Annotating Korean Demonstratives

Sun-Hee Lee
Wellesley College
Wellesley, USA
slee6@wellesley.edu

Jae-young Song
Yonsei University
Seoul, Korea
jysong@yonsei.ac.kr

Abstract

This paper presents preliminary work on a corpus-based study of Korean demonstratives. Through the development of an annotation scheme and the use of spoken and written corpora, we aim to determine different functions of demonstratives and to examine their distributional properties. Our corpus study adopts similar features of annotation used in Botley and McEnery (2001) and provides some linguistic hypotheses on grammatical functions of Korean demonstratives to be further explored.

1 Introduction

Korean demonstratives are known to have two different functions: anaphoric and deictic reference. Anaphoric demonstratives refer to objects, individuals, events, situations, or propositions in the given linguistic context. Deictic demonstratives refer to physical objects, individuals, or positions (or regions) in the given situational context. Deictic variations commonly signal the speaker's physical distance from specified items. Previous literature on Korean demonstratives has focused on deictic functions in spoken Korean, but a comprehensive approach to their diverse linguistic functions is still lacking. This study examines distinct usages of Korean demonstratives in a spoken and a written corpus through the annotation of relevant linguistic features. Our annotation scheme and features are expected to help clarify grammatical functions of Korean demonstratives, as well as other anaphoric expressions.

English demonstratives show a binary distinction that depends on physical distance; there is a distinction between proximal forms (*this*, *these*, *this N*, *these Ns*) and distal forms (*that*, *those*, *that N*, *those Ns*). In contrast, demonstratives in languages like Korean and Japanese show a three-way distinction: proximal forms, speaker-centered distal forms, and speaker- and hearer-centered distal forms. For example, deictic demonstrative *i* refers to a proximal object relative to the speaker, *ku* refers to a distant object that is close to the hearer, and *ce* refers to a distant object that is far from both the speaker and the hearer. Thus, distinct usage of *ce* and *ku* is associated with how the speaker allocates the deictic center and contextual space, i.e., the speaker-centered space vs. the speaker- and the hearer–centered space. In contrast with deictic usage, previous studies (Chang, 1980; Chang, 1984) assumed that anaphoric demonstratives show only a two-way distinction between proximal forms *i* and distal forms *ku*. However, it is still controversial as to whether the boundaries between anaphora and deixis are clear cut. With our annotation scheme, we aim to capture the linguistic properties contributing to interpretations of demonstratives in Korean. In particular, we aim to determine whether different registers or genres contribute to different functions of demonstratives by comparing their usage in a spoken corpus and a written corpus.

In consideration of a future comparative analysis with English demonstratives, we have designed our annotation scheme by adopting Botley and McEnery's (2001) paradigmatic set of distinctive features for English demonstratives. However, the detailed annotation features have been revised according to language specific features of Korean.

2 Corpus Study

For data extraction, we used two Sejong tagged corpora including a 20,343 *eojeol* spoken corpus and 21,023 *eojeol* written corpus.[1] Each corpus is

[1] The term *eojeol* refers to a unit set off by spaces and corresponds to a word unit in English.

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 162–165,
Uppsala, Sweden, 15-16 July 2010.

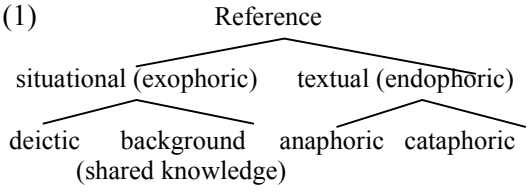
composed of four conversations/texts with approximately 5000 *eojeol*. The subcorpora of the spoken corpus are everyday conversations without assigned topics and those of the written corpus are three newspaper articles and part of a novel.

Compared to English, Korean demonstratives include more complex grammatical categories with morphological relations. The demonstrative forms *i*, *ku*, and *ce* combine with other words or morphemes and form complex words including nominals (e.g., *i-kes*: this+thing 'this'), adverbs (e.g., *ce-lehkey*: that+way 'that way'), adjectives (e.g., *ku-lehata*: it+is 'is so') and other lexical categories. Thus, it is difficult to determine if they all belong to the same category of demonstratives in Korean. In this study, demonstratives are restricted to words that contain *i*, *ku*, and *ce* maintaining a distinct referentialfunction of pointing. The selected demonstratives include adnouns (*i* N 'this N', *ku* N 'that N', *ce* 'that N'), pronouns (*i-es/i-ke* 'this', *ku-kes/ ku-ke* 'it', *ce-kes/ceke* 'that', *i-tul* 'these', *ku-tul* 'they' *ce-tul*), and locative pronouns (*yeki* 'here', *keki* 'there', *ceki* 'over there'). Although those forms have different lexical categories, strong similarities exist within the same morphological families, which we will refer to as *i* type, *ku* type, and *ce* type demonstratives. Our annotation work aims to extract a generalization of the fundamental usage of the three different types and to use that generalization for developing further research on various morphological variants containing *i, ku,* and *ce.*

2.1 The Annotation Scheme

In order to mark referential functions of Korean demonstratives, we first adopt Halliday and Hasan's (1976) classification of the different reference functions of demonstratives: exophoric vs. endophoric usage. We further divide exophora into deixis and background. While the former refers to a physical object or an individual (or location) in the situational context, the latter refers to certain shared information between the speaker and the hearer.

(1)



Six distinct features include "Lexical Category of a Demonstrative", "Endophoricity", "Exophoricity", "Syntactic Category of an Antecedent", "Phoric Type", and "Semantic Function of an Antecedent". The first five features are adopted from five features in Botley and McEnery's (2001) annotation work on English demonstratives.[2] The last feature (semantic function) has been added for future work annotating semantic information that facilitates anaphor resolution processes.

Lexical categories of Korean demonstratives in this study include four parts of speech: adnoun, pronoun, locative pronoun (functioning also as an adverb), and exclamatory expressions. While the first three categories show referential functions, the exclamatory expressions do not have reference. Instead, they are used as expressions conveying the speaker's emotion or state, e.g., embarrassment, confusion, hedging. We do not, however, exclude the possibility of linguistic connectivity between demonstrative and exclamatory forms. For instance, the distal demonstrative form *ce* tends to be used as a hedging expression in Korean. Our study includes exclamatory usage as an annotation feature.

Endophoricity refers to two different functions: anaphoric vs. cataphoric. Exophoricity refers to context based vs. deixis. According to Halliday and Hasan's classification in (1), demonstratives with referential function show two major usages: endophoric and exophoric. The first type takes its antecedent within the given text; the latter, within the given situation. Distinction between an anaphor and a cataphor depends on the position of the antecedent. When an endophor follows its antecedent, it is an anaphor; the other case is a cataphor. Demonstratives may have different types of antecedents syntactically. The corresponding values include nominals (including N or NP), clausals (including V, A, VP,

[2] As one of the reviewers pointed out, our study has some limitations as it only refers to two previous studies, Halliday and Hasan (1976) and Botley and McEnery (2001). Although we are aware of the other fundamental work including demonstratives in a broader range of referential expressions such as Gundel et al. (1993), Prince (1981), Nissim et al. (2004), etc., we choose to focus on Korean demonstratives because their exact grammatical functions have not been comprehensively studied in existing literature. In addition, developing a broader classification system for referential expressions in Korean is a challenging task from both theoretical and empirical perspectives; linguistic analyses of Korean nominal expressions must deal with controversial issues such as definiteness without articles, zero elements functioning as anaphors, unsystematic morphological marking of plurality and genericity, etc.

AP, etc.), and sentential elements (S or Ss for more than two sentences).[3]

The feature semantic function of an antecedent includes values of nominal entities, events, and propositions. This feature will be expanded into specified values such as event, process, state, and circumstances in our future study. Phoric type has been adopted from Botley and McEnery (2001) and refers to two distinct relations: reference and substitution. According to Halliday and Hasan, substitution is a relation between linguistic forms, whereas reference is a relation between meanings. The values of phoric type also include non-phoric such as exophora whose antecedents exist outside the text.

The annotation features and values we use are summarized in Table 1.

Feature	Value1	Value2	Value 3	Value4
Lexical Category (L)	AN (adnoun)	PR (Pronoun)	LPR (Locative pronoun)	EX (Exclamation)
Endophoricity (O)	A (anaphor)	C (cataphor)		
Exophoricity (X)	T (situational)	D (deictic)		
Syntactic Function (F)	NO (nominals)	CL (clausal)	S (sentential)	
Semantic Function (M)	N (entities)	E (event)	P (propositions)	
Phoric Type (H)	R (reference)	U (Substitution)	K (non-phoric)	

Table 1 Annotation Features and Possible Values

The initial results of inter-annotator agreement between two trained annotators are promising. Cohen's Kappa is 0.76 for the average agreement of six high level categories and it increases following a discussion period (K = 0.83, K=2)[4].

3 Results

We identified 1,235 demonstratives in our pilot study. The distributions of demonstratives were significantly different between the spoken and the written corpora. Table 2 shows the raw frequencies in the spoken and the written corpora for each combination of feature and value outlined in Table 1. The raw frequencies are supplemented with the log likelihood in order to show the significance for frequency differences in the two corpora in Table 2. Each demonstrative is followed by a two-character code separated by underscore. The first character denotes the feature and the second the value. For example, the first item *kulen* 'that (kind of)' whose lexical category (L) is adnoun (AN) mostly appeared in the spoken corpus and not in the written corpus.[5]

Feature	S	W	LL
kulen_L_AN	183	14	177.7
kulen_H_R	178	14	171.3
kulen_O_A	163	14	152.4
kuke(s)_L_PR	202	38	128.5
kuke(s)_H_R	187	38	112.5
ku_L_EX	114	9	109.6
i_O_A	6	105	104.0
kuke(s)_O_A	172	38	97.0
kulen_F_NO	69	2	82.4
ike(s)_H_K	68	3	75.7
ike(s)_X_D	63	2	74.3

Table 2 Frequency of Demonstrative Features

Whereas 931 demonstratives appeared in the spoken corpus, only 304 appeared in the written corpus. The distributions of three different types of demonstratives are listed in Table 3.

Types	Total Frequency	Written		Spoken	
		Freq.	%	Freq.	%
i	398	176	56	222	44
ku	773	128	17	645	83
ce	64	0	0	64	100
Total	1235	304	25	931	75

Table 3 Distribution of Three Demonstrative Types

The spoken corpus and the written corpus show different preferences for *i*, *ku*, and *ce* types.

Written: *i* (58%) > *ku* (42%) > *ce* (0%)
Spoken: *ku* (69%) > *i* (24%) > *ce* (7%)

Whereas *ku* demonstratives are preferred to corresponding *i* demonstratives in the spoken corpus, *i* demonstratives are preferred in the written cor-

[3] Although the syntactic category of an antecedent can be differentiated in a more sophisticated way using phrasal categories such as NP, VP, AdvP, etc. (as well as lexical categories), this will render the annotation process nearly impossible unless one uses a corpus with syntactic annotation, such as treebanks. Thus, we use simplified syntactic information such as nominal, clausal, and sentential.

[4] The agreement rate was calculated for each six high level categories separately and then averaged. The syntactic function has the lowest agreement rate even after the discussion (K=0.76). This is due to complex properties of Korean demonstratives with unclear boundaries between exclamatory expressions and other lexical categories.

[5] In Table 2, the log likelihood scores show that the usage of *kulen* is significantly different in the spoken and the written corpus. The log-likelihood scores in Table 2 are significant at a 99 percent confidence level with 1 degree of freedom if they are greater than 6.6. We only show a partial frequency list here due to the space limitations.

pus. This fact is associated with the linguistic function of *ku* that represents a speaker's desire to anchor interpersonal involvement with the hearer by actively inviting the hearer's voluntary understanding of the target referent. In contrast, *i* demonstratives imply that the speaker (writer) intends to incorporate the hearer (reader) within the proximal cognitive distance. In terms of annotation features, our findings are summarized as follows.

Lexical category: In both the written and spoken corpora, adnominal demonstratives are more frequently used than pronouns or locative pronouns. Demonstrative forms used as intensifiers, hedges, or personal habitual noise have been marked as exclamatives. Annotators have found that it is often difficult to clearly distinguish them from adnominal demonstratives.

Endophoricity: Our written corpus does not include any cataphors, whereas the spoken corpus shows 61 cases (cf. 523 anaphors). This fact seems to be related to the speaker's discourse strategy of intending to call the discourse participants' attention by placing an endophoric element before its antecedent.

Exophoricity: Exophoric usage of demonstratives in the written corpus is very limited. Only 17 cases were found (6 deixis vs. 11 context-based). In the spoken corpus, exophoric usages occur more frequently across three types of demonstratives. The deictic usage dominates the context-based usage (151 deixis vs. 79 context-based). As noted in previous literature, *ce* demonstratives mainly appear in deictic context, where its antecedent is visible or exists in the given situation. There seems to be a constraint of deictic usage of *ce* involving physical existence or visibility (or cognitive awareness) of an entity in addition to distance. This hypothesis needs to be further investigated with additional data.

Syntactic and Semantic Function: All three types of *i*, *ku*, and *ce* demonstratives refer to nominal entities as their antecedents. Although *i* and *ku* demonstratives are also used to refer to clausals and sentential elements, only a few examples of *ce* replace clausal or sentential elements. Another notable point is that *i* and *ku* demonstratives refer to clausal or sentential elements (corresponding to events or propositions) more frequently than nominal entities in both spoken and written corpora. 59% of the antecedents of *i* demonstratives (56% for *ku* type) in the written corpus are clausals or sentential elements, whereas 53% of the antecedents of *i* type (69% for *ku* type) are in the spoken corpus. This result needs to be tested on a larger corpus in our future study.

Phoric Type: In our annotated corpus, we only found referential examples, not substitutional cases. Exophoric examples are marked as non-phoric. In the written corpus, referential demonstratives are predominant (285 cases) and a small number of non-phoric cases are observed (18 cases). In the spoken corpus, referential demonstratives are more frequent (590 cases), whereas non-phoric cases have been more observed than in the written corpus (198 cases).

3 Conclusion

In this paper we presented a corpus-based study on Korean demonstratives. Six annotation features were used to mark complex linguistic functions of demonstratives. Using spoken and written corpora, we compared different usages of Korean demonstratives and showed that their usages are different depending on the registers of spoken and written Korean.

In spite of the deictic functions of demonstratives highlighted in previous research, our study indicates that endophoric usage is more predominant. This hypothesis, as well as others in this study, will be tested with a large corpus in our future work. We also plan to incorporate more sophisticated exploitation on semantic types of antecedents. This information will be useful for resolving the meaning of anaphoric demonstratives.

References

Botley, Simon and Tony McEnery. 2001. Demonstratives in English. *Journal of English Linguistics*, 29(1): 7-33.

Chang, Kyung-Hee. 1980. Semantic Analysis of Demonstrative *i*, *ku*, *ce*. *Ehakyenku,*16(2):167-0184.

Chang, Seok-Jin 1984. Cisiwa Coung. *Hangul,* 186: 115-149.

Gundel, Jaeanette, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language*, 69(2):274-307.

Halliday, M.A.K. and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Min, Kyung Mo. 2008. *A Study on Reference Items in Korean*. Ph.D. Dissertation. Yonsei University.

Poesio, Massimo. 2004. The MATE/GNOME Scheme for Anaphoric Annotation, Revisited. In *Proceedings of SIGDIAL*. Boston.

Prince, Ellen. 1981. Toward a Taxonomy of Given-New Information. *Radical Pragmatics*: 223-255. Academic Press. New York.