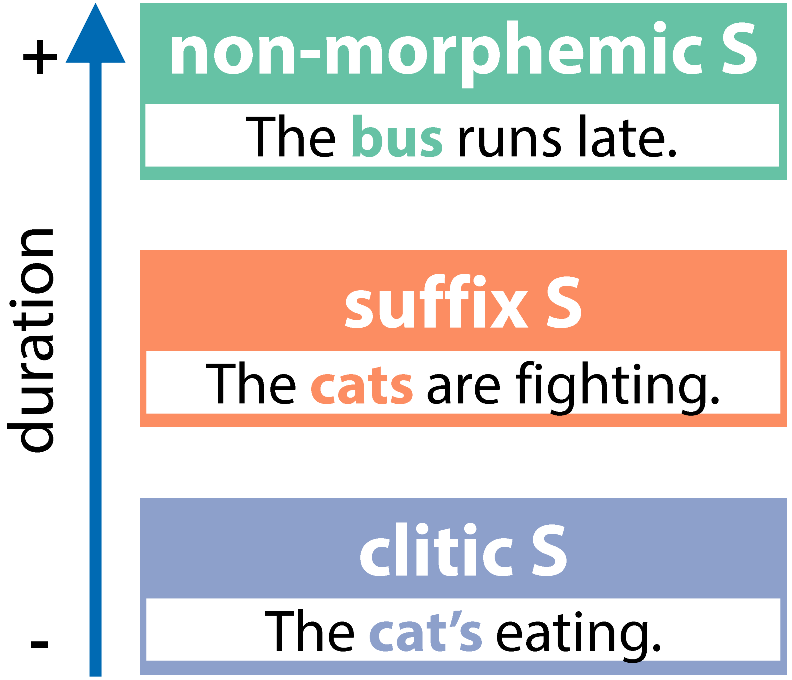
Durational differences of homophonous suffixes emerge from the lexicon:

Evidence from nonce words

Dominic Schmitz, Ingo Plag, Dinah Baer-Henney

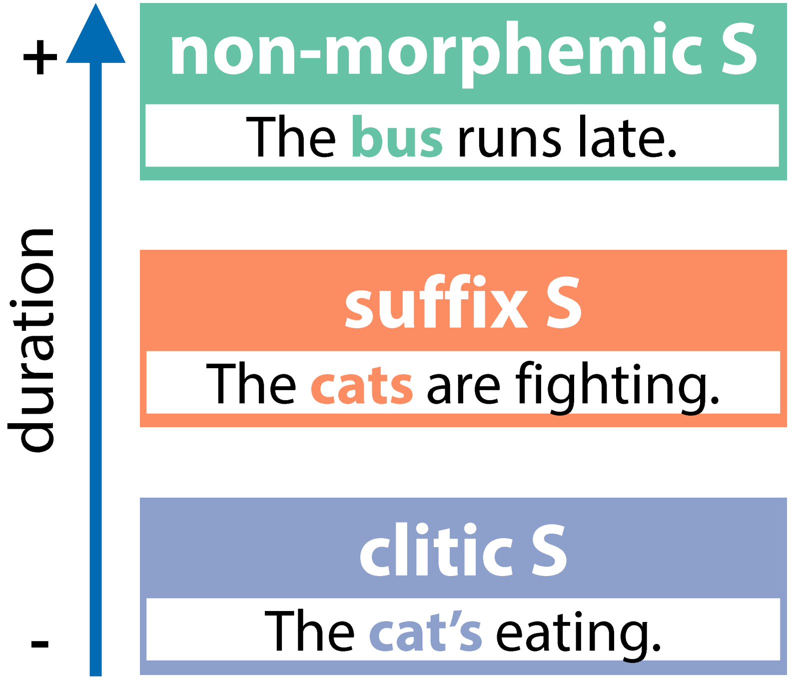
Starting point

- Zimmermann (2016)
- Plag et al. (2017)
- Tomaschek et al. (2019)
- Schmitz et al. (2020) on nonce words



Starting point

- Zimmermann (2016)
- Plag et al. (2017)
- Tomaschek et al. (2019)
- Schmitz et al. (2020) on nonce words



How do such differences come to existence?

Linear Discriminative Learning – LDL

Linear Discriminative Learning – LDL

- **LDL** (Baayen et al., 2018; 2019) describes a mathematical and computational model of the mental lexicon

Linear Discriminative Learning – LDL

- **LDL** (Baayen et al., 2018; 2019) describes a mathematical and computational model of the mental lexicon
- **form and meaning** can be mapped onto each other using linear networks

Linear Discriminative Learning – LDL

- **LDL** (Baayen et al., 2018; 2019) describes a mathematical and computational model of the mental lexicon
- **form and meaning** can be mapped onto each other using linear networks
- **LDL** takes **lexomes** as the basic units for lexical processing

Linear Discriminative Learning – LDL

- **LDL** (Baayen et al., 2018; 2019) describes a mathematical and computational model of the mental lexicon
- **form and meaning** can be mapped onto each other using linear networks
- **LDL** takes **lexomes** as the basic units for lexical processing
- each lexome is connected to a **semantic vector** containing the **association strengths** of its **lexome** with each of the other **lexomes**

Linear Discriminative Learning – LDL

- **LDL** (Baayen et al., 2018; 2019) describes a mathematical and computational model of the mental lexicon
- **form and meaning** can be mapped onto each other using linear networks
- **LDL** takes **lexomes** as the basic units for lexical processing
- each lexome is connected to a **semantic vector** containing the **association strengths** of its **lexome** with each of the other **lexomes**
- **lexomes** and their **association strengths** can then be used to obtain a number of **LDL measures**

How do we obtain LDL measures?

How do we obtain LDL measures?

1. From data to matrices

1. From data to matrices

- $\boxed{\text{data}}$

1. From data to matrices

- **data**
  - word forms, bases, affixes, and transcriptions

1. From data to matrices

- data
  - word forms, bases, affixes, and transcriptions

- real words (MALD, Tucker et al., 2018)

Word	Base	Affix	Transcription
meal	meal	NA	mil
meat	meat	NA	mit
students	student	PL	stjudHts
teacher	teacher	NA	tiJ@R

1. From data to matrices

- **data**
  - word forms, bases, affixes, and transcriptions

- real words (MALD, Tucker et al., 2018)

Word	Base	Affix	Transcription
meal	meal	NA	mil
meat	meat	NA	mit
students	student	PL	stjudHts
teacher	teacher	NA	tiJ@R

- pseudowords (Schmitz et al., 2020)

Word	Base	Affix	Transcription
bloups	bloups	NA	bl6ps
bloups	bloup	PL	bl6ps
pleeps	pleeps	NA	plips
pleeps	pleep	PL	plips

1. From data to matrices

[C]ue matrix

- contains the triphones of all **word forms**

	#mi	mil	il#	mit	it#
/mil/	1	1	1	0	0
/mit/	1	0	0	1	1
/stjudHt/	0	0	0	0	0
/tiJ@R/	0	0	0	0	0

1. From data to matrices

[S]emantic matrix

- there is a number of options when it comes to semantics, i.e. whether to use real (Chuang et al., 2020) or simulated (Baayen et al., 2018) semantics for parts of or all data
- today:
  - Simulated semantic vectors for real words and/or pseudowords
    → real and/or pseudowords contain some sort of semantics

1. From data to matrices

[S]emantic matrix

- contains semantic vectors for all **word forms**

	classroom	college	cook	eat	vegetable	PL
/mil/	0.003	0.0005	0.9	0.8	0.7	0.2
/mit/	0.0006	0.0002	0.8	0.9	0.5	0.04
/stjudHt/	0.9	0.8	0.05	0.1	0.005	0.7
/tiJ@R/	0.8	0.8	0.09	0.003	0.02	0.5

1. From data to matrices

[S]emantic matrix

- contains semantic vectors for all **word forms**

	classroom	college	cook	eat	vegetable	PL
/mil/	0.003	0.0005	0.9	0.8	0.7	0.2
/mit/	0.0006	0.0002	0.8	0.9	0.5	0.04
/stjudHt/	0.9	0.8	0.05	0.1	0.005	0.7
/tiJ@R/	0.8	0.8	0.09	0.003	0.02	0.5

1. From data to matrices

[S]emantic matrix

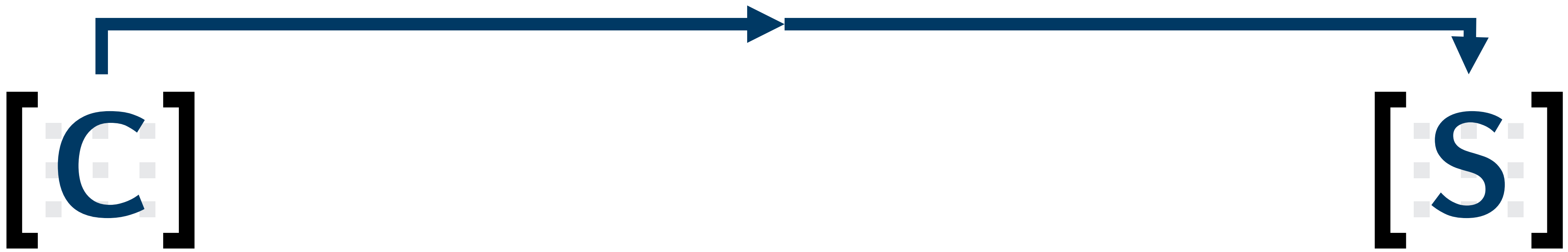
- contains semantic vectors for all **word forms**

	classroom	college	cook	eat	vegetable	PL
/mil/	0.003	0.0005	0.9	0.8	0.7	0.2
/mit/	0.0006	0.0002	0.8	0.9	0.5	0.04
/stjudHt/	0.9	0.8	0.05	0.1	0.005	0.7
/tiJ@R/	0.8	0.8	0.09	0.003	0.02	0.5

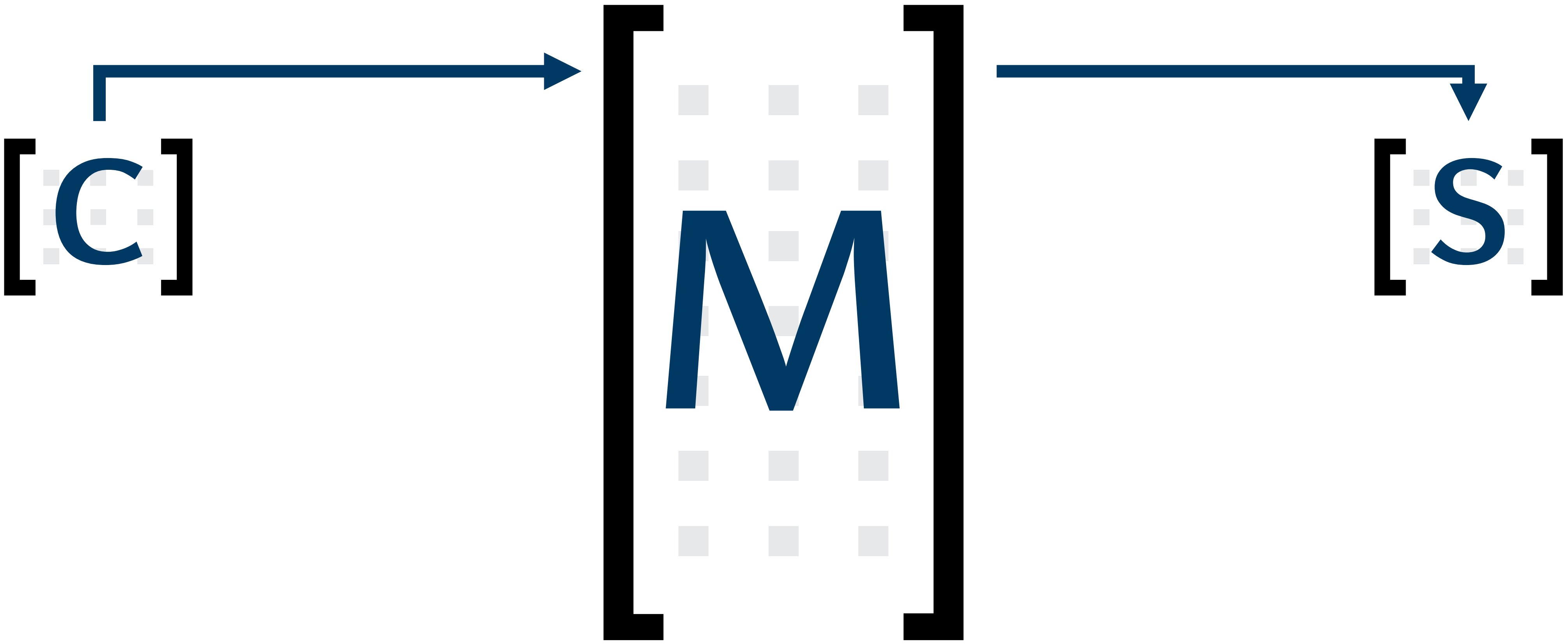
How do we obtain LDL measures?

1. From data to matrices
2. From matrices to comprehension & production

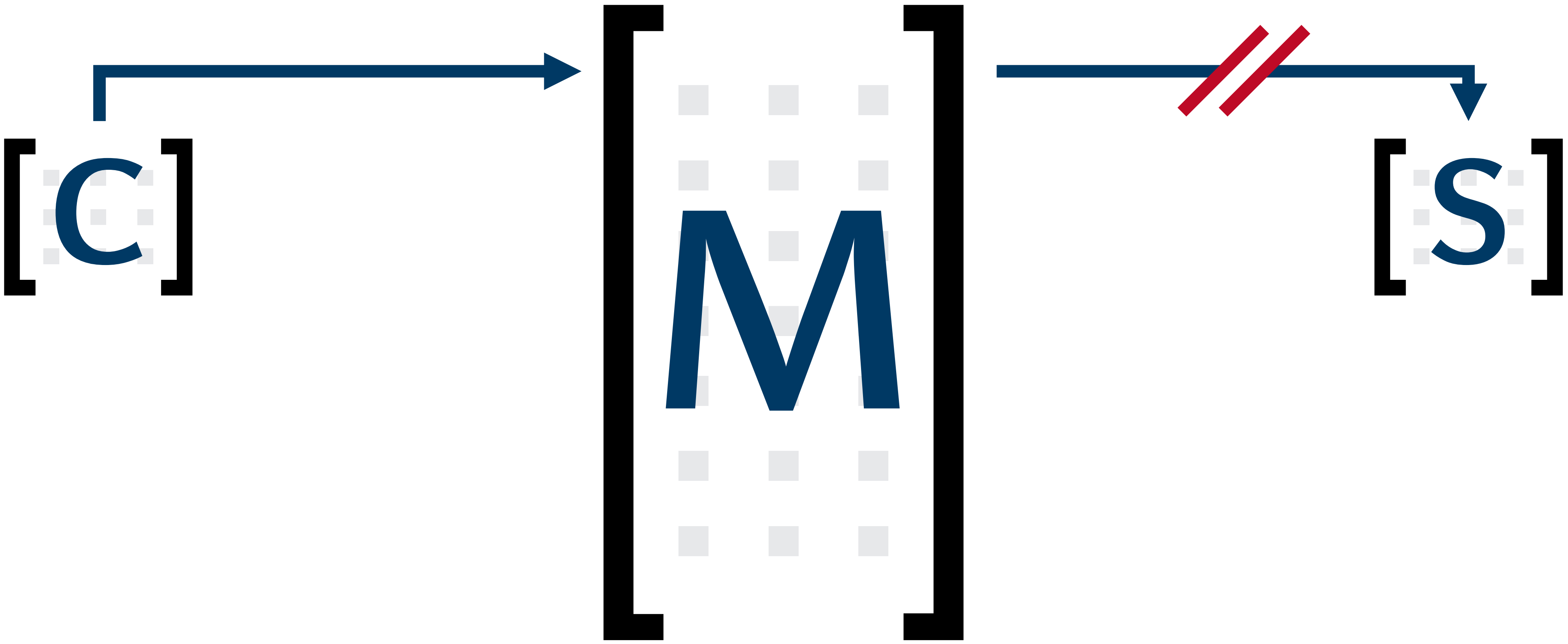
2. From matrices to comprehension & production



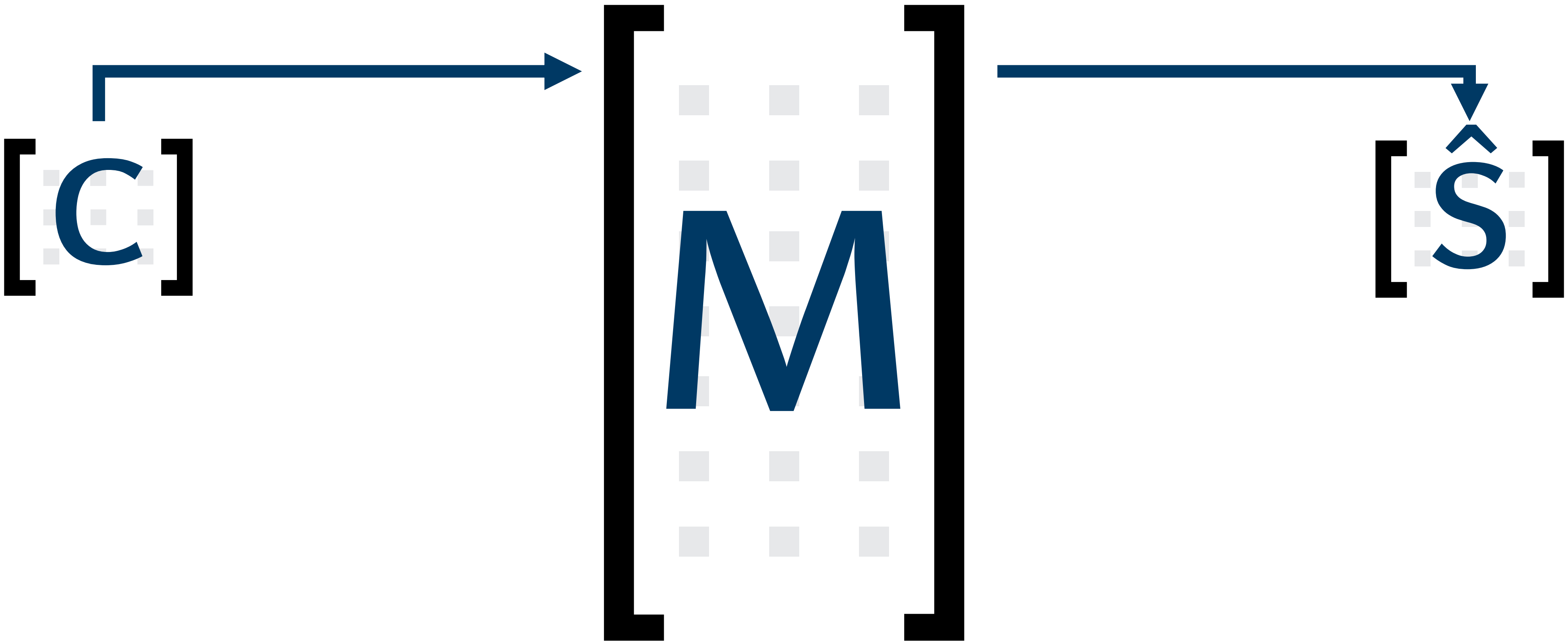
2. From matrices to comprehension & production



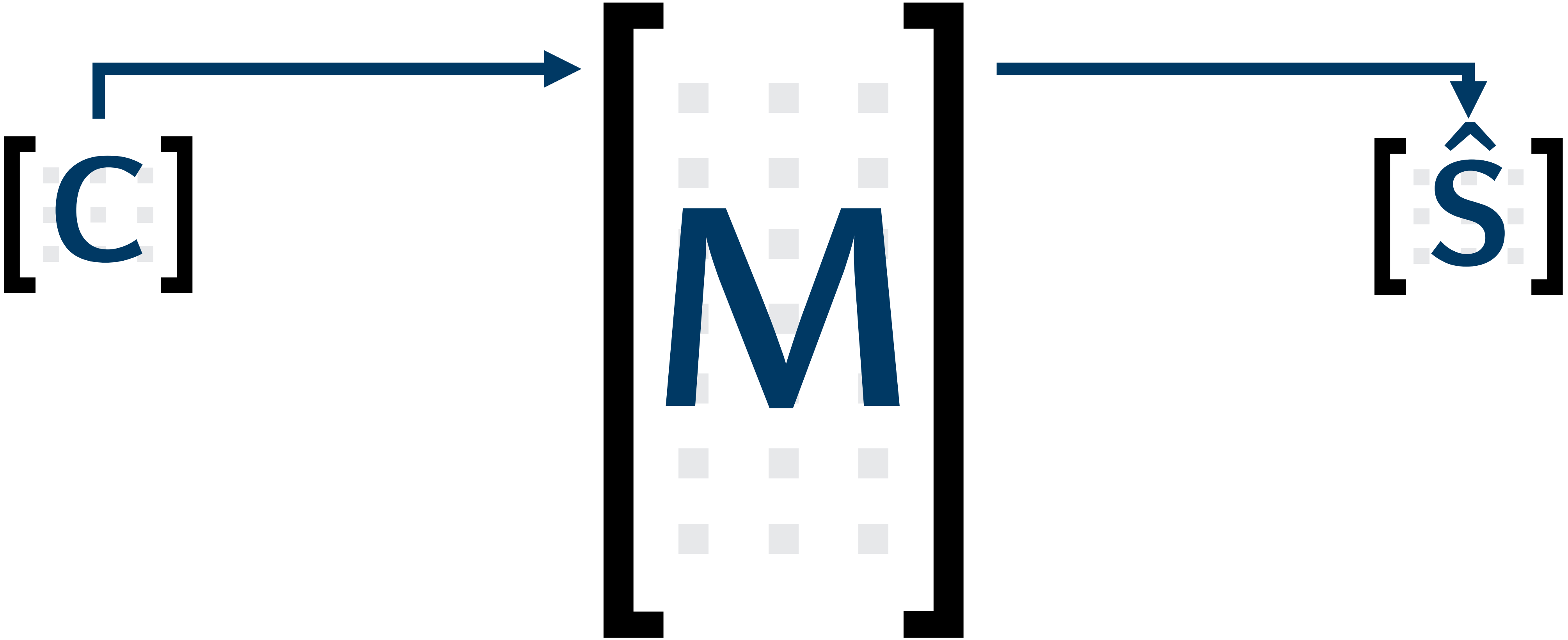
2. From matrices to comprehension & production



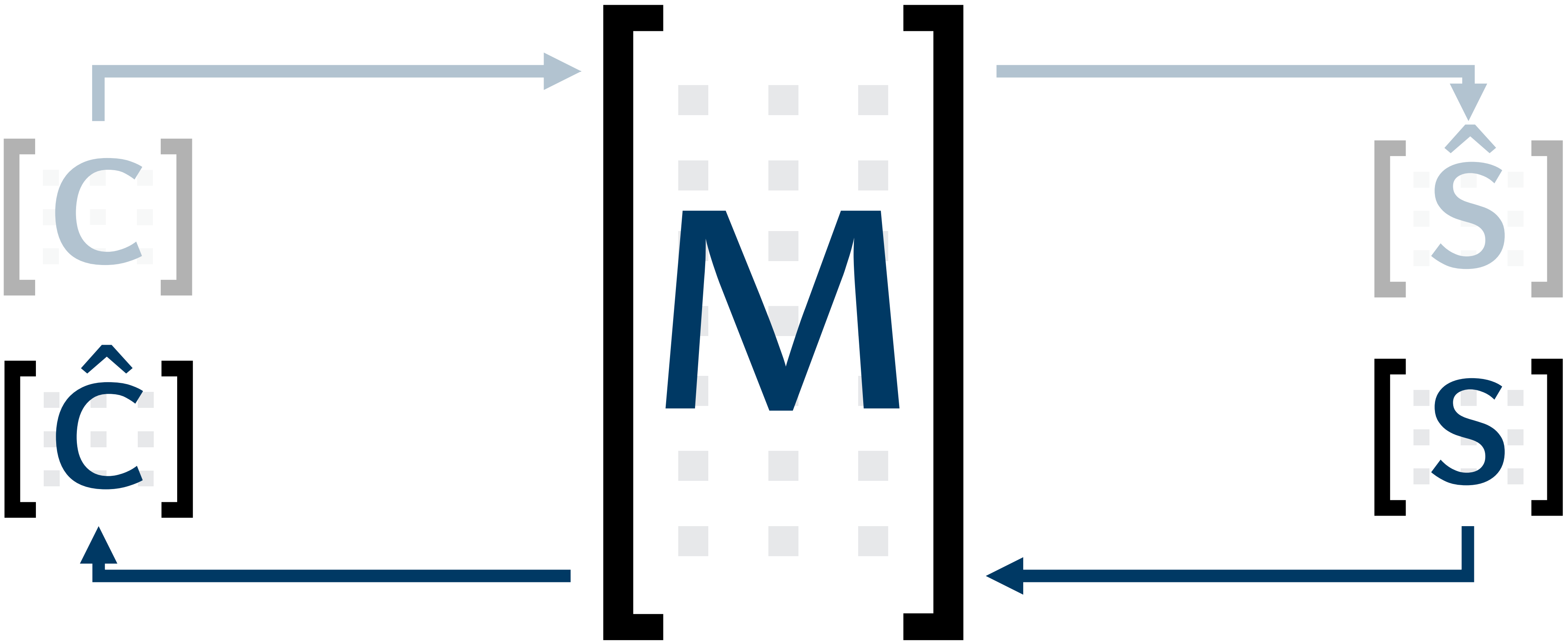
2. From matrices to comprehension & production



2. From matrices to comprehension & production



2. From matrices to comprehension & production



How do we obtain LDL measures?

1. From data to matrices
2. From matrices to comprehension & production
3. From comprehension & production to measures

4. From comprehension to measures

$[C] \rightarrow [M] \rightarrow [\hat{S}]$

&

$[S]$

&

data

comp measures

4. From production to measures

[S] → [M] → [Ĉ]

&

[C]

&

data

prod measures

Results

Results

a) Pseudowords

Results: Pseudowords

- mapping data: 48 pseudowords; 24 monomorphemic, 24 plurals

Results: Pseudowords

- mapping data: 48 pseudowords; 24 monomorphemic, 24 plurals
- Comprehension accuracy: 100%
- Production accuracy: 100%

Results: Pseudowords

- mapping data: 48 pseudowords; 24 monomorphemic, 24 plurals
- Comprehension accuracy: 100%
- Production accuracy: 100%
- LDL measures:

Results: Pseudowords

- mapping data: 48 pseudowords; 24 monomorphemic, 24 plurals
- Comprehension accuracy: 100%
- Production accuracy: 100%
- LDL measures:
  - checking relative variable importance and correlations, 1 LDL measure is found to be a significant predictor for /s/ duration :

    CORRELATIONS the correlation of the predicted path with the targeted semantic vector

Results: Pseudowords

- mixed effects regression model for the non-morphemic and plural /s/ duration data from Schmitz et al. (2020)

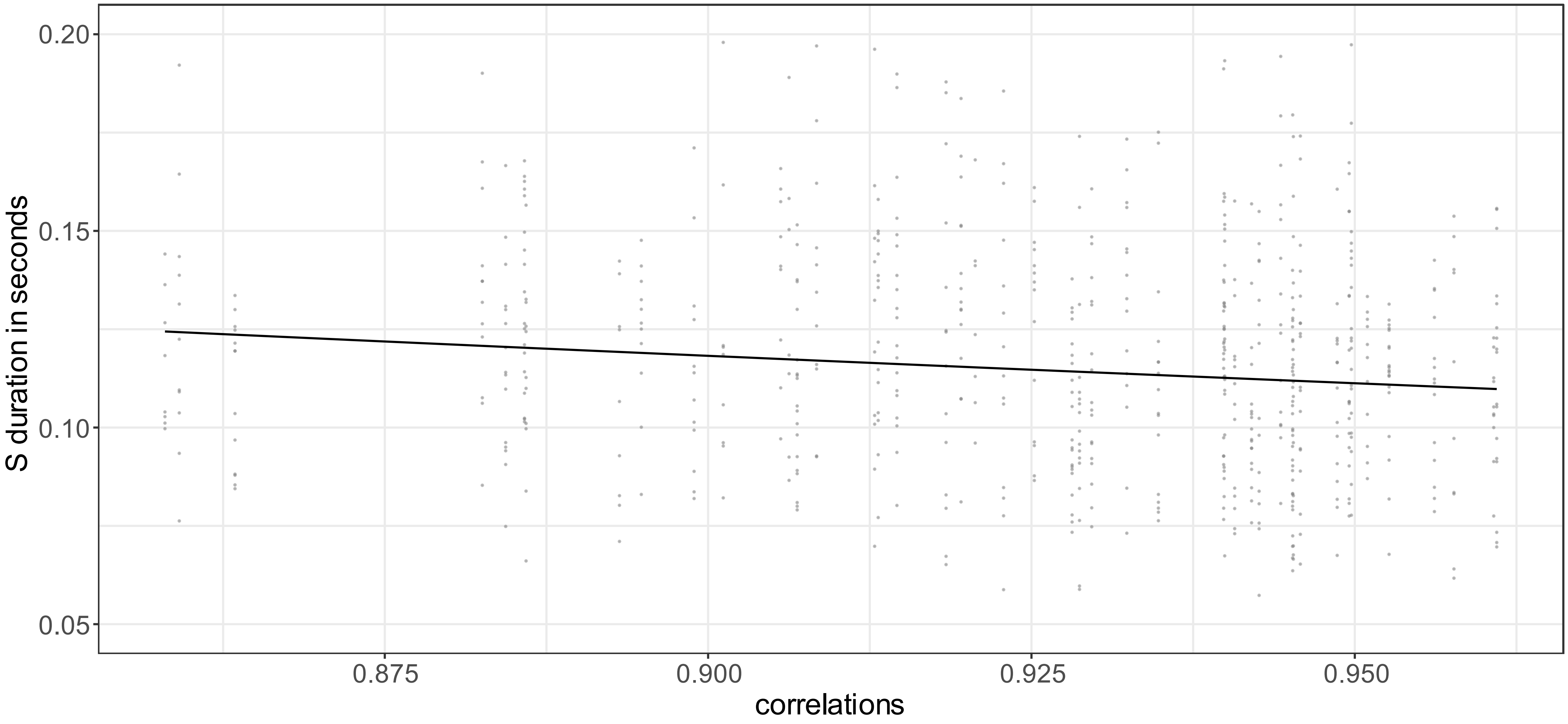
Results: Pseudowords

- mixed effects regression model for the non-morphemic and plural /s/ duration data from Schmitz et al. (2020)
- fixed effects (after exclusion of non-significant variables):
  - CORRELATIONS
  - PAUSEBIN — pause following the /s/: yes/no
  - FOLTYPE — phone following the /s/: approximant, fricative, etc.
  - SPEAKINGRATELOG — syllables per minute, log-transformed

Results: Pseudowords

- mixed effects regression model for the non-morphemic and plural /s/ duration data from Schmitz et al. (2020)
- fixed effects (after exclusion of non-significant variables):
  - **CORRELATIONS**
  - PAUSEBIN — pause following the /s/: yes/no
  - FOLTYPE — phone following the /s/: approximant, fricative, etc.
  - SPEAKINGRATELOG — syllables per minute, log-transformed
- random intercept:
  - SPEAKER

Results: Pseudowords



Results

a) Pseudowords

b) Real + Pseudowords

Results: Real + Pseudowords

- mapping data: 48 pseudowords; 24 monomorphemic, 24 plurals &

  8328 words; 6186 monomorphemic, 2094 with affixes (25 plurals)

Results: Real + Pseudowords

- mapping data: 48 pseudowords; 24 monomorphemic, 24 plurals &
  8328 words; 6186 monomorphemic, 2094 with affixes (25 plurals)
- Comprehension accuracy: 98.4%
- Production accuracy: 99.9%

Results: Real + Pseudowords

- mapping data: 48 pseudowords; 24 monomorphemic, 24 plurals &

  8328 words; 6186 monomorphemic, 2094 with affixes (25 plurals)

- Comprehension accuracy: 98.4%
- Production accuracy: 99.9%
- LDL measures:

Results: Real + Pseudowords

- mapping data: 48 pseudowords; 24 monomorphemic, 24 plurals &
  8328 words; 6186 monomorphemic, 2094 with affixes (25 plurals)
- Comprehension accuracy: 98.4%
- Production accuracy: 99.9%
- LDL measures:
  - checking relative variable importance and correlations, 1 LDL measure is found to be a significant predictor for /s/ duration:

    PATH_SUM the summed support for the predicted path

Results: Real + Pseudowords

- mixed effects regression model for the non-morphemic and plural /s/ duration data from Schmitz et al. (2020)

Results: Real + Pseudowords

- mixed effects regression model for the non-morphemic and plural /s/ duration data from Schmitz et al. (2020)
- fixed effects (after exclusion of non-significant variables):
  - **PATH_SUM**
  - PAUSEBIN
  - FOLTYPE
  - SPEAKINGRATELOG
- random intercept:
  - SPEAKER

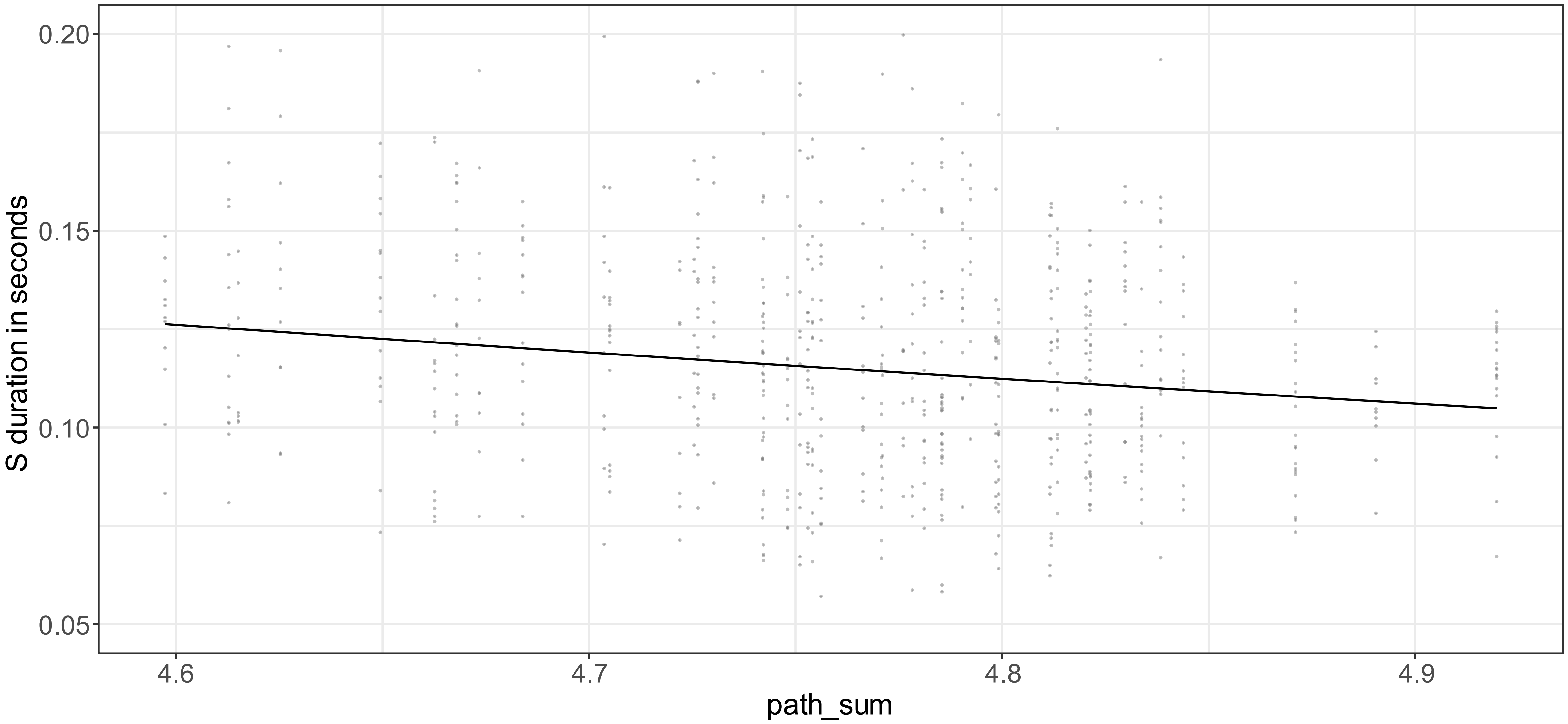
Results: Real + Pseudowords

- mixed effects regression model for the non-morphemic and plural /s/ duration data from Schmitz et al. (2020)
- fixed effects (after exclusion of non-significant variables):
  - **PATH_SUM** the summed support for the predicted path:

    $\xrightarrow{1} \#bl \xrightarrow{1} bli \xrightarrow{1} lip \xrightarrow{1} ips \xrightarrow{1} ps\#$

    $\sum = 5 \longrightarrow$ high support
  - PAUSEBIN
  - FOLTYPE
  - SPEAKINGRATELOG
- random intercept:
  - SPEAKER

Results: Real + Pseudowords



Discussion

correlations	path_sum	/s/ duration
high	high	short
low	low	long

Discussion

	correlations	path_sum	/s/ duration
plural	high	high	short
monomorphemic	low	low	long

Discussion

	correlations	path_sum	/s/ duration
plural	high	high	short
monomorphemic	low	low	long

- remaining questions:
  - *Why are predicted paths of plurals more correlated to their targeted semantic vectors?*
  - *Why is the certainty in plurals higher than in monomorphemic words?*

Conclusion

Conclusion

- Some LDL measures appear to be predictable for differences in /s/ durations, thus durational differences in word-final /s/ appear to emerge from the lexicon

Conclusion

- Some LDL measures appear to be predictable for differences in /s/ durations, thus durational differences in word-final /s/ appear to emerge from the lexicon
- However, further steps are necessary
  - use more data for mapping
  - use real semantics for real words, and derived semantics for pseudowords
  - analyse LDL measures not only for predicting /s/ durations in pseudowords but also for real words

Thank you!

References

Baayen, R. H., Chuang, Y. Y., & Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon 13*, 232-270.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The Discriminative Lexicon: A Unified Computational Model for the Lexicon and Lexical Processing in Comprehension and Production Grounded Not in (De)Composition but in Linear Discriminative Learning. *Complexity*.

Chuang, Y.-Y., Vollmer, M. L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2020). The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods*.

Li, H., Leonard, L., & Swanson, L. (1999). Some differences between English plural noun inflections and third singular verb inflections in the input: The contribution of frequency, sentence position and duration. *Journal of Child Language 26*, 531-543.

Plag, I., Homann, J., & Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics 53*, 181-216.

Plag, I., Ben Hedia, S., Lohmann, A., & Zimmermann, J. (2019). An <s> is an <s'>, or is it? Plural and genitive-plural are not homophonous. To appear in Körtvélyessy, L. & Stekauer, P. (Eds.) *Complex Words*. Cambridge: Cambridge University Press.

Schmitz, D., Baer-Henney, D., & Plag, I. (2020). *The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords*. Manuscript submitted for publication.

Seyfarth, S., Garallek, M., Gillingham, G., Ackermann, F., & Malouf, R. (2017). Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience*, 1-18.

Tomaschek, F., Plag, I., Baayen, R. H., & Ernestus, M. (2019). Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning. *Journal of Linguistics*, 1-39.

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadic, F., & Sims, M. (2018). The massive auditory lexical decision (mald) database. *Behavior Research Methods*, 1-18.

van de Vijver, R., & Baer-Henney, D. (2014). Developing biases. *Frontiers in Psychology 8*, 1-8.

Walsh, T., & Parker, F. (1983). The duration of morphemic and non-morphemic /s/ in English. *Journal of Phonetics 11*, 201-206.

Zimmermann, J. (2016). Morphological status and acoustic realization: Findings from NZE. In Carignan, C. & Tyler, M. D. (Eds.) *Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology (SST-2016)*, Parramatta, Australia, 6-9 December 2016. Canberra: ASSTA, 201-204.