https://doi.org/10.5392/IJoC.2019.15.1.039

Differentiation of Aphasic Patients from the Normal Control Via a Computational Analysis of Korean Utterances

HyangHee Kim
Graduate Program in Speech-Language Pathology, Department & Research Institute of Rehabilitation Medicine, Yonsei University College of Medicine, Seoul, Korea

Ji-Myoung Choi, Hansaem Kim
Interdisciplinary Graduate Program in Linguistics and Informatics, Institute of Language and Information Studies, Yonsei University, Seoul, Korea

Ginju Baek, Bo Seon Kim
Graduate Program in Speech-Language Pathology, Yonsei University, Seoul, Korea

Sang Kyu Seo
Department of Korean Language and Literature, Institute of Language and Information Studies, Yonsei University, Seoul, Korea

ABSTRACT

Spontaneous speech provides rich information defining the linguistic characteristics of individuals. As such, computational analysis of speech would enhance the efficiency involved in evaluating patients' speech. This study aims to provide a method to differentiate the persons with and without aphasia based on language usage. Ten aphasic patients and their counterpart normal controls participated, and they were all tasked to describe a set of given words. Their utterances were linguistically processed and compared to each other. Computational analyses from PCA (Principle Component Analysis) to machine learning were conducted to select the relevant linguistic features, and consequently to classify the two groups based on the features selected. It was found that functional words, not content words, were the main differentiator of the two groups. The most viable discriminators were demonstratives, function words, sentence final endings, and postpositions. The machine learning classification model was found to be quite accurate (90%), and to impressively be stable. This study is noteworthy as it is the first attempt that uses computational analysis to characterize the word usage patterns in Korean aphasic patients, thereby discriminating from the normal group.

Key words: *Aphasia, Automatic Speech Analysis, Computational Analysis, PCA, Machine Learning.*

1. INTRODUCTION

Unlike scripted speech, spontaneous speech is unprepared, with many challenges during production, and frequently includes hesitation, disfluencies, and word retrieval problems. As such, analysis of spontaneous speech may yield abundant information regarding a person's speech characteristics, and lead to a better understanding of the underlying nature of a person's speech and language impairment [1], particularly in aphasic patients with neurological disorders [2], [3]. Aphasia can be classified according to the level of fluency into non-fluent aphasia and fluent aphasia. Non-fluent aphasia is characterized by agrammatism consisting mainly of content words, with a lack of or the faulty use of function words. 'Agrammatism' is defined as "a disorder of sentence production involving the selective omission of function words and some grammatical endings on words" [4], and thus, the utterances of patients with agrammatism may often be called 'telegraphic speech' [5]. On the other hand, fluent aphasia can present excessive or unnecessary function and content words. Several previous studies have reported the differences in word class usages between aphasic and non-aphasic groups. For example, Ahmed et al. [6] detected different uses of pronouns and verbs over the progress of 15 patients with Alzheimer's disease, while Jarrold et al. [7] found different proportions of nouns, pronouns, adjectives and verbs between semantic dementia and normal control groups. These analyses could facilitate an accurate and

* *Corresponding author, Email: h.kim@yonsei.ac.kr*
Manuscript received Aug. 24, 2018; revised Oct. 29, 2018; accepted Nov. 20, 2018

objective comparison of language usages between the persons with and without aphasic symptoms.

Comparative linguistic analysis using computational methods requires collecting utterance data produced by subjects in question, i.e. subjects with and without aphasia in this case. There are a variety of ways to have aphasic patients produce connected speech, one of which is a word definition task. A 'word definition task' is a simple and efficient task in which one actively defines a given word in a free and spontaneous manner. As this task has the effect of imposing some constraints on the topics and themes covered, the task format itself controls for topic diversity to some extent. Without such constraints, what the subject wants to say could have influenced their language production such that comparison on the same ground will not be possible. The task is thus regarded as a more semantically controlled task, and can be used for the purpose of examining the aspect of word usage in neurologically impaired individuals [8], [9]. Utterances from a word definition task will clearly show the usage patterns of content words and function words by appropriately combining both in sentence structures.

Traditional methods of analyzing utterances have been tedious and time-consuming because they require detailed transcription as well as manual analysis of each utterance. The manual analysis of utterances can also pose the problems of consistency and subjectivity. To overcome this, computational analysis of language has recently been conducted, though only a few studies that employ this approach have been published thus far. Brown et al. [10] and Bryant et al. [11] used the automatic analysis tool to measure the propositional density in spontaneous speech based on parts-of-speech tags. The tools were reliable compared to human raters [10], and the propositional density of this program was utilized to provide an index of information content of aphasic discourse i.e., informativeness [11]. Computational "data-driven" approaches such as in [7], [12], [13] have also been adopted to identify the characteristics of discourse performance of patients with language impairment. For these studies, spontaneous utterances were collected, processed, and comparatively analyzed using computational statistical and machine learning algorithms. For example, a PCA (Principal Component Analysis) algorithm was used to pick out the lexical and syntactic difference between patients with semantic dementia and normal controls [12]. Machine learning algorithms were also utilized to successfully classify 3 patients and 10 normal persons through spontaneous discourse transcripts [14]. Jarrold et al. [7] also applied a machine learning algorithm to short speech samples (10 min.) from 4 types of dementia in differentiating the groups. Another machine learning-based study on the classification between patient group and healthy group is Fraser et al. [13], in which they collected and examined short narratives obtained from a picture description task. They used a logistic regression model with linguistic (morphological, syntactic, and psycholinguistic) and acoustic as variables and achieved the accuracy of over 80% after feature selections.

This pilot study is within the context of computational data-driven approach to detecting and differentiating persons with and without linguistic impairments. The goal of this study is thus to identify differential linguistic characteristics between aphasic patients and normal subjects through computational analysis of their Korean-specific morphological uses in spontaneous speech resulting from a word definition task. Morpho-syntactic information of spontaneous speech tells us how words are used and how they are related to each other. Morpho-syntactic specificity varies across languages, and their analysis adds valuable evidence that can be used to enrich speech-language clinical practice for a given language. English and other Western languages, for instance, express grammatical categories for gender, case, person, tense, number and other elements through the morphological changes of substantive or predicate words themselves, or through inflection. In Korean, on the other hand, grammatical forms such as postpositions or word endings are combined in the stems of substantive or predicate words to express their functions in the sentence, forming an 'eojeol' (a basic building block of a Korean sentence) that is separated by spaces before and after. In that regard, aphasia may manifest language-specific phenomena – in other words, aphasic speech characteristics may vary as a function of specific linguistic features [15].

In the course of this work, we hope to enhance the level of understanding of the speech characteristics of Korean aphasic patients, and to lay the groundwork for automatic diagnosis of aphasic cases based on linguistic symptoms.

2. METHODS AND PROCEDURES

2.1 Participants

The study was carried out on a total of 20 participants: 10 individuals with aphasia (AP) (5 females, 5 males) aged between 19 and 79 years (M = 51.6; SD = 18.2) and 10 non-aphasic control subjects (NC) (5 females, 5 males) aged between 22 and 76 years (M = 50.5; SD = 17.7). The number of 20 participants may not be large enough to lead to a broad generalization yet, but it is large enough to get practical and clinical implications and thus suggest directions for computational approach to automatic diagnosis of language deficit problems such as aphasia as in previous pilot studies [1], [9], [15]. It is demonstrated by the clear separation in linguistic patterns between the two groups from across all the three methodologies (as in the section 2.5) which is consistent and reliable.

The patients were diagnosed with aphasia due to cerebral infarction and were free from visual and auditory deficits. Aphasia types that the patients presented included anomic, Broca's, conduction, mixed transcortical and Wernicke's aphasia. The mean (±SD) aphasia quotient (AQ) was 72.6 (±16.2), and ranged from 45.8 to 91.9. The inclusion criteria of the normal control group were 1) no neurological/neuropsychiatric history; 2) no visual and auditory deficits; and 3) within the normal limits on the Korean-Montreal Cognitive Assessment (K-MoCA). The details of participants are presented in Table 1. Both groups were matched in terms of age, gender, and education. Mann-Whitney U-test revealed that age (p = .986) and education (p = .315) were not different between the two groups. All participants provided an informed written consent to participate in this study before the experiment. The textual data for analysis has

been anonymized by replacing names with ID numbers and converting the original text into a bag-of-words model after part-of-speech tagging.

Table 1. Demographic information of aphasic patient group and age-, gender-, and education-matched normal control group

ID	Age / Sex	Edu (yr.)	Type	POT (yr.;m.)	AQ	MoCA
AP1	49/F	16	Anomic	0;6	75.8	n/t
AP2	19/F	9	Broca's	1; 1	73.9	n/t
AP3	62/F	6	Mixed transcortical	0;2	45.8	n/t
AP4	76/F	12	Anomic	9; 8	80.8	n/t
AP5	79/F	6	Anomic	0;11	82.4	n/t
AP6	41/F	12	Conduction	4; 9	82.1	n/t
AP7	37/M	16	Anomic	4;10	91.9	n/t
AP8	48/M	16	Wernicke's	0;3	38.1	n/t
AP9	61/M	6	Anomic	15; 1	75.4	n/t
AP10	44/M	16	Conduction	0;6	80.1	n/t
NC1	49/F	12	-	-	-	24
NC2	22/F	12	-	-	-	29
NC3	61/F	9	-	-	-	23
NC4	73/F	12	-	-	-	26
NC5	76/F	12	-	-	-	24
NC6	41/M	16	-	-	-	26
NC7	31/M	16	-	-	-	30
NC8	48/M	16	-	-	-	25
NC9	64/M	16	-	-	-	24
NC10	40/M	16	-	-	-	28

* ID: subject id, Edu: education, Type: type of aphasia, POT: post onset time, AQ: aphasia quotient (out of 100), MoCA: Korean-Montreal Cognitive Assessment (out of 30), n/t: not tested

2.2 Tasks

Participants were asked to say whatever came to their mind when a given word was presented to them: a word definition task. In previous studies, the word definition task has been administered to investigate the effects of word types to be defined on the speech production, such as the role of concreteness and abstractness of words in utterance production of university students [16] and the consequence of the neighborhood density and frequency levels of words on the definitions of children with specific language impairment [17] and the different behaviors of an aphasic patient when given the task to activate many competing verbal responses (word definitions) and a task to activate few response options (picture description) [18]. The task had also been given to aphasic patients to find out how much lexical-semantic knowledge about words they have in "a straightforward way" [19]. In this study, however, a word definition task is adopted to facilitate an utterance production in a simple and efficient fashion as described above in Introduction, not as a way to examine the effects of the words on the conceptuality of the speech of aphasic patients. Assigning constraints on the task of utterance production in terms of contents and forms make it possible to gather comparable data between the speech groups.

For this study, concrete and abstract ten nouns [9] were provided to all participants: watermelon, pharmacy, electric fan, train, rabbit, jealousy, music, excursion, joke, friendship. The first five concrete nouns were chosen based on *The Florida Semantic Battery* [20] to represent semantic class, definitional class, animacy/inanimacy, and image. The abstract ones were chosen with their abstractness and clearness of the semantic boundaries and features taken into consideration. A time limit of 30 seconds was set for each word.

2.3 Language pre-processing

The audio recordings of the utterances were first transcribed orthographically to represent the spoken words and related spoken phenomena as faithfully as possible, without removing filler interjections (e.g. *um, ah*), repetitions, pauses, and false starts. The transcripts were then automatically converted into the cleaned dataset by attaching special tags to non-words such as repetition and false starts. Some dialect words and non-standard forms affecting the performance of the part-of-speech tagger were manually normalized into standard forms. All the words in the transcripts were assigned with their parts-of-speech tags, and some errors in parts-of-speech tagging were corrected semi-automatically in the post-edit stage. The grammatical tagging was conducted with a widely used Korean morphological analyzer, *UTagger*, based on the Sejong tagset. For detailed tagset, refer to the Appendix 1. An excerpt ("Watermelon" from patient AP4) of an untagged original transcript and its tagged version is shown below (the English in parentheses is the literal translation of the Korean utterances).

Original transcript

수박은 넝쿨에 널어갖고 있는데 수박이지

su-pak-ŭn nŏng-k'ul-e nŏl-ŏ-kach-ko iss-nŭn-te su-pak-i-chi

(Watermelons are on the vine, it is a watermelon)

Tagged text

수박/NNG 은/JX 넝쿨/NNG 에/JKB 널/VV 어/EC 갖/VX 고/EC 있/VX 는데/EC 수박/NNG 이/VCP 지/EF

2.4 Linguistic features

This research is focused on the word usage patterns across the two groups, rather than on acoustic characteristics. Accordingly, the majority of linguistic features chosen as variables are part-of-speech tagged words, and some non-word phenomena to gauge utterance fluency such as repetitions and the use of filler interjections. The 27 linguistic features presented in Table 2 are grouped into three categories: 1) *word usage pattern* (20 features); 2) *word usage [frequency] level* (3 features); and 3) *utterance fluency* (4 features). This means that we concentrate only on the surface structure of language

production, not underlying deep structure and human-mediated and complicated error patterns.

Table 2. Linguistic features (Caps: variable names)

1. Word usage pattern
– WORDS.NUM (Number of words)
– NOUNS (Number of nouns)
– VERBS (Number of verbs)
– NN_TO_VV (Noun-verb ratio)
– NN_RATIO (Noun ratio)
– DEMON_NP (Number of demonstratives: NP type)
– DEMON_MM (Number of demonstratives: MM type)
– DEMON_VA (Number of demonstratives: VA type)
– DEMON_ALL (Number of demonstratives: all types)
– ETM (Number of adnominal endings)
– ADJS (Number of adjectives
– ADVS (Number of adverbs)
– PRONS (Number of pronouns)
– PRON_TO_NOUN (Pronoun-noun ratio)
– FWS (Number of functional words)
– NUMERALS (Number of numerals)
– JASAS (Number of postpositions)
– SEONOMALS (Number of prefinal endings)
– CONNECT_EOMI (Number of connective endings)
– EOMALS (Number of sentence final endings)
2. Word usage [frequency] level
– FREQ.ALL.NORM (Frequency of all words)
– FREQ.NN.NORM (Frequency of all nouns)
– FREQ.VV.NORM (Frequency of all verbs)
3. Utterance Fluency
– REP (Number of repetition)
– TTR (Type-Token Ratio)
– WORD.LENGTH (Word length)
– FILLER (Number of fillers)

Firstly, the category of 'word usage pattern' evaluates how the two groups' usages of words such as nouns (NOUNS), verbs (VERBS), and postpositions (JOSAS) differ from each other. Among the 20 features of word usage patterns, 4 syntactic complexity indicators are included in the category; adnominal ending (ETM), function words (FWS), connective endings (CONNECT_EOMI), and sentence final endings (EOMALS). Adnominal ending as a subcategory of endings is used to form a modifier clause from adjectives or verbs immediately before a noun (phrase), allowing a clause to be embedded into another clause to formulate a complex sentence. Function words in Korean include all types of endings, postpositions, and connective adverbs. Korean verbs and adjectives require endings to act as appropriate grammatical constituents, and postpositions determine the role played by nouns, pronouns, and numerals in relation to other words, while connective adverbs function approximately as equivalents of the English conjunctions. Connective endings also play the same role as conjunctions in English and connect more than two clauses to form a complex sentence. Among the function words, sentence final endings should be attached to all the sentence closings with verbs and adjectives to signify the exact intended meaning. The frequency distribution of connective endings and sentence final endings could be an indicator of how complete or fragmentary utterances are.

In this study, of the 20 features of word usage patterns, demonstratives are subcategorised into 4 types, NP type (DEMON_NP), MM type (DEMON_MM), and VA type (DEMON_VA), and their combined all types (DEMON_ALL). These respectively signify demonstrative pronouns (e.g., as in 'This is an apple.' / '이것은 사과이다.'), demonstrative adjectives modifying nouns (e.g., as in 'This apple is red.' / '이 사과는 빨갛다.'), and demonstrative adjectives acting as predicates (e.g., as in 'The situation is like this.'/ '상황이 이렇다.'). On the other hand, English has two types of demonstratives: demonstrative pronouns and demonstrative adjectives.

Secondly, for the category of 'word usage [frequency] level,' it is gauged how common and familiar the words used by each group are, because a person with word retrieval deficits is likely to use more familiar words to them. We compared the frequency of all words (all words (FREQ.ALL.NORM), all nouns (FREQ.NN.NORM) and all verbs (FREQ.VV.NORM)) used by the participants with a norm frequency table. The norm frequency table was compiled by extracting frequency information from the *Sejong Corpus*, one of the largest and most balanced publicly accessible Korean language corpora constructed by the government-funded academic consortium. We found the corpus to be suitable for this purpose, given the size and diversity of texts it contains. If a mean frequency of nouns used by one person is significantly higher than that of another person, it may be an indicator that the former uses easier and more general words compared to the latter, meaning that the former is likely to rely on more general words in situations where more concrete and specific words would be suitable.

Finally, the category of 'fluency' was measured according to repetition (REP), type-token ratio (TTR), word length (WORD.LENGTH), and fillers (FILLER). Repetition denotes the number of word or phrase repetitions divided by the number of words. TTR is to measure the level of variety of words used by a speaker, whereas word length, measured as the mean number of characters in each word, signifies the morphological complexity of words used by a speaker. Filler interjections (e.g. *um, ah, eh*) are also one of the characteristics indicative of fluency level. The two features can be said to be proportional to the speaker's utterance fluency.

2.5 Analysis procedures

Using the 27 linguistic features automatically extracted, computational analysis was conducted to identify the most salient features distinguishing the patient group from the normal control group. We first executed PCA to look at how the 20 participants are positioned in a two-dimensional space and how they group together. PCA helps show hidden or unseen structures of data to the naked eye through dimension reduction techniques, facilitating the exploration of complex data. From the two principal components that account for the

majority of the variance, the most important linguistic features correlated with the components are extracted. After that, statistical tests were performed on every linguistic feature to identify the most relevant ones in distinguishing the two groups. In the process, it will be re-confirmed whether the features extracted in the PCA have explanatory power in differentiating the two groups. Finally, a classification model based on machine learning algorithms was built with only the features selected out of the entire feature set in the previous statistical analyses. Machine learning is a computational modelling technique that lets a computer learn from observed data and make decision about new data without human intervention. In a machine-learning classification task, a computer recognizes and learns patterns and their related categories from the data and predicts to which of the categories a new case belongs. In this study, the Bayesian logistic regression algorithm, a subtype of logistic regression, was chosen due to its resistance to data sparsity. The performance of the model was measured by prediction accuracy. The accuracy in a classification task such as the logistic regression is a measure of the ratio of true positives (when a positive case is correctly predicted to be a positive) and true negatives (when a negative case is correctly predicted to be a negative) to all prediction trials. For example, a true positive in this model is the prediction trial where a normal person is correctly classified as a normal person. A true negative is the prediction where 'not a normal person', i.e. a patient is successfully classified as a patient. False positives and false negatives are misclassifications. The higher the accuracy rate is, the better the model is.

3. OUTCOMES AND RESULTS

The language processing for the analysis resulted in 4,696 words for the patient group and 11,207 words long text for the healthy control group (Table 3). The mean number (469.6) of words produced by an aphasic patient was substantially smaller than what a non-aphasic subject produced (1120.7). The least and the most number of words produced in each group were 165 (AP9) and 1070 (AP7), in the patient group, and 431 (NC2) and 2177 (NC8), in the normal control group. In order to set off the impact of the difference in text length between the groups and among the individuals, the values of the variables in the Table 2 were obtained by calculating relative frequency, which means the frequencies used in this study were adjusted against the varying amounts of speech not to bias the results.

Table 3. Number of words in the two groups

Patient group		Normal group	
ID	Number of words	ID	Number of words
AP1	447	NC1	1,844
AP2	254	NC2	431
AP3	326	NC3	1,506
AP4	478	NC4	911
AP5	169	NC5	778
AP6	820	NC6	992
AP7	1,070	NC7	640
AP8	375	NC8	2,177
AP9	165	NC9	610
AP10	592	NC10	1318
Total	4969	Total	11207
Mean (±SD)	497 (±290.21)	Mean (SD)	1,121 (±574.64)

3.1 Overall distribution of the linguistic features

The overall distributions of the 24 linguistic features (excluding the 3 frequency level-related features) across the patient and the reference groups are presented in Fig. 1. This displays the comparative distributions of the linguistic features produced by the subjects, in which each of the plots contrasts the distributions of each linguistic feature between the patient (the right boxes in dark grey) and the reference (the left boxes in light grey) groups.

The heights of boxes in the plots indicate how widely the feature is distributed over the two groups, and the horizontal lines inside the boxes point to the medians of the frequencies. Nouns (NOUNS) and verbs (VERBS), for example, are shown to be used more variably by the patients compared to the normal controls (CV (patients) = 0.2 > CV (normal controls) = 0.158 for nouns; CV (patients) = 0.269 > CV (normal controls) = 0.107 for verbs). The horizontal lines of the verbs, on the other hand, indicate that the median frequency of the verb usage by the two groups is more similar to each other, as compared to the noun usage. As a contrast, the median frequency of 'JOSAS' was lower in the patient group, whereas that of 'EOMALS' was higher in the same group. As mentioned previously, there was a smaller number of words produced by the patient group (see the 'words.num' graph in the 4th row of Fig. 1).

3.2 Differential features between the patient and normal control groups

In accordance with Fig. 1, the PCA plot in Fig. 2 indicates that on dimension 2 (from the perspective of the *y* axis), not on dimension 1, the subjects split into two groups clearly with the two barycenters separated away. This suggests that the linguistic features correlated with dimension 2 will contribute more significantly to the group division, and that the features on the dimension 2 will be the distinctive characteristics distinguishing between the two groups. Between the two groups, on the other hand, the spread of patient data points (APs) shows a less condensed form, which means that the extent to which the linguistic features converge together is lower in the patient group than in the control group.

Among the linguistic features correlated with dimensions 2 are frequency level of all words (FREQ.ALL.NORM), sentence final endings (EOMALS), function words (FWS), numerals (NUMERALS), frequency level of verbs (FREQ.VV.NORM), postpositions (JOSAS) and demonstratives (DEMON_MM). Patient group tend to use significantly more final sentence endings, and less function words, numerals and postpositions, and vice versa, for the normal control group. Patients also tend to deploy more frequent, i.e. more common and easily retrievable, words in contrast to the normal controls.

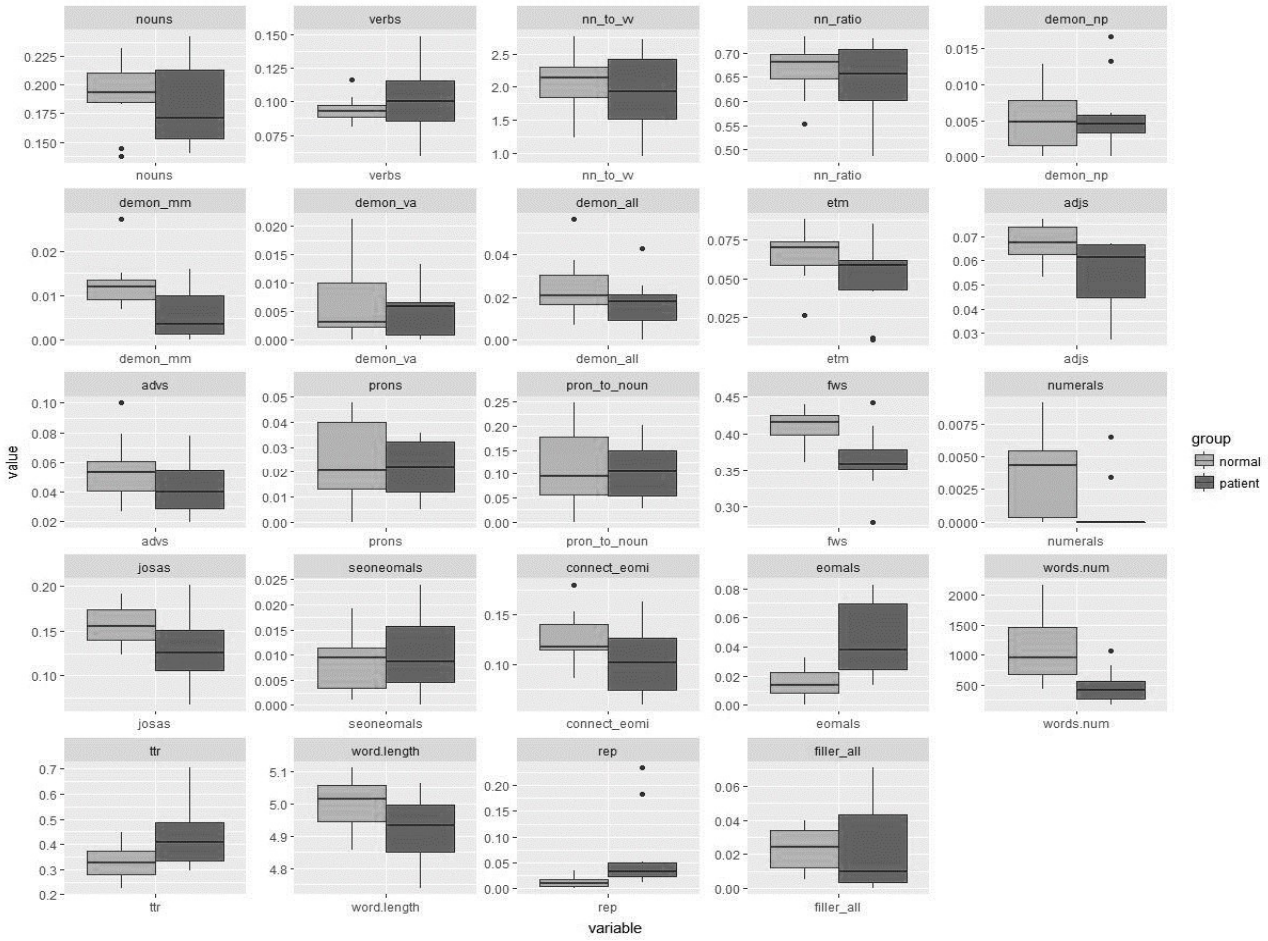

Fig. 1. Graphical presentation of the relative usage of 24 linguistic features between the aphasic patient and normal control groups (The titles of each subgraph are the variable names in Table 2.)

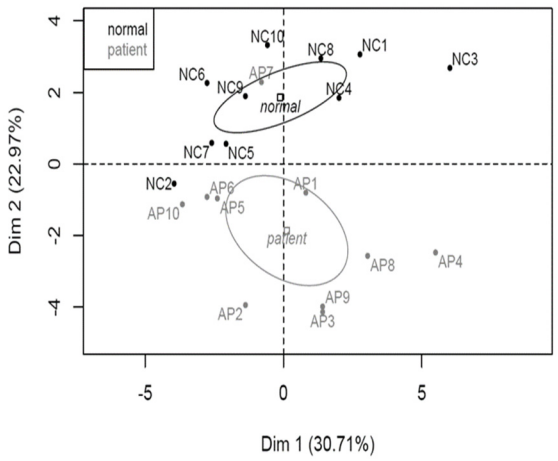

Fig. 2. PCA plot of individual participants of aphasic patient (AP: in grey) and normal control (NC: in black) groups (Two ellipses indicate the barycentres of the two groups.)

To validate the PCA results and the discriminatory features derived from dimension 2, statistical tests were conducted to find out which individual feature is statistically significant in distinguishing the two groups. Given the small number of samples (20 samples in total), permuted *t*-test and *U*-test, rather than ordinary *t*-test, were applied to all the 24 features excluding three frequency level features. To the three frequency level variables that are about frequency comparison, the *Smirnov-Kolmogorov* test was applied. The threshold of 0.05 was set for both the tests to determine if the features are to be accepted as distinctive between the groups. Table 4 presents the nine significant features: demonstratives (DEMON_MM), function words (FWS), numerals (NUMERALS), sentence final endings (EOMALS), frequency level of all words (FREQ.ALL.NORM), adjectives (ADJS), frequency level of nouns (FREQ.NN.NORM), repetitions (REP), and number of words (WORDS.NUM). The first 5 of the 9 features are also variables correlated with dimension 2 of the aforementioned PCA results. This finding suggests that the distributional patterns of formal linguistic features, such as function words, and the frequency level of words, not the content words like nouns and verbs, successfully divide the participants into the groups with and without aphasia.

Table 4. Significant features from statistical analysis

Variable	*U*-test		Permuted *t*-test (**Kolmogorov-Smirnov* test)	
	W	*p*-value	Diff (**D*)	*p*-value
WORDS.NUM	13	0.0039	-651.1	0.0028
DEMON_MM	18	0.0147	-0.00702	0.0122
ADJS	20	0.0232	-0.01302	0.0304
FWS	18	0.0147	-0.04569	0.0078
NUMERALS	24	0.0349	-0.00286	0.0353
EOMALS	87	0.0039	0.02926	0.0028
FREQ.ALL.NORM*	89	0.0021	0.7	0.0123
FREQ.NN.NORM*	90	0.0015	0.9	0.0002
REP	90	0.0028	0.05392	0.0006

3.3 Machine learning-based classification between the patient and normal groups

The features identified from the statistical tests are likely to be critical elements in predicting (or classifying) whether or not an individual is going to be diagnosed as having aphasia based on her/his language use. For this prediction, a logistic regression analysis was carried out with the selected features as variables.

As a feature set to be input into the classification modelling, the features under the category of word usage patterns were chosen that are correlated with dimension 2, and whose statistical values are found to be significant in the statistical tests. As numerous missing values with the variable of numerals (NUMERALS) were found, in particular from the aphasic patients (8 out of 10 persons), it was decided to use the postpositions (JOSAS) instead that are used much more than numerals, though with somewhat higher *p*-values ($p = 0.0684$) than the threshold of $p < 0.05$. The resulting feature set consequently consists of four variables of function words (FWS), sentence final endings (EOMALS), demonstratives (DEMON_MM), and postpositions (JOSAS).

The classification model was built using the Bayesian logistic regression, whose classificatory performance is assessed in terms of accuracy. The fit was bootstrapped 100 times to guarantee the stability of the model performance. The higher the accuracy rate is, the more relevant the selected features can be said to be to the prediction of the aphasic symptoms. The resulting model achieved a 90% success rate in distinguishing the patients from non-patients, far above the baseline of 50%, with 2 wrong predictions occurring out of 20 prediction trials. Table. 5 shows the result of class predictions. In the model, one patient (AP7) was misclassified as the other class with a patient class probability of 21.9%, and one non-aphasic (NC2) was assigned to the aphasic group with a normal class probability of 29.8%. These (un-)successful predictions are forecast in Fig. 3, where two persons (AP7 and NC2) are positioned together with the members of the other group. The possible reason why the two subjects were misclassified will be detailed in the next Discussion section.

Also of note is that among 10 persons on the left part of the Table. 5 (i.e. the patient group), 5 persons are correctly classified with more than 90% probability, whereas the other four persons are correctly classified with less than 80% probability ($SD = 0.14$ excluding the AP7). On the right side (i.e. the normal control group), only one has been classified as non-aphasic with more than 90% probability and six persons are classified with more than 80% probability ($SD = 0.097$ excluding the NC2). This suggests that the patients show more prominent characteristics of their own in terms of the four functional word categories compared to the other group. The higher level of irregularity (SD = 0.14 > SD = 0.097) found in the patient group, on the other hand, may be related to external factors such as the type of illness and education level, which is beyond the scope of this research.

Table 5. Class prediction by the model (90% accuracy)

Patient group	Prediction	Normal group	Prediction
AP1	patient	NC1	normal
AP2	patient	NC2	patient
AP3	patient	NC3	normal
AP4	patient	NC4	normal
AP5	patient	NC5	normal
AP6	patient	NC6	normal
AP7	normal	NC7	normal
AP8	patient	NC8	normal
AP9	patient	NC9	normal
AP10	patient	NC10	normal

To assess the stability of the model, two supplementary models were set up, one of which was constructed by incorporating two more variables into the initial model, and the other by removing a variable from the model. Into the first supplementary model to test the completeness of the original model, two variables were added that did not belong to word usage patterns but correlated with dimension 2 of the PCA model – namely, the frequency level of all words (FREQ.ALL.NORM) and frequency level of verbs (FREQ.VV.NORM). The two predictors were not found to improve the overall performance of the initial model. The model achieved the same accuracy rate (90%) with the same two persons (AP7 and NC2) misclassified, though the mean probability of patient candidates being correctly classified as patients rose marginally from 78.1% to 84.6%, suggesting that the frequency levels of words as a variable could play a minor role in distinguishing the two groups.

The second supplementary model, on the other hand, was constructed by removing one of the variables from the original model, which was the sentence final endings (EOMALS), a word class with the highest statistical significance and the highest correlation coefficient with group differentiation. It could be considered to be a predictable feature in that persons with aphasia tend to produce fragmentary sentences more frequently, and therefore, use more final endings. If the performance of the model would not deteriorate much, the original model with the linguistic features could be said to work well and stably. So, the removal of this predictable feature could provide a way to test the robustness of the original classification model. It turned out that the new model achieved an accuracy of 85%, with only three out of 20 persons misclassified. Two non-aphasic subjects (NC2 and NC5) and one aphasic patient (AP7) were misclassified. Without the

sentence final endings, the most important predictor, the model has successfully distinguished the two groups, operating on the other three functional word features.

The analyses have confirmed that functional linguistic features such as sentence endings, demonstratives, conjunctions and case markers are more crucial than content-related words (e.g. nouns and verbs), which are traditionally thought to be more significant factors, in distinguishing persons with aphasia from non-aphasic persons.

4. DISCUSSION

This study is noteworthy in that it is the first attempt to identify linguistic features, i.e. word usage patterns, to distinguish an aphasic patient from a non-aphasic subject, and to test how successfully the statistical and machine learning models based on the features can separate the two subject groups with and without aphasia.

4.1 General aphasic characteristics of the patient group

From the results, it was not surprising to observe two main characteristics of the aphasic patient group which have long been reported as aphasic symptoms: a reduced amount of speech and broad spectrum of performance variance across patients.

Firstly, the aphasic group's reduced amount of speech production was demonstrated by the use of less than half of the mean number of total words compared with that produced by the normal group. None of the aphasic patients exceeded the mean number of total words produced by the normal control group, regardless of the type and severity of aphasia. The aphasic phenomenon might also be aptly described as a word retrieval problem. Word retrieval deficits are widespread in aphasia [21], irrespective of the aphasia type. The big gap in the frequency level of words (i.e., at the 'word usage [frequency] level' category) between the two groups could be indicative of a word retrieval problem, resulting in the frequent use of easier and more familiar words by the aphasic group compared to the control group [22]. Another reasonable explanation for a smaller number of total words produced in the aphasic group could be the long pause duration and speech rate during the utterances observed in aphasia [23]. They argue that long pauses may be characterized as an index of the internal cognitive processes associated with sentence planning.

Secondly, performance variance across patients against the normal controls was observed. The variance phenomenon was exemplified by the graphical presentation of the relative usage of linguistic features between the two groups (see Fig. 1). Each patient may differ in various factors such as types of aphasia, severity of disorders, post onset time, size and locus of lesion, and education/literacy level, which may result in quantitative and qualitative variance among patients [24].

4.2 Function words category as strong classification features

The main finding from the current computational analysis of Korean utterances in terms of word usage pattern is that the *functional* linguistic features, such as sentence final endings (EOMALS), postpositions (JOSAS) and demonstratives (DEMON_MM), are more crucial in distinguishing aphasic patients and healthy controls than the content-related words (e.g. nouns, verbs, adjectives). These strong classification features between the two groups are less related to 'what' a speaker says than to 'how' she/he says it.

Firstly, the use of sentence final endings is much more frequent in the patient group despite the reduced overall number of words produced. In fact, patients with anomia frequently interrupt the utterance by terminating with sentence final endings, which diminishes the levels of sentence completeness [25], [26]. As for the sentence final endings, there are some important facts regarding Korean word usage. Korean is a verb-salient language [15] in which Korean verbs must have endings even in their expressive form of a root, located at the end of a sentence. Specifically, the grammatical structure of a Korean sentence consists of 'subject + object + verb' (e.g., 나는 사과를 샀다.' instead of 'I bought an apple' in English) where the verb follows the object. Unlike case markers which specify the role of a noun in a sentence, sentence final endings are attached to the end of the verb and finalize sentences.

Secondly, while sentence final endings were more frequently used features in the patient group, the normal control group more frequently used postpositions (JOSAS) appended to nouns, pronouns, and numerals. Less usage of postpositions in the aphasic group may be an indicator of syntactic deficits, especially in non-fluent aphasia. It can change the meaning of a sentence or utterance when we use postpositions in a different and wrong way. For example, case markers, which account for the majority of the postpositions (see Appendix 1), are very important elements in Korean language in that they determine the functional and relational roles of nouns in a sentence. If the roles of the subject and object need to be switched in a sentence in English, e.g., 'The man liked the woman,' the locations of the subject and the object must be interchanged, as in 'The woman liked the man.' However, in Korean, the nominative case marker needs to be replaced with the objective case marker and vice versa, while maintaining the locations of the subject and object. For instance, the nominative case marker '가' and the objective case marker '를' are used interchangeably in order to switch the thematic roles of the subject versus the object without changing the loci in a sentence, '남자가 여자 를 좋아했다' → '남자를 여자 가 좋아했다 '. Therefore, inappropriate uses and frequent omissions of case markers will obscure the relationship between grammatical constituents like the subject, the object, and the verb within a sentence or an utterance, and confuse the listener about what the speaker intends to say. In fact, a study has shown that analysis of morpho-syntactic feature of case markers of Korean has aided in differentiating a mild cognitive impaired group from a normal control group [27].

Thirdly, the more frequent use of demonstrative adjectives modifying nouns by the control group compared with the patient group can also be indicative of differences in two groups' uses of morpho-syntactic and/or semantic functions. In

this study, we trichotomized the types of demonstratives according to Korean linguistic features. A demonstrative adjective proceeds to modify a substantive indicating its specificity, whereas the use of a demonstrative pronoun conveys unspecificity of what is referred to by a specific noun. Even though no clear group difference is observed in the use of demonstrative pronouns in this study, they are produced more frequently by various clinical groups, who use them to substitute less specific words, than normal control group [28]. It could be inferred accordingly that the utterance of the patient group is vaguer in referring to words in context compared to the control group.

The significance of the functional word classes as separating indicators between the two groups was demonstrated in the successful classification modelling. The classification model built on the features of function words, sentence final endings, demonstratives, and postpositions was able to identify the aphasic patients with a 90% success rate. The robustness of the model, i.e. the relevance of the selected linguistic features, was demonstrated by the fact that there was little change in performance even when more features were added or a feature removed. As for the two misclassified cases in Fig. 2 and Table 5, the aphasic subject of AP7 displayed the mildest severity of aphasia with AQ of 91.9 and the largest number of spoken words (i.e., 1070). On the contrary, the normal control of NC2 was misclassified as an aphasic subject possibly due to having the smallest number of spoken words (i.e., 431) and frequent use of sentence final endings (EOMALS), which is revealed by a detailed analysis of feature rates of the 20 participants. The issue of these borderline errors would be resolved as larger and more diverse utterance data are available according to a clearer and more refined scheme.

4.3 Limitations

Despite the significant value of this paper, there is a limitation to our study. Most importantly, the heterogeneous nature of aphasic subjects with the relatively small number of subjects in each aphasia type adds the complication of increasing subject variance. Further research on a larger cohort of patients will be needed to explore ways in which linguistic features are linked to the specific types and levels of severity of aphasia.

5. CONCLUSION AND IMPLICATIONS

The study is significant in that it has provided crucial directions for clinical research in evaluating the spontaneous utterances of aphasia. The computational linguistic analysis employed in this study has proved quite accurate in identifying aphasic patients by producing output of a linguistic description of the patient group.

ACKNOWLEDGEMENTS

This work was supported by a National Research Foundation of Korea Grant, funded by the Korean Government (MOE) (NRF-2009-361-A00027). This paper was presented at the 10th Biennial Conference of the Asia-Pacific Society of Speech-Language & Hearing (APSSLH), held in Narita, Japan on 17th ~19th September, 2017.

REFERENCES

[1] A. Rofes, A. Talacchi, B. Santini, G. Pinna, L. Nickels, R. Bastiaanse, and G. Miceli, “Language in individuals with left hemisphere tumors: is spontaneous speech analysis comparable to formal testing?,” Journal of Clinical Experimental Neuropsychology, vol. 40, no. 7, 2018, pp. 722-732.

[2] A. M. Cohen-Goldberg, J. Cholin, M. Miozzo, and B. Rapp, “The interface between morphology and phonology: Exploring a morpho-phonological deficit in spoken production,” Cognition, vol. 127, no. 2, May. 2013, pp. 270-286.

[3] E. Schönberger, S. Heim, E. Meffert, P. Pieperhoff, P. da Costa Avelar, W. Huber, F. Binkofski, and M. Grande, “The neural correlates of agrammatism: Evidence from aphasic and healthy speakers performing an overt picture description task,” Frontiers in Psychology, vol. 5, Mar. 2014, p. 246.

[4] M. L. Kean, *Agrammatism*, Academic Press., London, 1985.

[5] H. Goodglass, “Agrammatism in aphasiology,” Clinical Neuroscience, vol. 4, no. 2, 1997, pp. 51-56.

[6] S. Ahmed, A. M. F. Haigh, C. A. de Jager, and P. Garrard, “Connected speech as a marker of disease progression in autopsy-proven Alzheimer’s disease,” Brain, vol. 136, no. 12, Oct. 2013, pp. 3727-3737.

[7] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, “Aided diagnosis of dementia type through computer-based analysis of spontaneous speech,” Proc. Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2014, pp. 27-36.

[8] A. J. Astell and T. A. Harley, “Accessing semantic knowledge in dementia: evidence from a word definition task,” Brain and Language, vol. 82, no. 3, Sep. 2002, pp. 312-326.

[9] S. R. Kim, S. Kim, M. J. Baek, and H. Kim, “Abstract word definition in patients with amnestic mild cognitive impairment,” Behavioral Neurology, 2015, Article ID 580246.

[10] C. Brown, T. Snodgrass, S. J. Kemper, R. Herman, and M. A. Covington, “Automatic measurement of propositional idea density from part-of-speech tagging,” Behavior Research Methods, vol. 40, no. 2, May. 2008, pp. 540-545.

[11] L. Bryant, E. Spencer, A. Ferguson, H. Craig, K. Colyvas, and L. Worrall, “Propositional idea density in aphasic discourse,” Aphasiology, vol. 27, no. 8, Aug. 2013, pp. 992-1009.

[12] P. Garrard and R. Forsyth, “Abnormal discourse in semantic dementia: A data-driven approach,” Neurocase, vol. 16, no. 6, Nov. 2010, pp. 520-528.

[13] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, “Linguistic features identify Alzheimer’s disease in narrative speech,”

Journal of Alzheimer's Disease, vol. 49, no. 2, Jan. 2016, pp. 407-422.
[14] P. Garrard, V. Rentoumi, B. Gesierich, B. Miller, and M. L. Gorno-Tempini, "Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse," Cortex, vol. 55, Jun. 2014, pp. 122-129.
[15] J. E. Sung, G. DeDe, and S. E. Lee, "Cross-linguistic differences in a picture-description task between Korean- and English-speaking individuals with aphasia," American Journal of Speech-Language Pathology, vol. 25, no. 4S, Dec. 2016, pp. S813-S822.
[16] J. R. Hanley, R. P. Hunt, D. A. Steed, and S. Jackman, "Concreteness and word production," Memory & cognition, vol. 41, no. 3, Apr. 2013, pp. 365-377.
[17] E. Mainela-Arnold, J. L. Evans, and J. A. Coady, "Explaining lexical–semantic deficits in specific language impairment: The role of phonological similarity, phonological working memory, and lexical competition," Journal of Speech, Language, and Hearing Research, vol. 53, no. 6, Dec. 2010, pp. 1742-1756.
[18] G. Robinson, J. Blair, and L. Cipolotti, "Dynamic aphasia: an inability to select between competing verbal responses?," Brain: a journal of neurology, vol. 121, no. 1, Jan. 1998, pp. 77-89.
[19] L. K. Tyler, H. E. Moss, and F. Jennings, "Abstract word deficits in aphasia: Evidence from semantic priming," Neuropsychology, vol. 9, no. 3, Jul. 1995, p. 354.
[20] A. M. Raymer, L. M. Maher, M. L. Greenwald, M. Morriset, L. J. G. Rothi, and K. M. Heilman, *The Florida Semantic Battery: Experimental Edition*, Gainesville, Florida, 1990.
[21] R. Grima and S. Franklin, "Usefulness of investigating error profiles in diagnosis of naming impairments," International Journal of Language and Communication Disorders, vol. 52, no. 2, Mar. 2017, pp. 214-226.
[22] A. K. Kittredge, G. S. Dell, J. Verkuilen, and M. F. Schwartz, "Where is the effect of frequency in word production? Insights from aphasic picture-naming errors," Cognitive Neuropsychology, vol. 25, no. 4, Jun. 2008, pp. 463-492.
[23] G. Angelopoulou, D. Kasselimis, G. Makrydakis, M. Varkanitsa, P. Roussos, D. Goutsos, I. Evdokimidis, and C. Potagas, "Silent pauses in aphasia," Neuropsychologia, vol. 114, Jun. 2018, pp. 41-49.
[24] R. M. Lazar, A. E. Speizer, J. R. Festa, J. W. Krakauer, and R. S. Marshall, "Variability in language recovery after first-time stroke," Journal of Neurology, Neurosurgery and Psychiatry, Sep. 2007, pp. 530-534.
[25] S. Andreetta, A. Cantagallo, and A. Marini, "Narrative discourse in anomic aphasia," Neuropsychologia, vol. 50, no. 8, Jul. 2012, pp. 1787-1793.
[26] D. Fromm, M. Forbes, A. Holland, S. G. Dalton, J. Richardson, and B. MacWhinney, "Discourse characteristics in aphasia beyond the Western Aphasia Battery cutoff," American Journal of Speech-Language Pathology, vol. 26, no. 3, Aug. 2017, pp.762-768.
[27] J. M. Hyun, J. E. Sung, J. H. Jeong, H. J. Kang, and H. J. Kim, "Effects of syntactic complexity on a case marker processing task in people with mild cognitive impairment," Communication Sciences and Disorders, vol. 18, no. 1, Mar. 2013, pp. 35-46.
[28] K. C. Fraser, J. A. Meltzer, N. L. Graham, C. Leonard, G. Hirst, S. E. Black, and E. Rochon, "Automated classification of primary progressive aphasia subtypes from narrative speech transcripts," Cortex, vol. 55, Jun. 2014, pp. 43-60.

HyangHee Kim
She received a Ph.D. in Communicative Disorders and Sciences from University of Wisconsin-Madison, USA in 1994. Currently, she is a professor at the Graduate Program in Speech-Language Pathology, Yonsei University College of Medicine, Korea. Her main research interests include aging and neurologic communication disorders and dysphagia.

Ji-Myoung Choi
He is a Ph.D. candidate in the Interdisciplinary Graduate Program in Linguistics and Informatics, after receiving an M.A. degree in Linguistic Information Science from Yonsei University. His research interests include stylometry, document classification, and language modelling based on statistical and machine learning methods, as well as corpus linguistics and lexicography.

Hansaem Kim
She received a Ph.D. in Interdisciplinary Graduate Program in Linguistics and Informatics from Yonsei University, Korea in 2005. Currently, she is a professor at the Institute of Language and Information Studies of Yonsei University. Her main research interests include construction of language resource, corpus linguistics, and lexicography.

Ginju Baek
She received a Master's degree in Speech-Language Pathology from Yonsei University in 2017. Currently, she works as a speech-language pathologist at the Rehabilitation Hospital, Yonsei University Health System. Her main research interests include developmental delay and neurologic communication disorders.

Bo Seon Kim

She received a Ph.D. in Graduate Program in Speech-Language Pathology from Yonsei University, Seoul, Korea in 2017 after receiving an M.A. degree in Linguistics from University of Delaware, USA in 2009. Since receiving a Ph.D., she has worked as a speech language pathologist. Her research interests include discourse, aging, and neurologic communication disorders.

Sang Kyu Seo

He received a Ph.D. degree in Korean historical grammar (16th century) in 1992 from Yonsei University, Korea. He is a professor at the Department of Korean Language and Literature at Yonsei University. His research interests include spoken/written corpus construction, lexicography

APPENDIX 1. Part-of-speech scheme in Korean (Sejong tagset)

Word class	Subcategory	Tag
Noun	common noun	NNG
	proper noun	NNP
	bound noun	NNB
Pronoun	Pronoun	NP
Numeral	Numeral	NR
Verb	Verb	VV
Adjective	Adjective	VA
Auxiliary verb	auxiliary verb	VX
Copula	positive copula	VCP
	negative copula	VCN
Adnominal	Adnominal	MM
Adverb	general adverb	MAG
	connective adverb	MAJ
Interjection	Interjection	IC
Postposition	nominative case marker	JKS
	complement case marker	JKC
	deterministic case marker	JKG
	objective case marker	JKO
	adverbial case marker	JKB
	vocative case marker	JKV
	quotative case marker	JKQ
	auxiliary particle	JX
	connective particle	JC
Ending	prefinal ending	EP
	final ending	EF
	connective ending	EC
	Nominal ending	ETN
	Adnominal ending	ETM
Prefix	Derivational prefix of noun	XPN
Suffix	Derivational suffix of noun	XSN
	Derivational suffix of verb	XSV
	Derivational suffix of adjective	XSA
Word root	Root	XR

APPENDIX 2. All feature rates of the 20 participants (Figures indicate relative frequency except for the variable "WORDS.NUM". WORDS.NUM figures represent raw number of words per subject.)

GROUP	ID	WORDS.NUM	NOUNS	VERBS	NN_TO_VV	NN_RATIO	DEMON_NP	DEMON_MM	DEMON_VA	DEMON_ALL	ETM	ADJS	ADVS	PRONS	PRON_TO_NOUN
patient	AP1	447	0.18345	0.10738	1.70833	0.63077	0.00447	0.00671	0.00671	0.0179	0.06264	0.05145	0.04474	0.03356	0.15464
patient	AP2	254	0.15748	0.05906	2.66667	0.72727	0.00394	0.00394	0.01181	0.01969	0.01181	0.06693	0.01969	0.01969	0.11111
patient	AP3	326	0.17485	0.11656	1.5	0.6	0.00307	0.00307	0.00613	0.01227	0.06442	0.02761	0.02454	0.02761	0.13636
patient	AP4	478	0.14017	0.14854	0.94366	0.48551	0.01674	0.00209	0.00628	0.0251	0.01046	0.0272	0.03556	0.03556	0.20238
patient	AP5	169	0.2426	0.11243	2.15789	0.68333	0	0	0	0	0.04142	0.06509	0.05917	0.01183	0.04651
patient	AP6	820	0.16707	0.06829	2.44643	0.70984	0	0.00122	0	0.00122	0.04756	0.06707	0.07805	0.00488	0.02837
patient	AP7	1070	0.22243	0.09346	2.38	0.70414	0.00467	0.01121	0.00561	0.0215	0.05981	0.05794	0.02897	0.0243	0.09848
patient	AP8	375	0.14667	0.09333	1.57143	0.61111	0.01333	0.016	0.01333	0.04267	0.05867	0.06667	0.05333	0.03467	0.19118
patient	AP9	165	0.15152	0.12727	1.19048	0.54348	0.00606	0.01212	0	0.01818	0.08485	0.04242	0.05455	0.01212	0.07407
patient	AP10	592	0.22466	0.08277	2.71429	0.73077	0.00507	0	0.00338	0.00845	0.05912	0.06588	0.02872	0.00676	0.0292
normal	NC1	1844	0.1372	0.09111	1.50595	0.60095	0.00868	0.01139	0.0103	0.03037	0.07267	0.06996	0.10033	0.04393	0.24251
normal	NC2	431	0.2181	0.09281	2.35	0.70149	0	0.00696	0	0.00696	0.08817	0.07657	0.03712	0	0
normal	NC3	1506	0.14409	0.1162	1.24	0.55357	0.00797	0.02722	0.02125	0.05644	0.02656	0.05312	0.0591	0.04781	0.24913
normal	NC4	911	0.18332	0.10318	1.7766	0.63985	0.00439	0.01317	0.01976	0.03732	0.05598	0.06806	0.0494	0.04061	0.18137
normal	NC5	778	0.20051	0.08098	2.47619	0.71233	0.00514	0.009	0.00257	0.01671	0.07455	0.07712	0.0617	0.01671	0.07692
normal	NC6	992	0.23185	0.08367	2.77108	0.73482	0.00101	0.01512	0.00302	0.01915	0.06754	0.05343	0.0494	0.0131	0.0535
normal	NC7	640	0.21406	0.09844	2.1746	0.685	0	0.0125	0	0.0125	0.06563	0.06719	0.02656	0.00781	0.03521
normal	NC8	2177	0.18879	0.09279	2.03465	0.67047	0.00735	0.01378	0.00919	0.03032	0.05145	0.06615	0.07901	0.03813	0.16802
normal	NC9	610	0.19672	0.09344	2.10526	0.67797	0.00328	0.00984	0.00328	0.01639	0.07213	0.07541	0.0377	0.01311	0.0625
normal	NC10	1318	0.19044	0.08801	2.16379	0.68392	0.0129	0.00759	0.00228	0.02276	0.08801	0.06146	0.05615	0.02428	0.11307

GROUP	ID	FWS	NUMERALS	JOSAS	SEONEOMALS	CONNECT_EOMI	EOMALS	FREQ.ALL.NORM	FREQ.NN.NORM	FREQ.VV.NORM	WORD.LENGTH	TTR	REP	FILLER_ALL
patient	AP1	0.36689	0	0.12081	0.00224	0.12975	0.02685	0.40153	0.039	0.11769	4.95973	0.4877	0.02614	0.00224
patient	AP2	0.27953	0	0.06693	0.01575	0.06693	0.08268	0.41357	0.00615	0.20692	4.90551	0.48425	0.0412	0.07087
patient	AP3	0.33436	0	0.10123	0.01534	0.05828	0.07055	0.53417	0.05324	0.14248	4.84356	0.36503	0.05233	0.0092
patient	AP4	0.41004	0	0.12762	0.00418	0.16318	0.07113	0.41655	0.04047	0.11961	4.73849	0.37238	0.03823	0.01046
patient	AP5	0.35503	0	0.12426	0.00592	0.15976	0.01775	0.48093	0.03459	0.06112	5.02367	0.70414	0.02874	0
patient	AP6	0.35366	0	0.14756	0.00732	0.0939	0.03537	0.35684	0.0383	0.15659	4.96829	0.32317	0.01792	0.02927
patient	AP7	0.44206	0.00654	0.20187	0.01589	0.11495	0.02336	0.31808	0.03847	0.11264	5.00561	0.29533	0.02104	0.0028
patient	AP8	0.34933	0	0.08533	0.024	0.096	0.04	0.48048	0.05189	0.17722	4.87733	0.44267	0.234	0.048
patient	AP9	0.38182	0	0.15152	0	0.06061	0.06667	0.578	0.07877	0.19612	4.81818	0.50303	0.01198	0.00606
patient	AP10	0.36318	0.00338	0.16385	0.01014	0.10642	0.01351	0.44431	0.03543	0.20481	5.0625	0.30912	0.18269	0.06926
normal	NC1	0.42245	0.00542	0.16974	0.01085	0.12419	0.00976	0.24665	0.03356	0.10606	4.93601	0.22397	0.01806	0.02603
normal	NC2	0.38515	0	0.14617	0.0116	0.08585	0.03248	0.43285	0.02571	0.14393	5.11137	0.4478	0	0.03944
normal	NC3	0.43891	0.00465	0.12351	0.01926	0.18061	0.02457	0.25558	0.03026	0.09608	4.85724	0.27756	0.01309	0.01793
normal	NC4	0.42261	0	0.14929	0.0011	0.15368	0.01207	0.29356	0.03048	0.1145	4.92645	0.32711	0.01512	0.02964
normal	NC5	0.36118	0	0.13496	0.009	0.11568	0.00643	0.36858	0.03394	0.12905	5.01671	0.34576	0.01763	0.03985
normal	NC6	0.40927	0.00403	0.19153	0.00302	0.10484	0.01613	0.28732	0.02514	0.11592	5.10181	0.3246	0.03405	0.02218
normal	NC7	0.40469	0.00156	0.17656	0.00313	0.11875	0.01719	0.36165	0.02801	0.14428	5.02969	0.38125	0	0.03594
normal	NC8	0.39596	0.00551	0.13826	0.01929	0.11713	0.02848	0.20006	0.02683	0.07551	5.01562	0.24116	0.00457	0.01011
normal	NC9	0.42459	0.0082	0.1623	0.00984	0.1459	0.0082	0.37449	0.02705	0.12379	4.96885	0.41639	0.00651	0.0082
normal	NC10	0.42792	0.0091	0.17527	0.00455	0.11457	0	0.25566	0.02747	0.1044	5.06525	0.29135	0.00603	0.00531