# Introduction

Gyu-Ho Shin
Palacký University Olomouc, Czech Republic

Corpora, or structured collections of authentic language use, have been serving as one of the important resources in language research. Literature on language acquisition/development joins this trend, revealing various aspects of developmental trajectories involving a first language (L1; e.g., Behrens 2006, Cameron-Faulkner et al. 2003, Stoll et al. 2009, Theakston et al. 2015) and second/foreign languages (L2; e.g., Biber et al. 2004, Ellis & Ferreira-Junior 2009, Gablasova et al. 2017, Kyle & Crossley 2017). Notably, however, this research practice often occurs with some major languages. This bias towards a limited range of languages renders it difficult to fully ascertain (i) whether the corpus-based/mediated research also holds for other languages and (ii) how this research practice addresses the multifaceted nature of language acquisition and development processes for lesser-studied languages. Korean, the working language in this journal, is one understudied language in the language acquisition/development literature. Although Korean corpora are growing in their quantity and types, systematic use of the existing corpora for this topic is not active. Moreover, in the context of the acquisition/development of linguistic knowledge, considering the increasing popularity of technology-based analysis of corpora found in major languages (e.g., Hawkins et al. 2020, Lu 2010, Kyle et al. 2017, Meurers & Dickenson, 2017, Warstadt & Bowman 2020), investigating whether and to what degree (large-scale) Korean corpora can be meaningfully utilised in conjunction with automatic data analysis is scant.

The current special issue of *Korean Linguistics* aims to complement these missing areas. The following four articles in this issue introduce some promising directions towards corpus-based/mediated research on the acquisition/development of Korean (as L1 or L2). All the studies are based on open-access corpora, which are sizable in quantity, and put emphasis on the reproducibility of procedures and results, which has remained unclear in the previous studies on this topic.

Shin (this issue) applies Natural Language Processing (NLP) techniques to the largest, open-access L1-Korean child language corpora (caregiver input and child production) in the CHILDES database. This study is motivated by the fact

https://doi.org/10.1075/kl.00001.int
*Korean Linguistics* 18:2 (2022), pp. 121–124. ISSN 0257-3784 | E-ISSN 2212-9731

that the currently available open-access NLP pipelines are not ideal for child language corpora and that the default data analysis programme offered by CHILDES does not support Korean. This study sets out its work in two ways: developing a Part-of-Speech (POS) tagger through a classic, simple neural-network algorithm (Perceptron), and developing a construction-identification scheme that automatically extracts core constructional patterns expressing a transitive event (active transitive; suffixal passive) from the given data. In each part, this study addresses potential challenges involving automatic data processing in Korean, with emphasis on language-specific aspects generating these challenges, and offers promising solutions that can alleviate or bypass the challenges. As the first methodological report on child language research in Korean in connection with NLP techniques, this study reveals the advantages and drawbacks of its approach in a balanced manner.

Kim (this issue) employs the Sejong spoken corpus, one of the representative open-access L1-Korean corpora, to explore the speech level shift from non-honorific to honorific contexts manifested by native speakers of Korean. For this purpose, this study manually extracts instances of casual conversations from that corpus, develops a three-tier transcription system to ensure quality information in each utterance, and applies a discourse-analysis approach to revealing how the target shift occurs in a conversation unit. This study's corpus-based analysis shows two major aspects in terms of the honorifics use in conversation. One is that the speech-level alternation occurs actively between the honorific and non-honorific boundaries. The other finding is that shifting the speech levels is affected more by affective/epistemic stance between interlocutors than by grammatical rules. Based on these findings, this study suggests the role of using corpora in supporting L2-Korean instruction, particularly capitalising the difference in usage contexts involving Korean honorifics between the corpus and learner textbooks.

Jung (this issue) compares large-scale, open-access L1-Korean corpora (Sejong written and spoken corpora) and L2-Korean learner textbooks used in tertiary-level instruction of Korean, with a core assumption that L2 textbooks provide L2 learners with essential/primary input for the target language. This study specifically focuses on how the locative function of the postposition *-ey* is manifested in relation to verb use. The two types of data are analysed in a semi-automatic manner, starting from automatic tokenisation and POS tagging and proceeding to manual inspection/extraction of instances relevant to the study. Combined with a corpus-internal analysis technique (keyness analysis), this study finds a key asymmetry between the Sejong corpora and the textbooks in terms of verb types co-occurring with the particular function of this postposition. Based on this asymmetric nature, this study argues that the L2-textbooks' emphasis on the usage contexts involving the locative function of *-ey* does not conform to how native speakers of Korean utilise the same function/postposition as attested in the L1-Korean corpora. This study further suggests the necessity of incorporating corpus-based accounts harmoniously into L2-Korean learning-teaching contexts, given the educational purposes in L2-Korean instruction.

Lee (this issue) employs an error-annotated L2-Korean writing corpus from the Korean Learner Corpus and identifies patterns of verb conjugation errors in learner production. In doing so, this study adopts an alternative description framework based on the paradigmatic relations amongst conjugated verb forms. Notably, this study adopts the concept of entropy to estimate the predictability of each class of verb conjugation. Instances of the production errors are extracted from a web-based engine and are submitted to a post-processing stage in which the errors relating to paradigm predictability are manually sorted. The result of this study reveals three major error types regarding verb conjugation, showing that L2-Korean learners tend to produce errors when a verb engages in a conjugated form which is frequently used or fairly predictable. Moreover, this study shows that the aforementioned trend seems to persist until learners achieve a high level of proficiency in Korean. Based on these findings, this study proposes some ideas on utilising the implications of the paradigmatic-relation-based frame of verb conjugation for L2-Korean instruction.

Taken together, the four studies in the special issue are expected to advance our current understanding of corpus-based/mediated research on the acquisition and development of Korean. These studies will also contribute to improving the research practice involving this area, particularly that at the interface of technology.

## References

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

Biber, Douglas, Susan Conrad & Viviana Cortes. 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

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

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

Gablasova, Dana, Vaclav Brezina & Tony McEnery. 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

Hawkins, Robert D., Takateru Yamakoshi, Thomas L. Griffiths & Adele E. Goldberg. 2020. Investigating representations of verb bias in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 4653–4663.

Kyle, Kristopher & Scott Crossley. 2017. Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing* 34.4. 513–535. https://doi.org/10.1177/0265532217712554

Kyle, Kristopher, Scott Crossley & Cynthia Berger. 2017. The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods* 50. 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Lu, Xiaofei. 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, Detmar & Markus Dickinson. 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

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

Theakston, Anna L., Paul Ibbotson, Daniel Freudenthal, Elena VM Lieven & Michael Tomasello. 2015. Productivity of noun slots in verb frames. *Cognitive Science* 39.6. 1369–1395. https://doi.org/10.1111/cogs.12216

Warstadt, Alex & Samuel R. Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data?. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 1737–1743.

## Address for correspondence

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