KoFREN: Comprehensive Korean Word Frequency Norms Derived from Large Scale Free Speech Corpora

Jin-seo Kim[1], Anna Seo Gyeong Choi[2], Sunghye Cho[3]
[1] School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA
[2] Department of Information Science, Cornell University, Ithaca, NY
[3] Linguistic Data Consortium, Department of Linguistics, University of Pennsylvania, Philadelphia, PA
jins0904@sas.upenn.edu, sc2359@cornell.edu, csunghye@ldc.upenn.edu

Abstract

Word frequencies are integral in linguistic studies, showing solid correlations with speakers' cognitive abilities and other critical linguistic parameters, including the Age of Acquisition (AoA). However, the lack of expansive speech data and a reliable part-of-speech (POS) tagger has obstructed the formulation of credible Korean word frequency norms. In this study, we unveil Korean word frequency norms (KoFREN), derived from large-scale spontaneous speech corpora (41 million words) that include a balanced representation of gender and age. We employed a machine learning-powered POS tagger, showcasing accuracy on par with human annotators. Our frequency norms correlate significantly with external studies' lexical decision time (LDT) and AoA measures. KoFREN also aligns with English counterparts sourced from $SUBTLEX_{US}$ - an English word frequency measure frequently used in the literature. KoFREN is poised to facilitate research in spontaneous Contemporary Korean and can be utilized in many fields, including clinical studies of Korean patients.

Keywords: Spontaneous speech, Word frequency, Crosslinguistic comparison, Contemporary Korean

1. Introduction

Word frequency is a significant linguistic variable in various fields, from assessing individuals' linguistic competence to identifying cognitive decline associated with neurodegeneration, including Alzheimer's disease. While a variety of English word frequency norms drawn from different sources are easily accessible (Kučera and Francis, 1967; R H. Baayen, R Piepenbrock, L Gulikers, 1995; Zeno et al., 1995; Balota et al., 2007; Brysbaert and New, 2009; Leech et al., 2014), word frequency norms drawn from large-scale spontaneous speech corpora are rarely available in minor languages, such as Korean. Additionally, current norms are either derived from subjective frequency estimates or not large enough to construct reliable norms (NIKL, 2003; Park, 2003; Park, 2004; Shin and Hill, 2016).

The establishment of Korean word frequency norms has been hindered by two major challenges. The first is the limited availability of publicly accessible, large-scale Korean speech corpora. The second challenge is the lack of a reliable part-of-speech (POS) tagger for the Korean language, which is essential for generating consistent tag sequences. The lack of a robust, automated POS tagger requires manual tokenization and tagging of linguistic data, which vastly limits the size of frequency measures that are currently available. This study introduces KoFREN: extensive Korean word frequency norms calculated from a significant volume of high-quality conversational speech corpora. To automate the frequency calculation, we first evaluated the performances of different Korean POS taggers that are publicly available and chose a machine-learning-powered POS tagger that delivered performance comparable to the accuracy of human annotators.

The best way to validate word frequency measures is to examine "how well they predict human processing latencies" (Brysbaert and New, 2009). For this reason, to validate our word frequency norms, we adopted reaction times for a lexical decision task in a recent study (Baek et al., 2024). We employed two sets of Age of Acquisition data (AoA; the average age at which children acquire a given word) to further validate our measures. AoA is a linguistic variable typically assessed by querying adults to recall the age at which they acquired specific words (Brysbaert and Biemiller, 2017). Despite relying on participants' memory, AoA remains a robust predictor of linguistic performance, strongly correlated with word frequency (Brysbaert and Biemiller, 2017; Cho et al., 2021). In this study, we incorporated Korean AoA data acquired from observational studies of young Korean children (Kwak and Pae, 2011; Frank et al., 2017) and AoA data acquired from surveying the adult participants in the lexical decision study (Baek et al., 2024), and examined their correlations with our KoFREN norms.

Lastly, we correlated our word frequency metrics against those derived from the $SUBTLEX_{US}$ corpus (Brysbaert and New, 2009), which we used as a benchmark for word frequency norms in spontaneous American English. The frequency norms of

LREC-COLING 2024, pages 9926–9931
20-25 May, 2024.

$SUBTLEX_{US}$ have demonstrated superiority over traditional benchmarks in forecasting participants' performance in linguistic tasks, such as lexical decision time (Brysbaert and New, 2009).

2. Methods

2.1. Data

To calculate Korean word frequency, we combined three distinct corpora of spontaneous, free speech from children (FS_{child}) (NHN Diquest, 2020a), young adults (FS_{young}) (NHN Diquest, 2020c), and elderly adults (FS_{old}) (NHN Diquest, 2020b), all of which are freely available on a website (www.ai-hub.or.kr) upon request by Korean nationals for research purposes. Each data set included audio files and their corresponding transcripts of approximately 3000 to 4000 hours of spontaneous speech from each group of speakers on a range of real-life conversation topics. These corpora were developed under an initiative led by the Ministry of Science and Information, Communication, and Technology in South Korea to facilitate AI-related research and development. Table 1 provides the details of each corpus, including the corpus size in hours, demographic characteristics of the speakers, and the number of total words and syllables. All three age groups consisted of an equal number of male and female speakers.

	FS_{child}	FS_{young}	FS_{old}
Speakers	1000	2000	1000
Age range	3-10	11-59	60+
Audio	3000 hrs	4000 hrs	3000 hrs
Words	11.2M	17.9M	11.8M
Syllables	28.9M	48.5M	31.0M

Table 1: Summary of the FS_{child}, FS_{young}, and FS_{old} data sets. Audio sizes are in hours, and word and syllable counts are in millions.

We compiled a data set of approximately 41 million words exclusively from fully transcribed and spelling-checked, spontaneous speech data. Brysbaert and New (2009) previously asserted that a corpus should comprise more than 16 million words to reliably capture the frequency of commonly used words and less commonly used words (i.e., fewer than 10 per million). We note that the size of our data set is about 2.5 times larger than the minimum word count suggested by Brysbaert and New (2009). Additionally, these words are derived from transcriptions of spontaneous, conversational speech. Thus, we expect our word frequency measures to be more representative of spontaneous, contemporary Korean in comparison to preexisting norms that employed text-based sources, including books, news articles, YouTube comments, blogs, or Wikipedia articles.

2.2. Tokenization

We selected a lemma count approach when calculating word frequency since Korean is an agglutinative language, where different morphemes combine without changing their forms to add grammatical meanings. Unlike the conventional word form frequencies (WF) approach, where different inflected forms of a word are counted individually, the lemma frequency approach counts each morpheme of a word separately. For example, a Korean word root *meok-*, which translates to "to eat" can appear in various inflected forms, such as $meok_{eat}$-da_{ending} ('to eat'), $meok_{eat}$-eot_{past}-da_{ending} ('ate'), or $meok_{eat}$-$neun_{progressive}$-da_{ending} ('is eating'). Under the lemma frequency approach, the frequency count for *meok-* would represent the combined occurrences of all these forms. To implement the lemma approach, we first tokenized all sentences in our data set using three different POS taggers (see Section 2.2.1). Since some tokenized words were monosyllabic homonyms (e.g., *-neun* as a topic-case marker vs. as a present progressive tense marker), we also assigned a part-of-speech (POS) tag to each word to differentiate these homonyms.

2.2.1. Tokenization and POS tagging evaluation

To automatically tokenize and assign POS tags to approximately 41 million words with high accuracy, we tested tokenizers and POS taggers from several Korean NLP programs that were available online and compared their performances. We initially implemented the Kkma (Lee et al., 2010), Komoran (Shin, 2013), and Hannanum (Choi, 1999) taggers from the KoNLPy package (Park and Cho, 2014), which offered Python wrappers for these taggers. The Hannanum tagger was excluded from the final analysis since it could only distinguish between nine POS tags, which were significantly fewer than the number of distinct tags produced by the other two. Instead, we included the Bareun tagger (Baikal.ai, 2023) in our analysis, which was trained using a cutting-edge deep-learning transformer model.

To evaluate the tokenizers, we randomly selected 50 sentences from the FS_{young} corpus, and an expert linguist generated gold-standard POS data by manually tokenizing and tagging the selected sentences. The average sentence from the sample contained 5.72 words (SD=3.09), with a mean syllable count of 16.04 (SD=9.01). The ground truth data comprised 580 unique tokens. The three most prevalent POS tokens were general nouns (n=90), verbs (n=71), and conjunctions (n=56). Subsequently, we ran the three NLP programs to

tokenize the words and tag their POS categories automatically. We calculated the accuracy of each tokenizer by dividing the number of correctly tokenized words by the total number of tokens. The accuracy of the POS tags was assessed through the word error rate of the predicted tag sequences, utilizing the JiWER library in Python (Vaessen, 2018). All automated preprocessing scripts were executed using Python 3.10.

2.2.2. Word frequency calculation

We used the Bareun tagger to process the spontaneous speech data set we compiled, as it exhibited superior performance in both tokenization and POS tagging compared to other taggers (see results in Section 3.1). First, each sentence underwent automated preprocessing to remove alphabets, numbers, and special characters, ensuring only Korean characters were retained. We then tokenized words in all sentences and assigned their respective POS tags. Word frequency norms were established by counting the unique combinations of a word and its associated POS tag.

Brysbaert and New (2009) introduced English word frequency norms derived from a large-scale corpus of subtitles of movies and television series ($SUBTLEX_{US}$). These norms have shown superior performance in predicting task outcomes, such as lexical decision and word naming, compared to traditional measures like those from Kučera and Francis (1967). For crosslinguistic comparisons, we scaled our word frequency norms to align with the size of the $SUBTLEX_{US}$ corpus, which included 49.7 million words in total, by multiplying the raw Korean word counts by the ratio of the total word count in $SUBTLEX_{US}$ to the total word count in our dataset. We also calculated the $\log_{10}$ values of these adjusted frequencies of Korean words to facilitate a comparative analysis with the word frequency distributions from $SUBTLEX_{US}$ (Brysbaert and New, 2009).

Our KoFREN norms comprise 96,339 unique combinations of tokens and POS tags. KoFREN is divided into four distinct datasets: a comprehensive measure using aggregated data from the FS_{child}, FS_{young}, and FS_{old} corpora, as well as those derived separately from each corpus to broaden its applicability and enable group comparisons based on speaker age. The KoFREN datasets and the Python scripts used for computing and evaluating the word frequencies are available here.

2.2.3. Word frequency evaluation

We assessed our word frequency measures using three additional data sets. First, Baek et al. (2024) performed a large-scale word recognition study on 497 Korean speakers aged between 20 and 60. This web-based study collected subjects' reaction times from a lexical decision task and their self-reported age-of-acquisition (AoA) for 120 Korean nouns of varying frequencies. We computed the mean reaction times and AoA across all subjects for each test word and measured their correlations with the $\log_{10}$ frequencies from the FS_{young} corpus and the aggregated corpora.

Additionally, we employed a children's observational age of acquisition (AoA) dataset for Korean words (Kwak and Pae, 2011) derived from Wordbank (Frank et al., 2017). This data set evaluates children's knowledge of each word across various developmental stages. We determined the mean AoA of each word by deploying the 'fit_aoa ()' function from the wordbankr package (Frank et al., 2017) in R, setting the proportion at 0.8. This calculation yielded the average age at which 80% of the children recognized a given word. Although the AoA dataset comprised 1,023 Korean words, a significant portion comprised onomatopoeic expressions. We selected 50 concrete words for our analysis, correlating these AoA norms with word frequencies extracted from KoFREN.

Finally, we selected nouns, verbs, and adjectives with concrete word meaning (concreteness > 4; Brysbaert et al., 2014) from the English word frequncy norms in $SUBTLEX_{US}$, and correlated their word frequencies with their Korean equivalents. We excluded English words with multiple potential Korean translations (and vice versa) from the analysis. For all datasets, we employed Pearson's correlation for log frequencies and Spearman's correlation for raw word frequencies.

3. Results

3.1. Evaluating tokenizer performance

Table 2 shows the performance of each tokenizer across the 50 sampled sentences. The Bareun tokenizer consistently outperformed the others in all evaluation metrics; thus, we employed the Bareun tokenizer for word frequency calculation.

3.2. Evaluating the frequency measures

3.2.1. Korean Lexical Decision Task Data

Figure 1 illustrates the correlations between the average reaction times of Korean speakers during a visual lexical decision task (Baek et al., 2024) and $\log_{10}$ word frequencies within the aggregated corpora. We observed a significant, moderate correlation between our word frequencies and average reaction times. As expected, less frequently used words took longer to recognize, while frequently used words were recognized significantly faster. (r = -0.5394, p-value < 0.001).

3.2.2. Korean AoA Data

We evaluated our frequency norms using two age-of-acquisition datasets: the self-reported AoA measures from adults (Baek et al., 2024) and

Metrics	**Bareun**	Kkma	Komoran
Acc_{Mean}	**0.936**	0.743	0.809
Acc_{STD}	0.171	0.250	0.198
Acc_{Total}	0.948	0.781	0.847
WER_{Mean}	0.115	0.655	0.432
WER_{STD}	0.183	0.272	0.292
WER_{Total}	0.091	0.586	0.353

Table 2: Performance of the Bareun, Kkma, and Komoran tokenizers. $Acc_{Mean, STD}$: Mean and standard deviation of tokenization accuracy per sentence, Acc_{Total}: Overall tokenization accuracy across the 50 sentences, $WER_{Mean, STD}$: Mean and standard deviation accuracy of POS tags per sentence in WER ; WER_{Total}: Overall POS tagging accuracy across the 50 sentences in WER.

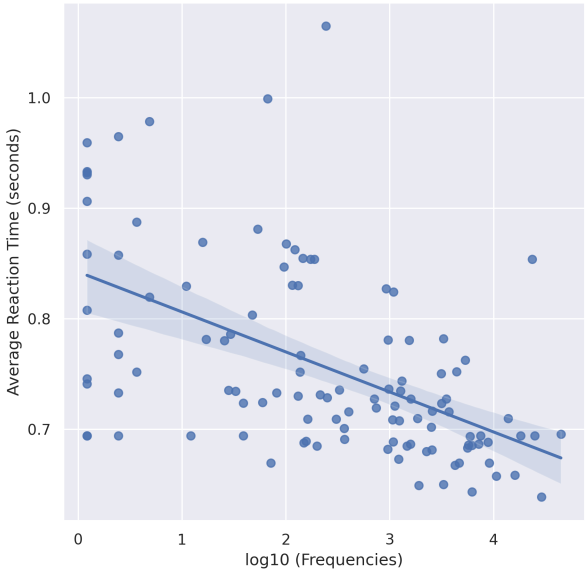

Figure 1: Correlations between average reaction time and log_{10} word frequencies from the aggregated corpora.

those derived from an observational study of Korean-speaking children (Kwak and Pae, 2011). Figure 2 illustrates the correlations of each AoA dataset with our log_{10} word frequencies. For the adult AoA dataset, words acquired at an earlier age significantly correlated with higher word frequencies from the FS_{young} corpus (r = -0.5113, p-value < 0.001; Fig.2A). The AoA measures exhibited moderately stronger correlation with the aggregated word frequencies (r = -0.5620, p-value < 0.001).

Similarly, children's AoA dataset exhibited a moderate, negative correlation with word frequencies from the FS_{child} corpus (log_{10} WF: r = -0.5163, p-value < 0.001; Fig.2B). Specifically, the AoA measures showed a stronger correlation with word frequencies taken only from the FS_{child} corpus compared to the aggregated frequencies (log_{10} WF: r = -0.4320, p-value = 0.002; Figure not shown), suggesting that word frequencies from the age-selective FS_{child} corpus more robustly captured speech patterns of children than those derived from the entire KoFREN dataset.

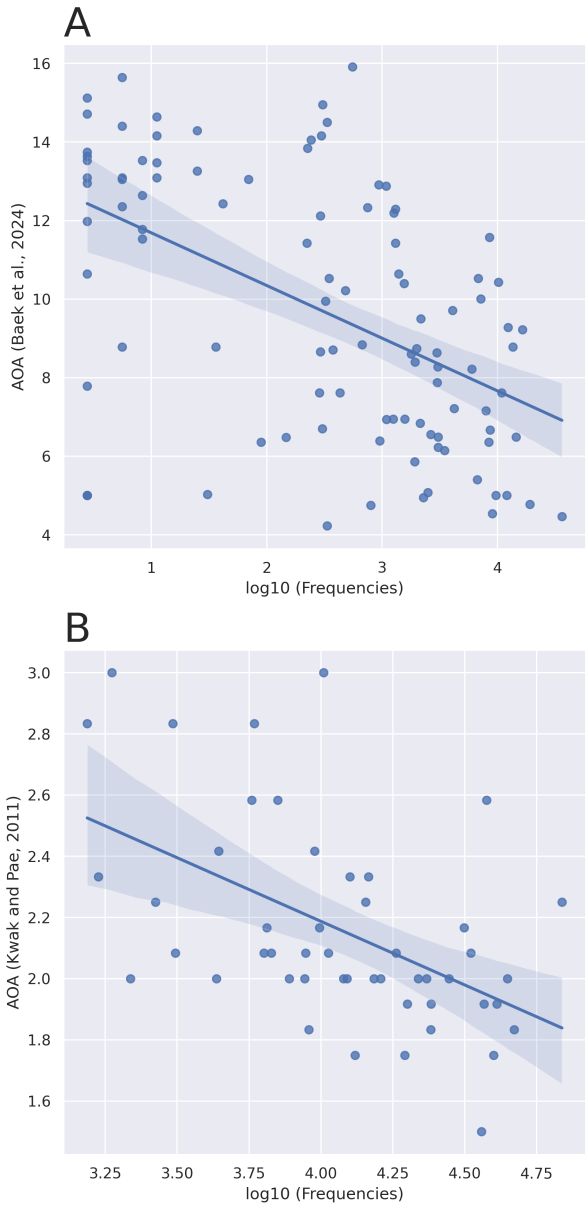

Figure 2: **A.** Correlations between adults' self-reported AoA measures and log_{10} word frequencies from FS_{young}; **B.** Correlations between Kwak and Pae (2011)'s observation based AoA measures and word frequencies from FS_{child}.

3.2.3. Comparison to the English word frequency from $SUBTLEX_{US}$

We compared 150 words from $SUBTLEX_{US}$ with their Korean equivalents from KoFREN. Both raw

and $\log_{10}$ frequencies exhibited moderate, positive correlations, aligning with our expectations ($\log_{10}$ WF: r = 0.6522, p-value < 0.001; raw WF: *rho* = 0.6623, p-value < 0.001). Figure 3 illustrates the correlations between the $\log_{10}$ frequencies of words in koFREN and their English counterparts in $\text{SUBTLEX}_{\text{US}}$.

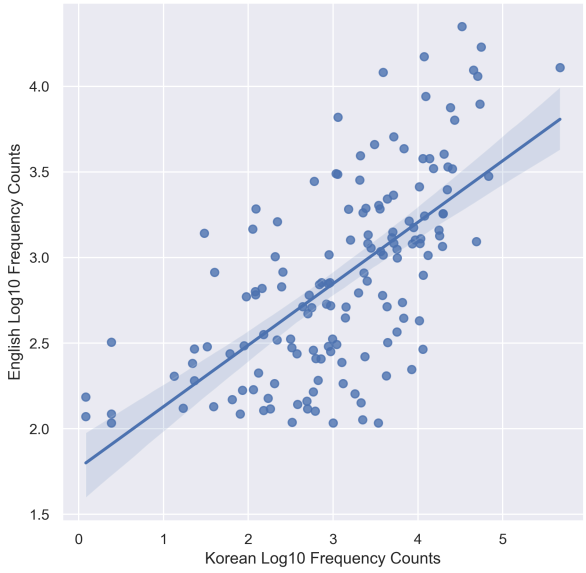

Figure 3: Correlations between $\log_{10}$ frequency norms taken from $\text{SUBTLEX}_{\text{US}}$ and KoFREN.

4. Discussion

In this study, we introduced KoFREN – the Korean word frequency norms derived from large-scale, annotated spontaneous speech corpora, offering insights into the intricacies of naturally spoken Contemporary Korean. Our norms provided a more reliable representation of spoken, natural language than those based on books, internet comments, or news articles. We reduced age or gender biases by selecting corpora that included equal numbers of male and female speakers with a wide age range. We have also validated the reliability and accuracy of our frequency norms by demonstrating statistically significant correlations with gold standard lexical decision time and age-of-acquisition (AoA) measures from multiple external datasets.

We offered aggregated frequency counts from all corpora and separate word frequency norms from each FS corpus so that researchers could choose a dataset that fits their needs. AoA measures from the observational study (Kwak and Pae, 2011) were more strongly correlated with word frequencies derived only from children's speech data than those from the entire dataset. This further validated that our word frequency norms were sensitive to different age groups.

Our comparative analysis with the $\text{SUBTLEX}_{\text{US}}$ corpus was constrained to a limited word subset. Frequent and concrete English words often had multiple translations in Korean, precluding direct one-to-one comparisons. In our examination of the $\text{SUBTLEX}_{\text{US}}$ corpus, we also noted that terms rarely used in everyday conversations like "heroin" and "warriors" exhibited frequency counts as high as commonly used terms like "shrimp" and "refrigerator." This seems to be because the word frequency norms in English were derived from movie subtitles, which often encompass themes of drugs and crime, thereby inflating the frequency of related terms. Our data do not extend to such thematic areas and reflect genuine word frequency in everyday communication.

Brysbaert and New (2009) suggests three critical variables when evaluating the quality of a frequency measure: the size and the register of the corpus and the frequency measure used. Our data set was comparable to the $\text{SUBTLEX}_{\text{US}}$ corpus, which was much larger than the minimum size required (16 million words). Also, we used corpora of spontaneous, conversational speech to derive word frequencies; therefore, our word frequency norms are a better approximation of real-life Contemporary Korean. Lastly, we used a lemma approach in calculating word frequency norms. Although Brysbaert and New (2009) did not find a significant difference between lemma-based and word-form-based approaches in English, it might have been because English has a relatively weak inflection system. Korean is an agglutinative language with a clear and systematic difference between verb roots and inflectional morphemes, so we believe the lemma approach we employed here works better for Korean. We hope our word frequency norms will serve as a reference dataset in future research in various fields, including psychology, linguistics, education, and medicine.

5. Bibliographical References

Hyunah Baek, Peter C Gordon, and Wonil Choi. 2024. Effects of age and word frequency on korean visual word recognition: Evidence from a web-based large-scale lexical-decision task. *Psychology and Aging*.

Baikal.ai. 2023. Bareun.ai. https://bareun.ai/. Accessed: October 2023.

David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis,

James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. 2007. The english lexicon project. *Behavior research methods*, 39:445–459.

Marc Brysbaert and Andrew Biemiller. 2017. Test-based age-of-acquisition norms for 44 thousand english word meanings. *Behavior research methods*, 49:1520–1523.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Sunghye Cho, Naomi Nevler, Natalia Parjane, Christopher Cieri, Mark Liberman, Murray Grossman, and Katheryn AQ Cousins. 2021. Automated analysis of digitized letter fluency data. *Frontiers in Psychology*, 12:3112.

Key-Sun Choi. 1999. Hannanum tokenizer. https://kldp.net/hannanum/. Accessed: October 2023.

Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. 2017. Wordbank: an open repository for developmental vocabulary data. *Journal of child language*, 44(3):677—694.

H. Kučera and W. N. Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

K. Kwak and S. Pae. 2011. *Korean MacArthur-Bates Communicative Development Inventories* (*K M-B CDI*). Mindpress, Seoul.

Dong-Joo Lee, Jong-Heum Yeon, In-Beom Hwang, and Sang-Goo Lee. 2010. Kkma: A tool for utilizing sejong corpus based on relational database. In *Proceedings of the Korean Information Science Society Conference*. Korean Institute of Information Scientists and Engineers.

Geoffrey Leech, Paul Rayson, et al. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Eunjeong L. Park and Sungzoon Cho. 2014. Konlpy: Korean natural language processing in python. In *Proceedings of the 26th Annual Conference on Human and Cognitive Language Technology*, Chuncheon, Korea.

Tae Jin Park. 2003. Subjective frequency estimates of korean words and frequency effect on word recognition. *The Korean Journal of Experimental Psychology*, 15(2):349–366.

Tae Jin Park. 2004. Investigation of association frequency and imagery value of korean words. *The Korean Journal of Experimental Psychology*, 16(2):237–260.

Junsoo Shin. 2013. Komoran tokenizer. https://docs.komoran.kr. Accessed: October 2023.

Sangeun Shin and Katya Hill. 2016. Korean word frequency and commonality study for augmentative and alternative communication. *International Journal of Language & Communication Disorders*, 51(4):415–429.

Nik Vaessen. 2018. Jiwer - a python package for word error rate calculation. https://github.com/jitsi/jiwer. Accessed: October 2023.

Susan Zeno, Stephen H Ivens, Robert T Millard, and Raj Duvvuri. 1995. *The educator's word frequency guide*. Touchstone Applied Science Associates.

6. Language Resource References

NHN Diquest. 2020a. *Korean Children Free Speech database*. NHN Diquest, 1.2.

NHN Diquest. 2020b. *Korean Elderly Adults Free Speech database*. NHN Diquest, 1.2.

NHN Diquest. 2020c. *Korean Young Adults Free Speech database*. NHN Diquest, 1.2.

NIKL. 2003. *Contemporary Korean Word Frequency Measures*. National Institute of Korean Language (NIKL).

R H. Baayen, R Piepenbrock, L Gulikers. 1995. *CELEX2 lexical database*. Linguistic Data Consortium.

7. Acknowledgment

This study was supported by the Alzheimer's Association (AARF-21-851126) and the Penn Data Driven Discovery Initiative.