Jay Kaiser

Dickinson L415

Final Project

16 December 2015

1. Introduction

In the Korean language, words are marked using postposition morphological function units called particles that mark general grammatical and semantic information, each representing a distinctive grammatical categorization. For the most part, these mark the grammatical case of nouns: 이/가[1] *i/ka* marks the nominative case, 을/를 *ul/lul* marks the accusative case, 의 *ui* marks the genitive case, 으로/로 *ulo/lo* marks the instrumental case, 에/에서 *ey/eyse* marks the locative case, etc.

However, in languages like Korean and Japanese there can be found also another particle known as the topic particle. Marked in Korean with 은/는 *un/nun* (and with は *wa* in Japanese), the topic particle marks the topic of the sentence, a concept which can be, and often is, distinct from the nominative subject. Though for most instances of its usage, this particle indicates general or factual information regarding a specific topic or contrasts of one topic from another, the topic particle's usage extends far past this to types of implicit sentence information that can be hard to parse for a native speaker of English like myself.

[1] In this paper, all English transliterations of Korean are transcribed using Yale Romanization.

Moreover, the topic particle provides a particular type of nuanced information that, while not grammatically essential, requires mastery if one is to become fully fluent in Korean.

Furthermore, the topic particle is unique from the other particles as well in several ways. The topic particle, unlike the other particles, can be appended and stacked onto other particles. On the locative/instrumental particle 으로/로 *ulo/lo*, the dative particles 한테 *hanthey* and 에게 *eykey*, and locational particles like 에 *ey* and 에서 *eyse*, etc., the topic particle can be appended to make entire phrases the topics of the sentence. In contrast to this, instead of appending the topic particle to the nominative particle or the accusative particle, they are replaced by the topic particle entirely. Despite this, no lack of information occurs to a native Korean speaker. Moreover, multiple topics can appear in a single sentence, allowing for multiple topics each with extraneous nuanced information that can be difficult to deconstruct for native English speakers. For example:

(a)	어제는	나는	밥은	먹었다.
	ecey	na-nun	pap-un	mek-ess-ta.
	yesterday-Topic	I-Topic	food-Topic	eat-Pst-Ind
	‘I ate food yesterday.’			

In example (a), three words of the sentence all hold the topic particle. Because of this, despite the fact that the sentence’s meaning is explicitly very simple, it’s meaning actually semantically translates to something closer to: ‘I (as

opposed to someone else) ate food (as opposed to something else) yesterday (as opposed to some time else).

Although Korean particles provide crucial grammatical information to a sentence, they can be, and often are, entirely excluded in informal contexts. Often in informal speech, most particles are removed altogether, yet regardlessly, grammatically the sentence is still perfectly understandable to native speakers of the language. For many instances grammaticality of the sentence is maintained solely because any other interpretation of the sentence would prove nonsensical; for example sentences (b) and (c) have the same meaning, despite the lack of particles in (c):

(b)	어제	내가	밥을	먹었다.
	ecey	nay-ka	pap-ul	mek-ess-ta.
	yesterday	I-Nom	food-Acc	eat-Pst-Ind
	‘I ate food yesterday.’			
(c)	어제	나	밥	먹었다.
	ecey	na	pap	mek-ess-ta.
	yesterday	I(-Nom)	food(-Acc)	eat-Pst-Ind
	‘I ate food yesterday.’			

Any other interpretation of the subject and object in example (c) would prove bizarre to parse, no matter the language; here, an alternative way to parse (c) would be “Food ate me yesterday.” So in instances like this, despite the elision of the particles, the sentence’s meaning is both sensical and natural for a speaker.

Further complicating this issue, even though particles can often be omitted, there are many instances in which they are necessary, and their removal would prove to cause the sentence to be grammatically incorrect; in contrast, there also exist few instances where a particle must be omitted, or else the sentence would feel awkward and unnatural. However, these instances usually are entirely semantic in nature or within idiomatic expressions, respectively. These many instances can prove extremely difficult for a non-native learner of Korean to comprehend and utilize both fully and accurately.

2. Thesis

The topic particle only further confounds one's understanding of this grammatical system, as its usage seemingly extends semantically beyond the usage of the other, more regular particles. Because the topic particle appears more to be optionally added to elements of a sentence rather than obligatorily omitted from them in certain instances, the topic particle proves to be an enigma in comparison to the other many particles of Korean.

Therefore, working to gain a more foundational understanding of the topic particle and its utilization could prove useful in several ways. Firstly, finding regular instances of topic particle usage could allow non-native learners of Korean a more standardized approach to particle acquisition, as the current teaching system for this topic deviate more or less of a basic overview of the topic particle's few regular aspects followed by the expectation that one must

master its correct usage through native-Korean exposure alone. Secondly, understanding which instances the topic of a sentence is necessary to mark explicitly could prove informative about topic-prominent languages (like Korean, Japanese, Vietnamese, ASL, etc) in general, providing insights to these languages currently unknown to non-topic-prominent language speakers.

Because of this, an analyzation of topic particle frequency and topic particle flexibility could provide very basic but very crucial steps toward further research into this complex category of Korean grammar. In this document, I will present the steps I took in working with the Korean particles to narrow down and better understand their usage and lack thereof. Though initially I had been focused on the concept of dropping the topic particle in informal spoken contexts, instead now I focused more about the idea of the topic particle's optionality in a sentence and whether or not it is possible to determine where and when it can and will occur.

3. Corpus Selection

To analyze these particles in natural context, a large corpus created by the Korea Advanced Institute of Science and Technology (KAIST), known appropriately as the KAIST corpus, was utilized; more specifically, a branch of this corpus called the 'High quality morpho-syntactically annotated corpus' was the focal corpus for this research. This sub-corpus, created in 2000, consists of over one million Korean "phrases" over 69 total documents, providing ideally

more than enough data for such a project. Moreover, this subcorpus has been hand-annotated to ensure maximum accuracy, a major reason for my choosing of it.

The format of this corpus is as follows: each word is given a new line, followed by a tab delineation and a morphological breakdown of the word, indicated by English-style part-of-speech tags with plus-signs marking the boundaries between appropriate lemmatized elements. All punctuation is marked separately from words, each on their own line. The only issue with this choice of corpus design can be found when a particle is attached to a parenthesized noun phrase; in instances like this, because punctuation is given its own line break, the particle is placed separately onto its own line. This proved difficult to work with ultimately, but thinking upon the issue I could not conceive of a better alternative for structure the corpus could have used to avoid this problem. However because of this, I had to ignore any topic particles appearing after punctuation (though the total number of them only added up 516 times total in the corpus of over one million words, so results are not terribly skewed because of it).

Particles in this corpus are indicated using j-tags. That is, any speech tag starting with the letter ‘j’ indicates some kind of particle. More specifically, the topic particles being analyzed here, along with various other particles, are marked using ‘jxc’. Nouns on the other hand are indicated using a variety of 79 different n-tags.

4. Complications and Their Resolutions

Inherently, multiple issues with the manner in which the KAIST corpus was encoded arose; insteading of being encoded with the more common UTF-8 scheme, this corpus was written in a solely Korean format. Therefore, when viewing the individual corpus files in an ordinary word document, the resulting output became incoherent mixes of characters with seemingly no pattern to their madness. When uploading the files into Antconc for analyzation, the resulting characters were rendered with double escape character ‘x’ unicode strings, with again seemingly random instances of Greek and Arabic letters, Latin-script letters with diacritics, Chinese characters, empty boxes, and most infuriatingly, certain Korean characters.

Because of this, for all manipulation of the corpus data, instead of directly utilizing Korean character strings, I had to instead search using the unicode bi-letter string depictions; this proved at first difficult, but through keeping a log of common characters and their respective encoding strings, searching could progress almost as normal. Luckily, upon utilizing a different program for viewing the files, the Korean characters appeared as usual, and any manipulation of the files still yielded correct rendering when opened in this different file viewer. Therefore, I could still analyze the corpus files and any output files I created that still contained Korean.

5. Methodology and Discussion

To acquire basic preliminary counts of how many instances there were of the topic particle in the corpus, the files were first plugged into AntConc, and searching there for the topic particle yielded 51,423 hits. However, I wanted to get the count using another program as well to verify this result. Through the creation of a simple Python[2] script, I was able to gain another count, though this one only amounted to 22,397 hits. This 29,026 hit difference in results proved very confusing, and I never came to understand the source of this finding. Therefore, unfortunately, I was ultimately forced to work with the Python scripting counts, since Python allowed me to do further searching using written Python scripts that Antconc did not.

It was upon closer manual inspection of the corpus files that I came across a seemingly anomalous structure: a topic particle appended to an adverb. This finding proved both confusing and alarming, as I had never even considered that the topic particle could be attached to anything but a noun, a nominalized verb, or another particle in a phrase. Immediately I constructed another Python script called WordTypeFinder.py to find, count, and output all instances in which the topic particle can occur. This resulted in 419 unique instances; however, it is

[2] All important scripts I constructed for this project that I feel could prove useful for anyone working with this corpus again will be uploaded into the Shareables folder on Oncourse with extensive commenting to allow for easy understanding of my philosophy in their construction. The one referenced here is called “FinalScript.py”.

important to note that a majority of these cases were compound nouns and compound noun phrases.

Even so, I edited the script to sort out the results and to find all instances that were neither a noun or compound particle phrase. This resulted in a total of 789 hits over 54 unique instances. Following this I again edited the script to find all instances where the topic particle did not appear. This yielded 885202 hits over 10,337 unique instances.

To sort these results, I created a script called Comparison.py that took the results from the previous script and sorted them into three categories: instances where the topic particle script always occurs, instances where the topic particle never occurs, and instances of both, with appropriate counts and a percentage comparison between the two. The instances where the topic particle always occurs proved untelling and unuseful. For these instances, almost none had more than one occurrence; therefore, I would claim that these are less likely specifically topic-prominent words and more likely just morphological word constructions that are unusual in nature.[3] Similarly, across the 10,328 unique instances where the topic particle could not appear, there were many cases with more likely unique morphological constructions than an actual useful result.[4]

---

[3] There was one anomaly in this list: npd+paa constructions had 40 counts for solely occurring with the topic particle. However, I chose not to look too far into this for the moment, and instead to focus on other counts. I will be looking at this more in the future, however.

[4] This list requires a much more thorough look through in the future to find commonalities among non-topic instances to explain for the absence of the topic particle in these constructions. The task appeared too daunting for now, which is why I did not go through them in this paper.

It was in the comparison.txt output file from this script though that useful results could be found. In this output file, each unique construction is listed per line, followed by the number of instances where the topic particle is appended to it and the number where it does not. Finally, the percentage of instances divided by non-instances is printed. The topic particle occurs in variable usage over 227 instances based on the findings of these results.

From this, I designed another script called PercentagesByType.py to account solely for the final morphological unit of each word. In Korean, because grammatical units stack to the right of the word, logically the final one before the topic particle would prove most beneficial to study and compare, as this one ultimately decides the grammatical function of the word in a sentence; therefore, this script sorted the counts by final part-of-speech tag per word and organized them so that they would appear together with similar other part-of-speech tags (e.g. noun n-tags would be listed together, particle j-tags, adverbial m-tags, etc). A new percentage was taken from each totalled count of instances with the topic particle divided by instances without, and then this data was manually averaged out by type of POS tag to create the chart below.

	Counts w/ Topic	Counts w/o Topic	With Topic / Total
Nouns	17,449	110,318	0.1366
Particle	8	50,062	0.00002
Adverbs	51	48,935	0.0010
Verbal Units	58	135	0.3005
Morpho. Units	3745	13,055	0.1634
Conjugation Units	135	54,900	0.0025
Punctuation	87	14,624	0.0059

From this chart, and especially from the percentages found from taking the count of topic-appended instances divided by the total counts including non-instances, several observations can be made. Firstly, roughly 13 percent of all noun instances that can have the topic particle will have it. This is expected, as nouns are the most prominent particle-marked sentence element; thus proportionally they are mostly likely, if the topic particle is to appear in a sentence, to utilize it. Secondly, it's very rare for the topic particle to be attached to an adverb. Therefore, my finding of that one adverb-particle phrase that led to a majority of this research is not common an occurrence. In fact, according to my findings, roughly only 0.1 percent of all adverbs have topic particles, so either this is not a grammatically usual or correct process, or instances where the topic particle should be used adverbially are especially uncommon. Finally, certain

verbal and conjugation units frequently utilize the topic particle, though as to which ones I do not know yet. These could ultimately be units that simply convert verbs into nouns or noun phrases, but until I fully comprehend the part-of-speech tags utilized here, I could not say for certain.

One result that proved confusing to me was the fact that so few topic particles appear following another particle (for example, after the locative particles 에/에서 *ey/eyse* and 으로/로 *ulo/lo* and the dative particles 한테/에게/께 *hanthey/eykey/kkey*). I had been under the impression that the appending of the topic particle onto these parts of a sentence was fairly common, so to find such a meager eight counts proving such feels inaccurate. More will have to be done to see where errors may have occurred in my scripts as to result in this odd finding.

6. Research to be Done

From these results, I have just scratched the surface of the topic particle, its usage in certain contexts and how often it occurs, and where it can possibly occur, less on a semantic basis but instead more on an underlying grammatical one. In order to do this, more scripts will have to be written to work through the corpus to find commonalities between the instances where the topic particle can appear versus those where it cannot. These exact instances could be found after careful searching through the corpus to understand which part-of-speech tags can appear together and which ones cannot, as well as after fully understanding the difference between each tag as well, as I could not find full documentation on

what each tag describes anywhere online, in a source specifically about the KAIST corpus..

Moreover, these results show to me less that the topic particle is irregularly placed and instead that the topic particle has at least a semi-regular usage; as one can find percentages of where the topic particle can appear, perhaps it is possible to create a program that can give a relative percentage as to how likely it is for a topic particle to occur in a given context. If this is possible, it could allow much greater comprehension of subtle elements in a Korean sentence, not only for non-native learners, but also for machine translation practices as well.

Essentially, much more work will need to be done in parsing through the corpus to find more accurate results and understanding the results I do end up finding. If I can master locating and understanding the contexts where the particles appear, it would provide a crucial foothold for me to work deeper into the data to find clearer instances of particle usage, and hopefully given enough data, I'd be able to create an experiment to be given to native speakers of Korean to consider when and where they feel the topic is necessary as compared to when a certain percentage may say it is necessary.

Finally, I'd like to be able to work with a larger corpus that contains a wider variety of data from native Korean speakers: data not only consisting of written documents, but also of spoken and read passages in a variety of contexts, both

formal and informal. Comparing findings found here and findings from these new sources could continue to provide critical insights about the topic particle and even about other special particles, like the additive particle 도 *to* whose function in the sentence mirrors locational placement to the topic particle but with an additive focus as compared to a contrastive one.

Ultimately, this research is merely a small, but critical, first step in a much more ambitious project that could cover much more than the topic particle. This will prove interesting, and hopefully it'll work to improve my own knowledge of Korean as well.

7. References

Choi, Key-Sun, Young S. Han, Young G. Han, and Oh W. Kwon. "KAIST Tree Bank Project for Korean: Present and Future Development." *Sharable Natural Language Resources* (1994): 7-14. Web.

Dickinson, Markus, Ross Israel, and Sun-Hee Lee. "Building a Korean Web Corpus for Analyzing Learner Language." *WAC-6 '10 Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop* (2010): 8-16. Web.

Dickinson, Markus, Ross Israel, and Sun-Hee Lee. "Developing Methodology for Korean Particle Error Detection." *IUNLPBEA '11 Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications* (2011): 81-86. Web.

Laleko O, Polinsky M. “Between syntax and discourse: Topic and case marking in heritage speakers and L2 learners of Japanese and Korean.” *Linguistic Approaches to Bilingualism*. In Press.

Lee, Hanjung. "Usage Probability and Subject–object Asymmetries in Korean Case Ellipsis: Experiments with subject case ellipsis." *Journal of Linguistics* (2015): 1-41. *Journal of Linguistics*.

Lee, Sun-Hee, Markus Dickinson, and Ross Israel. "Developing Learner Corpus Annotation for Korean Particle Errors." *LAW VI '12 Proceedings of the Sixth Linguistic Annotation Workshop* (2012): 129-33. Web.