi.org/10.21716/TKFL.44.5

Kim, S.J., & Kim, S.H. (2013). yeseng kyelhonimincauy kwueey nathanan tamhwaphyoci sayong yangsang yenkwu [A study on the use aspects of discouse markers appeared in spoken Korean language of marriage woman immigrants]. *The Journal of Linguistics Science*, 64, 25–46.

Kim, Y.-I. (2019). hankwuke kyocayuy '-i/hi/li/ki-' phitong tanwen pwunsekkwa kyoswu pangan ceysi [Analysis and Teaching Methods of the '-i/hi/li/gi-' Passive unit in Korean textbooks]. *Journal of Korean Language Education*, 30(1), 27–63. https://doi.org/10.18209/iakle.2019.30.1.27

Kim, W., & Ock, C.Y. (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.

Kwak, S.J. (2016). mikwuk nay tayhak haksupcatuluy cakmwun pwunsekul thonghan hankwuke swuktalto swucwunpyel pikyo yenkwu – yuthatay salyeyyenkwu [A comparative study of American university students' Korean proficiency by level through analysis of composition: A case study at the University of Utah]. *Teaching Korean as a Foreign Language*, 44, 23–51. https://doi.org/10.21716/TKFL.44.2

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication (Unpublished doctoral dissertation). Georgia State University, Atlanta.

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, 34(4), 513–535. https://doi.org/10.1177/0265532217712554

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior Research Methods*, 50(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing*, 26(2), 245–274. https://doi.org/10.1177/0265532208101007

Lee, I. (2011). kwukehakkaysel [Introduction to Korean linguistics]. Seoul: Hakyensa.

Lee, J. (2017). tayhaksayng cakmwuney nathanan ehwi tayangseng yenkwu – oykwukin yuhaksayngkwa hankwukinuy pikyolul cwungsimulo [A studyon lexical diversities in writing of university students – Focusing on the comparison of Koreans and foreign students]. *tayhakcakmwun*, 19, 61–91. https://doi.org/10.37736/kjlr.2017.03.19.61

Lee, S.-M. (2017). hankwuke haksupcauy malhakiwa ssukiey nathanan ehwi sayonguy congtancek yenkwu [A longitudinal study of vocabulary usage presented in speaking and writing of Korean learners]. *wulimalkul*, 74, 183–214. https://doi.org/10.18628/urimal.74..201709.183

Lee, S.-H., Dickenson, M., & Israel, R. (2016). Challenges of learner corpus annotation: Focusing on Korean Learner Language Analysis (KoLLA) system. *Language Facts and Perspectives*, 38, 221–251. https://doi.org/10.20988/lfp.2016.38..221

Lee, S.K. (2007). Effects of textual enhancement and topic familiarity on Korean EFL students' reading comprehension and learning of passive form. *Language Learning*, 57(1), 87–118. https://doi.org/10.1111/j.1467-9922.2007.00400.x

Lee, S.-A., & Choi, J.-T. (2013). hankwuke Verb_OntoNetuy selkyeywa kwuchwuk [Design and implementation of Korean Verb_OntoNet]. *Journal of Korean Institute of Information Technology*, 11(2), 161–167.

Li, C., & Thompson, S. (1981). *Mandarin Chinese: A functional reference grammar*. Berkeley, CA: University of California Press.

Liu, N. (2016). The structures of Chinese long and short *bei* passives revisited. *Language and Linguistics*, 17(6), 857–889. https://doi.org/10.1177/1606822X16660938

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Meurers, D. (2015). Learner corpora and natural language processing. In S. Granger, G. Gilquin, & F. Meunier (Eds.). *The Cambridge handbook of learner corpus research* (pp. 537–566). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.024

Meurers, D., & Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, 67(S1), 66–95. https://doi.org/10.1111/lang.12233

Miller, R., Mitchell, T., & Pessoa, S. (2016). Impact of source texts and prompts on students' genre uptake. *Journal of Second Language Writing*, 31, 11–24. https://doi.org/10.1016/j.jslw.2016.01.001

Nam, J., Kim, Y., & Kim, Y. (2016). L2 hankwuke mwune sanchwuleyseuy thongsa pokcapseng chukceng [Measuring syntactic complexity in L2 Korean writings]. *Korean Semantics*, 51, 21–56. https://doi.org/10.19033/sks.2016.03.51.21

Nam, Y. J., & Hong, U. P. (2014). L2loseuy hankwuke cayenpalhwa khophesuuy kwuchwukkwa hwalyong [Towards a corpus-based approach to Korean as a second language]. *The Journal of the Humanities for Unification*, 57, 193–220. https://doi.org/10.21185/jhu.2014.03.57.193

Park, E., & Cho, S. (2014). KoNLPy: swipko kankyelhan hankwuke cengpocheli phaissen phaykhici [KoNLPy: Korean natural language processing in Python]. *cey26hoy hankul mich hankwuke cengpocheli hakswultayhoy nonmwuncip*.

Park, H.-J., & Lee, M.-H. (2017). hankwuke haksupcauy ssuki theyksuthuey nathanan ungkyelsengkwa ungcipsenguy sangkwanpwunsek [Correlation analysis of cohesion and coherence in Korean as a second language student's writing]. *Wulimalkul*, 73, 133–157. https://doi.org/10.18628/urimal.73..201706.133

Park, J., Hong, J. P., & Cha, J. W. (2016). Korean language resources for everyone. In J. C. Park & J.-W. Chung (Eds.), *Proceedings of the 30th Pacific Asia conference on language, information and computation: Oral Papers* (pp. 49–58).

Park, Y.-H., & Lee, H.-W. (2014). hankwuke haksupcalul wihan hankwuke mwuncang kwuseng kyoyuk pangan yenkwu – cwungkwukin haksupcauy eswuney ttalun kulssuki olyu pwunsekul thonghaye [A study on effective teaching strategies for Korean language writers through error analysis]. *Studies in Linguistics*, 33, 159–174. https://doi.org/10.17002/sil..33.201410.159

Park, Y.-K., Kim, J.-M., Lee, S.-D., & Lee, H. A. (2017). oykwukin haksupcalul wihan mwunmayk kipan silsikan kwuke mwuncang kyoceng [Context Based Real-time Korean Writing Correction for Foreigners]. *Journal of KIISE*, 44(10), 1087–1093. https://doi.org/10.5626/JOK.2017.44.10.1087

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Petrov, S., Das, D., & McDonald, R. (2012). A universal part-of-speech tagset. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation* (pp. 2089–2096). European Language Resources Association.

Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2018). Universal dependency parsing from scratch. In D. Zeman & J. Hajič (Eds.), *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 160–170). Stroudsburg: Association for Computational Linguistics.

Römer, U., Roberson, A., O'Donnell, M. B., & Ellis, N. C. (2014). Linking learner corpus and experimental data in studying second language learners' knowledge of verb-argument constructions. *ICAME Journal*, 38(1), 115–135. https://doi.org/10.2478/icame-2014-0006

Ryu, S. (2017). hankwuke haksupcauy cakmwun calyoey nathanan cepsokpwusa sayong yangsang yenkwu – pinto cengpolul cwungsimulo [A Study on the use of Korean conjunctive adverbs in Korean learners by analyzing their writing – Focusing on frequency information]. *mwunpep kyoyuk*, 29, 143–168. https://doi.org/10.21850/kge.2017.29..143

Seo, S.-B. (2014). oykwukin yuhaksaynguy hankwuke ssuki olyu pwunsek – hakpwu cayhak yuhaksayng paykilcang cakmwunul taysangulo [A study on analysis of error patterns in Korean writing of international students – Focusing on essay writing contest for university students]. *wulimalkul*, 62, 127–157. https://doi.org/10.18628/urimal.62..201409.127

Siewierska, A. (2013). Alignment of verbal person marking. In M. Haspelmath, M. Dryer, D. Gil, & B. Comrie (Eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved at http://wals.info/chapter/100

Shin, G.-H. (forthcoming). Automatic analysis of caregiver input and child production: Insight into corpus-based research on child language development in Korean. *Korean Linguistics.*

Shin, G.-H. (2020). Connecting input to comprehension: First language acquisition of active transitives and suffixal passives by Korean-speaking preschool children. (Unpublished doctoral dissertation). University of Hawaiʻi at Mānoa, Honolulu.

Sohn, H. M. (1999). *The Korean language*. Cambridge: Cambridge University Press.

Song, J. J. (2015). Causatives. In L. Brown & J. Yeon (Eds.), *The handbook of Korean linguistics* (pp. 116–136). Oxford: John Wiley & Sons. https://doi.org/10.1002/9781118371008.ch6

Song, S., & Choe, J. W. (2007). Type hierarchies for passive forms in Korean. In S. Müller (Ed.), *Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar, Stanford Department of Linguistics and CSLI's LinGO Lab* (pp. 250–270). Stanford, CA: CSLI Publications.

Song, W. (2018). cwungkwukin chokup hankwuke haksupcauy kulssukiey nathanan cosa olyu yangsangkwa cito pangan yenkwu [A Study on the auxiliary word error pattern and guidance method in the writing of Chinese elementary Korean learners]. *cakmwunyenkwu*, 38, 119–147.

Straka, M., & Straková, J. (2017). Tokenizing, POS Tagging, lemmatizing and parsing UD 2.0 with UDPipe. In J. Hajič & D. Zeman (Eds.), *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 88–99). Stroudsburg: Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-3009

Sun, C. F., & Givón, T. (1985). On the so-called SOV word order in Mandarin Chinese: A quantified text study and its implications. *Language*, 61, 329–351. https://doi.org/10.2307/414148

Sung, M.-C., & Kim, H. (2020). Effects of verb–construction association on second language constructional generalizations in production and comprehension. *Second Language Research*. https://doi.org/10.1177/0267658320932625

Xiao, R. (2007). What can SLA learn from contrastive corpus linguistics? The case of passive constructions in Chinese learner English. *Indonesian JELT*, 3(1), 1–19.

Xiao, R., McEnery, T., & Qian, Y. (2006). Passive constructions in English and Chinese: A corpus-based contrastive study. *Languages in Contrast*, 6(1), 109–149. https://doi.org/10.1075/lic.6.1.05xia

Won, M., Wang, Y., Zhu, Y., & Wang, H. (2017). hankwuke haksupcauy ssukiey nathanan ehwi phwungyoto yenkwu-swuktalto chukceng tokwulosse ehwi phwungyoto chukceng kanungsengul cwungsimulo [A study of lexical richness of Korean learners' writing: The possibility of using lexical richness to measure language level]. *emwunlonchong*, 71, 33–55.

Yeon, J. (2015). Passives. In L. Brown & J. Yeon (Eds.), *The handbook of Korean linguistics* (pp. 116–136). Oxford: John Wiley & Sons. https://doi.org/10.1002/9781118371008.ch7

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., & Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In D. Zeman & J. Hajič (Eds.), *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 1–21). Stroudsburg: Association for Computational Linguistics.

Appendix A. Summary of recent learner corpus research on L2 Korean

Type	**Study**	**Focus**	**L1**†	**Proficiency**	**Data size**	**Automatic?**
Error analysis	Bang (2014)	Sino-Korean word	CHN	Advanced	Unclear (20 essays)	No
	Huh (2018)	Word order	CHN	Unclear	Unclear (129 essays)	No
	Lee, Dickenson, & Israel (2016)	Overall (annotation system)	Various	Beginner; Intermediate	10,038 eojeol (100 essays)	
	Park, Kim, Lee, & Lee (2017)	Error correction	Unclear	Unclear	425 eojeol	Yes
	Park & Lee (2014)	Word order	CHN	Intermediate	Unclear (162 essays)	No
	Seo (2014)	Various	Various	Unclear	Unclear (47 essays)	No
	Song (2018)	Case marking	CHN	Beginner	Unclear (105 essays)	No
Use of linguistic items	Cho & Park (2018)	Word (semantic similarity)	ENG	Beginner; Advanced	Unclear (16 essays)	Yes
	Choi (2018)	Adnominal clause	Various	Unclear	6,132 eojeol (28 essays)	No
	Cui & Wang (2018)	Auxiliary verb	CHN	Advanced	42,824 eojeol (240 essays)	No
	Kim & Kim (2013)	Discourse marker	Various	Unclear	44,694 eojeol (spoken only)	No
	Kim, Park, Kim, Kim, Choi, Suh, & Kwak (2016)	Grammar pattern (specified in learner textbook)	Various	Beginner; Intermediate; Advanced	133,785 eojeol (1,288 essays)	Yes

Type	Study	Focus	L1[†]	Proficiency	Data size	Automatic?
	Nam & Hong (2014)	Case marking	CHN	Beginner; Intermediate; Advanced	Unclear (spoken only)	Yes
	Park & Lee (2017)	Cohesive device	Various	Advanced	4,882 eojeol (31 essays)	Yes
	Ryu (2017)	Conjunctive Adverb	Various	Beginner; Intermediate; Advanced	27,405 eojeol (270 essays)	Yes
CAF	Kwak (2016)	–	ENG	Intermediate; Advanced	Unclear	Yes
	Nam, Kim, & Kim (2016)	syntactic complexity	Various	Beginner; Intermediate; Advanced	Unclear (55 essays)	No
	J. Lee (2017)	Lexical richness	JPN & CHN	Intermediate; Advanced	15,266 eojeol (30 essays)	Yes
	S. Lee (2017)	Lexical richness	CHN	Beginner	Unclear (35 essays) + spoken	Yes
	Won, Wang, Zhu, & Wang (2017)	Lexical richness	CHN	Intermediate; Advanced	4,523 eojeol (40 essays)	No

Note.
† CHN = Chinese; ENG = English; JPN = Japanese.

Address for correspondence

Gyu-Ho Shin
Palacký University Olomouc
Department of Asian Studies
Křížkovského 512/10
771 80 Olomouc
Czech Republic
gyuho.shin@upol.cz

Co-author information

Boo Kyung Jung
University of Pittsburgh
Department of East Asian Languages and Literatures
boj11@pitt.edu