



## 2. RELATED WORK

Studies on the utterances incorporating *in-situ wh*-particles have been done widely in the areas of syntax [16, 1, 9, 13] and prosody-semantics interface [6, 2]. Especially for Korean and Japanese, which are typical *wh-in-situ* languages, the variability of the *wh*-particles was handled within the topic of LF intervention [15]. Also, in [8], it was proposed that a *wh*-particle in embedded self-addressed questions be interpreted as an existential quantifier.

In a slightly different view, in [5], the *wh*-particles associated with negative polarity items were investigated, suggesting the circumstances where the intervention is canceled. In [14], the usage of *wh*-particles as an interrogative and indefinite NP is investigated in a pragmatic view, accompanying the interpretation of gray-zone cases as rhetorical ones. Another thing to note is that it suggests that ‘왜 (way, why)’ be interpreted as an exclamation. Taking this into account, we did not generate sentences that include *way*.

In the view of language acquisition, [4] found that L2 Korean learners have difficulties with interpreting *in-situ wh*-particles. This implies the necessity of disambiguation that incorporates syntax, semantics, and phonetics, to which this paper attempts to contribute via a corpus-based approach.

## 3. CORPUS GENERATION

In generating the corpus script, namely five factors were considered: *wh*-particles that initiate an utterance, *predicates* that convey the content, *reportive particles* that give the utterance evidentiality, *sentence enders* that possess potential to represent various intentions, and *politeness suffixes* which come just after the sentence ender to assign honorific mood to the sentence.

### 3.1. *wh*-particles

Among the six *wh*-particles, namely ‘누구 (nwukwu, who)’, ‘뭐 (mwe, what)’, ‘어디 (eti, where)’, ‘언제 (encey, when)’, ‘어떻게 (ettehkey, how)’, and ‘왜 (way, why)’, only the first five were utilized in constructing the corpus. This is because *way* is rarely used as a quantifier, except for some cases in child language. Instead of *way*, we used ‘몇 (meych, the number of)’, which is widely used as a quantifier for counting. For the purpose of variation, in some cases, nominative (NOM) or accusative cases (ACC) were attached to the *wh*-particles.

### 3.2. Predicates

Predicates largely depend on the *wh*-particle they are aligned with. For instance, *nwukwu* (*who*) harmonizes with the verbs that are related to interaction, such as *give* and *receive*. In contrast, *eti* (*where*) matches with the verbs concerning location, such as *come* and *go*. In selecting the verbs, we referred to the set of 5,800 frequently used lexicons, released by the National Institute of Korean Language (<https://www.korean.go.kr/>). Depending on the verbs, appropriate particles were agglutinated and the phrases that contain object/complement were inserted. In some circumstances, polarity items such as ‘좀 (com, bit)’ or ‘하나 (hana, a piece)’ were augmented to modify or restrict the implicature.

### 3.3. Reportive particles

The reportive particles (RPT) provide utterances with evidential mood. Usually ‘-대 (tay)’, ‘-래 (lay)’, and ‘-재 (cyay)’ are used for statements, commands, and hortatives [11]. The particles were selectively added considering the content.

### 3.4. Sentence enders

The sentence enders (SEs) with various roles are components that influence the sentence type and intention of the utterance. There are mainly two types of SEs; the first type is SEs with a fixed role, e.g., ‘-다 (ta)’ for declaratives and ‘-니 (ni)’ for interrogatives [11]. For these, the sentence type is fixed but the intention can vary regarding *wh*-intervention and rhetoricalness. The second type is the underspecified SEs whose feature is not fixed (e.g., ‘어 (e)’, ‘지 (ci)’). They have the potential to display various intention types depending on the prosody. Both types of SEs were utilized in the generation.

### 3.5. Politeness suffix

The politeness suffix (POL), ‘요 (yo)’, can be agglutinated to SEs and in most cases does not affect the functional variability of the sentence, except for rhetoricalness. For some SEs such as ‘지 (ci)’ or ‘야지 (yaci)’, the augmented form is modified to ‘쪄 (cyo)’. On the other hand, the utterances with SEs to which the politeness suffix is not attachable, such as ‘냐 (nya)’, were left without the politeness suffix. An example sentence incorporating the aforementioned concepts (3.1-5) is as follows:

(2) 뭐 좀 먹었대요      mwe com mek-ess-tay-yo  
(statement or *y/n Q*)      what bit eat-PST-RPT-POL



<i>Who</i>	1,895
<i>What</i>	877
<i>Where</i>	199
<i>When</i>	172
<i>How</i>	163
<i>How much</i>	246

<Table 1-a >

<i>Statement</i>	1,085
<i>Yes/no Q</i>	1,047
<i>Wh- Q</i>	849
<i>Rhetorical Q</i>	302
<i>Commands</i>	175
<i>Requests</i>	56
<i>Rhetorical C</i>	38

<Table 1-b >

S, YN, WH	424
S, YN	242
S, YN, WH, RQ	137
S, WH	74
S, YN, RQ	43
YN, WH	33

...

S, RQ	25
YN, WH, C	25
S, R	24
S, WH, RQ	24
S, YN, WH, C	23
YN, C	23
WH, R	22

... (total 32 cases)

<Table 1-c >

**Table 1:** (1-a) describes the statistics on *wh-* particle and (1-b) on the intention types. (1-c) describes some of the possible intention sets that are engaged in a single utterance, here for 1,292 sentences. The code is disclosed with the corpus.

## 5. DISCUSSION

### 5.1. Corpus specification

In the corpus construction, the first version of the sentence list was generated by the methodology explained in Section 3, and only the sentences that received the consensus of first three authors (native speakers of Seoul Korean dialect) were taken into account. In total, the corpus contains 3,552 utterances that fall into the seven classes of intention. All the utterances were recorded by two native Koreans, a male and a female. The speech corpus containing a total of 7,104 (= 3,552 \* 2) utterances are available on-line (<https://www.github.com/warnikchow/prosem>) as with the corpus.

The statistics on the corpus is presented in Table 1. Sorting by the *wh-* particles that initiate the sentences, the most were the sentences starting with ‘누구 (nwugu, *who*)’, and the least were the ones starting with ‘어떻게 (ettehkey, *how*)’. Sorting by the intentions, *S* accounted for the most of the corpus data and *RC* did the least, following the tendency displayed in the recent Korean corpus [3]. Sorting by the possible intentions from a single utterance, the number of cases ranged from 2 to 4. The possible cases are partially listed in Table 1-c, tagged with the quantity. *Wh-* intervention occurs in most cases where interpreting the particles as *wh-* is allowed, but not vice versa.

	<i>S</i>	<i>YN</i>	<i>WH</i>	<i>RQ</i>	<i>C</i>	<i>R</i>	<i>RC</i>
<i>Who</i>	547	544	446	202	112	26	18
<i>What</i>	294	283	186	64	32	14	4
<i>Where</i>	64	64	49	6	11	4	1
<i>When</i>	37	54	40	22	0	4	15
<i>How</i>	59	62	28	8	6	0	0
<i>How much</i>	84	40	100	0	14	8	0

**Table 2:** A frequency matrix on *wh-* particles and the intention types.

### 5.2. Analysis

For a more detailed analysis of the corpus, we performed a simple calculation that shows the correlation between the indices (Table 2). It is assumed that *wh-* intervention largely occurs among *how much*-sentences, considering the portion of *wh-* questions within. Also, commands starting with *how* are rare, due to the fact that a to-do-list [12] is usually recommended to convey a specific instruction.

Concerning rhetoricalness, it is notable that *how much* is scarce among the rhetorical sentences since in that case polarity items are accompanied, disambiguating the *wh-* intervention. Consequently, the non-rhetorical directives (*YN·WH* and *C·R*) dominate *how much* sentences. The portion of rhetorical directives scored the highest in *when-* sentences for both *RQ/C*; we roughly assume that *when-* questions have potential to be interpreted as ‘*have sbd ever ...*’ and *when-* commands usually act idiomatically.

## 6. CONCLUSION

In this paper, we proposed a construction scheme for a corpus that incorporates single-text utterances with multi prosody/intention. In the process, five constituents, namely *wh-* particles, predicates, evidentiality, sentence enders and politeness suffix were considered. We obtained 3,552 utterances from 1,292 sentences, with the major intention types of *statement*, *yes/no question* and *wh- question*. A set of recordings by the native speakers (total 7,104 instances) is disclosed as a pilot research and can be supplemented for industrial purpose.

We suggest the corpus to be used for spoken language understanding systems which require disambiguation of the utterances that may induce *wh-* intervention. Along with the multi-modal approaches as in [7], various statistics- or deep learning-based classification systems may be able to infer a proper intention for a given speech and transcript. Not only for the industry, but this study can also be utilized for the Korean language learners, especially for the acquisition of prosody.

## ACKNOWLEDGEMENT

This research was supported by Projects for Research and Development of Police science and Technology under Center for Research and Development of Police science and Technology and Korean National Police Agency funded by the Ministry of Science, ICT and Future Planning (PA-J000001-2017-101). Also, this work was supported by the Technology Innovation Program (10076583, Development of free-running speech recognition technologies for embedded robot system) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea). After all, the authors appreciate the helpful advices provided by Mijeong Song and Minhwa Chung.

## 7. REFERENCES

- [1] Aoun, J., Li, Y.-h. A. 1993. Wh-elements in situ: Syntax or If? *Linguistic Inquiry* 24(2), 199–238.
- [2] Baunaz, L. 2005. The syntax and semantics of wh in-situ and existentials: the case of french. *Leiden Papers in Linguistics* 2(2), 1.
- [3] Cho, W. I., Lee, H. S., Yoon, J. W., Kim, S. M., Kim, N. S. 2018. Speech intention understanding in a head-final language: A disambiguation utilizing intonation-dependency. *arXiv preprint arXiv:1811.04231*.
- [4] Choi, M. H. 2009. *The acquisition of wh-in-situ constructions in second language acquisition*. PhD thesis Georgetown University.
- [5] Choi, Y.-S. 2007. Intervention effect in korean wh-questions: Indefinite and beyond. *Lingua* 117(12), 2055–2076.
- [6] Dalrymple, M. Semantics, information structure, and prosody in lfg part i: Semantics in lfg part ii: Information structure in lfg part iii: Prosody, syntax and semantics.
- [7] Gu, Y., Li, X., Chen, S., Zhang, J., Marsic, I. 2017. Speech intention classification with multi-modal deep learning. *Canadian Conference on Artificial Intelligence*. Springer 260–271.
- [8] Jang, Y. 1999. Two types of question and existential quantification. *Linguistics* 37(5), 847–869.
- [9] Lai-Shen Cheng, L., Rooryck, J. 2000. Licensing wh-in-situ. *Syntax* 3(1), 1–19.
- [10] Lee, H. Y. 1999. An acoustic phonetic study of korean nuclear tones. *Journal of the phonetic society of Korea* 38, 25–39.
- [11] Pak, M. D. 2008. Types of clauses and sentence end particles in korean. *Korean Linguistics* 14(1), 113–156.
- [12] Portner, P. 2004. The semantics of imperatives within a theory of clause types. *Semantics and linguistic theory* volume 14 235–252.
- [13] Soh, H. L. 2005. Wh-in-situ in mandarin chinese. *Linguistic Inquiry* 36(1), 143–155.
- [14] Song, S. 2010. Pragmatic usage of wh-elements in korea. *Language Informatio* 10, 95–117.
- [15] Tomioka, S. 2007. Pragmatics of If intervention effects: Japanese and korean wh-interrogatives. *Journal of Pragmatics* 39(9), 1570–1590.
- [16] Watanabe, A. 1992. Subjacency and s-structure movement of wh-in-situ. *Journal of east Asian linguistics* 1(3), 255–291.