

Annotation of Spelling Errors for Korean Learner Corpora

Seok Bae Jang, Sun-Hee Lee, & Markus
Dickinson

AALL-09; Tempe, AZ; March 11, 2009

Introduction

- Learners frequently make spelling errors in Korean (Ko et al 2004)
 - Errors reflect aspects of linguistic knowledge (cf. Rimrott & Heift 2008, and references therein)
 - In Korean: frequent mismatches between a syllable and a character, due to phonological rules and morphological boundaries
 - Errors are computationally problematic (e.g., van Rooy & Schäfer 2002, de Felice & Pulman 2008)
- Goal: Provide taxonomy & annotation scheme for Korean learner spelling errors
 - Useful for automatic error diagnosis
 - Useful for feedback instruction in a Korean ICALL system

Overview

- Background on Korean
 - Korean writing system
 - Korean word formation
- Error taxonomy
- Corpus Study
 - Interannotator agreement with pilot corpus
 - Abilities of current spelling checkers
 - Spelling Error Analysis with 100 Learner corpus

Korean writing system

- Syllabic alphabet with one umjeol ‘syllable’, as a basic unit
- A syllable is composed of at most 3 components:
 - choseong, the first sound, e.g., ᄒ [h]
 - jungseong, the second sound(s), e.g., ㅏ [a]
 - jongseong, the last sound(s), e.g., ㅓ [k]
- Syllable is represented combinatorically, e.g., 학 [hak]
 - unlike the more linear representation for English

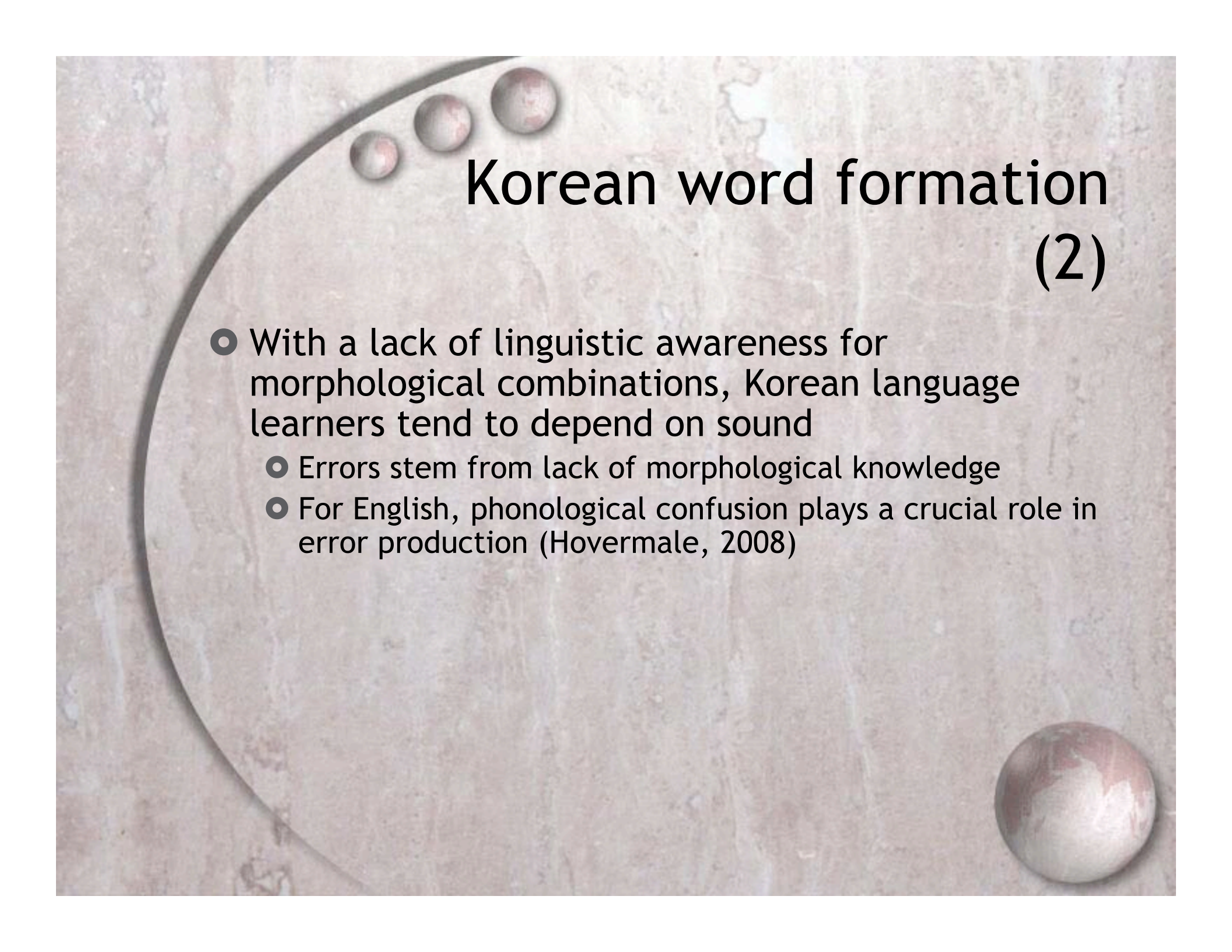
Korean writing system

Syllabic representation of the Korean writing system requires learners to acquire:

- Linguistic knowledge of exact syllable compositions
 - e.g., for 학기 [hakki] ‘semester’:
 - the first ㄱ [k] combines with the preceding vowel ㅏ [a]
 - the second ㄱ attaches to the following vowel ㅣ [i]
- Specific orthographic knowledge, including sound-letter relationships
 - e.g., [k] and [g] sounds both correspond to ㄱ
- Positional constraints of letter patterns
 - e.g., double consonant ㄴㅈ [nj]/[n] does not appear word initially

Korean word formation (1)

- As an agglutinative language, word formation in Korean has complex morphological combinations
 - Morphemic boundaries tend to be maintained in spite of application of a phonological rule.
- 잡-으시-었-겠-습니다 jap-eusi-eoss-kess-seupnita
 - Pronounced:
자브시어께씀니다 ja-beu-si-eo-kkey-sseum-ni-da
 - Composed of 5 morphemes & orthographically maintains the basic dictionary form for each morpheme



Korean word formation (2)

- With a lack of linguistic awareness for morphological combinations, Korean language learners tend to depend on sound
 - Errors stem from lack of morphological knowledge
 - For English, phonological confusion plays a crucial role in error production (Hovermale, 2008)

Classifying errors

- Spelling errors range from simple mistakes to more linguistically complex errors (Kukich, 1992)
 - Mistakes result from inattention or physical conditions, including typos
 - Systematic errors reflect a lack of linguistic knowledge & require more informative feedback
 - lack of phonological awareness: phoneme discrimination, identification, or segmentation
 - lack of morphological awareness: morpheme identification or segmentation (words, particles, inflected verbal endings)

Error taxonomy

- An annotated corpus analysis can show
 - actual range of spelling errors of Korean learners
 - how each type of error is related to linguistic knowledge
 - related to learner's native language
 - related to deficit of phonological/morphological knowledge of Korean (cf., e.g., Rimrott & Heift, 2008).
- We classify 5 categories of spelling errors:
 - phonological, morphological, typographical, incomprehensible, foreign word
- Similar to Hovermale (2008), except errors with foreign words are separately marked

Phonological errors

- Phonological errors are based on:
 - incorrect mapping between a sound & a letter
 - Consonant mismatch ㅂ-표-ㅃ; ㄱ-ㅋ-ㄲ; ㄷ-ㅌ-ㅈ
 - plain-tense-aspirated: [p, p', p^h]; [k, k', k^h]; [t, t', t^h]
 - ex. correct form - 예뻐니다 incorrect form-예ㅂ니다
 - plain-tense; [s, s']; [ts, ts']
 - ex. correct form - 날씨가 incorrect form-날시가
 - Vowel mismatch ㄴ-ㄴ; ㅈ-ㅈ
 - ex. correct form: 노래 incorrect form 너래
 - These types of errors are generally restricted to differences between Korean & English

Phonological errors

- Phonological distance between sounds can be remote

- consonants: ㄱ [k] vs. ㅍ [p]

example:

correct 외숙모 [wesungmo] vs. incorrect 외습모 [wesupmo]

- vowels: ㅏ [a] vs. ㅜ [eu]

example:

correct 나빠졌다 [nappajeotta] vs. incorrect 나쁘졌다 [nappeujeotta]

- This perceptual confusion is rare in native Korean

Morphological errors

- Morphological errors include:
 - failures of morpheme identification
 - Example. 먹+었+습니다 (ate)
correct form 먹었습니다 incorrect form 먹어습니다
 - Example. 맛[mat]+이[i] : [masi]
correct form 맛이 incorrect form 마시
 - double consonants 맑다[makta]
correct form 맑다는 incorrect 막다는
 - overgeneralizations
 - Example 것+이+에요 이+에 →예
correct form 것이에요 incorrect form 것이예요
- These errors are related to inflection, word syllabification, & syllable boundaries

Foreign word errors

- Foreign words in Korean borrowed from other languages often have non-predictable spelling
 - For example, the proper name *New York*
 - correct standard form: 뉴욕 *nyuyok*
 - commonly learner innovation: 뉴-요-크 *nyuyokhu*
- Closely related to phonological confusion
 - But hard to determine the exact match between a sound and a letter or identify the exact phonological rule

Distinguishing phonological & morphological errors

- Morphological variations caused by phonological rules have been treated as morphological errors

❖ Examples.

≡ **assimilation** : correct form 원래 incorrect form 월래

Phonological variants:

correct form 부모님**과** [kwa] incorrect form 부모님**와** [wa]

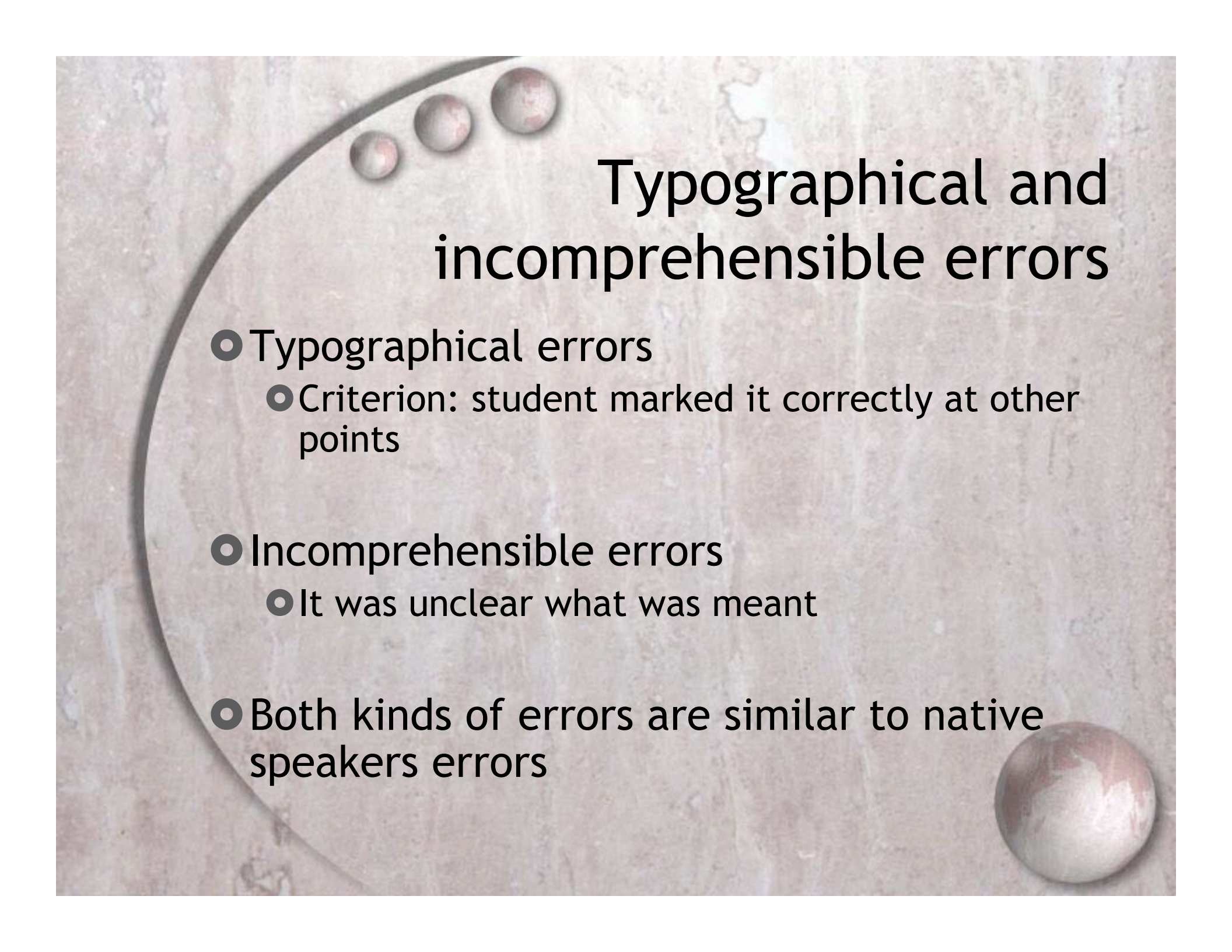
correct form 한국말**을** [eul] incorrect form 한국말**를** [leul]

- Sometimes phonological variation disappears and become morphological variations.

Example. Sound distinction loss among native speakers

ㅏ [æ̃] vs. ㅓ [e]

correct form 한**ㅏ**테 incorrect form 한**ㅓ**테



Typographical and incomprehensible errors

- Typographical errors
 - Criterion: student marked it correctly at other points
- Incomprehensible errors
 - It was unclear what was meant
- Both kinds of errors are similar to native speakers errors

Our approach

- Pilot study: Gather a small learner corpus of 10 people
 - 10 non-heritage intermediate learners with 1 article each
 - Test interannotator agreement for spelling annotation
 - Test accuracy of existing spelling checkers
- Actual annotation of data for 100 people
 - 25 heritage beginners, 25 non-heritage beginners
 - 25 heritage intermediates, 25 non-heritage intermediates

Interannotator agreement with pilot corpus

- We evaluated the Kappa statistic to measure interannotator agreement
 - 1) Error Type: $K=0.83$ / 1281 pairs
 - P for phonological, M for Morphological, etc.
 - 2) Correction: $K=0.73$ / 1281 pairs
 - Wrong word:Correct word pairs
 - e.g., 아~~름~~답습니다: 아~~름~~답습니다 ---> to be beautiful
 - 3) Feedback: $K=0.75$ / 1281 pairs
 - e.g., ~~ㅂ~~: ~~ㅁ~~ ---> consonant ~~ㅂ~~ (p) should be replaced with consonant ~~ㅁ~~ (m)
- All 3 scores show high correlation between 2 annotators: positive results for our guidelines

Spelling checking

- Spell checkers for Korean do not adequately handle learner errors
 - Hypothesis: more morphological in nature
 - Closely related with word-spacing errors
- Learner errors:
 - require different error diagnostic tactics
 - need to support feedback
- Two spelling checkers used
 - HWP (Korean Word Processor)
 - HAM v.5 (Hangeul Analysis Module)

Accuracy of spelling checkers with a Pilot Corpus

○ HWP (Korean word Processor)

	Raw Corpus	After Word-spacing Correction
Precision	0.717	0.927
Recall	0.551	0.551
F-measure	0.623	0.691

○ HAM v.5

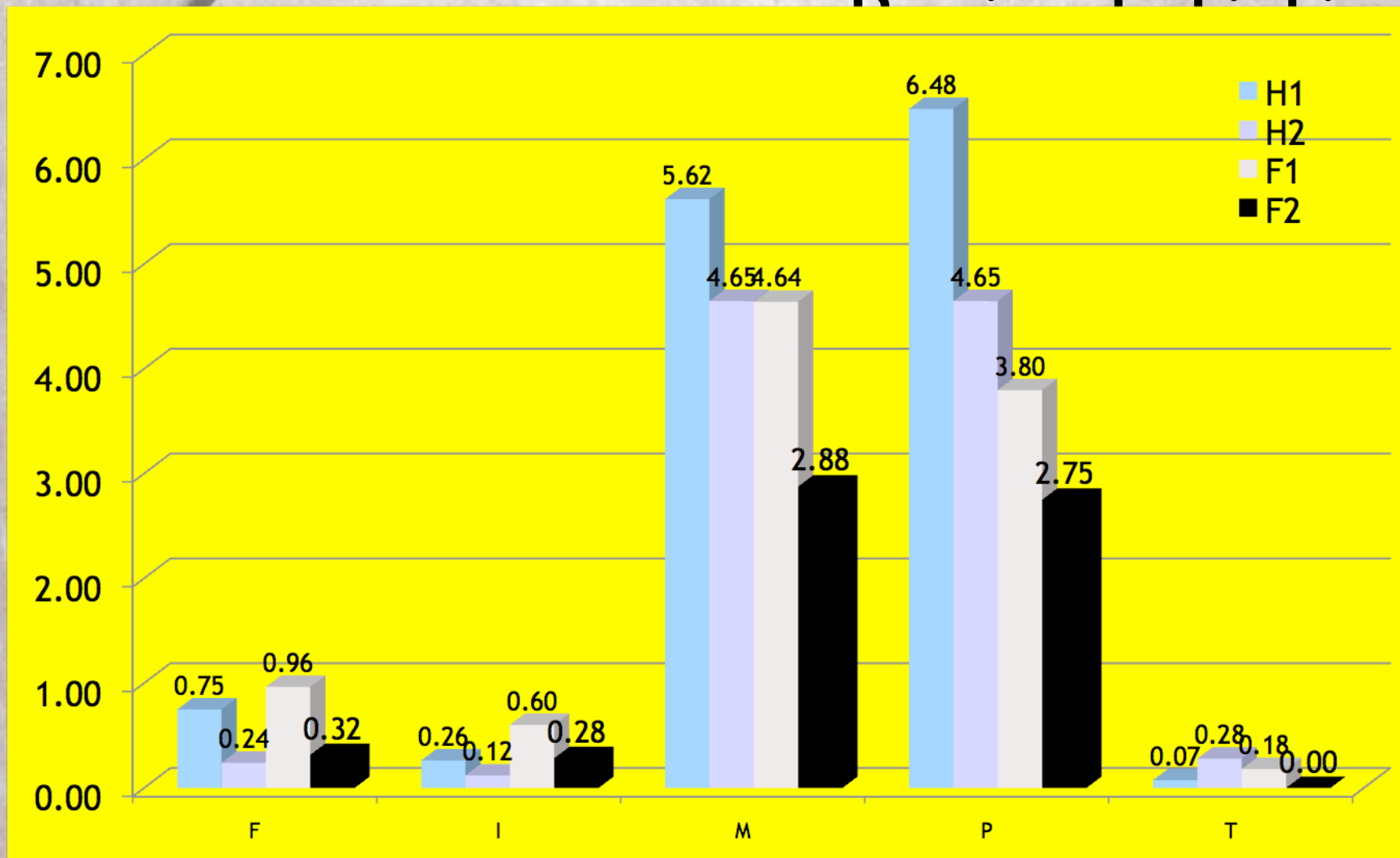
	Raw Corpus	After Word-spacing Correction
Precision	0.500	0.611
Recall	0.667	0.638
F-measure	0.571	0.624

Corpus of 100 Learners

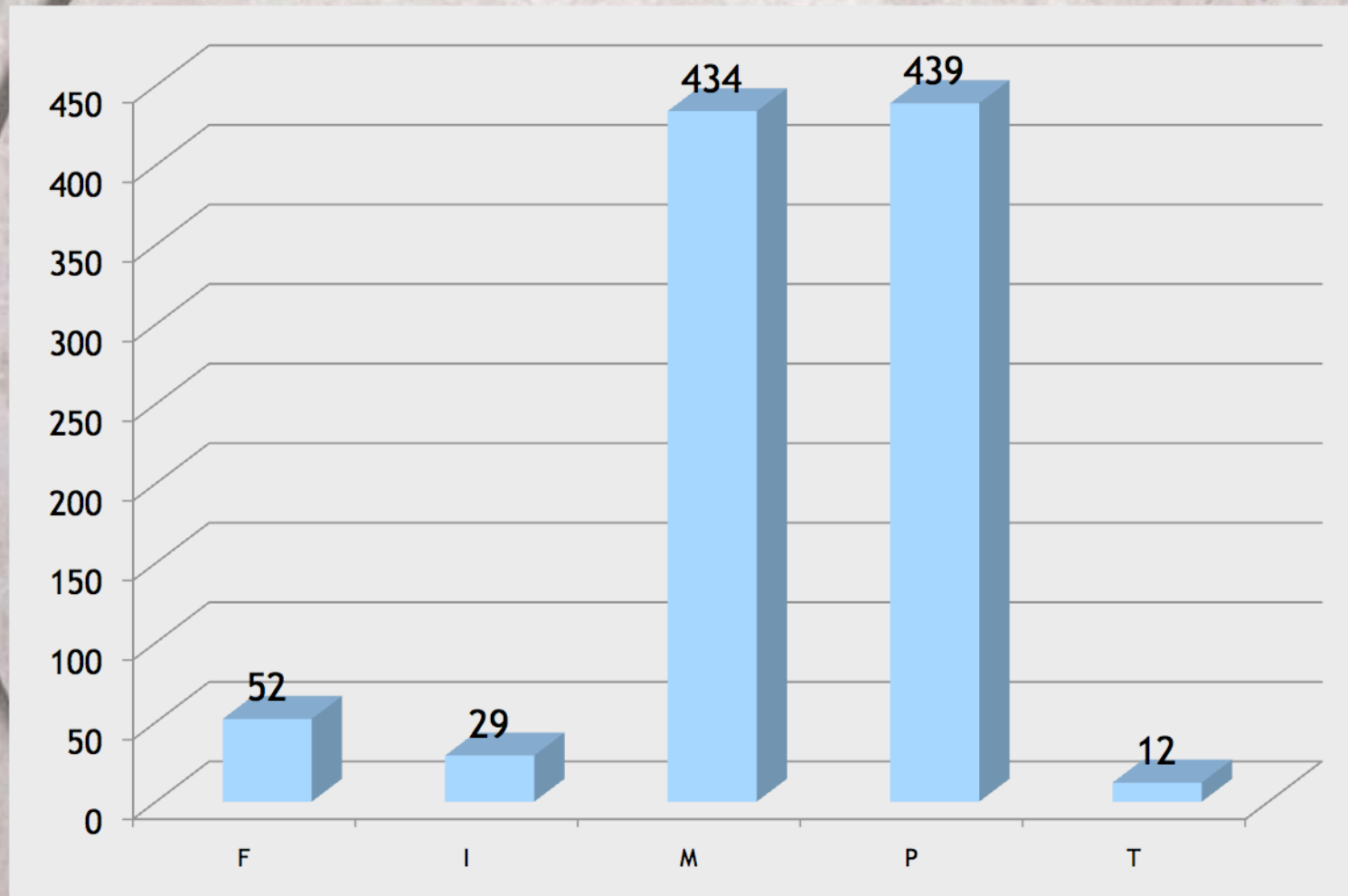
- Heritage Beginner: 2,669 words
- Heritage Intermediate: 2,496 words
- Non-heritage Beginner: 1,659 words
- Non-heritage Intermediate: 3,163 words

- 25 learners in each group

Corpus of 100 Learners



Corpus of 100 Learners



Analysis by background

	Heritage		Non-heritage		SUM	
	#	%	#	%	#	%
F	26	4.33	26	7.10	52	5.38
I	10	1.67	19	5.19	29	3.00
M	266	44.33	168	45.90	434	44.93
P	289	48.17	150	40.98	439	45.45
T	9	1.50	3	0.82	12	1.24
Total	600		366		966	

Analysis by language level

	Beginner		Intermediate		Sum	
	#	%	#	%	#	%
F	36	6.91	16	3.60	52	5.38
I	17	3.26	12	2.70	29	3.00
M	227	43.57	207	46.52	434	44.93
P	236	45.30	203	45.62	439	45.45
T	5	0.96	7	1.57	12	1.24
Total	521		445		966	

Spelling checking (HWP) on corpus of 100 learners

	H1		H2		F1		F2		Total	
	#	%	#	%	#	%	#	%	#	%
Checked w/ C	77	22.19	67	26.91	34	21.52	44	22.92	222	23.47
Checked w/o C	184	53.03	95	38.15	70	44.30	63	32.81	412	43.55
Not checked	86	24.78	87	34.94	54	34.18	85	44.27	312	32.98

Analysis of Phonological Errors

		H1	H2	F1	F2
# of Error Types	Consonants	17	24	17	30
	Vowels	25	34	24	21
	Total	42	58	41	51

Phonological Error Samples: Consonants & Vowels

	H1		H2		F1		F2	
	#	%	#	%	#	%	#	%
H : H	5	11.90	30	34.09	6	16.67	2	6.06
⊥ : ⊥	9	21.43	10	11.36	6	16.67	5	15.15
⊥ : H	4	9.52	1	1.14	4	11.11	1	3.03
Total		42.86		46.59		44.44		24.24

Challenges

- For automatic spelling checking:
 - Need to account for learner errors, by a better morpheme:phoneme mapping?
- For corpus annotation development:
 - Collect & annotate more data
 - Integrate annotation with other types of annotation
 - use multi-layered annotation (cf. Lüdeling et al. 2005)

Conclusions

- With this annotated corpus:
 - Can improve evaluation for particle error detection (Lee, Eom, & Dickinson 2009)
 - Can test several methods of spelling error detection to determine their effectiveness for each error type
 - methods involving POS tagging, string similarity, machine learning, etc.
- Future: conduct experiments to find out the most effective feedback for each type of spelling errors



Acknowledgements

Thanks to the following annotators:

- Sung-eun Choi (BYU)
- Jae-young Song (Wellesley)

Thanks for the support from:

- Brigham Young University
- Wellesley College

References

- De Felice, Rachele & Stephen Pulman (2008). A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of COLING-08*. Manchester.
- Hovermale, DJ (2008). Developing an Annotation Scheme for ELL Spelling Errors. In *Proceedings of MCLC-08*. East Lansing, MI.
- Ko, S., M. Kim, J. Kim, S. Seo, H. Chung, & S. Han (2004). *An analysis of Korean learner corpora and errors*. Hanguk Publishing Co.
- Lee, Chong Min, Soojeong Eom, & Markus Dickinson (2009). Towards Analyzing Korean Learner Particles. Talk given at *Automatic Analysis of Learner Language (AALL'09)*. Tempe, AZ.
- Rimrott, Anne & Trude Heift (2008). Classification Systems for Misspellings by Non-native Writers. Talk given at *Pre-CALICO Workshop on "Automatic Analysis of Learner Language"*. San Francisco, CA.
- van Rooy, Bertus & Lande Schäfer (2002). The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies* 20, 325-335.