JESSICA NIEDER & JOHANN-MATTIS LIST

MULTILINGUAL COMPUTATIONAL LINGUISTICS

UNIVERSITY OF PASSAU

# A COMPUTATIONAL MODEL FOR THE ASSESSMENT OF MUTUAL INTELLIGIBILITY AMONG CLOSELY RELATED LANGUAGES

# INTRODUCTION

- Speakers of a given language, e.g. German, can often partially understand speakers of other closely related languages, e.g. Dutch

# INTRODUCTION

- Speakers of a given language, e.g. German, can often partially understand speakers of other closely related languages, e.g. Dutch

Mein Name ist Jessica. Ich esse gerne Brot, trinke gerne Wasser und gehe gerne schwimmen.

# INTRODUCTION

- Speakers of a given language, e.g. German, can often partially understand speakers of other closely related languages, e.g. Dutch

Mijn naam is Jessica. Ik eet graag brood, drink graag water en ga graag zwemmen.

# INTRODUCTION

- Speakers of a given language, e.g. German, can often partially understand speakers of other closely related languages, e.g. Dutch

*'My name is Jessica. I like to eat bread, I like to drink water and I like to go swimming.'*

# INTRODUCTION

- Speakers of a given language, e.g. German, can often partially understand speakers of other closely related languages, e.g. Dutch

## = MUTUAL INTELLIGIBILITY

# INTRODUCTION

- Speakers of a given language, e.g. German, can often partially understand speakers of other closely related languages, e.g. Dutch

## = MUTUAL INTELLIGIBILITY

LINGUISTIC FACTORS

EXTRA-LINGUISTIC FACTORS

# INTRODUCTION

- Speakers of a given language, e.g. German, can often partially understand speakers of other closely related languages, e.g. Dutch

## = MUTUAL INTELLIGIBILITY

### LINGUISTIC FACTORS

- Lexicon
- Orthography
- Morphology
- Phonological similarities
- Modality: spoken vs. written

### EXTRA-LINGUISTIC FACTORS

# INTRODUCTION

- Speakers of a given language, e.g. German, can often partially understand speakers of other closely related languages, e.g. Dutch

## = MUTUAL INTELLIGIBILITY

### LINGUISTIC FACTORS

- Lexicon
- Orthography
- Morphology
- Phonological similarities
- Modality: spoken vs. written

### EXTRA-LINGUISTIC FACTORS

- Previous exposure
- Attitude towards target language

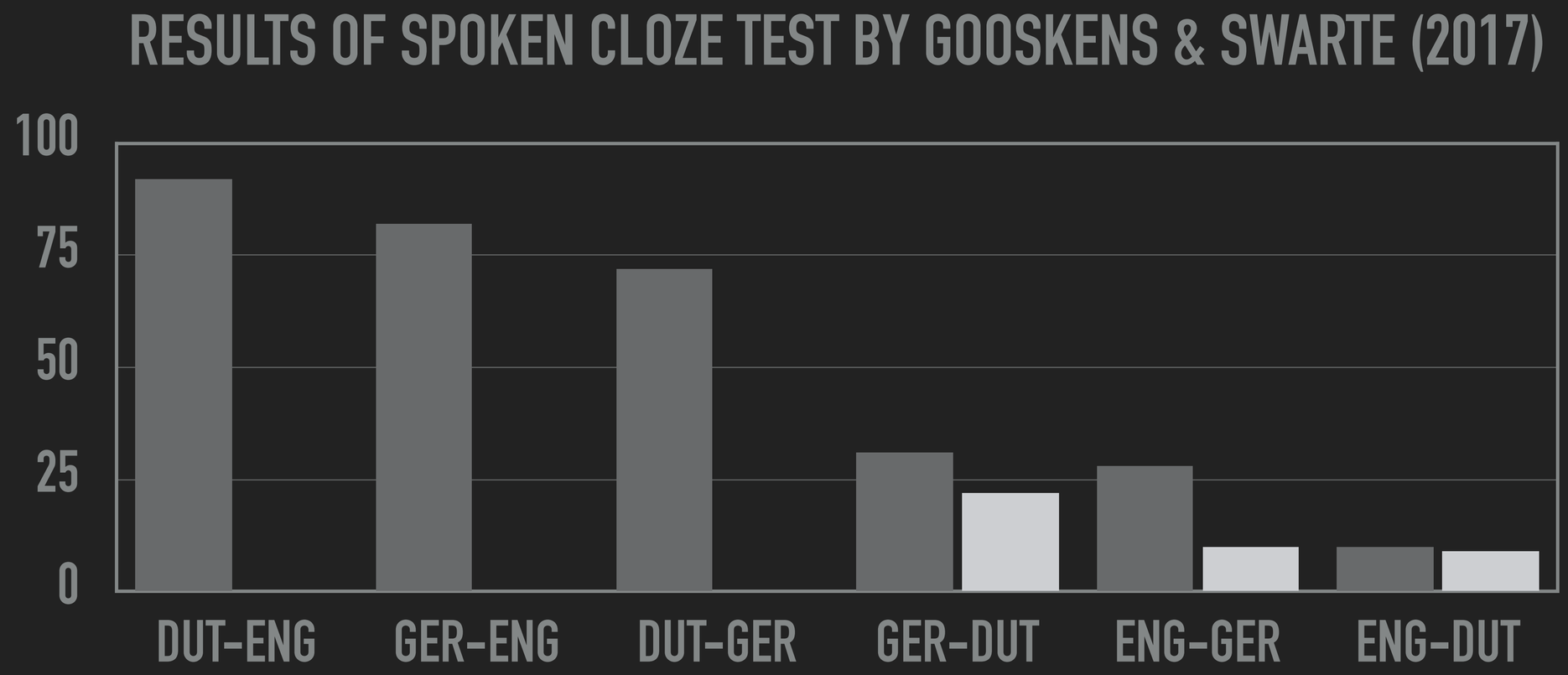
# PREVIOUS EXPERIMENTAL WORK

- Research on mutual intelligibility involves experimental studies with participants
- Gooskens and Swarte (2017): large-scale study on mutual intelligibility of Germanic languages using a spoken and written cloze test and language background questionnaires
- Total of 954 participants with 5 different native languages (Dutch, German, English, Swedish, Danish)

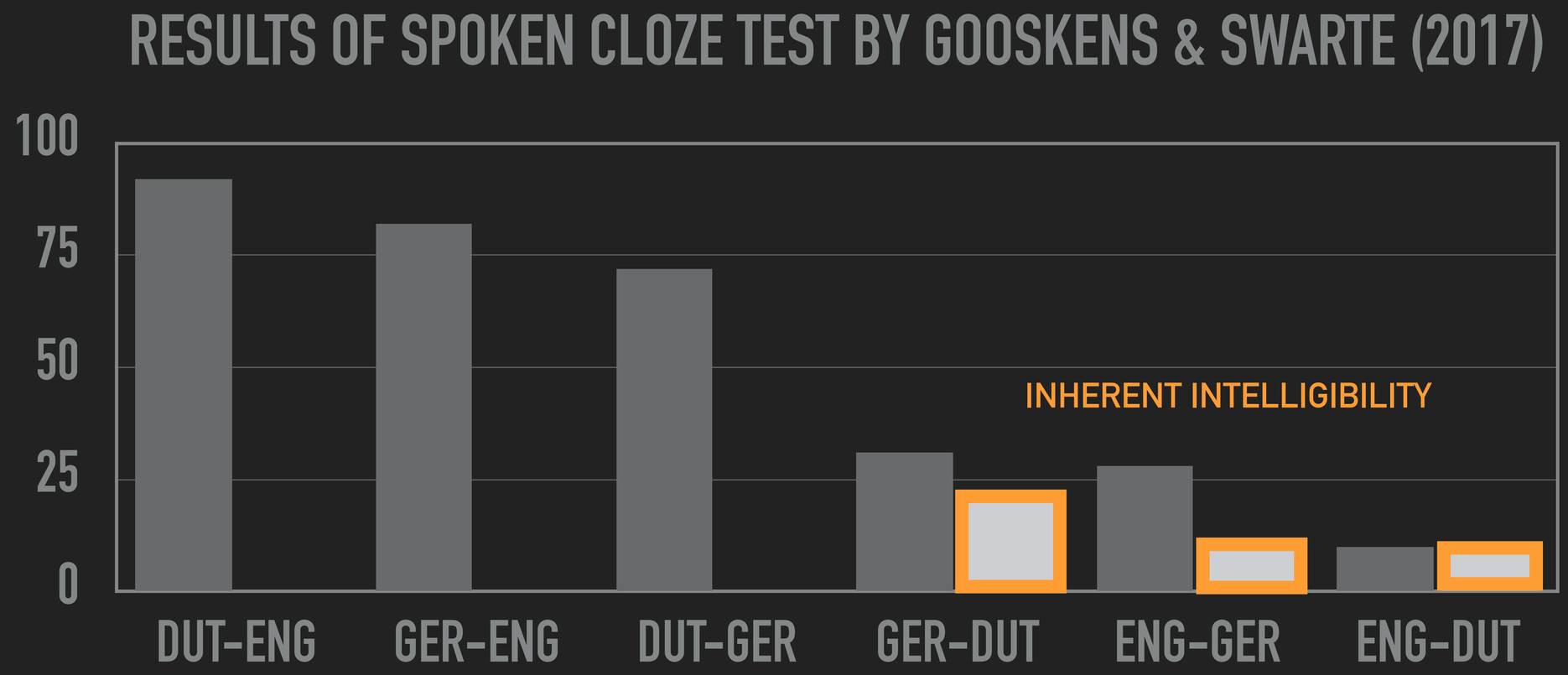
# PREVIOUS EXPERIMENTAL WORK

- Research on mutual intelligibility involves experimental studies with human participants
- Gooskens and Swarte (2017): large-scale study on mutual intelligibility of Germanic languages using a spoken and written cloze test and language background questionnaires
- Total of 954 participants with 5 different native languages (Dutch, German, English, Swedish, Danish)

# PREVIOUS EXPERIMENTAL WORK



# PREVIOUS EXPERIMENTAL WORK



# PREVIOUS EXPERIMENTAL WORK

- Previous language exposure strongest factor for mutual intelligibility scores
- When focusing on minimum of exposure (i.e. inherent intelligibility), lexical distances and orthographic distances are the most important factors

# THIS STUDY

- A computer-assisted method to assess mutual intelligibility in Germanic languages (Dutch, German, English)

# THIS STUDY

- A computer-assisted method to assess mutual intelligibility in Germanic languages (Dutch, German, English)
  - Uniform method that can be adapted to various languages
  - Testing human participants is a time- and resource-consuming effort
  - Finding participants with no exposure to another language is almost impossible
  - Testing linguistic factors only

# THE COMPUTATIONAL FRAMEWORK

- For testing mutual intelligibility computationally we need a model that is

# THE COMPUTATIONAL FRAMEWORK

- For testing mutual intelligibility computationally we need a model that is

  1) able to model word comprehension

# THE COMPUTATIONAL FRAMEWORK

- For testing mutual intelligibility computationally we need a model that is

  1) able to model word comprehension

  2) has a cognitively valid approach

# THE COMPUTATIONAL FRAMEWORK

- For testing mutual intelligibility computationally we need a model that is

  1) able to model word comprehension

  2) has a cognitively valid approach

## LINEAR DISCRIMINATIVE LEARNING

# THE COMPUTATIONAL FRAMEWORK

- For testing mutual intelligibility computationally we need a model that is

  1) able to model word comprehension

  2) has a cognitively valid approach

## LINEAR DISCRIMINATIVE LEARNING

- Based on Discriminative Lexicon framework by Baayen et al. (2019)
- Model of language processing exploring cognitive mapping mechanisms involved in language learning
- Provides method to model word comprehension

# THE COMPUTATIONAL FRAMEWORK

- Word comprehension in LDL: mapping of form onto meaning

# THE COMPUTATIONAL FRAMEWORK

- Word comprehension in LDL: mapping of form onto meaning

#D_DO_OG_G#

Animate, four legs, fur, floppy ears…

{0.1, 0.2, 0.11, 0.45…}

# THE COMPUTATIONAL FRAMEWORK

- Word comprehension in LDL: mapping of form onto meaning
- Implemented as multivariate regression using phonological form matrix **C** and semantic matrix **S**
- Association weight between cues are specified in training, model predicts semantic vector during testing
- Predicted vector used for comprehension accuracy = measures how well a form is understood

# PHONOLOGICAL CUES

- Cognate sets derived from Kluge (2002): total of 340 word forms in German with reflexes in Dutch and English
- Added phonetic transcriptions and phonetic alignments using EDICTOR (List, 2021)
- Reduced phonetic detail using Dolgopolsky sound classes (Dolgopolsky, 1986)
- two different representations of word forms: full forms and trimmed forms (=bare stems; Blum & List, 2023)

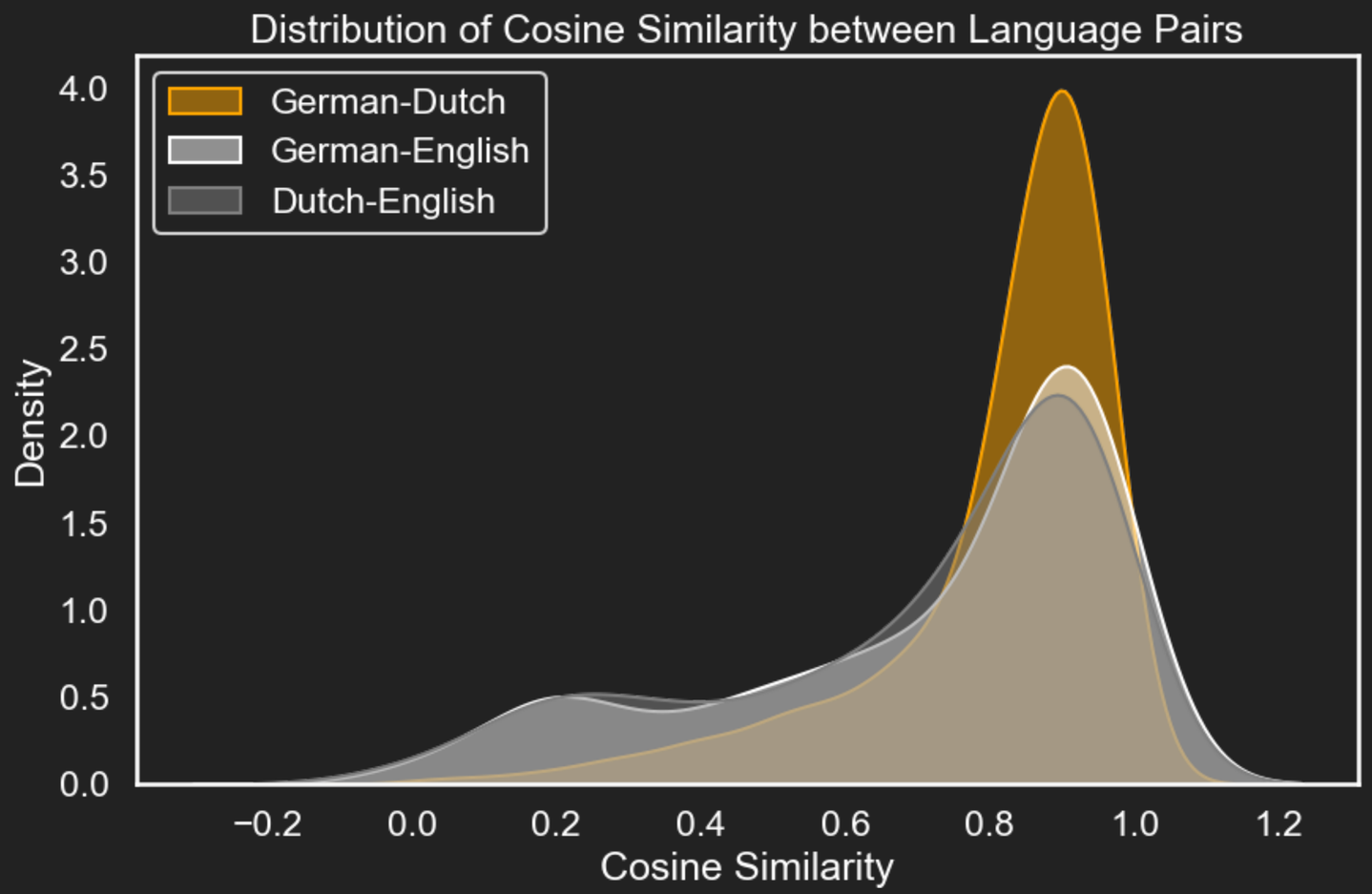
# PHONOLOGICAL CUES

- Cognate sets derived from Kluge (2002): total of 340 word forms in German with reflexes in Dutch and English, **e.g. drink vs. trinken**
- Added phonetic transcriptions and phonetic alignments using EDICTOR (List, 2021), **e.g. driŋk vs. triŋkən**
- Reduced phonetic detail using Dolgopolsky sound classes (Dolgopolsky, 1986), **e.g. T R V N K vs. T R V N K V N**
- Two different representations of word forms: full forms and trimmed forms (=bare stems; Blum & List, 2023) , **e.g. T R V N K vs. T R V N K V**

# MEANING REPRESENTATIONS

- As meaning representation we used the multilingual ConceptNet Numberbatch word embeddings version 19.08 from Speer et al. (2017)

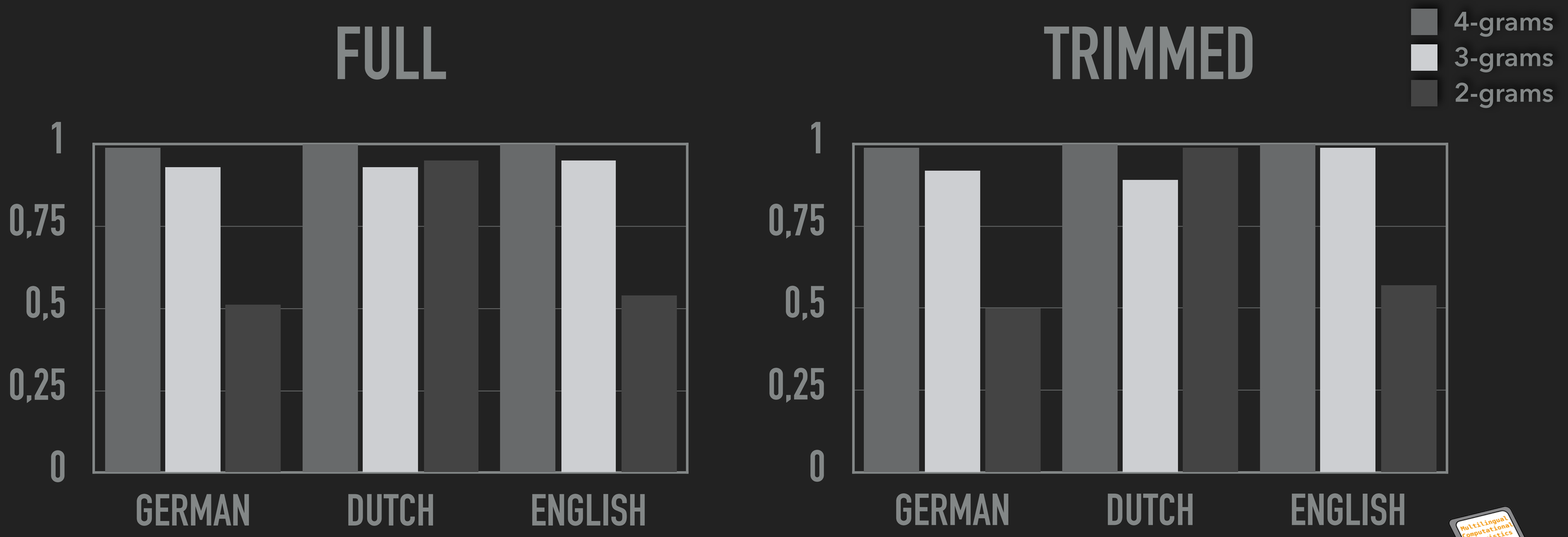
# MEANING REPRESENTATIONS – COSINE SIMILARITY



# LINEAR DISCRIMINATIVE LEARNING MODEL

- Step 1:
  - Evaluation on cognate data of individual languages
  - 339 forms in total due to a missing form in Dutch
  - 4-gram, 3-gram and 2-gram chunks of sound classes
  - Full word forms vs. Trimmed forms

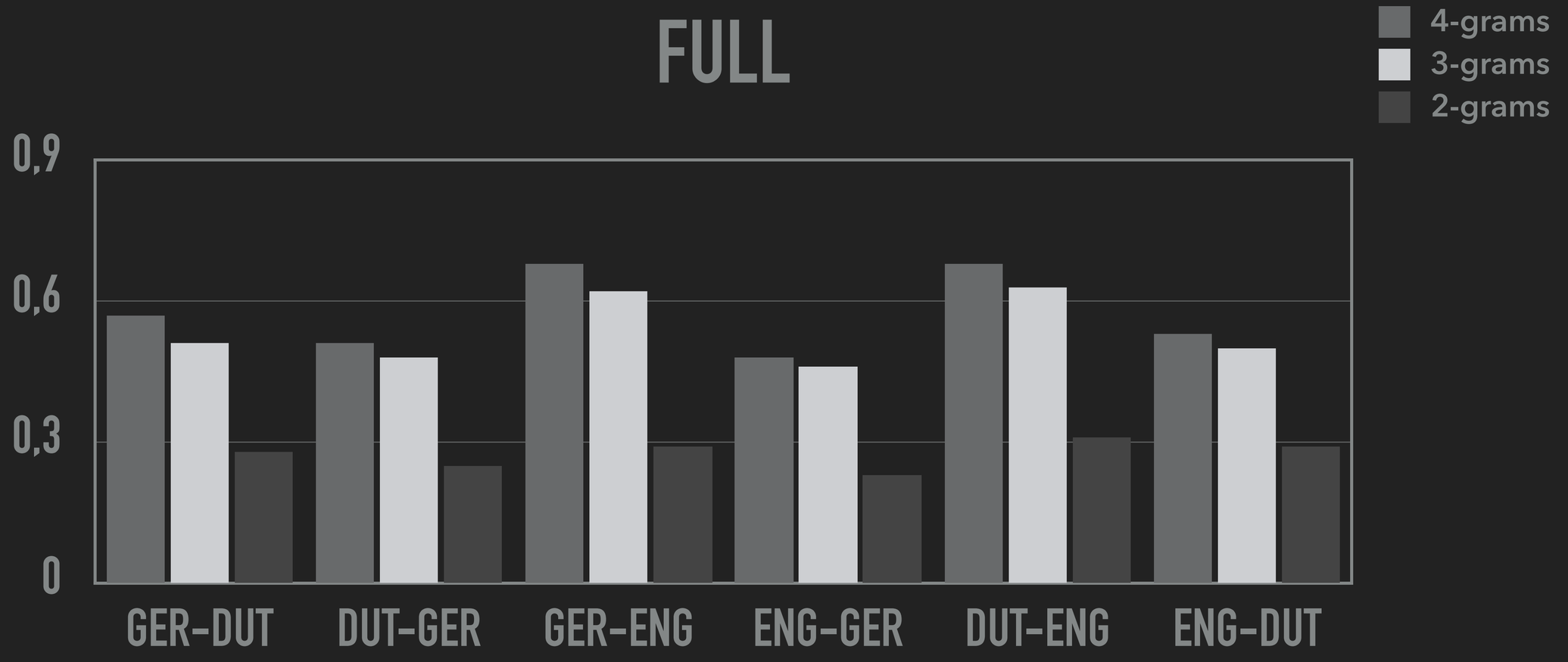
# RESULTS


FULL
TRIMMED
4-grams
3-grams
2-grams
1
0,75
0,5
0,25
0
GERMAN
DUTCH
ENGLISH
1
0,75
0,5
0,25
0
GERMAN
DUTCH
ENGLISH

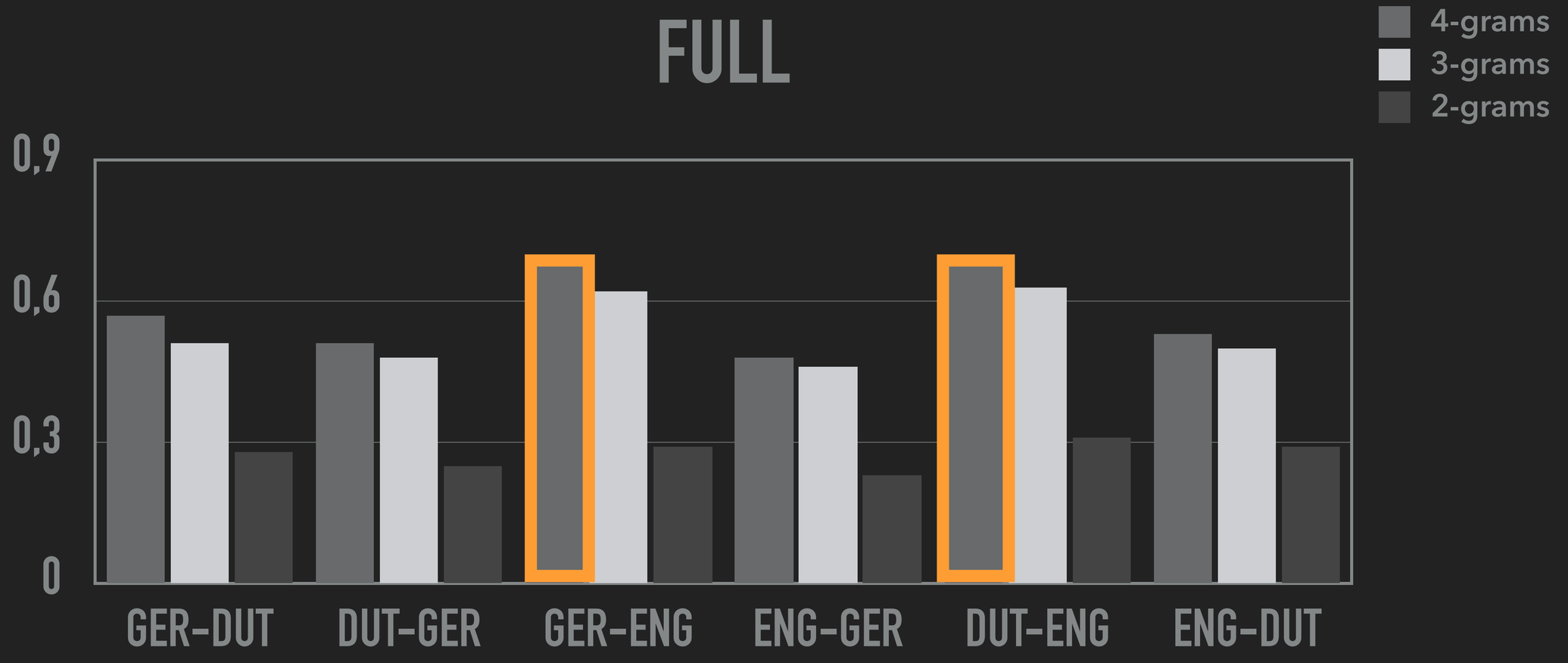
# LINEAR DISCRIMINATIVE LEARNING MODEL

- Step 2:
  - Evaluation on cognate data across languages
  - 339 forms in total due to a missing form in Dutch
  - 4-gram, 3-gram and 2-gram chunks of sound classes
  - Full word forms vs. Trimmed forms

# RESULTS

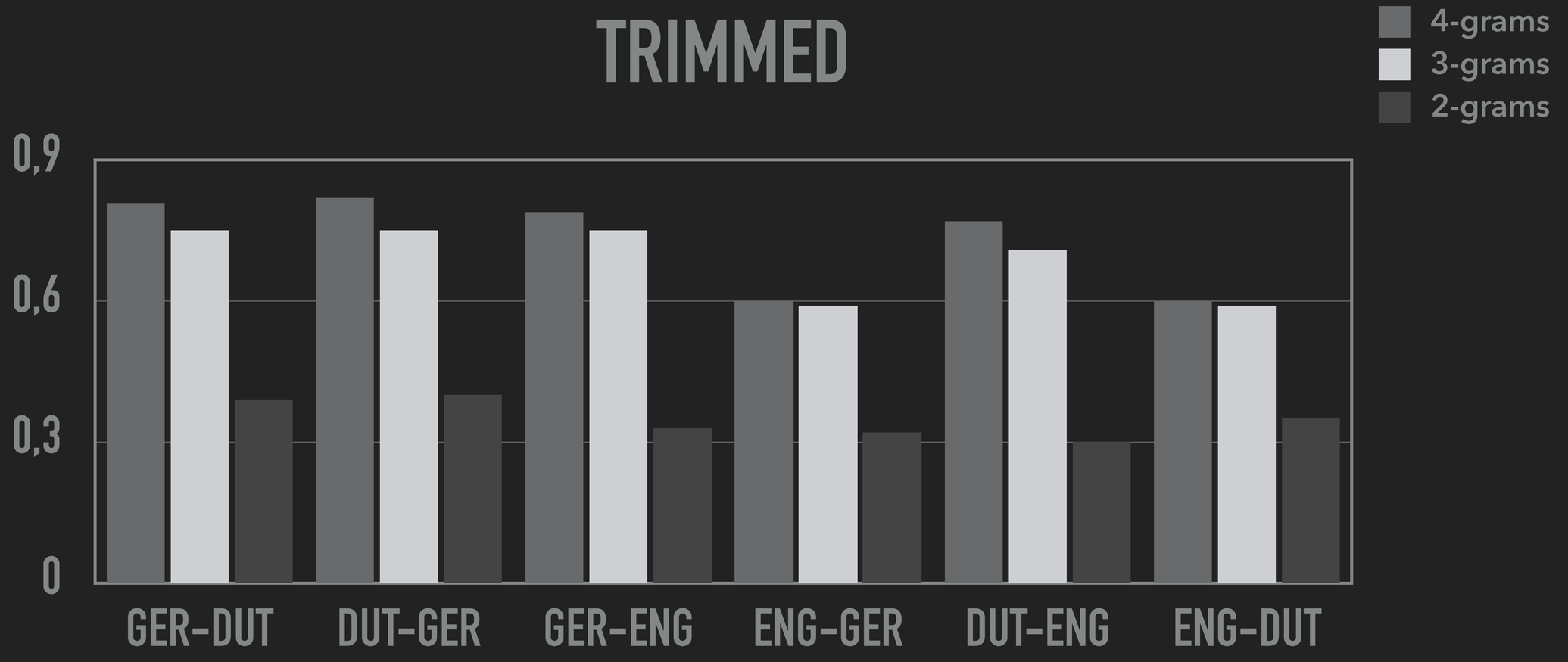

FULL
4-grams
3-grams
2-grams
0,9
0,6
0,3
0
GER-DUT
DUT-GER
GER-ENG
ENG-GER
DUT-ENG
ENG-DUT

# RESULTS

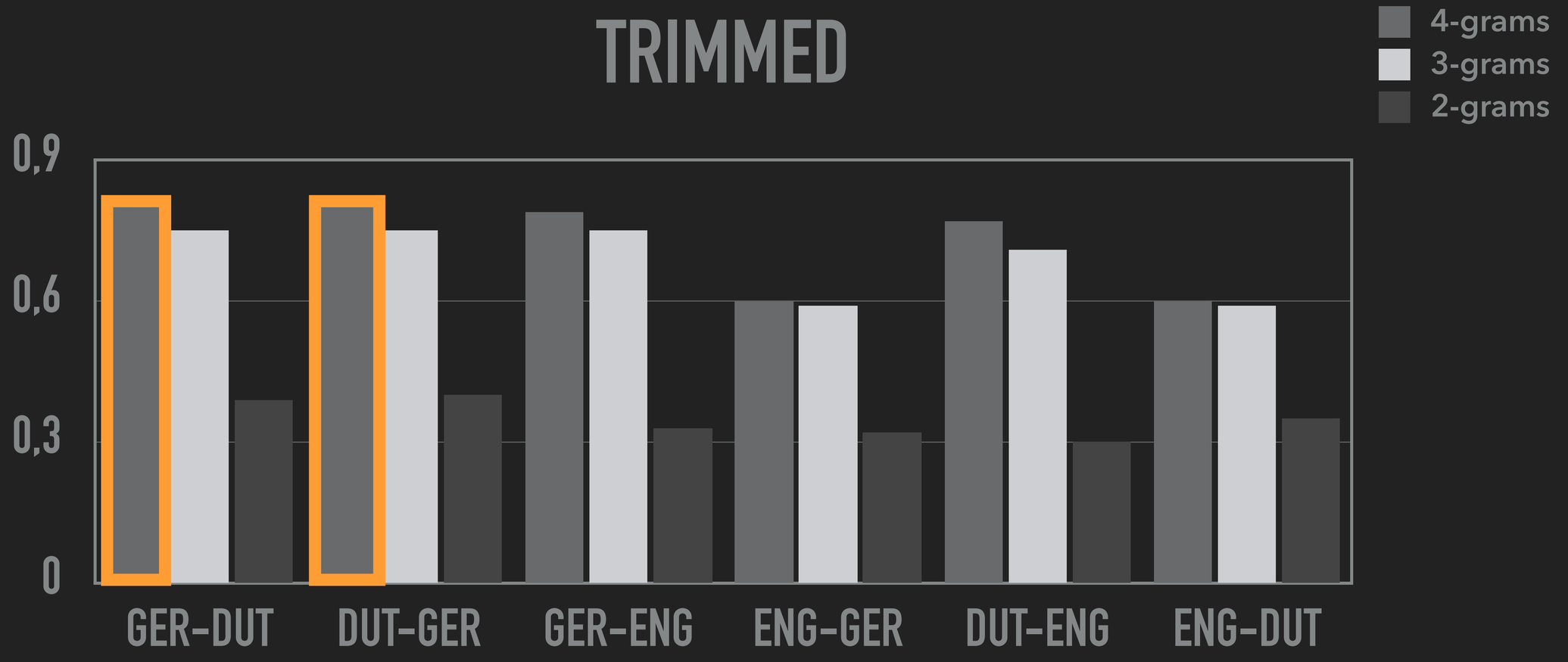



IN LINE WITH GOOSKENS & SWARTE (2017)

# RESULTS


TRIMMED
4-grams
3-grams
2-grams
0,9
0,6
0,3
0
GER-DUT
DUT-GER
GER-ENG
ENG-GER
DUT-ENG
ENG-DUT

# RESULTS


TRIMMED
4-grams
3-grams
2-grams
0,9
0,6
0,3
0
GER-DUT
DUT-GER
GER-ENG
ENG-GER
DUT-ENG
ENG-DUT

# DISCUSSION

- Best results with 4-grams and 3-grams
- Better results for trimmed than for full words
- Full word forms: best results Dutch-English, in line with Gooskens and Swarte (2017)
- Trimmed word forms: best results for Dutch-German
- English least advantageous native language in our and Gooskens and Swarte (2017)'s setting

# DISCUSSION

- Higher accuracy for German-English than German-Dutch (in line with Gooskens and Swarte, 2017) for full forms but opposite effect for trimmed
- Dutch-English better than Dutch-German for full forms but again opposite picture for the trimmed version

# DISCUSSION

- Higher accuracy for German-English than German-Dutch (in line with Gooskens and Swarte, 2017) for full forms but opposite effect for trimmed
- Dutch-English better than Dutch-German for full forms but again opposite picture for the trimmed version

**MORPHOLOGICAL KNOWLEDGE AFFECTS COMPREHENSION**

# DISCUSSION

- We present a computational approach to test mutual intelligibility across languages using LDL
- Our data shows similarities to human comprehension results, making it a useful tool to assess mutual intelligibility

# LIMITATIONS

- We tested 3 Germanic languages only, this needs to be extended to other languages and language families
- We tested cognate data only, this needs to be extended to non-cognate data

# THANK YOU, GRAZZI ĦAFNA, DANKE AND DANK JE WEL!

# REFERENCES

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, 2019. https://doi.org/10.1155/2019/4895891

Blum, F., & List, J.-M. (2023). Trimming phonetic alignments improves the inference of sound correspondence patterns from multilingual wordlists. Proceedings of the 5th Workshop on Computational Typology and Multilingual NLP, 52–64. https://aclanthology.org/2023.sigtyp-1.6.pdf

Dolgopolsky, A. B. (1986). A probabilistic hypothesis concerning the oldest relationships among the language families of Northern Eurasia. In V. V. Shevoroshkin (Ed. & Trans.), Typology, Relationship and Time: A collection of papers on language change and relationship by Soviet linguists (V. V. Shevoroshkin, Ed. & Trans.; pp. 27–50). Karoma Publisher.

Gooskens, C., & Swarte, F. (2017). Linguistic and extra-linguistic predictors of mutual intelligibility between Germanic languages. Nordic Journal of Linguistics, 40(2), 123–147. https://doi.org/10.1017/S0332586517000099

Kluge, F. (2002). Etymologisches Wörterbuch der deutschen Sprache (24th ed.). de Gruyter.

List, J.-M. (2021). Edictor. a web-based tool for creating, editing, and publishing etymological datasets. Max Planck Institute for Evolutionary Anthropology. https://doi.org/https://doi.org/10.5281/zenodo.4685130

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. Proceedings of the AAAI Conference on Artificial Intelligence 2017, 4444–4451. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972