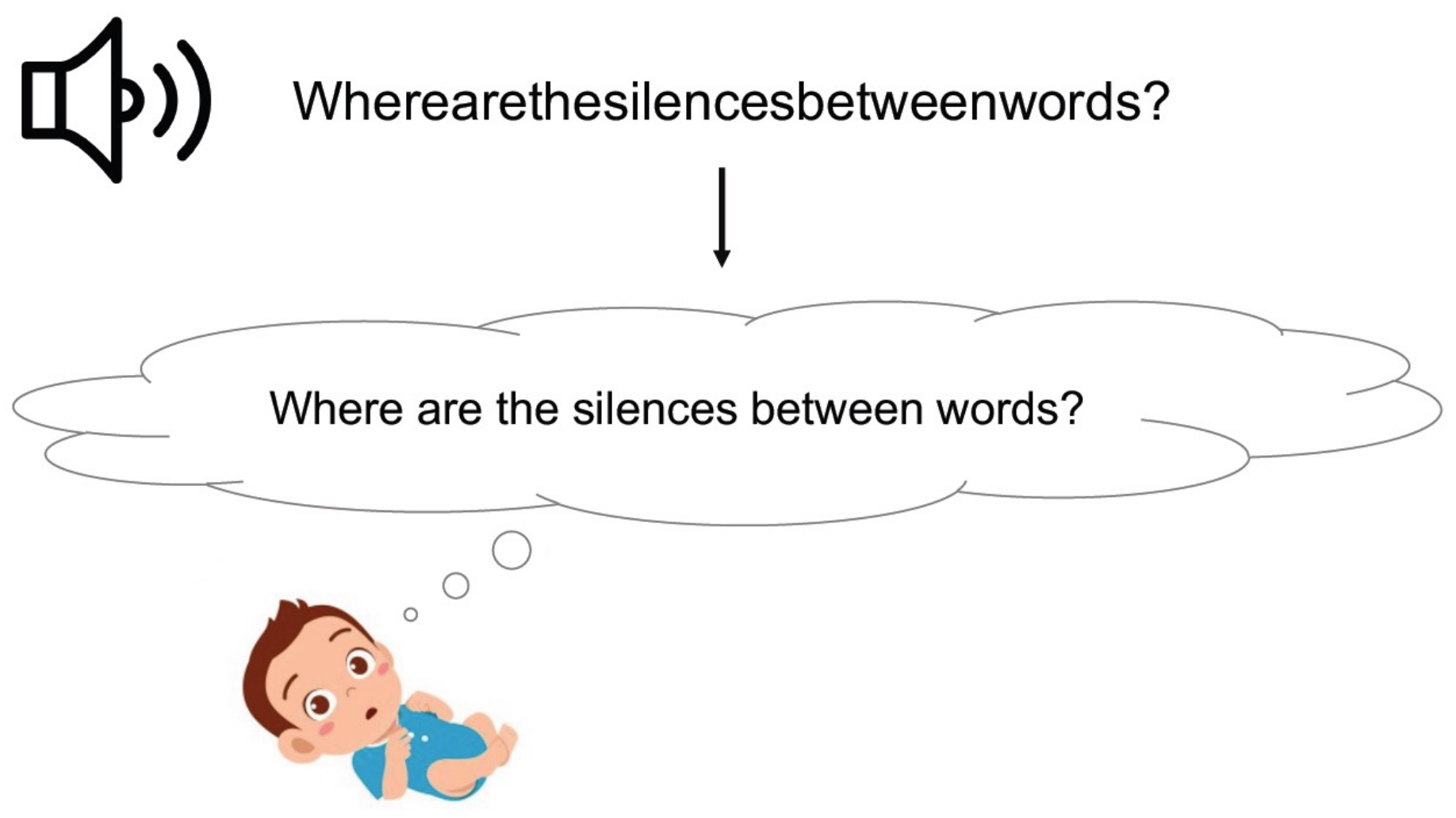
Statistical word segmentation in Korean child-directed speech

Seongmin Mun & Eon-Suk Ko
Chosun University

Introduction

- **First language acquisition:** A prerequisite for infants to build a lexicon for word learning is the ability to segment words out of the speech stream (e.g., Brent & Siskind, 2001; Jusczyk and Aslin, 1995).


Wherearethesilencesbetweenwords?

Where are the silences between words?

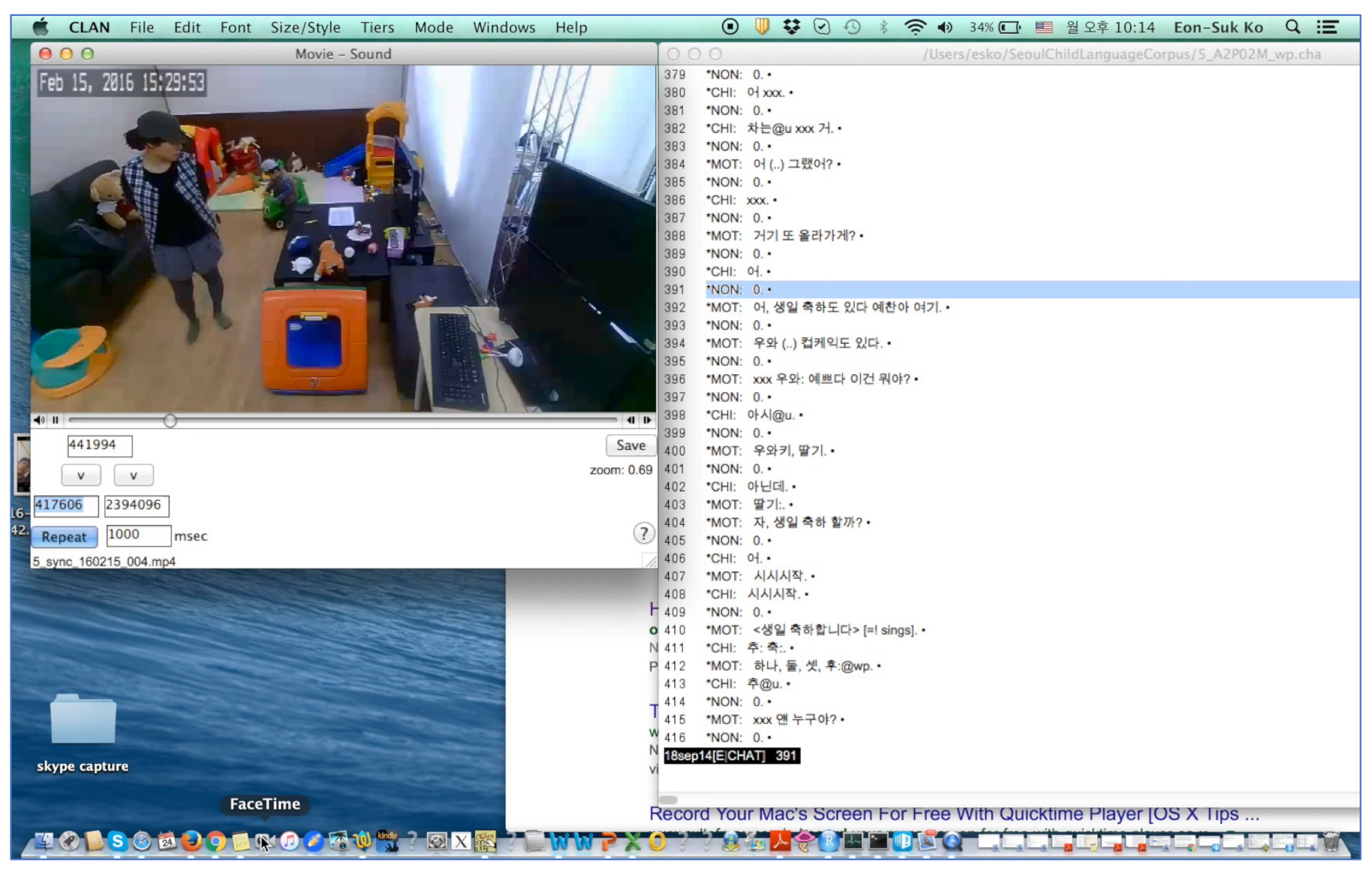
- **Background:** Behavioral studies suggest that infant's segments words more easily in CDS (child-directed speech) than ADS (adult-directed speech) (e.g., Fernald, 2000; Thiessen et al., 2005).
- Previous research on statistical segmentation:

Researches	Languages	Algorithms	CDS advantage?
Batchelder (2002)	English, Spanish, Japanese	1	Yes
Fourtassi et al. (2013)	English, Japanese	1	Yes
Ludusan et al. (2017)	Japanese	4	Yes
Cristina et al. (2018)	English	9	Not much
Loukatou et al. (2019)	French	17	Not much

- **Research question:** *Is there CDS advantages over ADS in the statistical segmentation of words in Korean?*

Methods

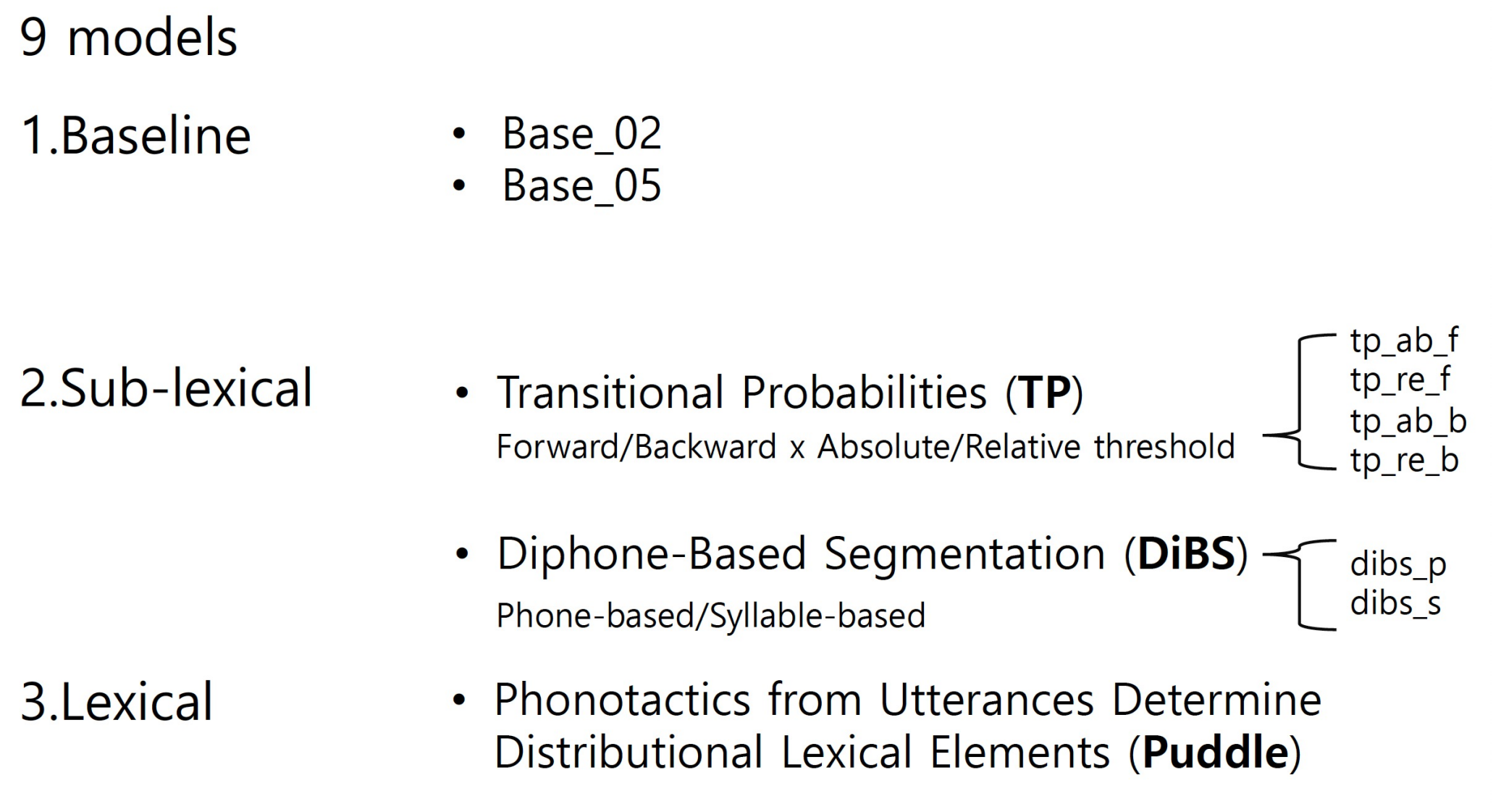
- **Data:** *Ko corpus* containing 35 mothers freely interacting with their own children for about 40 minutes. The same corpus also contains ADS in which the mother talks to their family members and experimenters for about 10 minutes(Ko et al., 2020).


CLAN File Edit Font Size/Style Tiers Mode Windows Help
34% 월 오후 10:14 Eon-Suk Ko
Movie - Sound
Feb 15, 2016 15:29:53
/Users/esko/SeoulChildLanguageCorpus/5_A2P02M_wp.cha
441994
Save
zoom: 0.69
417606
2394096
Repeat
1000
msec
5_sync_160215_004.mp4
skype capture
FaceTime
18sep14[E|CHAT] 391
Record Your Mac's Screen For Free With Quicktime Player [OS X Tips ...

```
379  *NON:  0. •
380  *CHI:  어 xxx. •
381  *NON:  0. •
382  *CHI:  차는@u xxx 거. •
383  *NON:  0. •
384  *MOT:  어 (..) 그랬어? •
385  *NON:  0. •
386  *CHI:  xxx. •
387  *NON:  0. •
388  *MOT:  거기 또 올라가게? •
389  *NON:  0. •
390  *CHI:  어. •
391  *NON:  0. •
392  *MOT:  어, 생일 축하도 있다 예찬아 여기. •
393  *NON:  0. •
394  *MOT:  우와 (..) 컵케익도 있다. •
395  *NON:  0. •
396  *MOT:  xxx 우와: 예쁘다 이건 뭐야? •
397  *NON:  0. •
398  *CHI:  아시@u. •
399  *NON:  0. •
400  *MOT:  우와키, 딸기. •
401  *NON:  0. •
402  *CHI:  아닌데. •
403  *MOT:  딸기:. •
404  *MOT:  자, 생일 축하 할까? •
405  *NON:  0. •
406  *CHI:  어. •
407  *MOT:  시시시작. •
408  *CHI:  시시시작. •
409  *NON:  0. •
410  *MOT:  <생일 축하합니다> [=! sings]. •
411  *CHI:  추: 축:. •
412  *MOT:  하나, 둘, 셋, 후:@wp. •
413  *CHI:  추@u. •
414  *NON:  0. •
415  *MOT:  xxx 앤 누구야? •
416  *NON:  0. •
```

Figure 1: The pictures show the environment of the apartment where the data were collected and the hand-coded transcriptions.

- **Statistical word segmentation models:** We used 9 word segmentation models through Python, by adapting functions provided by WordSeg (Bernard et al., 2019).


9 models

1.Baseline
- Base_02
- Base_05

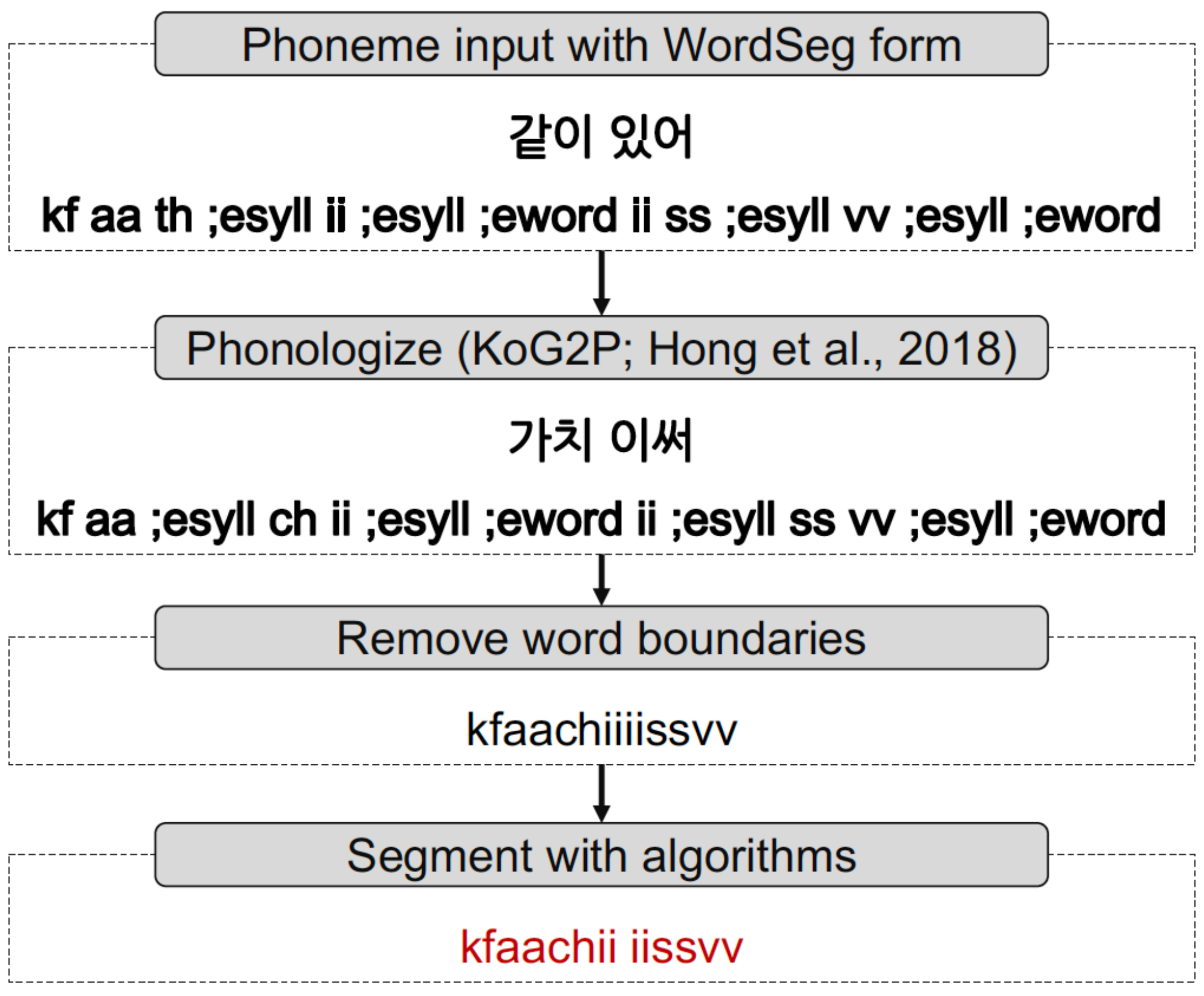
2.Sub-lexical
- Transitional Probabilities (**TP**)
  Forward/Backward x Absolute/Relative threshold
  tp_ab_f
  tp_re_f
  tp_ab_b
  tp_re_b
- Diphone-Based Segmentation (**DiBS**)
  Phone-based/Syllable-based
  dibs_p
  dibs_s

3.Lexical
- Phonotactics from Utterances Determine Distributional Lexical Elements (**Puddle**)

Figure 2: 9 word segmentation models that we used in this study.

Methods

- **Procedure:** We derived phonetic input from phonemic corpus by applying a set of phonological rules by using KoG2P (Hong et al., 2018). After then, we employed 9 word segmentation models through WordSeg (Bernard et al., 2019). Model performance was measured by comparing the word boundaries in the original input sentence with the word boundaries generated via each model.


Phoneme input with WordSeg form

같이 있어

kf aa th ;esyll ii ;esyll ;eword ii ss ;esyll vv ;esyll ;eword

Phonologize (KoG2P; Hong et al., 2018)

가치 이써

kf aa ;esyll ch ii ;esyll ;eword ii ;esyll ss vv ;esyll ;eword

Remove word boundaries

kfaachiiiissvv

Segment with algorithms

kfaachii iissvv

Figure 3: The overview of research process

Results

- **Characteristics of our CDS vs ADS data**

	phoneme					phonetic				
	Sylls	Tokens	Types	MATTR	Utts	Sylls	Tokens	Types	MATTR	Utts
ADS	24,088	11,012	3,227	0.909	2,544	24,088	11,011	3,215	0.909	2,544
CDS	144,609	63,887	8,818	0.837	22,203	144,615	63,826	8,770	0.837	22,203

Figure 4: Characteristics of the ADS and CDS portions of the corpus by phoneme input and phonetic input.

Feature	CDS	ADS	p
Word length (s)	1.68 (.11)	1.74 (0.16)	.101
Utterance length (s)	6.54 (.88)	9.21 (2.76)	2.671e-06 ***
% 1-w phrase	.33 (.06)	.33 (.12)	.77
MATTR	.84 (.07)	.91 (.03)	6.595e-07 ***
% hapaxes	.22 (.05)	.49 (.07)	< 2.2e-16 ***

Figure 5: Results of statistical analysis, t-tests measuring feature differences across CDS and ADS in phonetic form.

✓ The utterance length of ADS is longer than CDS.

✓ The MATTR (i.e., moving average type to token ratio) is high in ADS compared with CDS. This indicates that ADS has more types of words than CDS in a fixed window of 20 words.

✓ At the last, the proportion of hapaxes is high in ADS compared with CDS, which means that the portion of words that are used only one time in the corpus is higher in ADS than in CDS.

Results

- **Does CDS have a segmentation advantage over ADS?**

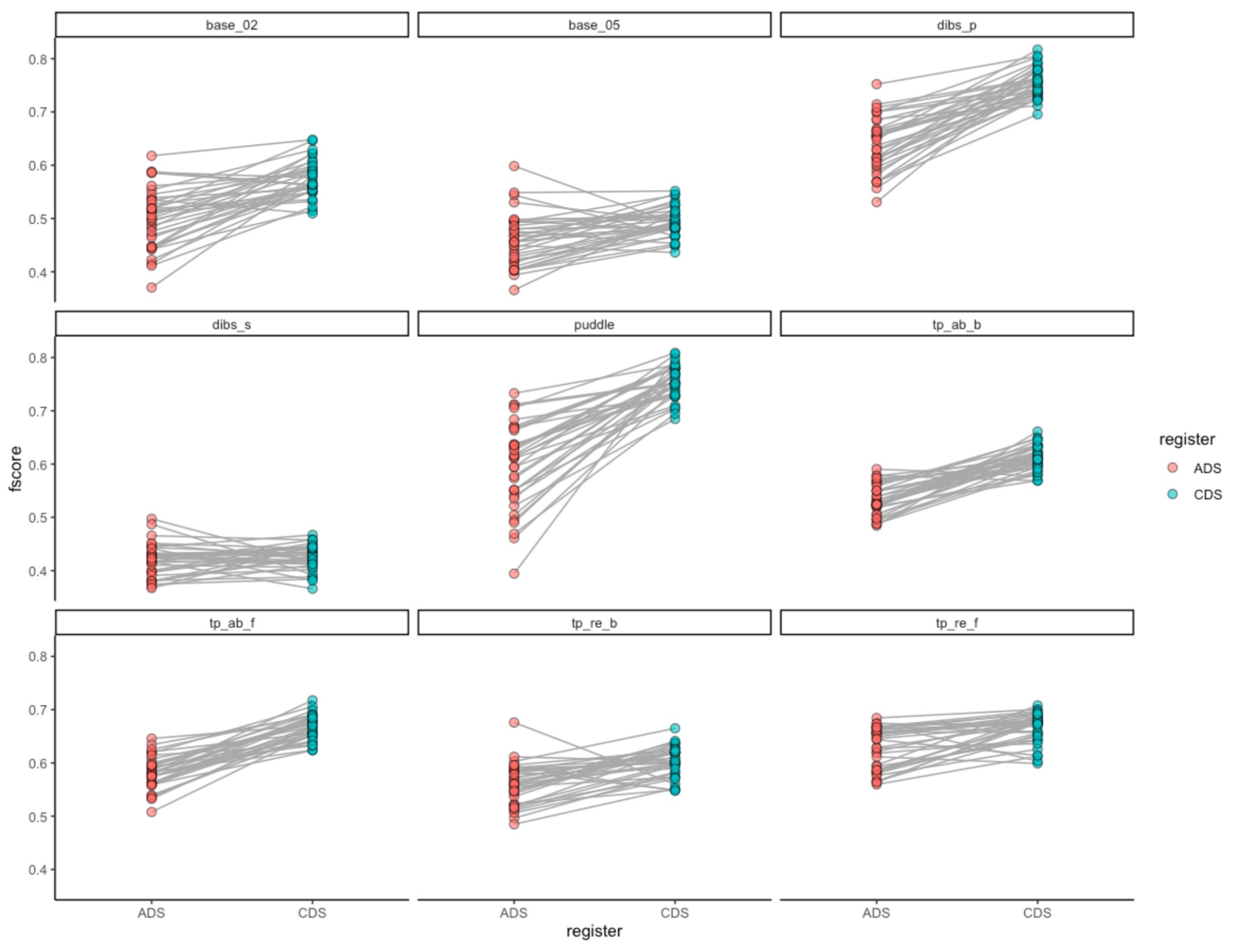


Figure 6: Token F-scores obtained by each algorithm for CDS and ADS.

- **Statistical modelling for CDS segmentation advantages**

Formula: f-score ~ register + unit + corpus size + (1+register|algo) + (1+register|baby)

factor	Estimate (β)	Std. Error	df	t value	p
register (CDS)	0.0888	0.021	24.9677	4.2318	0.0003 ***
unit (phonetic)	0.0014	0.0012	1173.0969	1.1634	0.2449
corpus size	.0	.0	36.4052	-1.5532	.129

Figure 7: Result of linear mixed effects regression models to statistically test the difference in the f0 ratio by registers and units

- **Which corpus properties have an effect on the segmentation advantages CDS?**

Formula: f-score ~ word length (s) + utterance length (s) + % hapaxes + % 1-w phrase+ MATTR + (1+register|algo)+(1+register|baby)

factor	Estimate (β)	Std. Error	df	t value	p
Word length (s)	-0.1059	0.0139	63.6124	-7.6212	0 ***
Utterance length (s)	-0.0093	0.0011	57.7701	-8.2175	0 ***
% hapaxes	-0.0295	0.0205	68.1422	-1.44	0.1545
% 1-w phrase	1.00E-04	0	35.1851	2.8841	0.0067 **
MATTR	-0.0347	0.0191	40.2442	-1.8166	0.0767 .

Figure 8: Result of linear mixed effects regression models to investigate the relationship between model performance and the corpus properties.

- **Conclusion & Discussion**

✓ CDS has a significantly greater advantage in word segmentation than ADS.

✓ Shorter word-length and utterance-length in CDS yields a greater F-ratio.

✓ A greater proportion of one-word phrases in CDS yields a greater F-ratio.

✓ A greater repetition ratio of repetition (MATTR) in CDS yields a greater F-ratio.

- **Future directions**

✓ Examine the role of sound symbolism and word play in segmentation.

✓ Control of corpus size with additional ADS corpus.

REFERENCES

- Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. Cognition, 83, 167–206.
- Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G., Larsen, E., Johnson, M., Fibla, L., Dupoux, E., Daland, R., Cao X. & Cristia, A. (2020). WordSeg: Standardizing unsupervised word form segmentation from text. Behav Res 52, 264–278. https://doi.org/10.3758/s13428-019-01223-3
- Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. Cognition, 81(2), B33–B44.
- Cristia, A., Dupoux, E. Ratner, N. & Soderstrom, M. (2019). Segmentability Differences Between Child-Directed and Adult-Directed Speech: A Systematic Test With an Ecologically Valid Corpus. Open Mind: Discoveries in Cognitive Science, 3, 13–22.a_00022
- Fernald, A. (2000). Speech to infants as hyperspeech: Knowledge-driven processes in early word recognition. Phonetica, 57, 242–254.
- Fourtassi, A., Börschinger, B., Johnson, M. & Dupoux, E. (2013). Why is English so easy to segment?. In Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL), 1–10.
- Hong, Y-S., Ki, K-S. & Gweon, G. 2018. Automatic Miscue Detection Using RNN Based Models with Data Augmentation. In Proc. Interspeech, 1646-1650.
- Jusczyk, P. W. & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. Cognitive Psychology, 29(1):1-23.
- Ko, E-S., Jo, J., On, K-W. & Zhang, B-T. (2020). Introducing the ko corpus of korean mother-child interaction. Frontiers in Psychology.
- Loukatou, G., Normand, M. & Cristia, A. (2019). Is it easier to segment words from infant-directed speech? Modeling evidence from an ecological French corpus. The 41st Annual Meeting of the Cognitive Science Society, 2186-2193.
- Ludusan, B., Mazuka, R., Bernard, M., Cristia, A., & Dupoux, E. (2017). The role of prosody and speech register in word segmentation: A computational modelling perspective. In Proceedings of the Annual Conference of the Association for Computational Linguistics (2): 178–183.
- Thiessen, E., Hill, E. & Saffran, J. (2005). Infant-directed speech facilitates word segmentation. Infancy, 7(1):53–71.

Presented at the ICIS 2022 (07-10 JULY 2022)

Corresponding Author: Seongmin Mun (*simon.seongmin.mun@gmail.com*)