MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

WWW.MPI.NL

A Corpus Analysis of the Word Segmentation Cues in German Child-Directed Speech

Katja Stärk[1,2], Evan Kidd[1,3,4] & Rebecca L.A. Frost[1]

1) Max-Planck-Institute for Psycholinguistics, Nijmegen, The Netherlands 2) IMPRS for Language Sciences, Nijmegen, The Netherlands 3) Research School of Psychology, The Australian National University, Canberra, Australia 4) ARC Centre of Excellence for the Dynamics of Language, Canberra, Australia

Please contact me:
Katja.Staerk@mpi.nl

@katjastaerk

LCICD 2019

Introduction

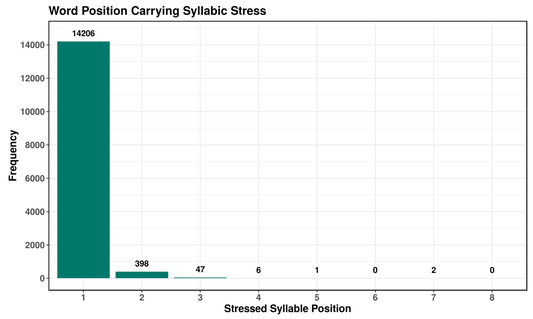
- One of the first challenges in language acquisition is speech segmentation
- Multiple cues have been found to aid segmentation[1,2]
  - → Stress[3]
  - → Highly frequent function words[4,5]
  - → Transitional probabilities[6,7]
- Availability and reliabitlity of cues differs between languages → RQ: which segmentation cues are available in German CDS?

Method

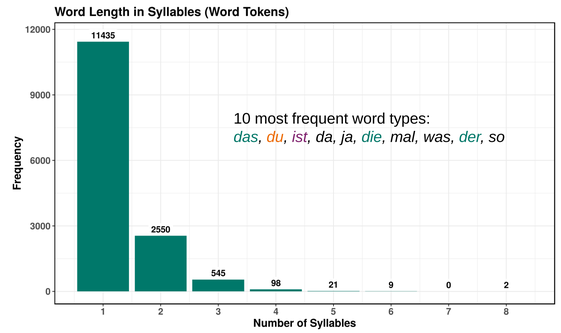
- Corpus analysis, performed on data from German CHILDES[8]
- Coded ~16,000 words (~4,000 utterances) = ca. 1 day of input[9]
- Coded for: *primary word stress*, *word length* (number of syllables), *word category* (function/content word), *syllable structure*, *TPs*

Results

- Stress: 97% of words contained word-initial stress (87% of unique word tokens (frequency-controlled), 86% excluding monosyllabic words)



- Word length: 78% monosyllabic, 17% disyllabic, (37% mono-, 40% disyllabic for unique word tokens)



- Word category: 62% of all words were monosyllabic function words; 45/50 most frequent words were function words (90%)
- Syllable structure: Varied depending on syllable position within words: intermediate syllables were often open, wheras initial and final syllables tended to be closed

	overall		initial		intermediate		final	
1	CVC	3777	CVV	3190	CV	338	CVC	3305
2	CVV	3482	CVC	3008	CVV	157	CVV	2633
3	CV	2993	CV	1324	CVC	120	CV	1748
4	CVVC	1449	CVVC	1318	CCVV	42	CVVC	1312
5	CVVV	1061	CVVV	1025	CCV	39	CVVV	1004

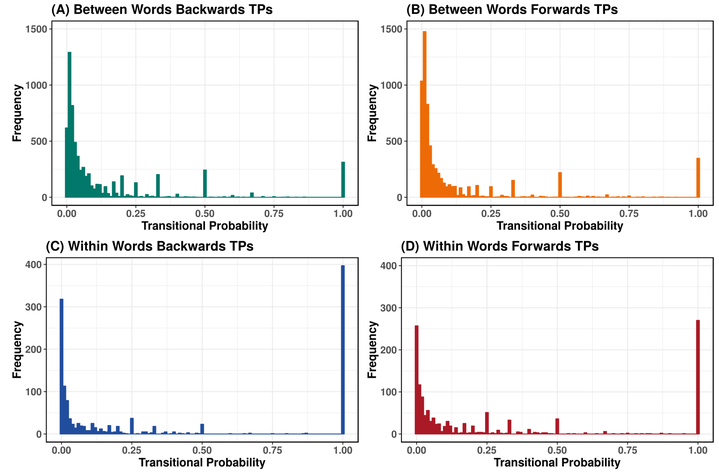
- Transitional probabilities (TPs): We found two main effects and an interaction, controlling for the frequency of the syllable pair

	Forwards TPs	Backwards TPs
Within-word	M = 0.296	M = 0.364
	(SD = 0.388)	(SD = 0.436)
Between-words	M = 0.097	M = 0.124
	(SD = 0.207)	(SD = 0.224)

→ TPs were higher within than between words (β = -0.222, SE = 0.006, t = -38.524, p < .001)

→ TPs were higher backwards than forwards (β = 0.025, SE = 0.003, t = 8.851, p < .001)

→ There was a larger difference within/between-words for the backwards TPs (β = -0.025, SE = 0.006, t = 8.851, p < .001)



Summary

- First corpus analysis on word segmentation cues in German
- Multiple segmentation cues are available
  - → Stress as almost perfect cue to wordhood in German CDS
  - → High proportion of high frequency function words, which may act as anchors during segmentation[4,5]
  - → Statistical cues to segmentation (TPs) are present but seem to be weaker than stress cue

REF: 1) Aslin et al. (1996). *Models of Word Segmentation in Fluent Maternal Speech to Infants*; 2) Lehiste (1970). *Suprasegmentals*; 3) Friederici, et al. (2007). *Curr. Biol.*; 4) Bortfeld et al. (2005). *Psych. Sci.*; 5) Frost et al. (2019). *JEP: LMC*; 6) Saffran et al. (1996). *Science*; 7) Aslin et al. (1998). *Psych. Sci.*; 8) MacWhinney (2000). *The CHILDES Project*; 9) Rowland et al., (unpubl.). *The Language 0-5 Project*