CALICO 2018 – Connecting CALL's Past to its Future
University of Illinois, Urbana-Champaign, May 29 – June 2

@fcornillie

# Automating elicited imitation for spoken practice in German L2: task design, speech recognition, and language models

Frederik Cornillie (University of Leuven & imec)
Dirk De Hertog (University of Leuven & imec)
Kristof Baten (Ghent University)

KU LEUVEN

GHENT UNIVERSITY

imec

# Spoken practice: what and why ?

spoken activities in a L2
that focus on <u>specific linguistic constructions</u>
and that involve a considerable amount of
<u>recycling, feedback</u>, and often <u>time pressure</u>,
with the goal of developing <u>explicit knowledge</u>
about these constructions
<u>as well as skills</u> in the L2

All you need is input

vs.

Output practice and feedback can aid noticing and automatization

the Krashen school

the interactionist school

# The relative effects of input and output practice

- Inconsistent findings:
  - Effects on comprehension:
    - Input practice ~ output practice (Morgan-Short & Bowden, 2006; Nagata, 1998; Salaberry, 1997; Toth, 2006)
    - Input practice > output practice (Benati, 2001; 2005; DeKeyser & Sokalski, 1996)
  - Effects on production:
    - Input practice ~ output practice (Benati; 2001; 2005)
    - Output practice > input practice (Dekeyser & Sokalski, 1996; Morgan-Short & Bowden, 2006; Nagata, 1998; Toth, 2006)
- Limitations:
  - (very) short treatments (1-6 hours) over short periods of time (1-7 days)
  - Only accuracy rates considered

→ No evidence of relative effects on automatization: transfer to communicative tasks?

# CALL to the rescue ? (a call from the past)

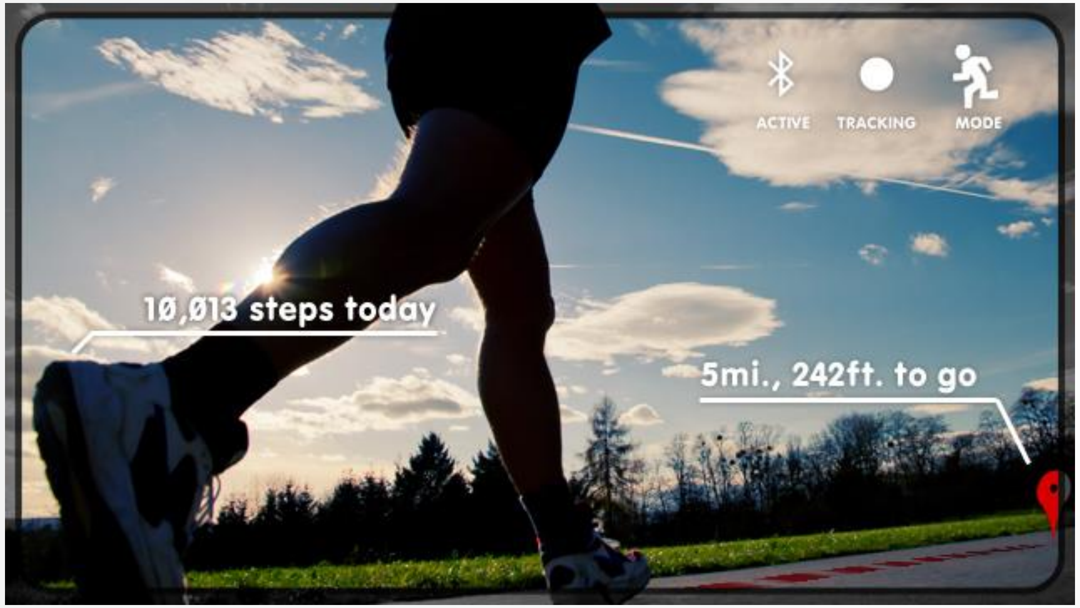
Research on practice [must be] very fine-grained, to allow for tracking of stimuli and responses in milliseconds […] while being longitudinal in nature […]
Perhaps new technology can solve this problem by allowing for massive data collection *and* sophisticated analysis at the fine-grained level *and* longitudinally, from many learners, without losing sight of the importance of individual differences.

## Robert DeKeyser

*Practice in a Second Language. Perspectives from Applied Linguistics and Cognitive Psychology* (2007)

# Data collection today

## in everyday apps




- longitudinal and massive
- uncontrolled environments
- updated and analyzed continuously
- valorized (e.g. for personalization)

## in SLA research

- typically no longer than a couple of weeks
- controlled environments
- write once, analyze once
- typically not valorized in learning environments

# But … big data is gaining traction in CALL

Reply | Reply List | Forward | Archive | Junk | Delete | More

From Stephen Fancsali <sfancsali@gmail.com>
Subject **[edm-announce] Jobs @ Duolingo: Multiple R&D Roles** 9/05/2017 22:28
To edm-announce@freelists.org

[on behalf of Burr Settles]

Duolingo is hiring several research and data scientist positions. We already have a strong group of interdisciplinary scholars focused on novel applied research at the intersection of machine learning, computational linguistics, and cognitive science, and we are expanding quickly: https://www.duolingo.com

Duolingo is the world's largest online language learning service with more than 150 million users worldwide. Our apps have been selected as Apple's iPhone App of the Year and Google's Best of the Best for Android multiple times. Our scientists split their time between primary research on new and unique problems (publication is encouraged), and translating these research findings into production systems that improve learning and engagement outcomes for millions.

Candidates interested in user modeling, educational data mining, applied machine
learning, NLP and/or speech processing should apply!

See the full advertisements for two positions below.

Cheers,



## 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM)

This challenge is in conjunction with the 13th BEA Workshop and NAACL-HLT 2018 conference.

## Introduction

As educational apps increase in popularity, vast amounts of student learning data become available, which can and should be used to drive personalized instruction. While there have been some recent advances in domains like mathematics, modeling **second language acquisition (SLA)** is more nuanced, involving the interaction of lexical knowledge, morpho-syntactic processing, and other skills. Furthermore, most work in NLP for second language (L2) learners has focused on intermediate-to-advanced students of English in assessment settings. Much less work has been done involving beginners, learners of languages other than English, or study over time.

This task aims to forge new territory by utilizing student trace data from users of Duolingo, the world's most popular online language-learning platform. Participating teams are provided with transcripts from millions of exercises completed by thousands of students over their first 30 days of learning on Duolingo. These transcripts are annotated for token (word) level mistakes, and the task is to predict what mistakes each learner will make in the future.

Novel and interesting research opportunities in this task:

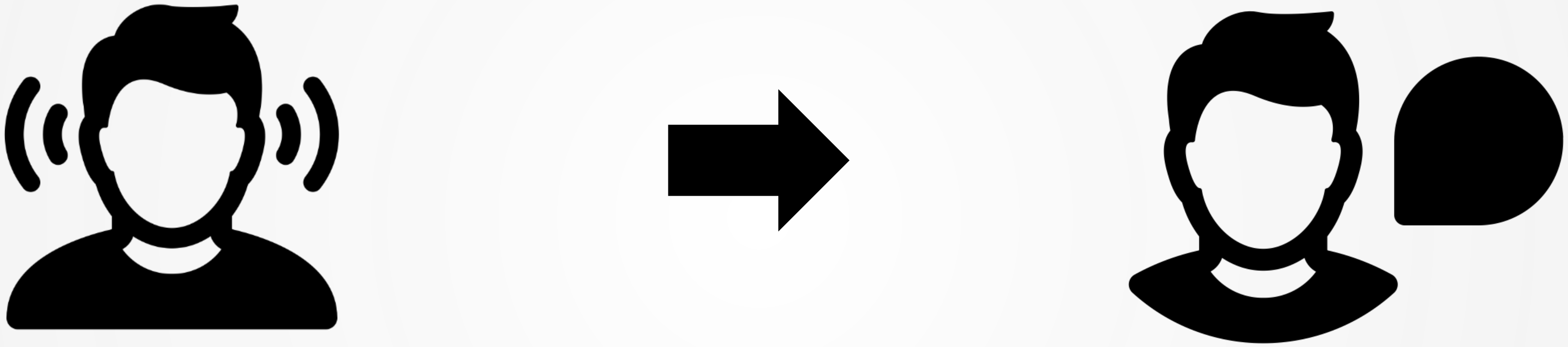
- There will be three (3) tracks for learners of **English**, **Spanish**, and **French**. Teams are encouraged to explore features which generalize across all three languages.
- Anonymized **user IDs** and **time data** will be provided. This allows teams to explore various personalized, adaptive SLA modeling approaches.
- The sequential nature of the data also allows teams to model language **learning** (and **forgetting**) over time.

# ORAL ELICITED IMITATION

# Oral elicited imitation: the basic task

stimulus

response



relatively short
and simple sentences

repeat
as exactly as possible

# Oral elicited imitation: cognitive processes

## stimulus

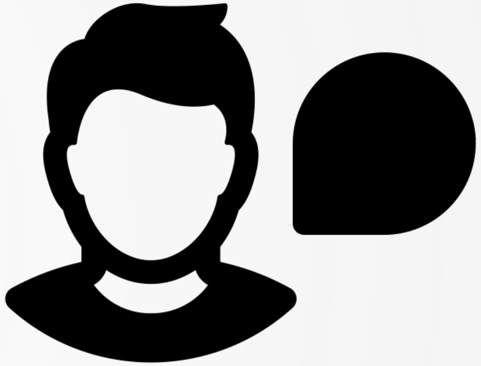
relatively short
and simple sentences

(target-language-like
or deviating)

SEMANTIC PROCESSING
→ erases memory of the form
(Erlam, 2006)

SYNTACTIC PROCESSING

## response



repeat
<u>and reconstruct</u>

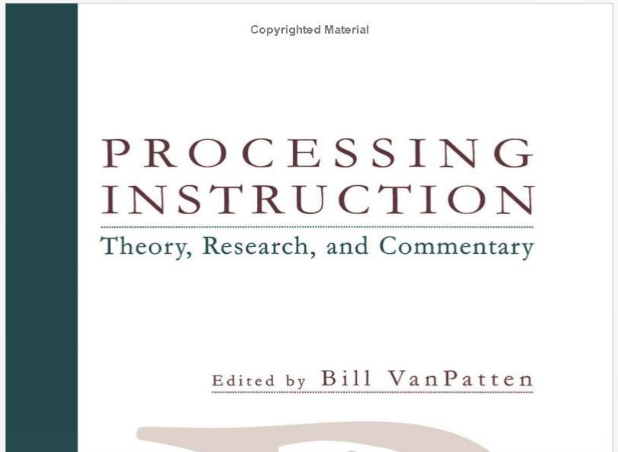
→ insight in the learner's
interlanguage system

# Oral elicited imitation in L2 assessment

- OEI can measure
  - oral proficiency (Tracy-Ventura, McManus, Norris, & Ortega, 2014)
  - implicit knowledge (e.g. Erlam, 2009)
  - automatized explicit knowledge (Suzuki & DeKeyser, 2015)
- The assessment task can be automated with speech recognition
  - (Cook, Mcghee, & Lonsdale, 2011; Graham, Lonsdale, Kennington, Johnson, & McGhee, 2008)

# Oral elicited imitation for output practice: issues for CALL

meaningful language processing
or mechanical parroting?




corrective feedback
in order to stimulate noticing

speech recognition technology
& language models for error diagnosis

# EMPIRICAL STUDY ON GERMAN L2

# The current study

Goal prepare task design, materials and technology
for a study on the relative effects of output practice in German L2

Research questions:

1. Does the design of the OEI task focus learners' attention on meaning?
   → task design
2. How accurately does state-of-the-art speech recognition transcribe the learners' production?
   → speech recognition
3. What was the nature of linguistic variation in the learners' production?
   → language models

# Materials: target constructions

stimulus

48 sentences
case marking and word order
in German L2

length 5-8 words
high-frequency vocabulary

- transitives – e.g. [The dog chases the man]
  *Der Hund verfolgt den Mann.*
  **Der Hund verfolgt der Mann.*
  *Den Mann verfolgt der Hund.*
  **Der Mann verfolgt der Hund.*
  topicalization

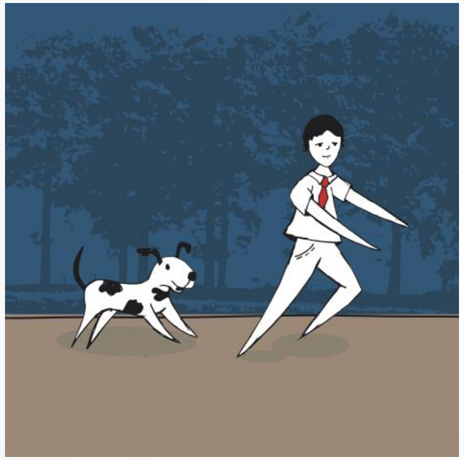
- ditransitives – e.g. [The teacher gives the headmaster flowers]
  *Die Lehrerin schenkt dem Direktor die Blumen.*
  **Die Lehrerin schenkt der Direktor die Blumen.*
  *Dem Direktor schenkt die Lehrerin die Blumen.*
  **Der Direktor schenkt die Lehrerin die Blumen.*
  topicalization

- prepositional phrases – e.g. [The man walks through NP]
  *Der Mann spaziert durch den Tunnel.*
  **Der Mann spaziert durch der Park.*

# Materials: task design

stimulus

picture matching response



spoken response

*Den Mann verfolgt der Hund.*
[The dog chases the man]

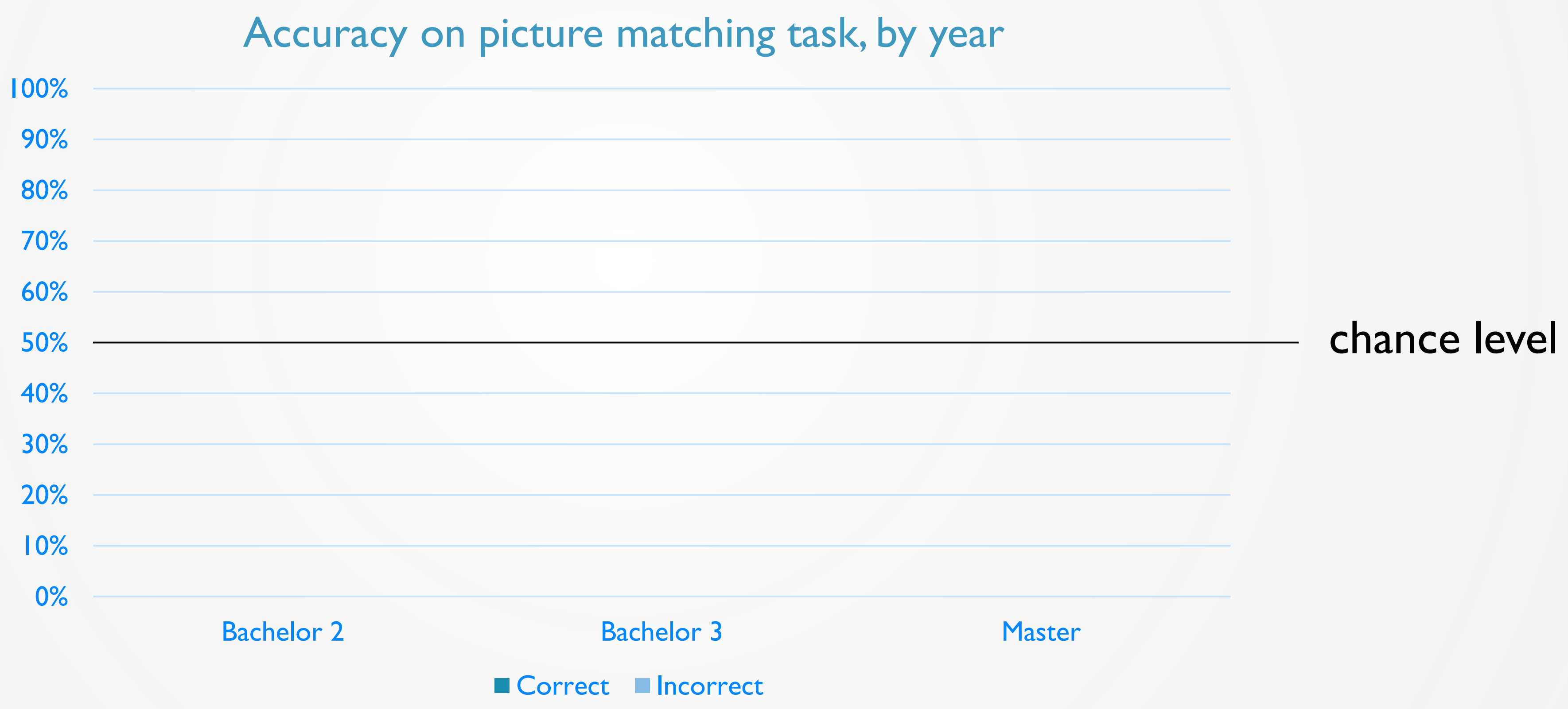
instruction:
“repeat
in as good German
as possible”

# Participants & data

- participants:
  - Flemish learners of German L2 (*N* = 36)
  - academic programme in Languages and Literature, Ghent University
    - 2nd bachelor (*N*=11)
    - 3rd bachelor (*N*=10)
    - master (*N*=15)
  - 18-23 years old
- data:
  - collected online (item order counterbalanced), using headsets
  - total of 1728 learner-item interactions:
    - 1728 picture-matching responses
    - 1487 spoken responses manually transcribed

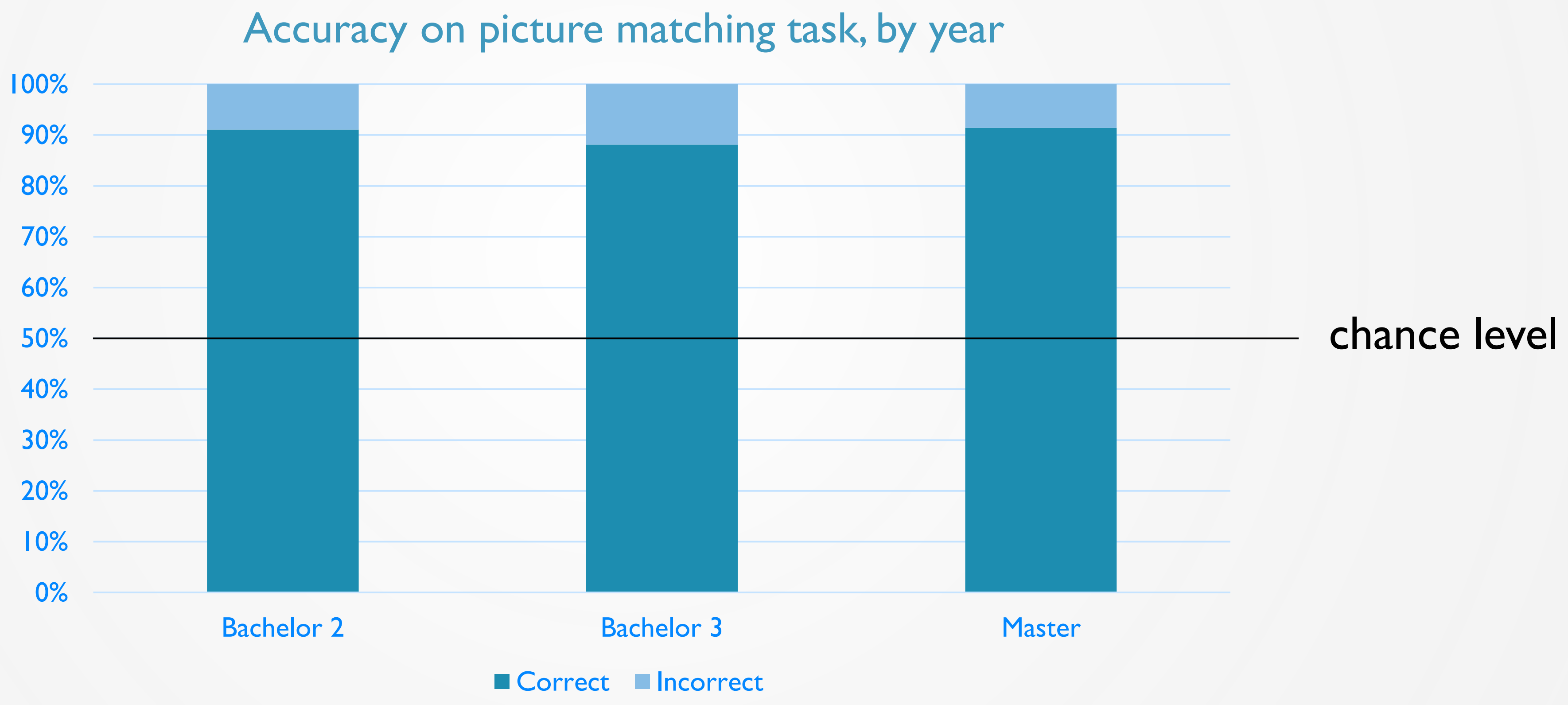
# Results for task design

Does the design of the task focus learners' attention on meaning?



# Results for task design

Does the design of the task focus learners' attention on meaning?




difference between groups: $F(2, 33) = 0.88$, $p = 0.42$

# Results for task design

Does the design of the task focus learners' attention on meaning?

Grammatical accuracy of production (correct picture matching responses only)

| | $N$ | Min | Max | Mean | $SD$ |
|---|---|---|---|---|---|
| Grammatical stimuli | 36 | 0.87 | 1 | 0.986 | .028 |
| Ungrammatical stimuli | 36 | 0.208 | 1 | 0.716 | .199 |

$r = 0.62, p < 0.001, N = 36$

→ reconstructive

# Results for speech recognition

**Tools**

Google Cloud

- easy API
- black box
- pay for what you use

CMUSphinx

- more tricky to set up
- open source
- pay for a server

**Implementations**

Google Cloud:

- out of the box

CMUSphinx:

- out of the box
- acoustic model
- language model
- language model & acoustic model

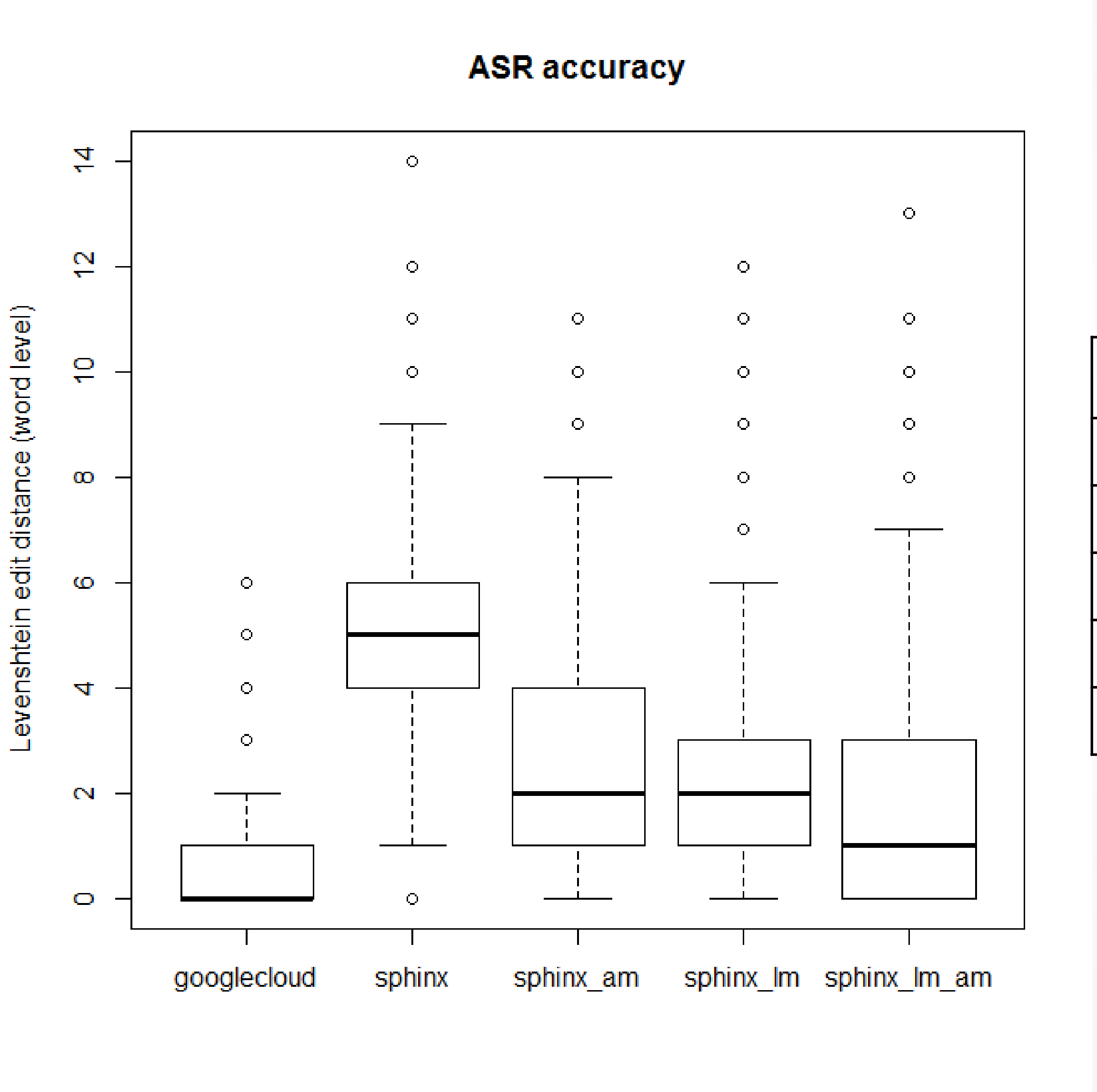
**Evaluation metric**

Levenshtein edit distance (word level)

| | | | | | | |
|---|---|---|---|---|---|---|
| den | Direktor | schimpfe | | Lehrerin | die | Blumen |
| den | Direktor | schenkt | die | Lehrerin | den | Blumen |

→ 3

# Results for speech recognition




| | Min | Max | Mean | Median | N |
|---|---|---|---|---|---|
| Google | 0 | 6 | 0.55 | 0 | 1487 |
| Sphinx | 0 | 14 | 4.70 | 5 | 1412 |
| Sphinx AM | 0 | 11 | 2.48 | 2 | 1410 |
| Sphinx LM | 0 | 12 | 2.23 | 2 | 1413 |
| Sphinx LM+AM | 0 | 13 | 1.87 | 1 | 1413 |

# Results for speech recognition

Google Cloud

Some other relevant findings:

- no error correction

  _der Vater zeigt *[den Sohn] die Brille_
  _der Mann ist gegen *[dem dem Baum] gefahren_
  _der Junge geht *[zu Bäcker]_
  _die Lehrerin schenkt dem Direktor *[den Blumen] die Blumen_

- possible quick win: improve recognition by prioritizing key vocabulary in the language model

  _der Polizist sucht den Becher (< Bäcker)_

  _die Lehrerin schenkt den Jagd aber (< Direktor) die Blumen_

# Results for language models (work in progress)

What was the nature of linguistic variation in the learners' production?

- Linguistic variation
  - Semantic — *Der Mann ist gegen den Baum gefallen (< gefahren)*
  - Morphological — **Die Lehrerin schenkt *den (< dem) Direktor den Blumen*
  - Syntactic — *Die Lehrerin schenkt dem Direktor die Blumen*<br>*< Dem Direktor schenkt die Lehrerin die Blumen*
  - Combinations — *Der Vater schenkt der Junge den Junge die Brille*<br>*< Dem Sohn zeigt der Vater die Brille*
- Variation due to cognitive processes
  - Self-correction — *Das Mädchen kommt aus der Shop - dem Shop*
  - Disfluencies — *Der Doktor verklauf verkauft dem Clown das Buch*
  - Multiple repetitions — *Die Frau gibt den Mann den Apfel. Die Frau gibt dem Mann den Apfel.*

# Discussion and next steps

- OEI as implemented in this study has potential as a practice task
  - Picture matching simulated meaningful language processing
  - Google Cloud speech API handled non-native German speech relatively well
- Limitations:
  - Advanced students > role of working memory?
  - Controlled setting
  - Meaning-focus could be stronger
  - Google Cloud Speech API is a black box
- Next steps:
  - Develop language models for error correction
  - Increase the meaning-focus of the task, e.g. individual sentences form a coherent story

# The future of research on CALL practice ?

open data

open tools and technologies

real collaboration academics - industry

# Thank You !

@fcornillie


## Acknowledgements

- German native speaker stimuli recorded by Carola Strobl
- Drawings created by Fridl Cuvelier
- Data collected by Wouter Vanacker
- Icons created by Gregor Cresnar and Oksana Latysheva from Noun Project



Frederik Cornillie (University of Leuven & imec)

Dirk De Hertog (University of Leuven & imec)

Kristof Baten (Ghent University)