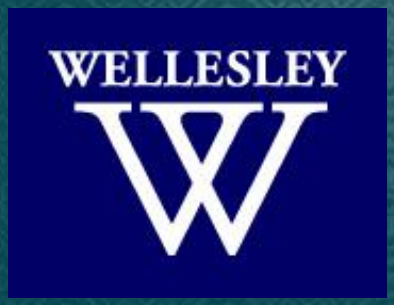
Annotating Particle Realization and Ellipsis in Korean

Linguistic Annotation Workshop 2012

July 13, 2012

Lee, Sun-Hee & Song, Jae-Young

Wellesley College Yonsei University



Goals of Study

- Provide a novel scheme for annotating the Korean particles while determining relevant issues of annotation and providing solutions.
- Evaluate how register variation contributes to the distributions of Korean particles
- Identify some linguistic factors involving particle ellipsis.
- Provide useful resources for linguistic analysis, Korean language learning, and NPL processing

Particle in Korean

- Subject, Object and Other Particles

(1) 오늘-은 민아-가 교실-에서 점심-을 먹-어
Onul-un Mina-ka kyosil-eyse cemsim-ul mek-e
Today-TOP M-SUBJ classroom-in lunch-OBJ eat-END
'Mina eats lunch in the classroom today'

은(*un*)/는(*nun*)
: Topic Marker

가(*ka*)/이(*i*)
: Subject
Agent
Focus (?)

에서(*eyso*)
: *Locative*

을(*ul*)/를(*lul*)
: Object
Theme
Focus (?)

Particle Ellipsis in Korean

- Subject, Object and Other Particles

(1') 오늘-Ø 민아-Ø 교실-에서 점심-Ø 먹-어
Onul-Ø Mina-Ø kyosil-eyse cemsim-Ø mek-e
Today-TOP M-SUBJ classroom-in lunch-OBJ eat-END
'Mina eats lunch in the classroom today'

은(*un*)/는(*nun*)
: Topic Marker

가(*ka*)/이(*i*)
: Subject
Agent
Focus (?)

에서(*eyse*)
: *Locative*

을(*ul*)/를(*lul*)
: Object
Theme
Focus (?)

Why Are Particles Important in Korean?

- Theoretical Linguistics and NLP:

  To determine grammatical or semantic (also pragmatic) functions of nominals; Syntactic, semantic, and discourse analysis

- Language Learning: Particle errors are one of the most frequent errors that Korean learners generate.

  Ko et al. (2004) - Error analysis with 100,000-*word* learner corpus:

Lexical Errors	>	Particles	>	Misspelling	>	Verbal Endings
(28.3%)		(24.4%)		(20.8%)		(16%)

Cf. Compare English preposition error percentage of 13.5% in the Cambridge Learner Corpus (Leacock et al. 2010)

Relevant Background

Classification of Korean Particles in Korean linguistics
(Nam, 2000; Lee, 2006)

- Case Particles
  - Structural Case: Subject (*ka/i; kkeyse*), Object (*ul/lul*)
  - Inherent Case: Dative (*eykey, hantey, kkey*), Goal (*lo/ulo, kkaci*), Locative (*ey, eyse*), Instrument (*lo, ulo*), etc.
- Auxiliary Particles
  - Topic Markers: *un/nun*
  - Particles with lexical meanings: *cocha* 'even', *to* 'also', *man 'only'*,

  N.B. These particles can combine with other particles except subject and object case particles.

- Conjunctive Particles: *wa/kwa, ina/na, itunci/tunci,* etc.

  e.g. Boston-*KWA* New York (Boston and New York)

Recovering Missing Particles

- Essential for determining accurate grammar relations :
  Computational processes of parsing, discourse analysis, machine translation, etc.
- This process excludes auxiliary particles as candidates due to their unpredictable distributions.
- Validity of recovering zero forms: Controversial whether a particle is deleted or originates as a zero form.

✓ It is important that a missing particle corresponds to a particular case particle and its identification is crucial for determining the grammatical and semantic function of the bare nominal.

Findings of Previous Research

- Hong et al.(1998): More dropping of subject case particles

	Class I Case Realization		Class II Bare NP (Dropping)		Class III Delimiter Replacement		Class IV Error		Total	Deletion rate	
	#	%	#	%	#	%	#	%	#	#	%
Subject	388	65.9	**43**	**7.3**	**154**	**26.1**	4	0.7	589	197	33.4
Object	359	72.5	**68**	**13.7**	**62**	**12.5**	6	1.3	495	130	26.2

- Kim & Kwon(2004): More dropping of object case particles

Pattern	Case Marker	Realization	Dropping	Total
Subject	이/가 *i/ka*	79.82% (1527)	20.18%(386)	100% (1913)
Object	을/를 *ul/lul*	54.51% (731)	45.49%(610)	100% (1341)

Data and Annotation Frame

- 100,128 *Ecel* Balanced Corpora from Sejong Tagged Corpora.
  (*Ecel*: similar to word unit but space-based)

Spoken Language Corpora	Written Language Corpora
50,097 *Ecel*	50,031 *Ecel*
100,128 *Ecel*	

- Balanced spoken and written corpora of 4 different registers

The Composition of Our Corpora

Type	Registers		# of Files	Size
Spoken	Private	Everyday Conversations (E)	7	12,504
		Monologues (M)	6	12,502
	Public	TV Debates & Discussions (D)	6	12, 547
		Lectures & Speeches (L)	6	12, 526
Written	Personal Essays (PE)		6	12, 510
	Novels (N)		6	12, 505
	Newspaper Articles (P)		6	12, 511
	Academic Textbooks (A)		6	12, 505

- A total of 49 different files were selected to make a balanced corpora.
- Approximately 2,000 *Ecel* were selected from each file.

Annotation Process

1) Manually corrected relevant errors in segmentation and morpheme tags before performing annotation

2) Identified all the nominal categories in the corpora that can combine with particles using morpheme tags

3) Annotated particles and determined their categories using the tag set and four annotation features, namely, particle_realized, particle_realized_type, particle_dropped, and particle_dropped_type.

Extra features : predicate and predicate type at the same level of a sentence with a bare nominal and light verb information and also comment (note) section for further discussion.

Our Tag Set of Particles

- **CASE:**

Subject (S): *ka/i*
Subject Honorific (SH): *keyse*
Object (O): *ul/lul* Genitive (G): *uy*
Dative (D): *ey/eykey* 'to', *hanthey* 'to'
Dative Honorific (DH): *kkey* 'to'
Complement (C): *ka/i*
Adverbial Case (B):
Time (BT): *ey* 'in, at'
Location (BL): *ey* 'to', *eyse* 'from'
Instrument (BI): *lo/ulo* 'with'
Direction (BD): *lo/ulo* 'to, as'
Source (BS): eyse 'from', *eykey*(*se*) 'from', *hanthey*(*se*) 'from' , *pwuthe* 'from', *ulopwuthe* 'from', *eysepwuthe* 'from'
Goal (BG): *ey* 'to', *kkaci* 'to'
Accompany (BA): *wa/kwa* 'with', *hako* 'with', *ilang/lang* 'with'

Vocative (V): *a/ya*
Comparative (R): *pota* 'than', *mankhum* ' as~as', etc.

- **Auxiliary (Discourse/Modal)**:
  Topic (T): *un/nun/n*
  Auxiliary (A): *to* 'also', *man* 'only, *mata* 'each', *pakkey* ('only'), chelem 'like', *mankhum* 'as much as', etc.

- **Conjunction** (J): *wa/kwa* 'and', *hako* 'and', *ina/na* 'or', *itunci/tunci*'or' *ilang/lang* 'and', etc.

Annotation Features and Sample

Origianl Eojeol	Morpheme Analysis	Particle_Realized	P_R_Type	Particle_Dropped	Particle_Type	Predicate	Predicate_Type	Light verb case	Light verb	comments
</u>										
<u who=P1>										
<s n=00034>										
자	자/IC									
여러분께서	여러분/NP+께서/JKS	께서/JKS	SH							
강의	강의/NNG			를	O	신청하다	PT			
신청	신청/NNG							을	신청하다	
하실	하/VV+시/EP+ㄹ/ETM									
때에	때/NNG+에/JKB	에/JKB	BT							
미리	미리/MAG									
인터넷	인터넷/NNG			을	O	통하다	PT			
통해서	통하/VV+ㅕ서/EC									
아마	아마/MAG									
무슨	무슨/MM									
내용을	내용/NNG+을/JKO	을/JKO	O							
할	하/VV+ㄹ/ETM									
것인가	것/NNB+이/VCP+ㄴ가/EF			P						
참고는	참고/NNG+는/JX	는/JX	T							
어~	어/IC									
하셨으리라고	하/VV+시/EP+ㅕㅆ/EP+으리/EP+라고/EC									
봅니다.	보/VV+ㅂ니다/EF+./SF									
</s>										
<s n=00035>										
어~	어/IC									
거기에서	거기/NP+에서/JKB	에서/JKB	BL							
별로	별로/MAG									
변동<pause>된	변동/NNG+되/XSV+ㄴ/ETM							이	변동되다	
사항	사항/NNG			E						N 없이
없이	없이/MAG									
<trunc-iu>										
</s>										
</u>										

Unpredictable Cases of Particle Ellipsis

[1] Genitive Case '*uy*'

- The generative *uy* tends to disappear after a complement nominal of a verbal noun

e.g. 영화의/Ø 촬영
yenghwa-uy/Ø chwalyeng
Movie-GEN filming
'filming of a movie'

- Whereas *uy* appears after a subject nominal of a verbal noun

e.g. 존의/?*Ø 우승
John-uy/ ?* Ø wusung
John-GEN winning
'John's winning'

Unpredictable Cases of Particle Ellipsis

[2] Particles in Light Verb Constructions

- Light verb constructions:
  Verbal noun + light verb(*hata*/*toyta*/*sikita*)

e.g. Silhyen(accomplishment) + *hata*/*toyta*/*sikita*
('accomplish/to be accomplished/to make it accomplish')

i) *Silhyen-**ul** hata* (accomplishment-OBJ do)
*Silhyen-**i** toyta* (accomplishment-SBJ become)
*Silhyen-**ul** sikhita* (accomplishment-OBJ make)

ii) ? John-i kkum-**ul** shlhyen-**ul** hayssta
J-SBJ dream-OBJ accomplishment-OBJ did
'John accomplished his dream'

Unpredictable Cases of Particle Ellipsis

[3] Optional Particles with Bound Nouns (or Defective Nouns)

- Bound nouns tend to combine a certain type of particle.

  tey ('place'), *ttay* ('time'), *swu* ('way'), *ke(s)* ('thing'), *cwul* ('way'), *check* ('pretense') etc.

e.g. 학교-에서 공부할 수(-가) 있다

hakkyo-eyse kongpwuha-l swu(-ka) issta

school-at study-REL way(-SBJ) exist

'It is possible to study at school'

Unpredictable Cases of Particle Ellipsis

[4] Mandatory Non-occurrences of Particles: Compounds, Idioms or Formulaic Expressions

Noun Compound:

e.g. [*palcen*+Ø(*-*uy*) *keyhwoyk*+Ø(*-*uy*) *pokose*]
'development plan report'.

Formulaic Expressions:

e.g. *kes-(*kwa)*+ *kathta* (thing-(*with) + similar) 'seem'
ke-Ø + *aniya* (thing + isn't) 'isn't it?'
ne-Ø + *ttaymwun* (N+ reason) 'because of you' etc.

Annotation Features for Bare Nominals

- L: Non-occurrence of a particle in light verb constructions
- N: Non-occurrence of a particle after a nominal that forms a compound with the following nominal
- E: Non-occurrence of a particle based upon lexical or morpho-syntactic constraints

- P: Predicate nominals combining with copula *ita*. It also marks a nominal standing alone without *ita*, as answering utterance
- ER: Errors including a repeated nominal by mistake or an incomplete utterance

Annotating Particle Ellipsis

- Annotation Principles of missing particles:

1) Annotate only obligatory case particles and conjunctive particles but exclude auxiliary (discourse/modal) particles.

2) Instead of selecting a single best particle, present a set of multiple candidate without preference ranking. (Lee et al. 2012)

3) Annotate stacked particles as single units without separating them into smaller particles.

Inter-Annotator Agreement

- 5,000 *Ecel* corpus with 466 nominals that appear without particles
- Two experienced annotators; manually annotated the data separately and cross-examined each other's annotation
- Agreement = 91.23% for the specific particles (Cohen's Kappa):
- Reasons for high agreement:

  Highly-trained annotators & a stable set of guidelines

Corpus Analysis

Spoken Corpora		E	M	D	L	Total
Particle Realization		2081	2853	3334	3672	11940
Predicate Nominals (P)		741	590	742	757	2830
Zero Particles	Ellipsis	843	395	237	185	1660
	Compounds (N)	320	297	350	411	1378
	Optional (E)	796	735	841	802	3174
	Light Verb (L)	308	190	482	410	1390
	Vocative (V)	24	3	6	20	53
Errors		82	36	41	43	202
Written Corpora		PE	N	P	A	Total
Particle Realization		4707	4715	4603	4928	18953
Predicate Nominals (P)		593	600	393	612	2197
Zero Particles	Ellipsis	98	86	165	12	361
	Compounds (N)	406	104	1941	728	3179
	Optional (E)	996	1125	1492	712	4325
	Light Verb (L)	361	437	965	917	2680

Particle Realization vs. Ellipsis

Spoken	Conversation	Monologue	Discussion	Lecture	Total
Realized	71%	88%	93%	95%	88%
Ellipsis	29%	12%	7%	5%	12%
Written	Essay	Novel	News	Academic	Total
Realized	98%	98%	97%	99.7%	98%
Ellipsis	2%	2%	3%	0.3%	2%

Findings

- Low case ellipsis rates across two corpora
- Significant difference between the spoken and the written corpora (χ^2=851.78, p <.001)
- Significant genre factor:
  Particle ellipsis in everyday conversations is significantly more frequent than in monologues, debates, or lectures using a Bonferroni adjusted alpha level of .008 per comparison (.05/6). ($\chi^2(1)$=266.64, p<.001; $\chi^2(1)$=571.19, p<.001; $\chi^2(1)$=746.93, p<.001).

  Cf. Particle ellipsis between debates and lectures ($\chi^2(1)$=11.72, p<.001).

Particle Realization vs. Ellipsis

Spoken	Conversation	Monologue	Discussion	Lecture	Total
Realized	71%	88%	93%	95%	88%
Ellipsis	29%	12%	7%	5%	12%
Written	Essay	Novel	News	Academic	Total
Realized	98%	98%	97%	99.7%	98%
Ellipsis	2%	2%	3%	0.3%	2%

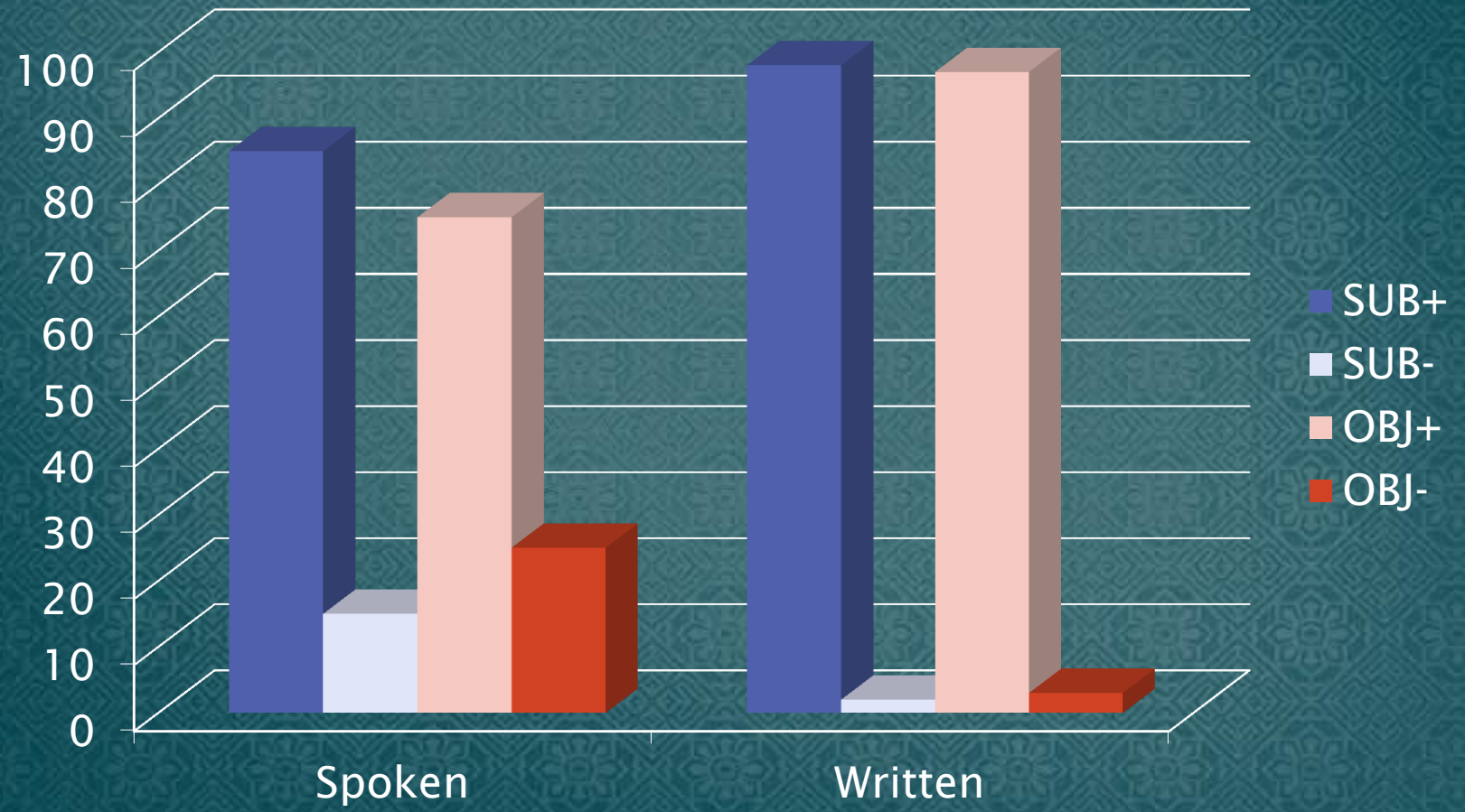
Particles (Spoken Corpora)

Particles	Spoken				
	Conversation	Monologue	Discussion	Lecture	Total
SUBJ +	63% (539)	88% (776)	93% (927)	95% (848)	85% (3090)
SUBJ –	37% (318)	11% (97)	7% (67)	5% (48)	15% (530)
OBJ +	51% (398)	73% (535)	85% (698)	89% (771)	75% (2402)
OBJ –	49% (389)	27% (198)	15% (121)	11% (92)	25% (800)
CONJ +	92% (57)	68% (54)	90% (89)	98% (137)	88% (337)
CONJ –	8% (5)	32% (26)	10% (10)	2% (3)	12% (44)
OTHERS +	81% (549)	90% (634)	95% (859)	97% (1174)	92% (3213)
OTHERS –	19% (131)	10% (74)	4% (39)	3% (42)	8% (286)

Particles (Written Corpora)

Particles	Written				
	Essay	Novel	News	Academic Text	Total
SUBJ +	97% (743)	97% (840)	92% (635)	99.7% (588)	98% (2806)
SUBJ –	3% (25)	3% (24)	3% (18)	0.3% (2)	2% (69)
OBJ +	94% (967)	95% (1066)	99% (1050)	99% (1026)	97% (4109)
OBJ –	5% (56)	5% (53)	1% (13)	1% (9)	3% (131)
CONJ +	100% (133)	100% (113)	97% (226)	99.7% (276)	99% (748)
CONJ –	0% 0	0% (0)	3% (7)	0.3% (1)	1% (8)
OTHERS +	99% (1778)	99.5% (1739)	93% (1680)	100% (2173)	98% (7370)
OTHERS –	1% (17)	0.5% (9)	7% (127)	0% (0)	2% (153)

Distribution of Subject/Object Particles: Spoken vs. Written Corpora(%)



Findings

- Significant object particle dropping in the spoken corpora (χ^2 =797.03, p<.001) & consistently higher than the subject particle ellipsis at each register.

- Genre variation: more case dropping for less formal corpora e.g. everyday conversations: 49% object particle elided & 37% subject particle elided

- In parallel to case particles, more dropping of conjunctive particles and other case particles in the spoken corpora

Linguistic Properties

- Definiteness and Specificity

Kim(1991): A case particle is likely to be dropped when the preceding noun is definite or specific.

i) ku haksayng-i/-Ø na-lul chacawa-ss-e
that student-SBJ/Ø I-OBJ visit-PAST-END
'That student visited me'

ii) etten haksayng-i/*Ø na-lul chacawa-ss-e
some student-SBJ/Ø I-OBJ visit-PAST-END
'Some student visited me'

Linguistic Properties

- Familiarity/Background

e.g. tampay-ʔlul/-Ø cwu-seyyo
cigarette- OBJ-Ø give-IMPERATIVE
'Please give me cigarette.'

- Salience:

e.g. i) philyohan ke-l hanato mos tule, na-Ø
necessary-REL thing- OBJ anything not take, I-Ø
'I cannot take anything that is necessary'
ii) wuli sensayngnim-**Ø (?ul)**, ne alla?
our teacher-Ø (OBJ) you know
'Do you know our teacher?'

Conclusion

- Overemphasized particle ellipsis in the spoken corpora
- Register variation factor only in spoken corpora: More particle ellipsis in formal dialogues

  N.B. Formality per se is not the deciding factor, but a partially related factor
- More object particles ellipsis than subject particle ellipsis
- Particle ellipsis and semantic/pragmatic constraints
- Usability of our corpora for linguistic analysis, language learning, and NLP processing

Further Works

- Run error detection software on our corpus to verify the consistency of our annotation (Dickinson and Meurers, 2003)
- Double-check consistency of annotation and release the corpus with annotation guideline
- More sophisticate linguistic analysis of the annotated corpora

Reference

- Markus Dickinson and Detmar Meurers. 2003. Detecting Errors in Part-of-Speech Annotation. Proceedings of the 10th Conference of European Chapter of the Association for Computational Linguistics (EACL-03). Budapest, Hungary.
- John Fry. 2001. Ellipsis and 'wa'-marking in Japanese conversation. Doctoral Dissertation. Stanford University.
- Paul Hopper and Sandra A. Thompson. 1984. The Discourse Basis for Lexical Categories in Universal Grammar. Language, 60: 703-752.
- Kun-hee Kim and Jae-il Kwon. 2004. Korean Particles in Spoken Discourse-A Statistical Analysis for the Unification of Grammar. Hanmal Yenku, 15: 1-22.
- Jae-il Kwon. 1989. Characteristic of Case and the Methodology of the Case Ellipsis, Language Research, 25(1): 129-139.
- Hyo Sang Lee and Sandra A. Thompson. 1989. A discourse account of the Korean accusative marker. Studies in Language, 13: 105-128
- Hanjung Lee. 2006. Parallel Optimization in Case Systems: Evidence from Case Ellipsis in Korean. Journal of East Asian Linguistics, 15: 69-96.
- Sun-Hee Lee. 2006. Particles (*Cosa*). Why Do We Need to Reinvestigate Part of Speeches? (in Korean): 302-346.
- Sun-Hee Lee, Markus Dickinson, and Ross Israel. 2012. Developing Learner Corpus Annotation for Korean Particle Errors. In Proceedings of the Sixth Linguistic Annotation Workshop (this volume). Jeju, Korea
- Minpyo Hong, Kyongjae Park, Inkie Chung, and Ji-young Kim. 1998. Elided Postpositions in Spoken Korean and their Implications on Center Management, Korean Journal of Cognitive Science, 9(3): 35-45.

감사합니다.

Thank you!