

Systemes de dialogue pour l'apprentissage des langues : typologie des systemes et mesure des effets

Serge Bibauw

CENTAL, UCLouvain

ITEC, imec research group at **KU Leuven**

Universidad Central del Ecuador

Séminaire du CENTAL

Louvain-la-Neuve, 21 novembre 2019

KU LEUVEN

UCLouvain



Dialogue systems for language learning: typology of systems and measurement of effects

Dialogue systems for language learning

Terms, fields and definition

Rationale

Typology of systems

Types of dialogue-based CALL systems

Technological approaches in research and industry

Past effectiveness

Meta-analysis of previous effectiveness studies

Evaluation of *LanguageHero*

Measuring effects on L2 development

Challenges and opportunities



Dialogue systems for language learning: typology of systems and measurement of effects



► Dialogue systems for language learning

Terms, fields and definition

Rationale

Typology of systems

Types of dialogue-based CALL systems

Technological approaches in research and industry

Past effectiveness

Meta-analysis of previous effectiveness studies

Evaluation of *LanguageHero*

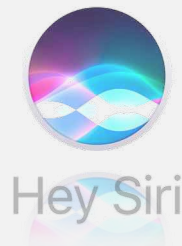
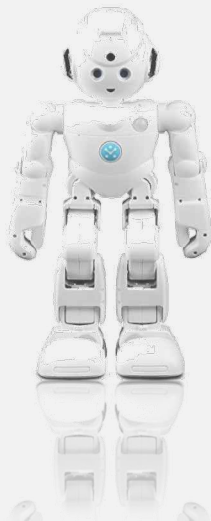
Measuring effects on L2 development

Challenges and opportunities

Dialogue systems for language learning

Language learning through **dialogues** with **automated** agents

(chatbot, talking robot, automated personal assistant,
conversational agent, non-player character in videogames...)



Dialogue systems for language learning

A dispersed and fragmented field

Studies scattered among different domains/traditions,
under many different terms:

intelligent tutoring systems, chatbots, conversational agents, spoken dialogue systems, virtual worlds, serious games, robot-assisted language learning (RALL), ASR-based CALL, computer-assisted pronunciation training (CAPT)...

Only partial literature reviews

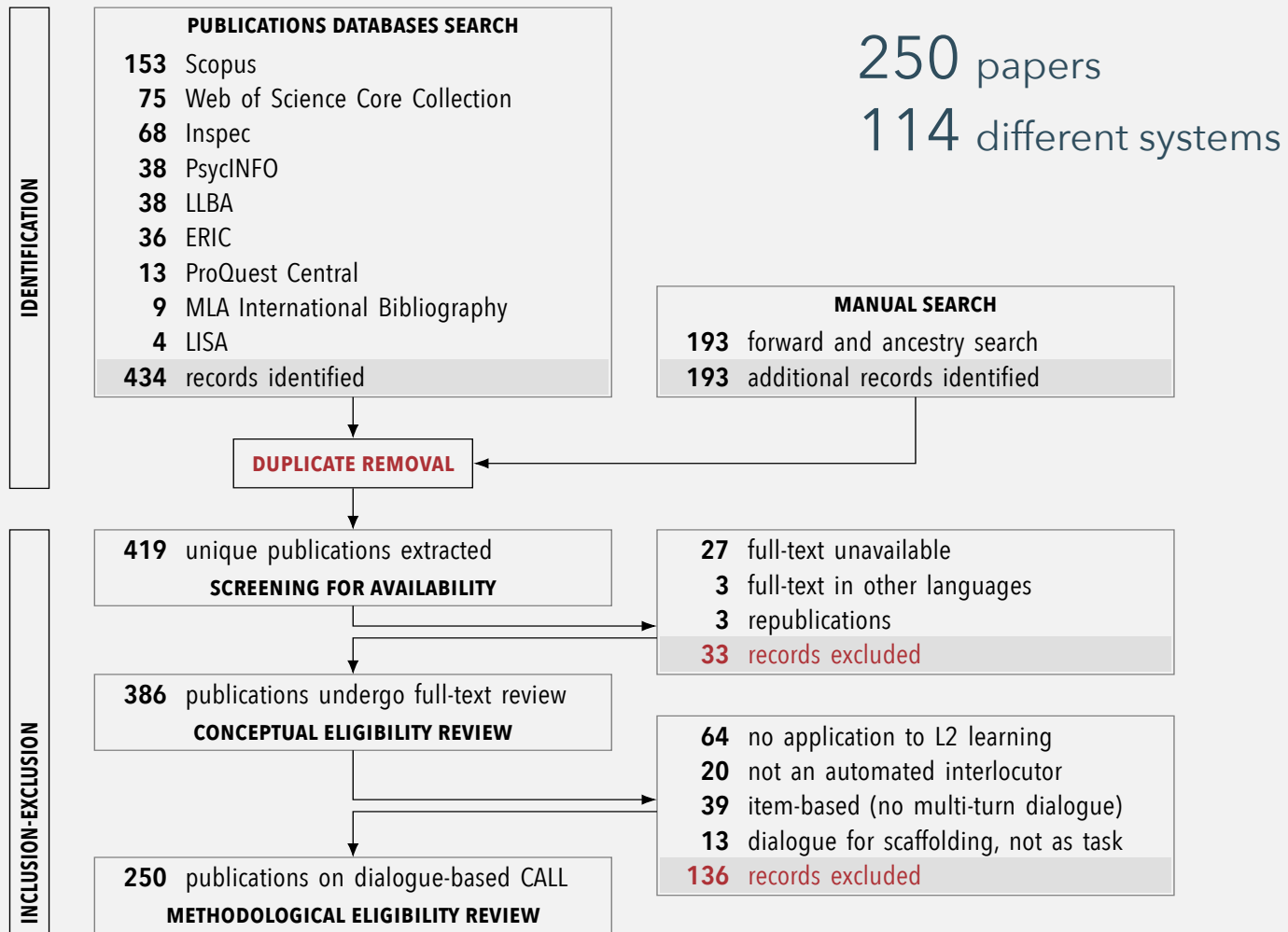
(Wachowicz & Scott, 1999; Eskenazi, 2009; Golonka et al, 2014)

→ Small clusters of research, low mutual awareness,
no established research community, short-lived projects

→ NLP-based efforts underestimate instructional challenges;
CALL-based efforts underestimate NLP challenges

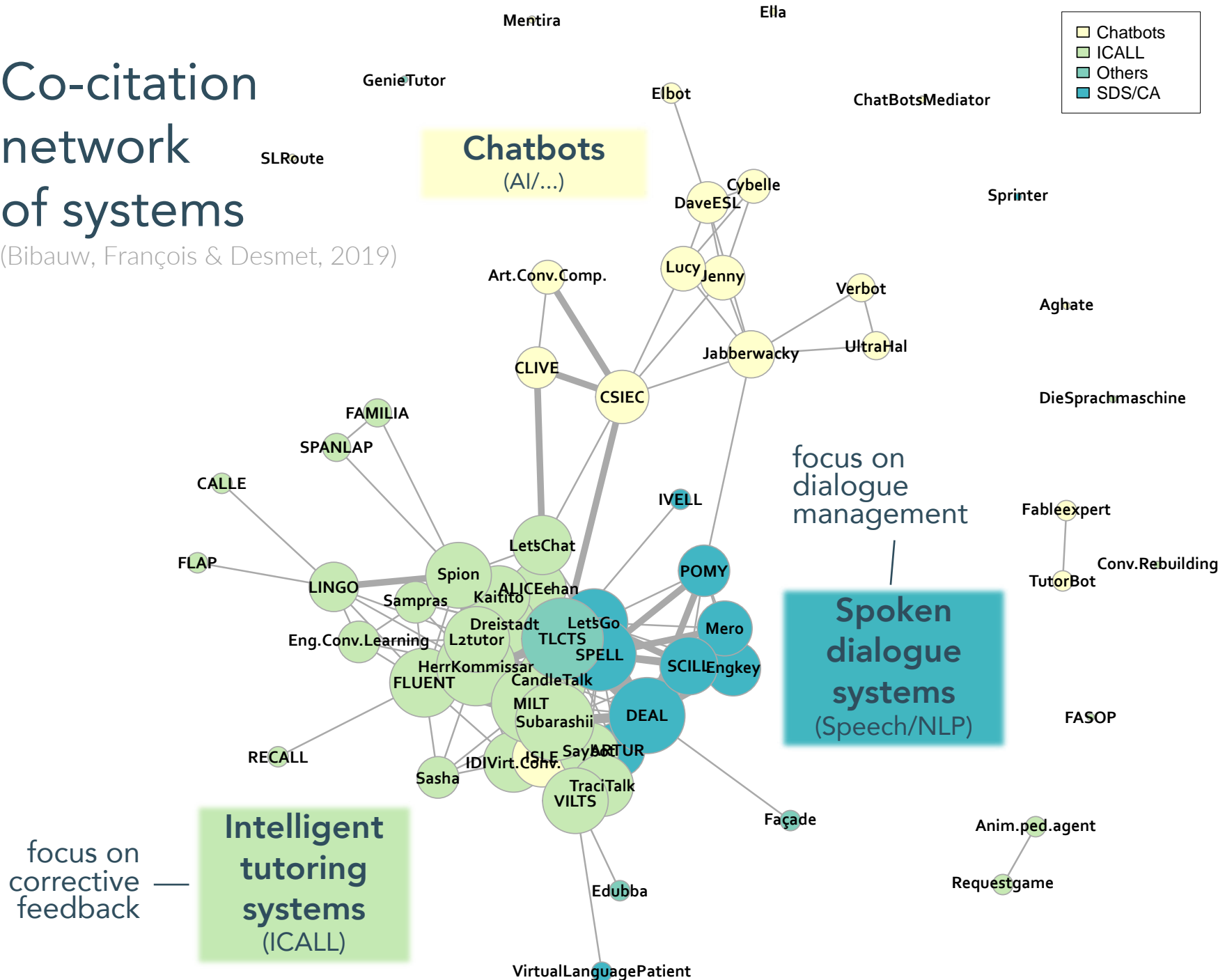
Dialogue systems for language learning

Research synthesis



Co-citation network of systems

(Bibauw, François & Desmet, 2019)



Dialogue systems for language learning

(Bibauw, François & Desmet, 2019)

Any application or system allowing

to maintain a **dialogue**

[immediate, synchronous interaction]

[written or spoken]

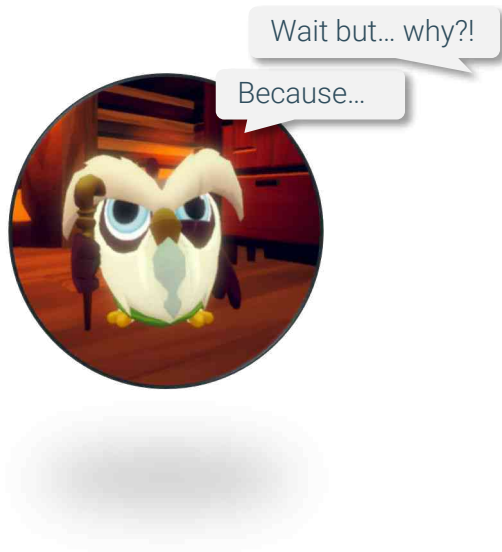
with an **automated agent**

[chatbot, talking robot, automated personal assistant, conversational agent, non-player character in a video game...]

[tutorial CALL (≠ computer-mediated communication)]

for **language learning** purposes.

Dialogue systems for language learning: typology of systems and measurement of effects



Dialogue systems for language learning

Terms, fields and definition

► Rationale

Typology of systems

Types of dialogue-based CALL systems

Technological approaches in research and industry

Past effectiveness

Meta-analysis of previous effectiveness studies

Evaluation of *LanguageHero*

Measuring effects on L2 development

Challenges and opportunities

Dialogue systems for language learning

Rationale (1)

10

Assumption: **meaningful practice** → L2 proficiency development

Many learning contexts: lack of occasions for meaningful L2 practice

Automated agents can compensate for the absence of human interlocutors

“Virtual immersion” (Ellis & Bogart, 2007)

Also in MOOCs and online learning contexts (Read, 2014)

Interactionist perspective to second language acquisition
(Long, 1996)

Negotiation of meaning (Pica, 2013), *pushed output* (Swain, 2005)

Visible transcript promotes *noticing* (Lai & Zhao, 2006)

Practice → **Proceduralisation** by automatizing (DeKeyser, 2007)

Dialogue systems for language learning

Rationale (2)

11

Some advantages over human interlocutors

Always available, ubiquitous

Endless patience, allowing for repetition (Fryer & Carpenter, 2006)

Low-anxiety environment → ↗ willingness to communicate (Ayedoun, Hayashi & Seta, 2015)

Fully controllable learning environment

Opportunities for fully monitored conditions for empirical research on interaction (Hegelheimer & Chapelle, 2000)

Dialogue systems for language learning: typology of systems and measurement of effects



Dialogue systems for language learning

Terms, fields and definition

Rationale

Typology of systems

- ▶ Types of dialogue-based CALL systems
- Technological approaches in research and industry

Past effectiveness

Meta-analysis of previous effectiveness studies

Evaluation of *LanguageHero*

Measuring effects on L2 development

Challenges and opportunities

Typology of systems (Bibauw, François & Desmet, 2019)

Continuum of constraints

Explicit

Constraints on meaning

Implicit



Form-focused systems

Goal-oriented systems



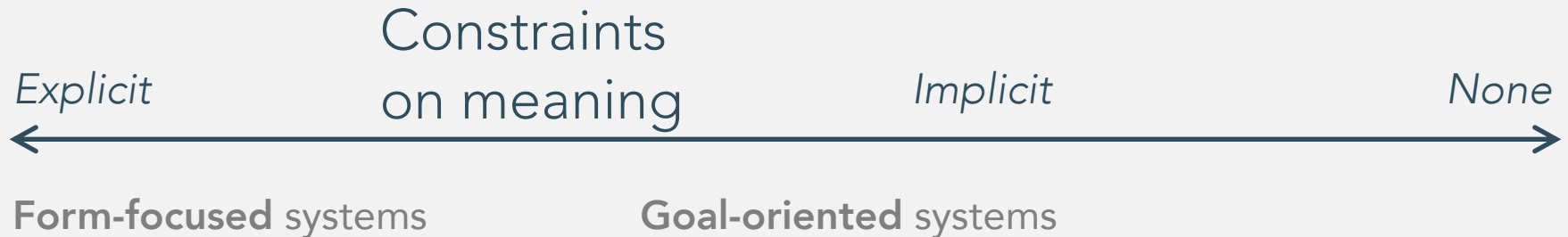
CALL-SLT (Baur, Rayner & Tsourakis, 2014)



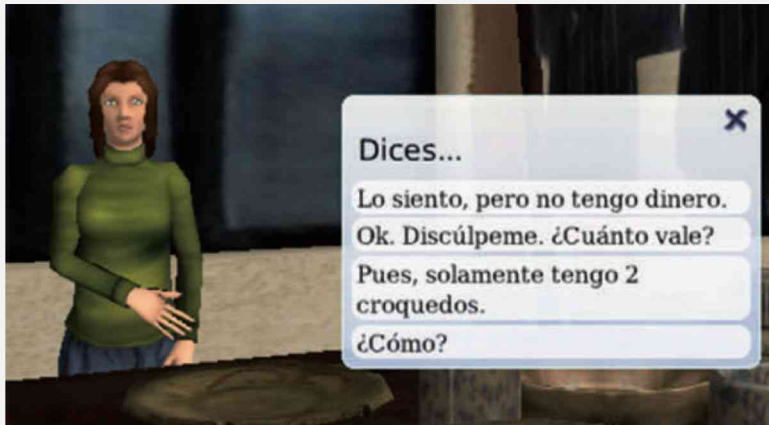
SPELL (Morton, Gunson & Jack, 2012)

Typology of systems (Bibauw, François & Desmet, 2019)

Four types of dialogue-based CALL systems



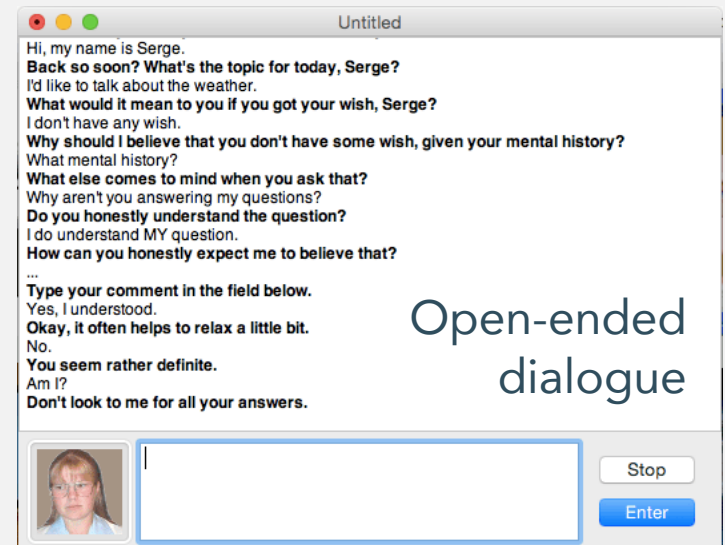
Narrative systems



Croquelandia (Sykes, 2008)

Branching dialogue
Pre-set form

Reactive systems



ELIZA (Weizenbaum, 1966)

Typology of systems (Bibauw, François & Desmet, 2019)

Form-focused / Goal-oriented



Form-focused systems

Explicit constraints on meaning:
gap-filling, predetermined answers

Focus of forms

Limited interactivity:
mostly corrective feedback

No dialogue management:
pre-scripted dialogue

Goal-oriented systems

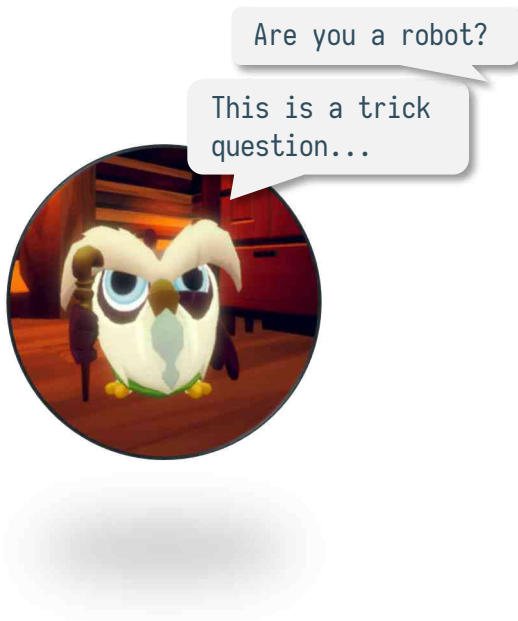
Contextual constraints on meaning:
interactional task and context

Focus on meaning/form

High interactivity:
conversation influenced by user

Advanced dialogue management:
→ high-level NLP required

Dialogue systems for language learning: typology of systems and measurement of effects



Dialogue systems for language learning

Terms, fields and definition

Rationale

Typology of systems

Types of dialogue-based CALL systems

- ▶ Technological approaches in research and industry

Past effectiveness

Meta-analysis of previous effectiveness studies

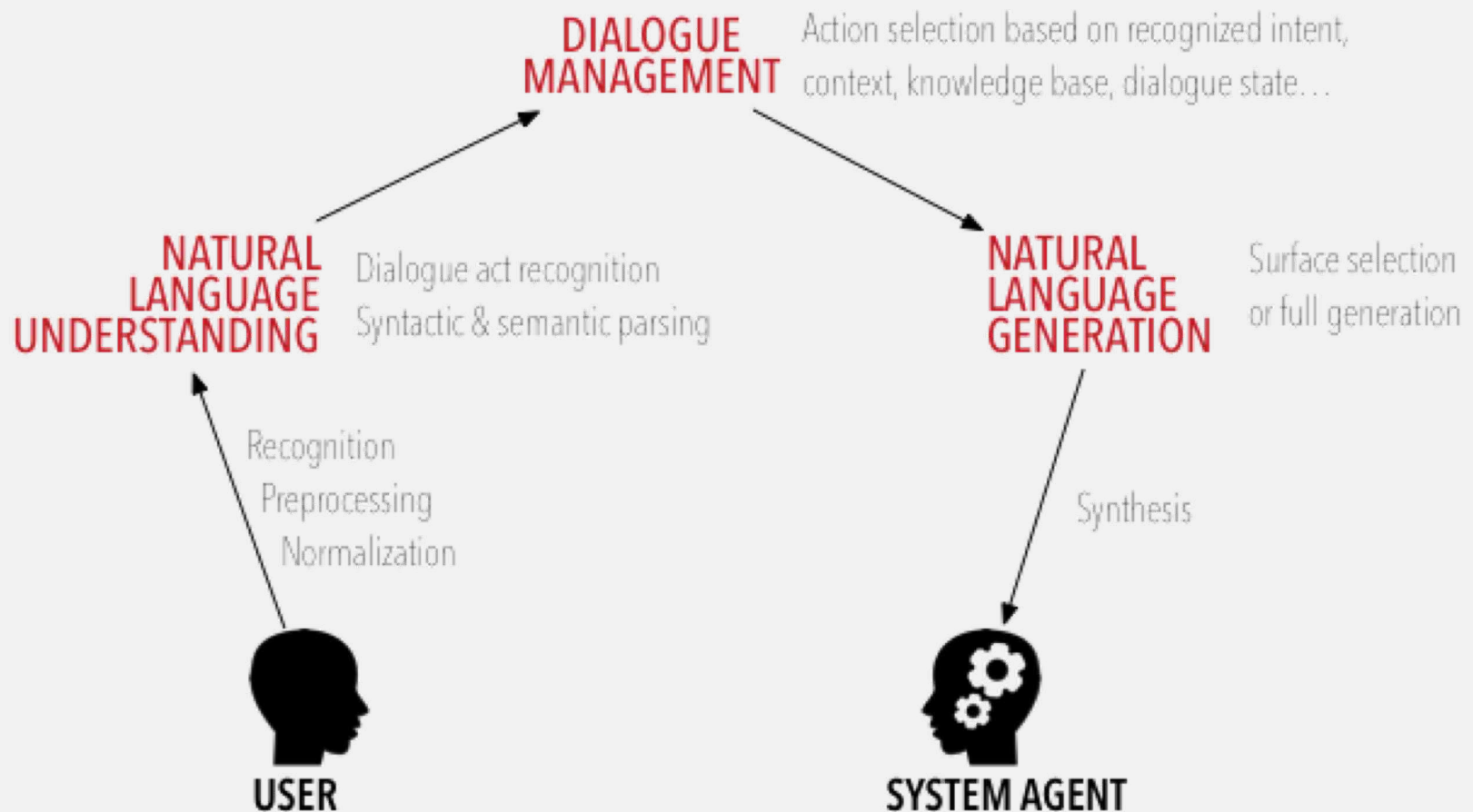
Evaluation of *LanguageHero*

Measuring effects on L2 development

Challenges and opportunities

Dialogue systems

Technological approaches



Dialogue systems

Technological approaches

Reactive systems (chatbots):

Rules-based approach

Research on dialogue systems:

Fully **data-driven** approaches

(Goal-oriented) systems in production:

Hybrid, *ad-hoc* approaches

Technological approaches

Handcrafted rules-based approach

Markup language for 'fast' manual rules writing

AIML (Wallace, 2003) (*Alice, Pandorabots*)

```
<category>
  <pattern>
    WHAT IS A CIRCLE
  </pattern>
  <template>
    <set_it>A circle</set_it> is the set of points
    equidistant from a common point called the center.
  </template>
</category>
```

ChatScript (Wilcox)

RiveScript (Petherbridge)

Very high number of rules

Many avoidance strategies as fallback

Disappointing

Technological approaches

Data-driven approaches in research

Deep learning (neural net) approaches

Based on very large corpora, restricted to certain domains (*Switchboard, Ubuntu Dialogue Corpus...*)

Promising results on mostly open-ended dialogue since 2015

- Pipeline vs. End-to-End methods
- Generative models vs. Retrieval-based methods

Still in need of standardised evaluation methods

See Serban et al, 2018, doi:[10.5087/dad.2018.101](https://doi.org/10.5087/dad.2018.101);
Chen et al, 2017, arXiv:1711.01731v2

Technological approaches

Hybrid, *ad-hoc* approach in production

Fully data-driven approaches not reliable enough for production.

Using data-driven NLU:

Intent recognition (dialogue act identification)
(Named) entity recognition

→ Commercial and open source platforms for NLU:
Rasa NLU, DialogFlow, Wit.ai, Microsoft LUIS, IBM Watson...

Mostly handwritten dialogue management and pre-scripted responses.

Concrete case of dialogue system

LanguageHero, dialogue-based game for French

Codeveloped with Leuven-based start-up Linguineo.

Prototype developed for Dutch-speaking teenage learners of French.

Task-based free conversational written interaction.

Logged in as sbibauw

Logout

Target language:

fr

Tutor language:

en

Interface language:

Réglages

Language Hero

Conversations:

Conversation 1: After the storm - Meet Sensei and find out what happened and where you are.

Meilleur score: 828

Conversation 2: Meet Baldog - Meet Baldog and ask him for help.

Meilleur score: 0

Conversation 3: The snails - Vincent - Get to know the snails family

Meilleur score: 426

Conversation 4: The snails - Angélique - Get to know the mother of the snails family

Meilleur score: 0

Conversation 5: The snails - Claudette - Get to know one of the triplets of the snails family

Meilleur score: 0

Conversation 6: Return to Baldog - Go back to Baldog and tell him his problem is solved.

Visit the world

Quit

Conversation: The snails - Vincent - Get to know the snails family



Contextualization

▶ *He: Bien le bonjour! Comment t'appelles-tu?*
✔ *You: bonjour je m'appelle Marco*
▶ *He: Enchanté de faire ta connaissance, Rinc! Rinc. Rinc. Rinc. Ne t'en fais pas, je ne suis pas fou. C'est juste que je répète ton nom pour ne pas l'oublier.*
You: Comment tu t'appelle?
He does not seem to have heard you...
You: Tu t'appelle coment?
He does not seem to have heard you...
✔ *You: Tu t'appelle comment?*
Correction: appelle - Vérifiez l'accord entre le pronom « Tu » et le verbe « appelle ».
Task accomplished: Good. That was what we were wondering about.
▶ *He: Moi, c'est Vincent. Elle, là-bas, c'est Angélique. Ça, c'est Delphine. Puis on a Georges dans le coin. Et évidemment, on ne peut pas oublier les triplées : Lisette, Claudette et Yvette. Oh! Et puis le petit là-bas, c'est Louis.*

Corrective feedback

Score: 405 ?
Friendship lvl0: Acquaintance

Gamification

🔥 Current task (2/30):
Say it is nice to meet them.

Microtasks to guide the conversation

Type or say your answer:

Type text..

Free written input

➔ Send your reply

🎤 Record your answer

? Disable help

⊗ End conversation

We can give you suggestions you can use to come up with an answer:

Scaffolding

LanguageHero

Instructional and technological approach

Dialogue guided by **microtasks/instructions**

“Ask what happened.”
“Tell B... you were actually
hoping he would help you.”

→ Give directions to the user

→ Higher predictability of the user intents (NLP)

Technologically, **hybrid system**:

- **Machine learning** for speech recognition and **intent recognition** (i.a. ~RASA NLU)
- **Parser-** and **rule-based** detection of task completion and dialogue management (i.a. ~*ChatScript*), as well as for corrective feedback provision.
- All possible responses pre-scripted.

Dialogue systems for language learning: typology of systems and measurement of effects



Dialogue systems for language learning

Terms, fields and definition
Rationale

Typology of systems

Types of dialogue-based CALL systems
Technological approaches in research and industry

Past effectiveness

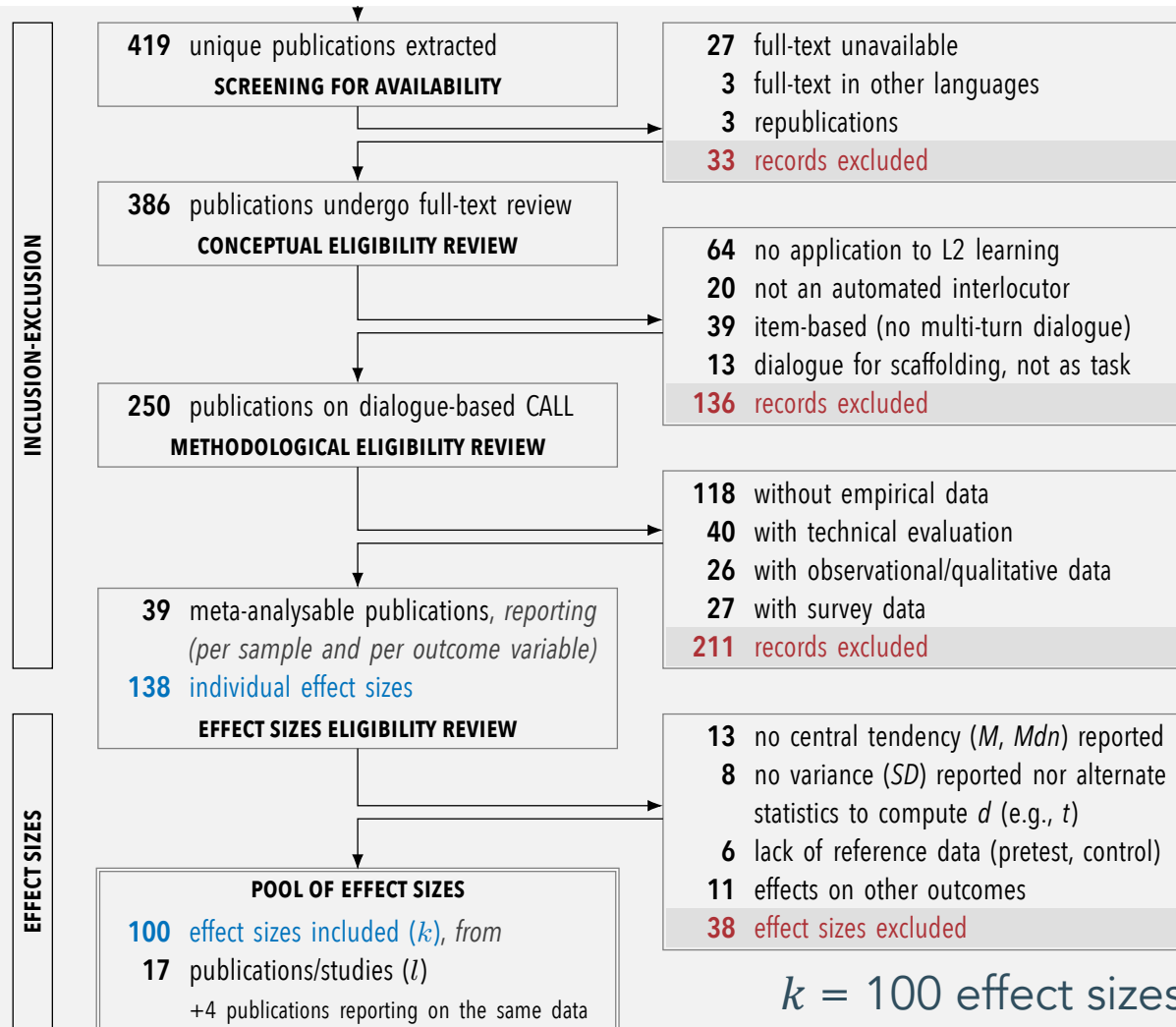
- ▶ Meta-analysis of previous effectiveness studies

Evaluation of *LanguageHero*

Measuring effects on L2 development
Challenges and opportunities

Meta-analysis of effectiveness studies

Inclusion of individual effect sizes



Meta-analysis: methods

Comparable effect size metrics

28

Morris & DeShon (2002) offer a comparable metrics across experimental designs (EC / PP / ECPP)

- **change metric** (aligned on *within*-group effect)
- **raw metric** (aligned on *between*-groups effect)

We selected the *raw* metric formula:

$$d_{PP} = J(df_{PP}) \left(\frac{M_{\text{post,E}} - M_{\text{pre,E}}}{SD_{\text{pre,E}}} \right)$$

$$d_{ECPP} = J(df_{ECPP}) \left(\frac{M_{\text{post,E}} - M_{\text{pre,E}}}{SD_{\text{pre,E}}} - \frac{M_{\text{post,C}} - M_{\text{pre,C}}}{SD_{\text{pre,C}}} \right)$$

Meta-analysis: methods

Multilevel modeling

(Van den Noortgate & Onghena, 2003)

Publications report multiple outcome measures (e.g., vocabulary and morphology tests) or multiple sampling groups (e.g., proficiency levels)

Traditional meta-analysis techniques allow only one (independent) effect size per study, but losing thus all the information on distinct implementations.

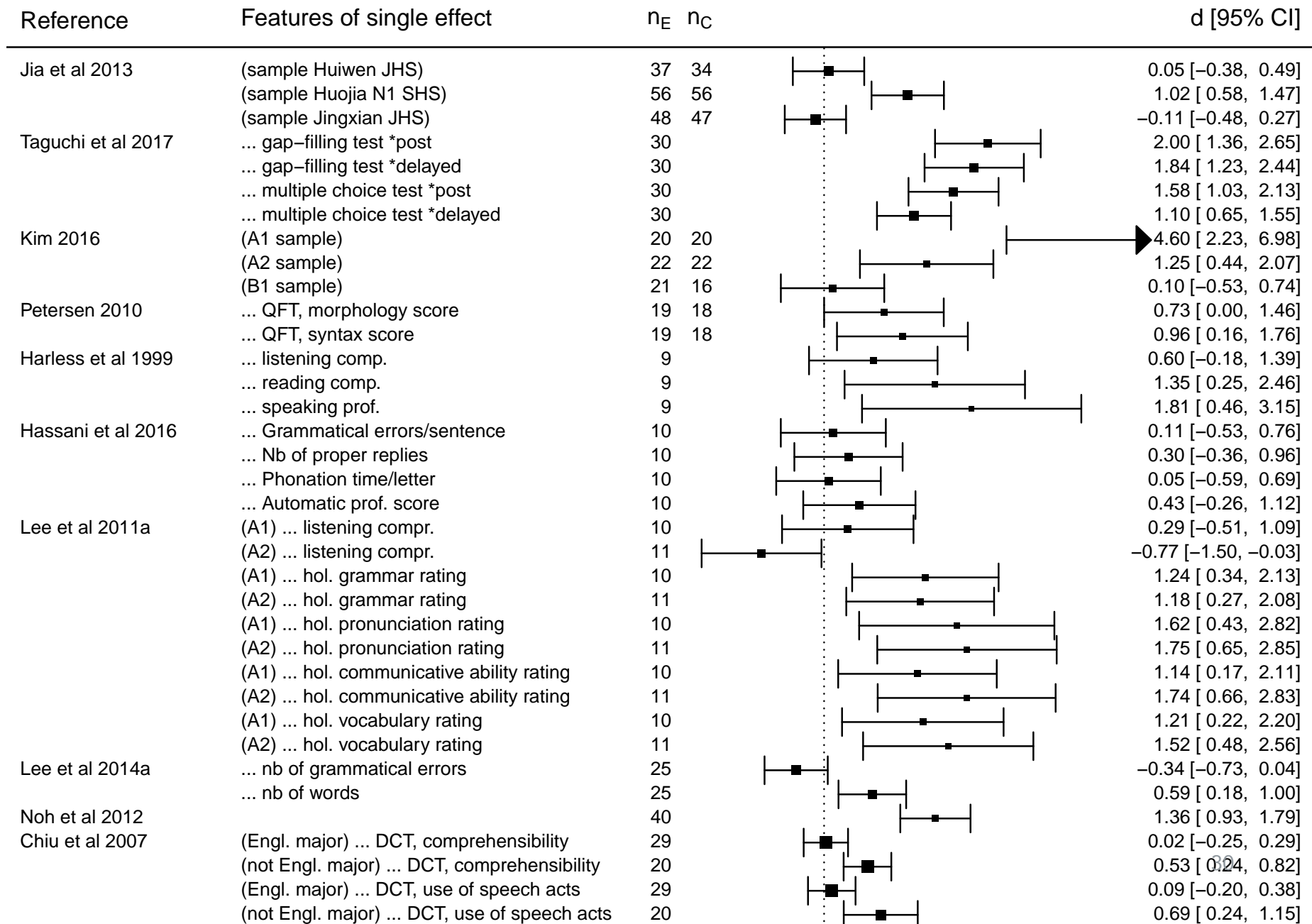
⇒ Including all the variation without “fooling” the model with non-independent measures:

Multilevel modelling:

aggregates **multiple effects per study**, by adding an intermediate level of *within-study* variation.

Table 1: Levels of multilevel meta-analytic model

	Level	Number of clusters/items	Source of variance
1	Samples	$k = 96$ ($n = 803$)	Random sampling variance
2	Effects sizes	$k = 96$	Variation within study
3	Studies	$l = 17$	Variation between studies



Meta-analysis: results

Summary effect

General effectiveness of dialogue-based CALL
for L2 proficiency development ($k = 96$):

$$d = 0.602^{***}$$

$$95\% \text{ CI} = [0.373, 0.831]$$

= Medium effect (Plonsky & Oswald, 2014)

FYI, if converted/computed as *change* metrics:

$$d_{\text{change}} = 0.658^{***} [0.414, 0.901]$$

Immediate effect only (no delayed posttests, $k = 73$):

$$d_{\text{raw}} = 0.627^{***} [0.390, 0.863]$$

Meta-analysis: results

Summary effect compared to CALL/SLA

32

Global effect close to the median of **meta-analyses in CALL/SLA**

(Plonsky & Oswald, 2014)

- \gtrsim game-based learning ($d = .53$, Chiu et al, 2012)
- \lesssim CALL in general ($d = .84$, Plonsky & Ziegler, 2016)

Consistent with effect of **face-to-face interaction** (Mackey & Goo, 2007) and SCMC.

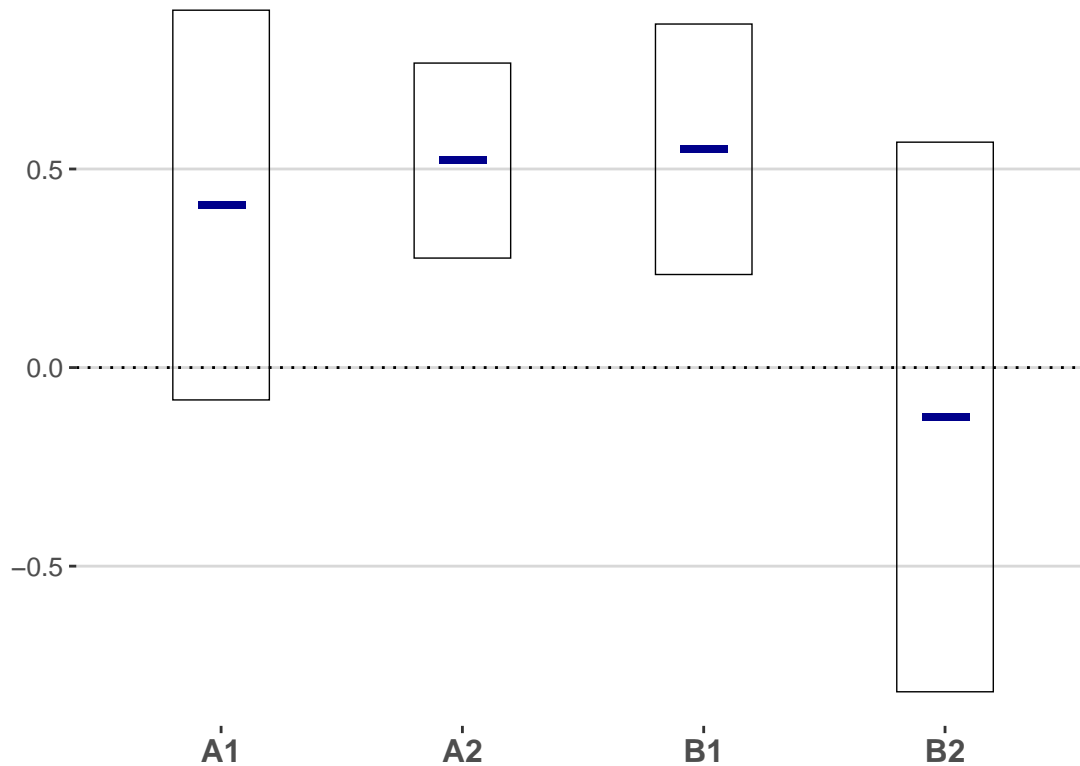
- \lesssim F2F interaction ($d = .75$, Mackey & Goo, 2007)
- \lesssim SCMC (Ziegler, 2015; Lin, 2015)

Slightly **inferior** to the above (although within 95% CI), but logical:

- Human interlocutors remain the gold standard!
- Outcome variables often very ambitious (holistic proficiency...) and treatment duration often very reduced ($\leq 3h$)

Meta-analysis: moderator analyses

Participants ▶ L2 proficiency

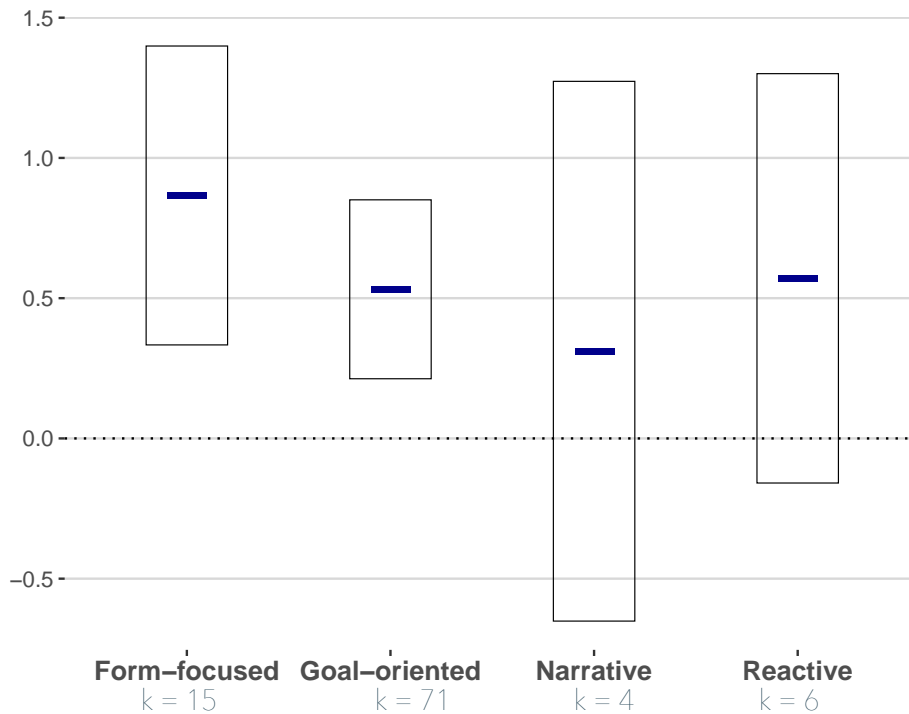


Mostly effective
for **A2-B1** learners.

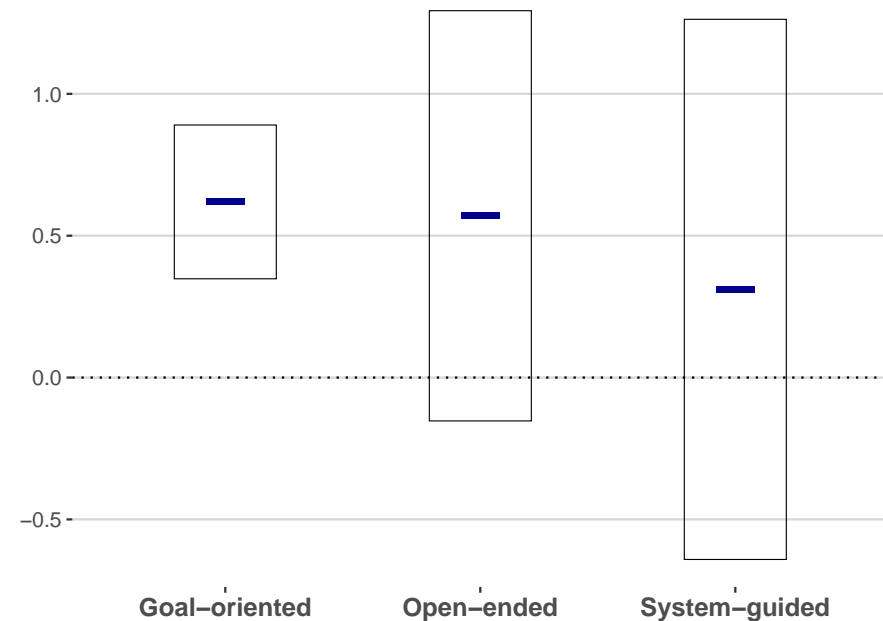
After consolidating
basic structures?

Meta-analysis: moderator analyses

System ▶ Type of system



Form-focused and goal-oriented systems confirm their effectiveness. Unclear difference though.

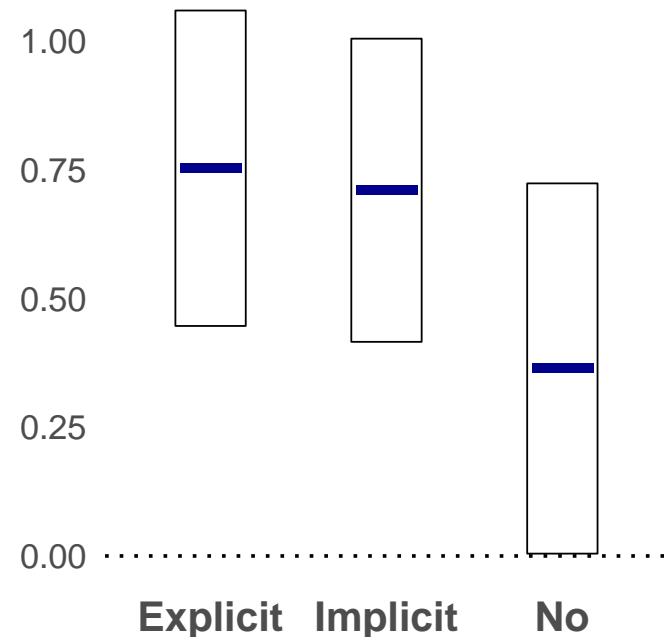


Meta-analysis: moderator analyses

System ▶ Corrective feedback

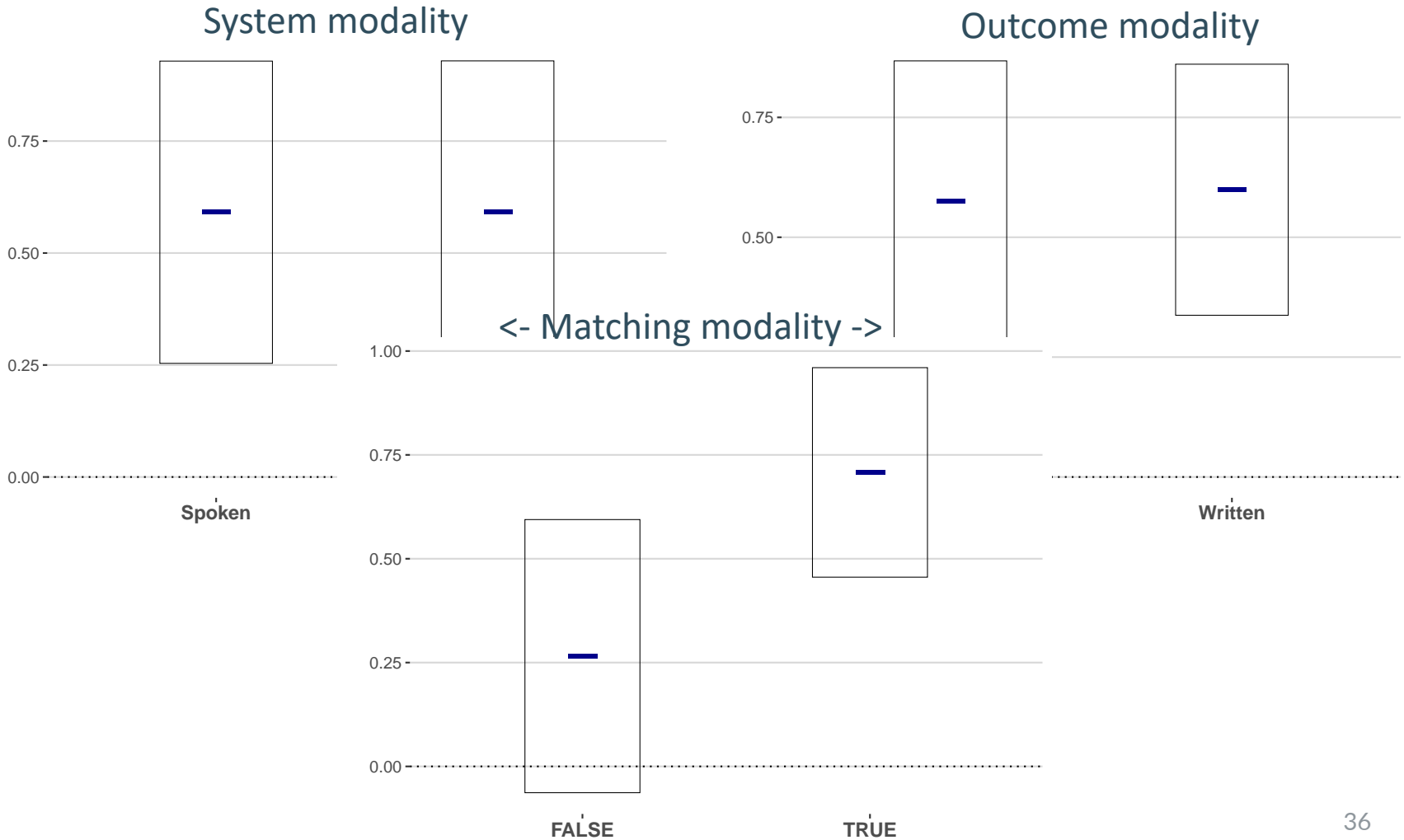
Consistently with what we know about corrective feedback, systems providing **feedback** are **much more effective**.

If binary (w/ vs. w/o CF):
 $QM_{(df = 1)} = 2.53, p = 0.111$



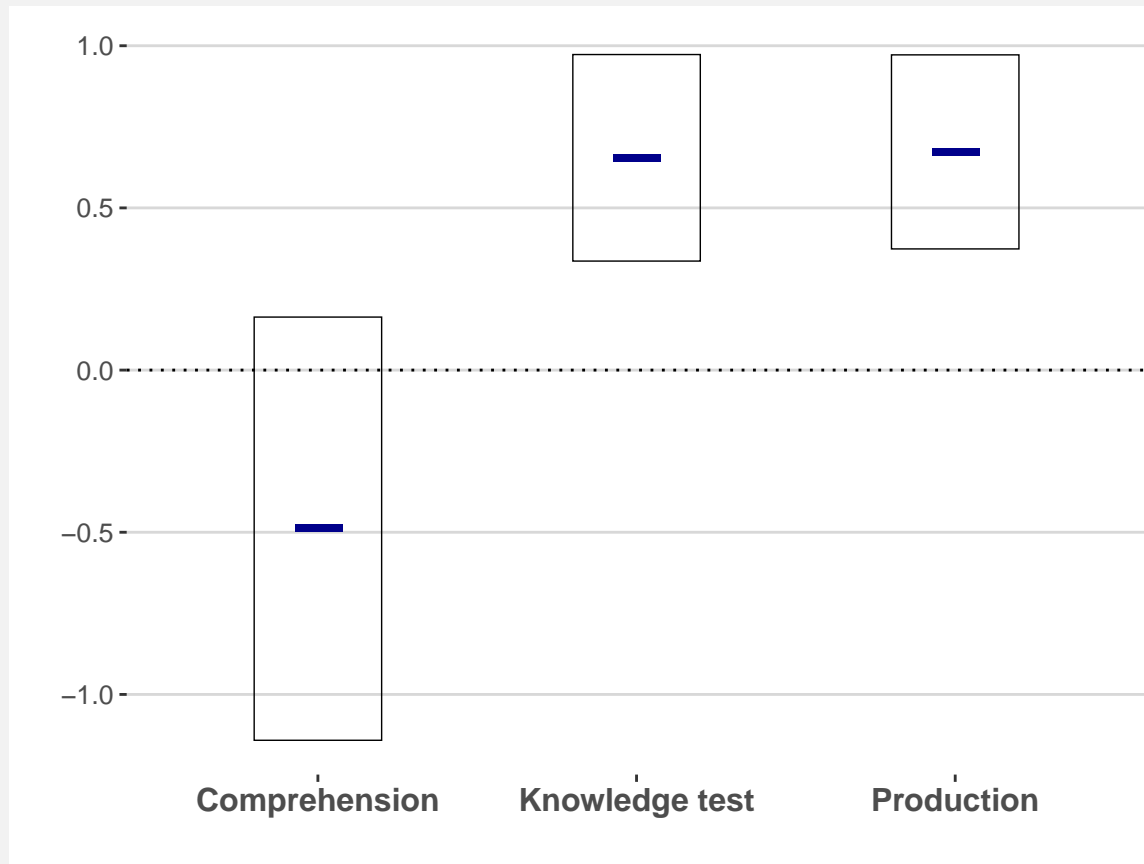
Meta-analysis: moderator analyses

Practice and outcome modality



Meta-analysis: moderator analyses

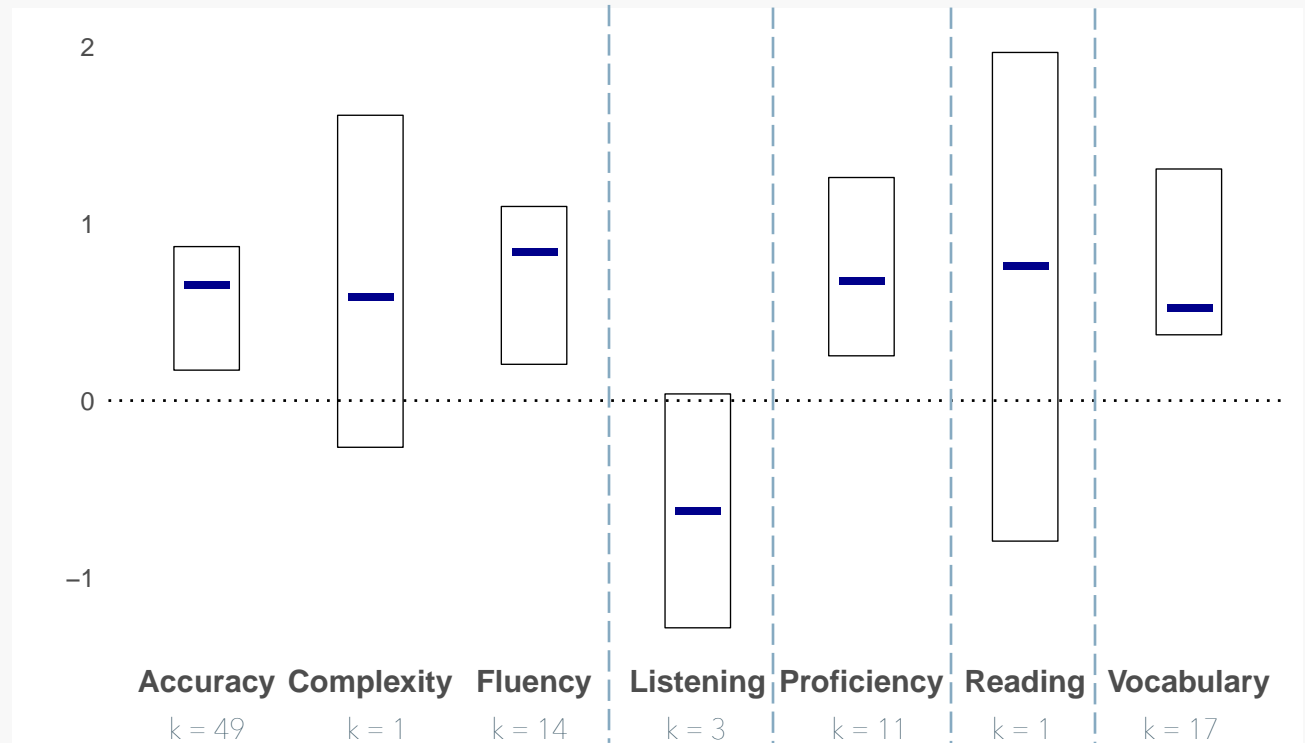
Outcome ► Dimensions



Meta-analysis: moderator analyses

Outcome ► Dimensions

More promising effects on **fluency** and possibly vocabulary



Meta-analysis: moderator analyses

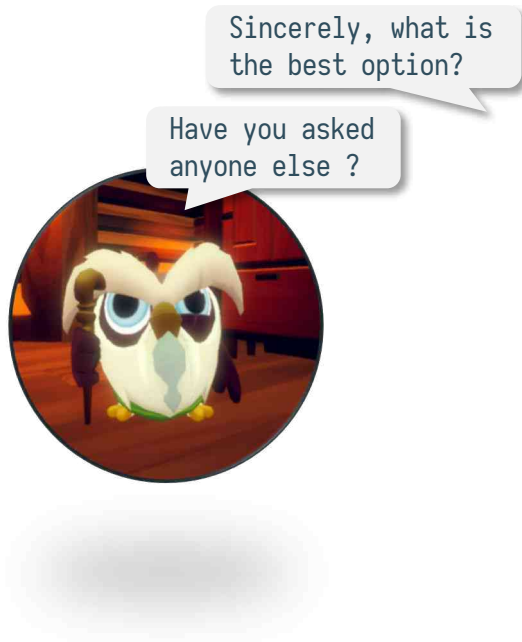
Insights for future research and development

Global effectiveness of dialogue-based CALL,
but too few studies to determine significant differences
between systems, interventions and outcomes.

Promising design and target characteristics:

- task-based / goal-oriented
but significantly different from form-focused ?
- with corrective feedback
- for beginner/low-intermediate learners
- for fluency and vocabulary development

Dialogue systems for language learning: typology of systems and measurement of effects



Dialogue systems for language learning

Terms, fields and definition
Rationale

Typology of systems

Types of dialogue-based CALL systems
Technological approaches in research and industry

Past effectiveness

Meta-analysis of previous effectiveness studies

Evaluation of *LanguageHero*

- ▶ Measuring effects on L2 development
Challenges and opportunities

Dialogue systems for L2 research

Research question

Technologically, it is considerably easier to “fake” the interaction by restraining/ignoring the learner, rather than offering full interactivity, freedom and contextual task completion. Are these technological developments worth it?

1. Do (more) interactive and emergent dialogue systems offer significantly better pedagogical opportunities for L2 development, in comparison with more constrained ones?

Responding it would also answer questions regarding what aspects of interactivity in general are really promoting language learning.

Intervention · Conditions

Interactive vs. static dialogue

Compare:

(A) fully interactive,
immediate/synchronous
dialogue system

(B) classic, asynchronous
dialogue completion task

Conditions with identical tasks,
input, output opportunities,
feedback and scaffolding.

The top screenshot, labeled 'dialogue system', shows a 3D environment with a character (an owl) and a chat window. The chat window contains the following text:

- Lui: Bonjour, petit hibou. Tu es enfin réveillé!
- Vous: Bonjour Monsieur
- Lui: Comment va-tu? Tu es tombé de très haut?
- Vous: Un peu.
- Lui: Ouf, c'est une bonne nouvelle! Ton aile avait l'air très mal-en-point quand je t'ai amené ici.
- Vous: Ou je suis?
- Lui: Tu es dans le Grand Monastère des Montagnes éternelles, mon petit.

The bottom screenshot, labeled 'dialogue completion', shows the same 3D environment and character, but with a list of prompts and input fields. The prompts are:

- Lui: Bonjour mon petit. Je vois que tu te réveilles enfin. (← Greet the old owl.)
- Lui: Comment vas-tu ? (← Say how it is going.)
- Lui: Ah, je suis soulagé ! Ton aile avait l'air très mal-en-point quand je t'ai amené ici. (← Ask where you are.)
- Lui: Tu es dans le Grand Monastère des Montagnes éternelles, mon petit. (← Ask who he is.)
- Lui: Je suis Sensei, le moine. Et toi, comment t'appelles-tu ? (← Tell your name.)
- Lui: Enchanté ! (← Ask what happened.)

Each prompt is followed by an input field labeled 'Ecris ta réponse ici'. A 'Terminer' button is visible at the bottom right of the interface.

Conversation: The snails - Vincent - Get to know the snails family



He: Bien le bonjour! Comment t'appelles-tu?



You: bonjour, je m'appelle Rinc



He: Enchanté de faire ta connaissance, Rinc! Rinc. Rinc. Rinc. Ne t'en fais pas, je ne suis pas fou. C'est juste que je répète ton nom pour ne pas l'oublier.

You: Coment tu t'apele?

He does not seem to have heard you...

Conversation: The snails - Vincent - Get to know the snails family



He: Bien le bonjour ! Comment t'appelles-tu ?



Je m'appelle Rinc



He: Enchanté de faire ta connaissance, Rinc! Rinc. Rinc. Rinc. Ne t'en fais pas, je ne suis pas fou. C'est juste que je répète ton nom pour ne pas l'oublier.



Coment tu t'apele?



He: Moi, c'est Vincent. Elle, là-bas, c'est Angélique. Ça, c'est Delphine. Puis on a Georges dans le coin. Et évidemment, on ne peut pas oublier les triplées : Lisette, Claudette et Yvette. Oh! Et puis le petit là-bas, c'est Louis.



Typ tekst..

Send your reply



Task: Say it is nice to meet them.



He: Ah, vraiment? C'est aussi ce que l'ours a dit! Mais après, il oublie nos prénoms et nous traite de limaces! Des LIMACES!? Tu imagines? Si tu es si content de nous connaître, alors tu peux me répéter nos prénoms? Ah! Tu vois! Tu t'en souviens pas, hein?! Désolé, c'est pas de ta faute, petit, mais personne ne fait jamais attention à nous.



Typ tekst..

Send your reply



Quit

Methods

Population and group assignment

4 schools volunteered to participate, with 2-3 classes each:

$$N_{\text{clusters}} = 11$$

$$N_{\text{participants}} = 215 \text{ (208 complete cases)}$$

Random assignment of classes to 3 conditions (distr. equally across schools):

- **Dialogue System** (experimental): $n_{\text{D.Sys.}} = 81$
- **Dialogue Completion** ('baseline'): $n_{\text{D.Compl}} = 79$
- **Control** ('business-as-usual') $n_{\text{control}} = 49$

Flemish 2nd year secondary school learners of French ($M_{\text{age}} = 13.4$ y.o.)

L1 = 95,3 % Dutch

L2 = French = first L2, $M = 3,1$ years of instruction, mostly at **A1** level

(M_{score} in productive vocabulary size test = 3.6/30 in 1K frequency band)

10 (near-)native speakers of French excluded (final N = 198)

Methods

Procedure

1-4 weeks,
depending
on school
schedule

All sessions
at school



Pretest

- Computer-delivered spoken interview
- Target vocabulary test
- Vocabulary size test

In-app session (max 50 min):

DSys / DCompl

In-app session (max 50 min) :

DSys / DCompl

In-app session (max 50 min) :

DSys / DCompl

Posttest

- Computer-delivered spoken interview
- Perceptions questionnaire
- Target vocabulary test

Methods · Instruments

Perceptions questionnaire (post)

Construct	Subdimensions	Items	α	Source/Theoretical framework
Perceived ease-of-use	Corrective feedback, Comprehensibility, Interface, Tasks	5 (7)	.67	Technology Acceptance Model (Davis 1989), partially from Cornillie et al (2013)'s translation (adapted)
Perceived usefulness	General usefulness, Corrective feedback, Hints, Tasks	11	.89	
Perceived interactivity	Immediacy, Control, Mutuality	11 (13)	.79	New scale developed
Perceived authenticity	General Academic Personal	6 (7)	.84	Perceived Authenticity of Writing Scale (Behizadeh & Engelhard 2014) (adapted)

e.g., PERCEIVED INTERACTIVITY: "Through my answers, I could really have an impact on the game."

PERCEIVED USEFULNESS: "I am less afraid to speak French now than I was before playing the game."

Target Vocabulary Test (1)

“Target” words and sequences seen and potentially produced inside the intervention: based on frequency of exposure across whole available content, selecting the most frequent lemmas and the most frequent formulaic sequences.

But no explicit target of instruction (no specific feedback, no glossing, no systematic presentation)
⇒ **Incidental learning only**

At pre- and post-test (identical, randomized order)

Methods · Instruments

Target Vocabulary Test (2)

- **Receptive** part (meaning recognition):

25 items

translation, as multiple choice

e.g., Potager: soep moestuin vriend potaarde

Ik weet het niet

soup vegetable garden friend potting soil

I don't know

- **Productive** part (in-context form recall):

25 items

gap-filling (L2 only) on formulaic sequences

e.g., Cet auteur a vraiment _ _ _ _ _ d'imagination : ses livres
sont très originaux !

*This author really has a lot of imagination: his books are
really special!*

Computer-delivered speaking interview

Automatized simultaneous speaking test

Individual, in-class & simultaneous,
with headset, in front of computer

24 questions

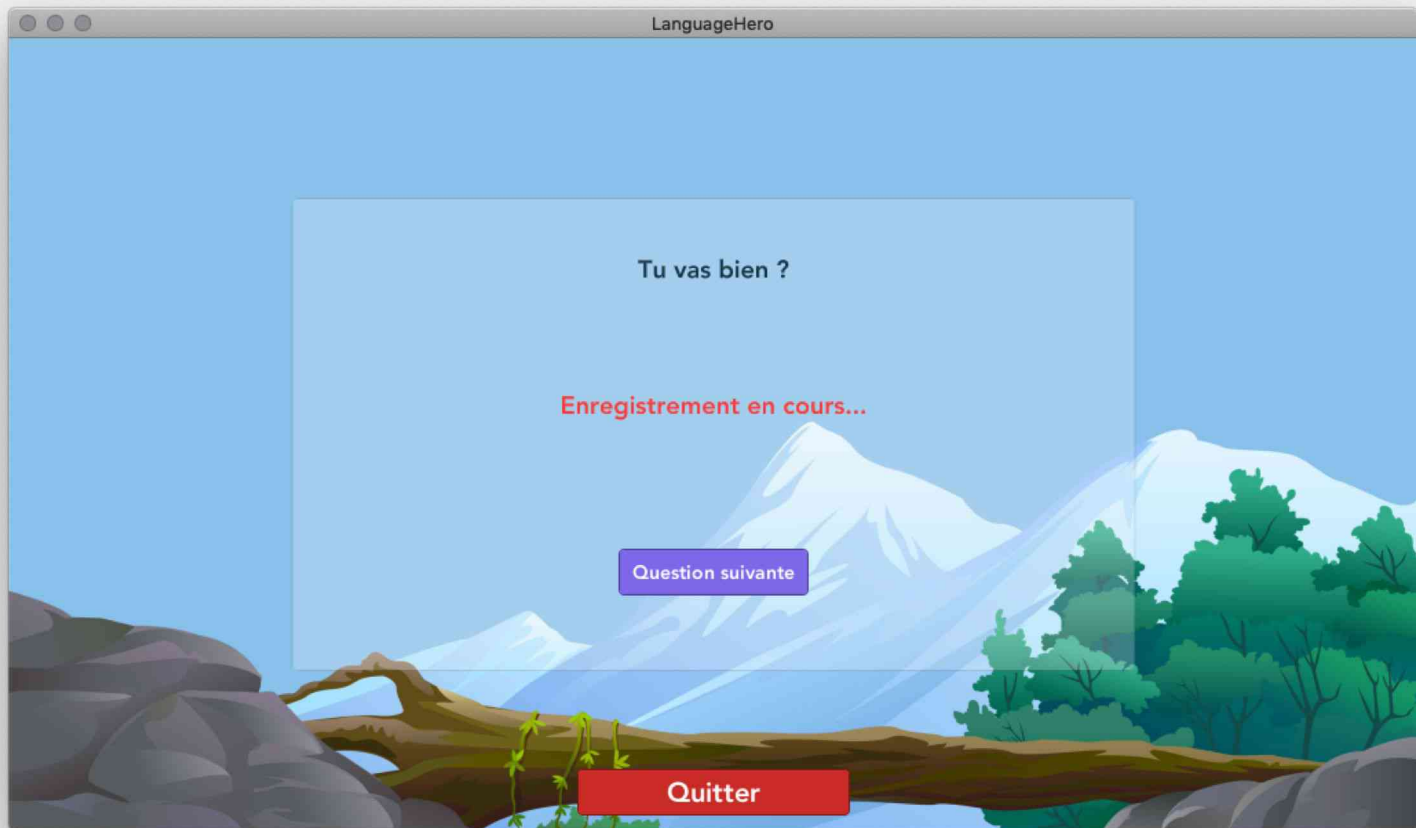
from basic ("How are you?") to questions targeting
specific communicative functions ("Can you
describe your French teacher?")

Question oral + written presentation,

then automatically starts recording,
30 sec limits or "Next question" button

Methods · Instruments

Computer-delivered speaking interview



Methods

Automated fluency metrics computation

±10 000 single audio files (N=208 * 24 questions * pre+post)

- Automated speech recognition (Google Cloud Speech-to-text) for transcription
- Manual correction of transcriptions + annotation of filled pauses, L1/LF use, meta-discourse, etc.
- Automated detection of pauses (Praat syllable nuclei detection script, de Jong & Wempe, 2009)
- Automated computation of syllables from transcript, with variations in pruning, and selection of measures that best predict proficiency level.

Methods

Fluency metrics

Speaking fluency (Segalowitz, 2010)

~~• Cognitive fluency~~

~~• Perceived fluency~~

• Utterance fluency (temporal/performance)

• Speed fluency

• speech rate, articulation rate, syllable duration, length of runs (syllables), duration of runs (sec)... (Bosker et al, 2013; Hilton, 2014; Kormos & Denes, 2004; Götz, 2013...)

• Breakdown/Pauses

• silent pause rate, silent pause duration... (Bosker et al, 2013; de Jong & Bosker, 2013; Kahng, 2014; Hilton, 2014...)

~~• filled pauses: not good differentiator (Cucchiari et al, 2002...), unrelated to other fluency measures (Segalowitz et al 2017)~~

~~• Repair fluency: not good differentiator of proficiency (Cucchiari et al, 2002; Revesz et al 2016; Saito et al 2018; Dumont, 2017...)~~

Combined
metric via
Principal
Component
Analysis

Using a silent pause threshold of 250ms (de Jong & Bosker, 2013; Préfontaine et al, 2016)

Differences of learners' behaviours

Pilot (2 classes in first school): “Discourse Completion Task” even more limited (no explicit validation of responses, no feedback, no scaffolding), to reflect the paper version of such a task

→ Strong attitudinal influence (DCT condition):

at session 2, a few learners asked “why are we doing this?”

at mid-session 3, multiple pupils stopped trying/working altogether

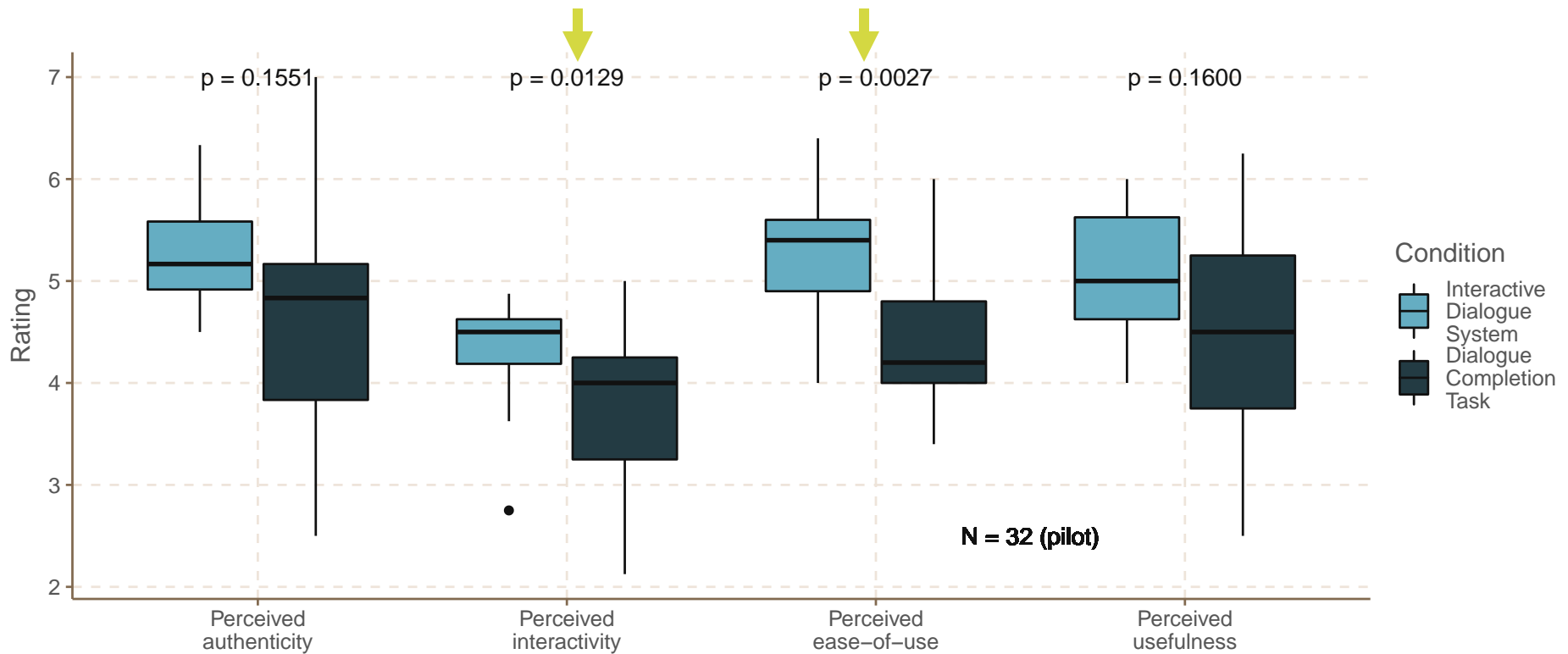
23.7% of messages containing “voluntary noise”

→ Raised ethical issues

⇒ Added **basic “correct/not” feedback** and **writing support afterwards** → essentially solved the issue

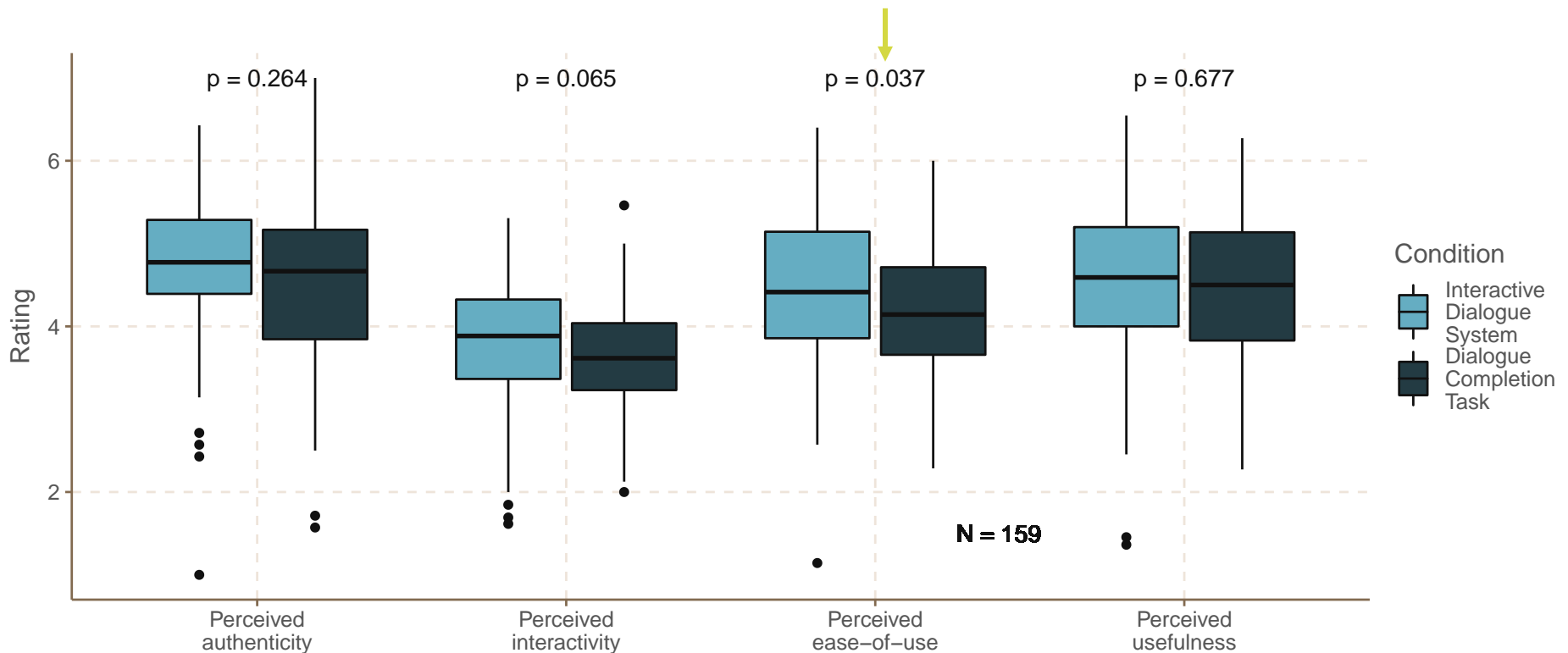
Experimental results

Differences of learners' perceptions (pilot only)



Experimental results

Differences of learners' perceptions



Differences of learners' perceptions

Feeling of interactivity within dialogue-based CALL game seem to be majorly influenced by the **basic feedback** received.

Goal vs. form-orientation

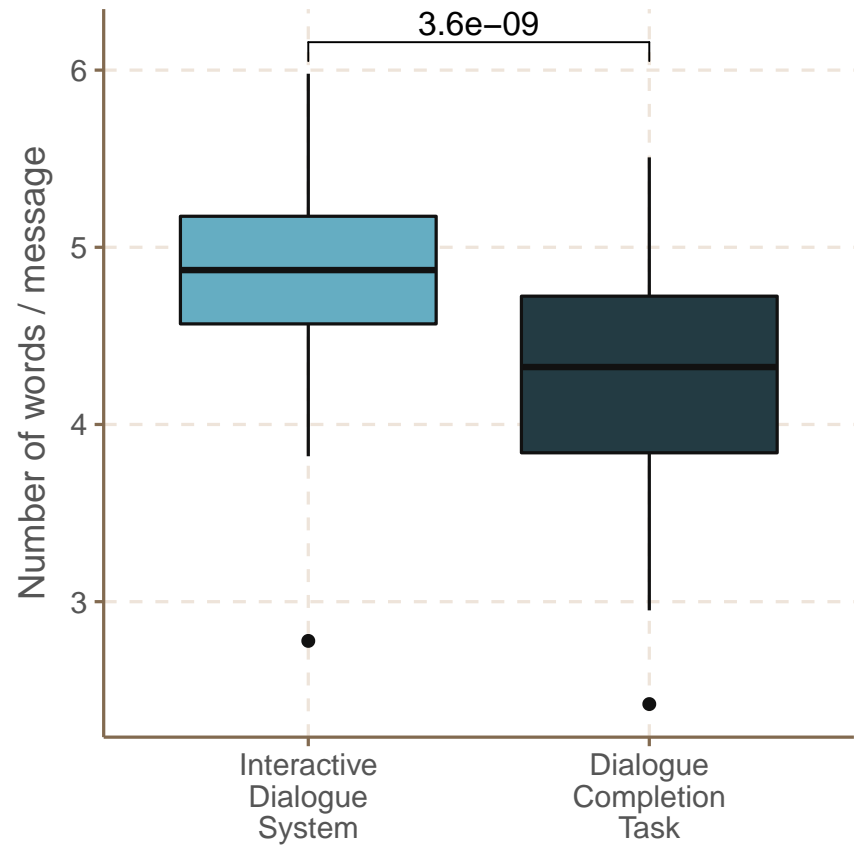
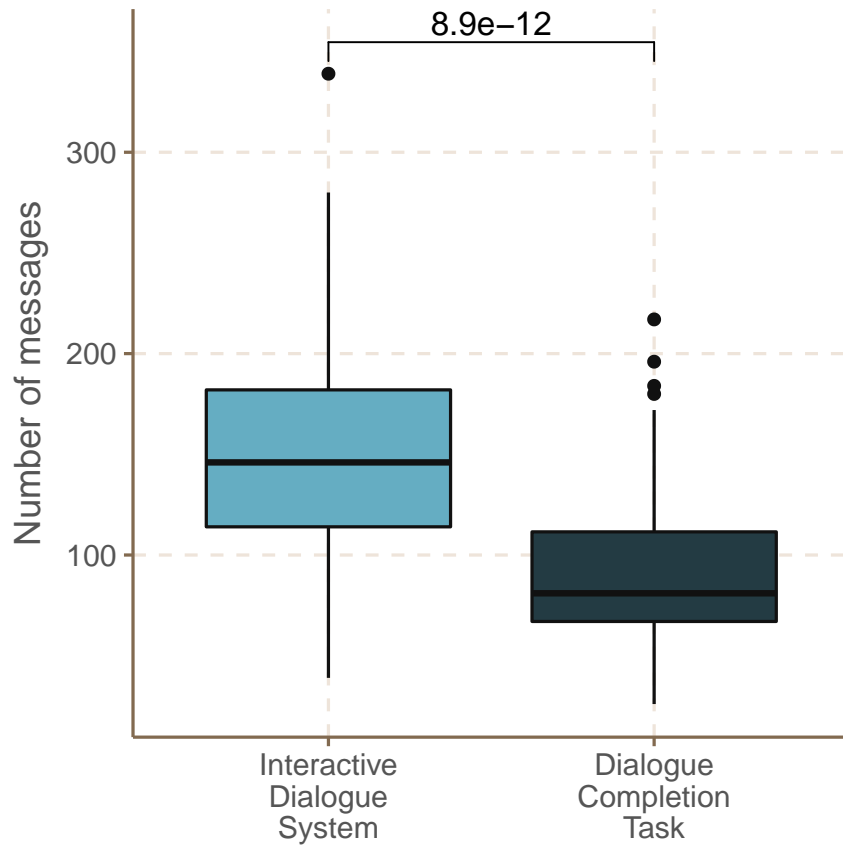
form-orientation behaviour/'exercise mindset'
among many participants from both conditions:

due to in-school experiment? age factor?
presentation of the instructions?

→ lack of perception of task goals as meaningful

Experimental results

Quantity of in-task production



Results

Receptive vocabulary

Very significant increase.

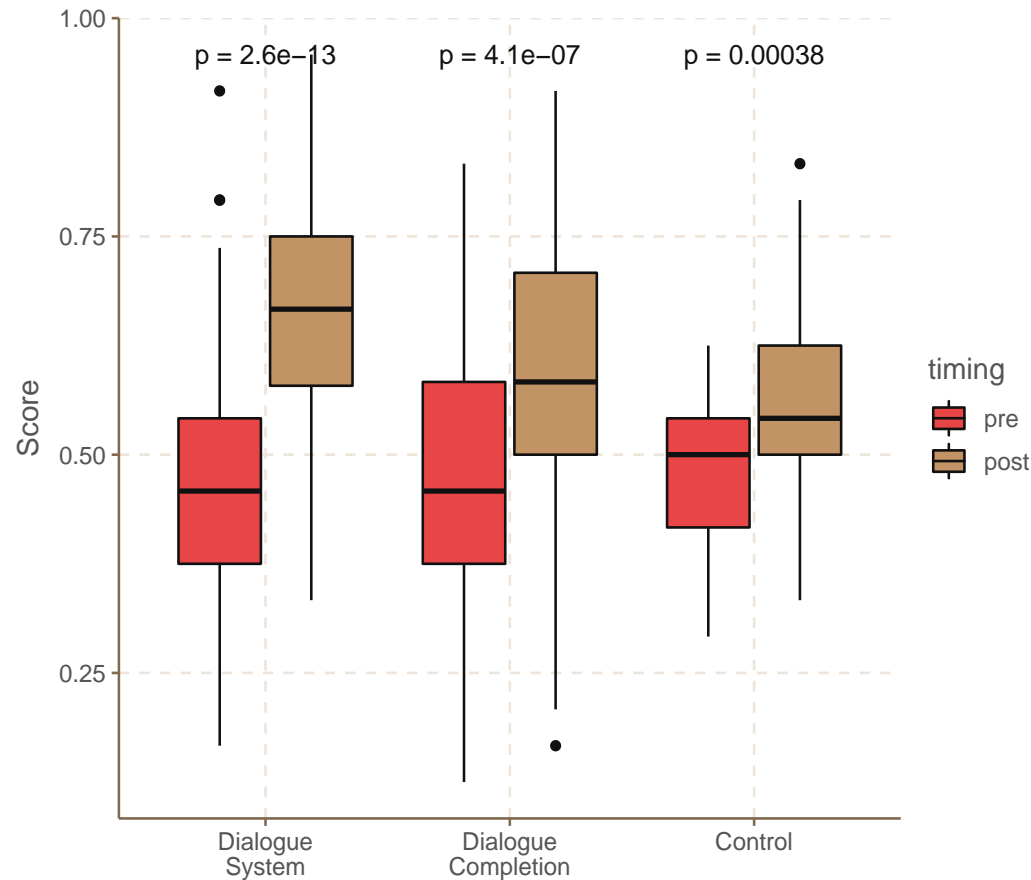
$$d_{\text{DSystem}} = 1.17^{***}$$

$$d_{\text{DCompletion}} = 0.80^{***}$$

$$d_{\text{DControl}} = 0.67^{***}$$

Considering the short treatment (2h),
clear difference between conditions.

$$d_{\text{DSys vs DCompl}} = 0.25^*$$



Results

Productive vocabulary

Less marked increase,
and much more difficult test.

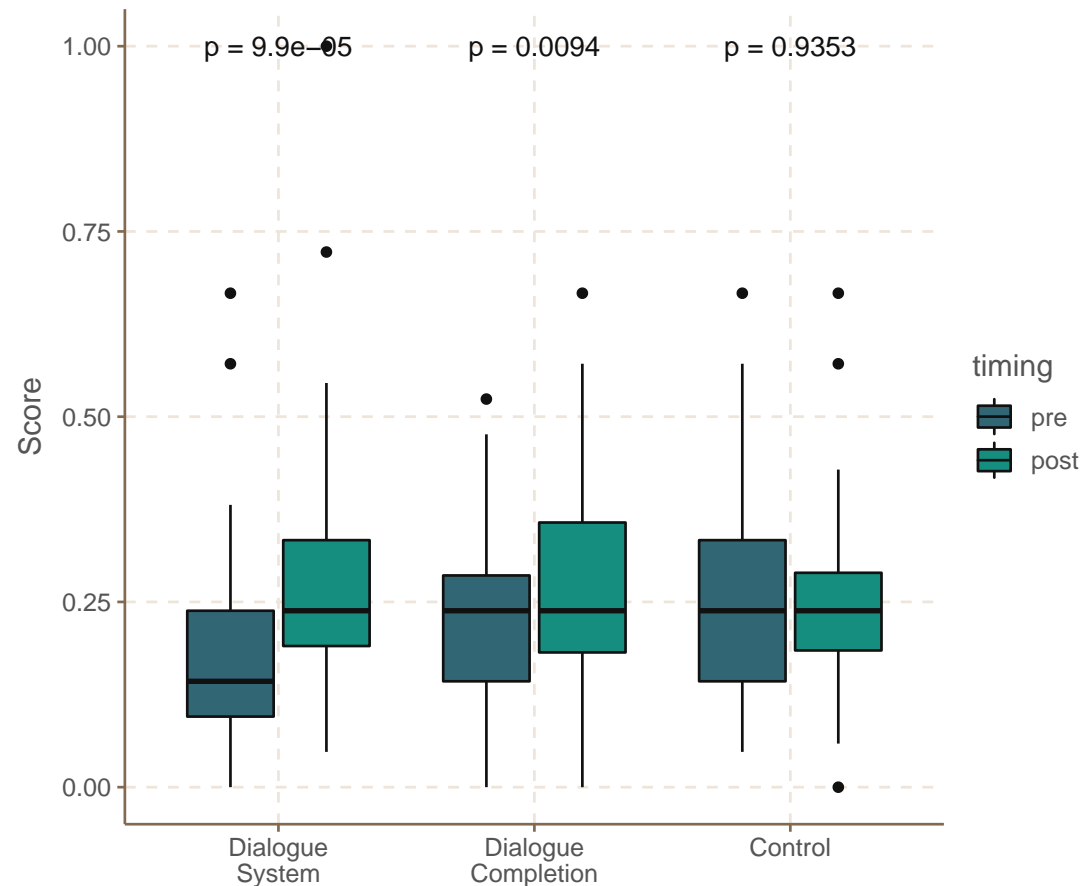
$$d_{\text{DSystem}} = 0.56^{***}$$

$$d_{\text{DCompletion}} = 0.39^{***}$$

$$d_{\text{DControl}} = 0.02 \text{ n.s.}$$

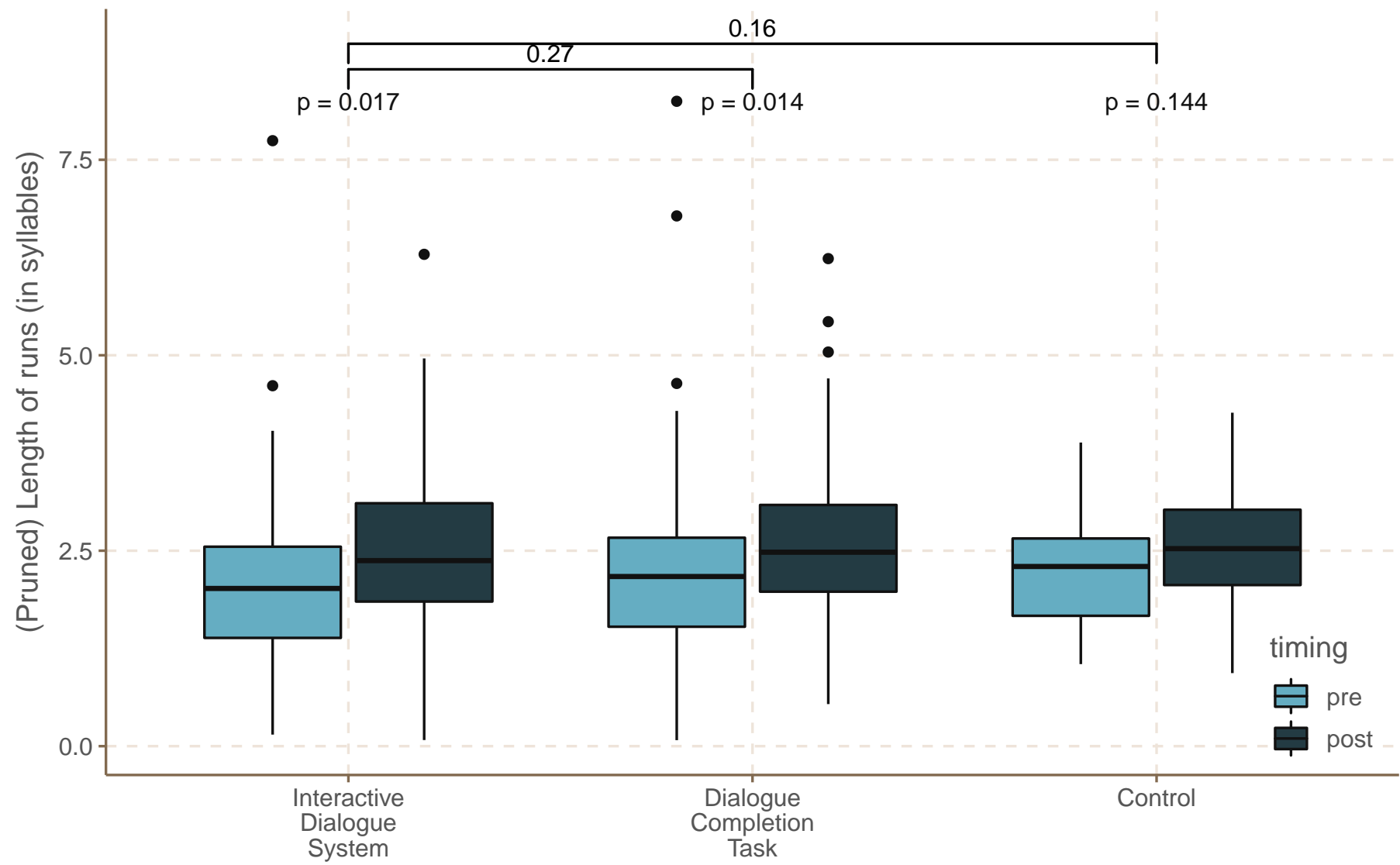
But here, no improvement in
control group and benefits of
practice are much clearer.

$$d_{\text{DSys vs DCompl}} = 0.23 \text{ n.s.}$$



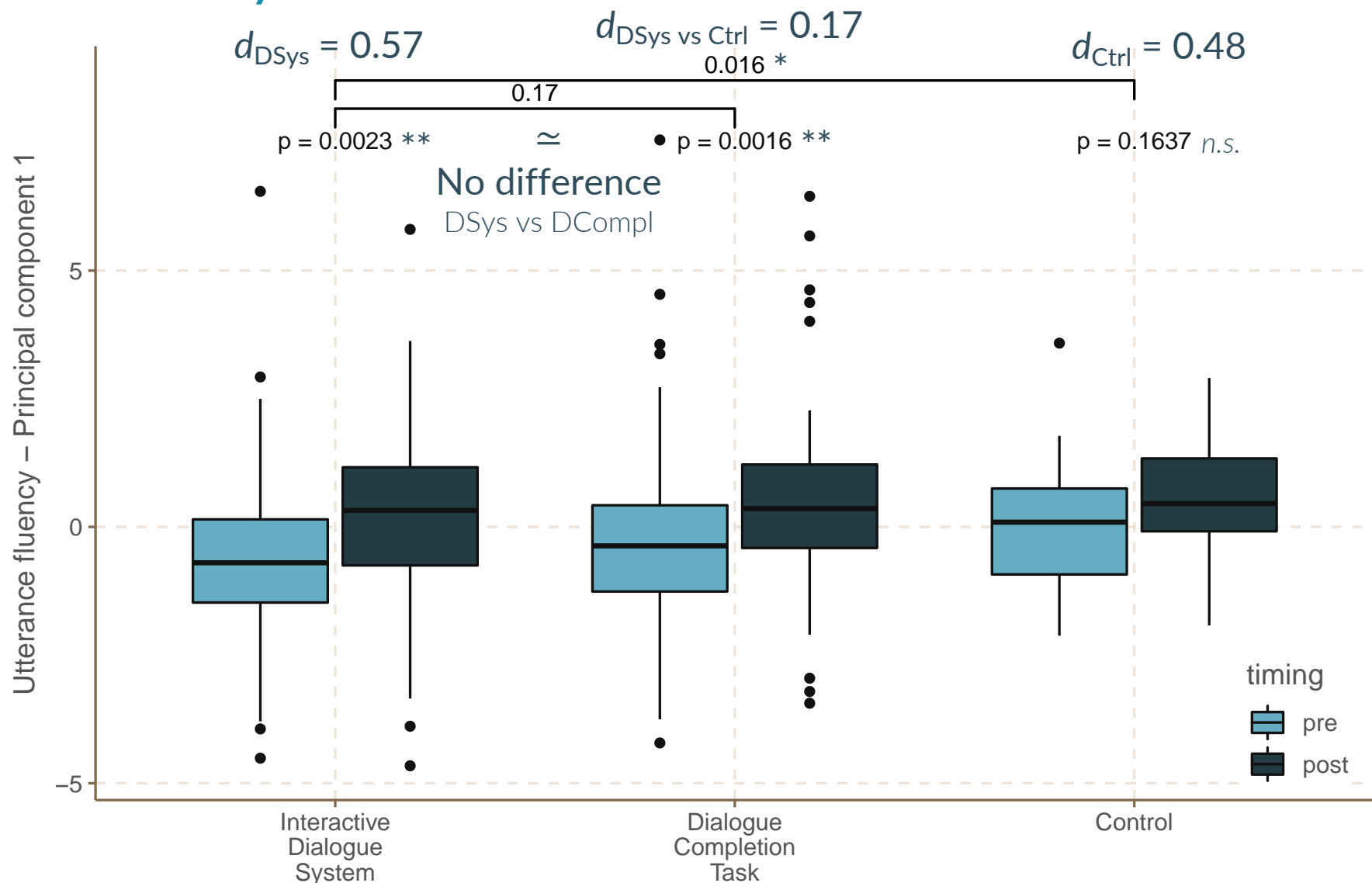
Results

Fluency



Results

Fluency



Discussion

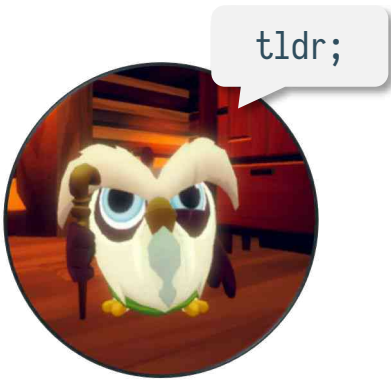
Fluency

Very small effect ($d_{\text{DSys vs Ctrl}} = 0.17$), when controlled for “base development” and training to the test effect,

but very **short treatment** (2h) → expected (effect on general L2 speaking proficiency by *written practice*)

No difference between DSys and DComp
⇒ In line with observations of form-orientation

Dialogue systems for language learning: typology of systems and measurement of effects



Dialogue systems for language learning

Terms, fields and definition

Rationale

Typology of systems

Types of dialogue-based CALL systems

Technological approaches in research and industry

Past effectiveness

Meta-analysis of previous effectiveness studies

Evaluation of *LanguageHero*

Measuring effects on L2 development

- ▶ Challenges and opportunities

Conclusions

Effects of dialogue-based CALL

Clear effect of dialogue-based CALL practice on L2 development, especially on **vocabulary** acquisition.

Very small effect on **fluency**

Still quite promising that possible to observe an effect on fluency on such a small timeframe.

+ Fine-grained evaluation of fluency metrics via automated comparison

⇒ Methodological innovation

Conclusions

Relative effects of interactivity

Limitation: Strong form-orientation/“exercise mindset” in many participants from both conditions:

Due to school context? age factor? presentation of the instructions?

→ Probably reduced the “interactivity” of the Dialogue system condition a lot.

Limited differences in perception

Small differences in receptive vocabulary learning

No difference in prod. vocabulary and fluency dev.

Perspectives

Dialogue systems for language learning

The question of interactivity and freedom vs. constraints remains open:

uncertainty regarding the pedagogical and motivational advantage of a goal-oriented, fully interactive dialogue system.

well possible that more beneficial to invest more time in pedagogical content and instructional design, and less in complex AI/NLP development (Bibauw, Van den Noortgate, François & Desmet, *under review*)

→ **Trade-off** technological/instructional development

Perspectives

Dialogue systems for language learning

Dialogue has yet to see the breakthroughs other NLP tasks have witnessed from deep learning. → Still much room for improvement (dialogue management, response generation/selection, evaluation...)

For language learning:

- To **compensate for the lack of human-human interaction** (native, teacher and peer interlocutors remain preferable)
- ‘**Constrained by design**’ route seems the most manageable (e.g., Duolingo Bots)
- Prefer it for **well-defined, signposted, conventional interactions** (not open-ended social chat)
- Needs extensive **corrective feedback** and **scaffolding**

Perspectives

Dialogue systems as an L2 research environment

Dialogue systems offer **fully controllable and reproducible interaction:** opportunities to monitor and to alter infinity of details.

Experimental testing (A/B testing) with different types of tasks, instructions, feedback, exposure, reactions...

Thank you!
Merci !
Dank u!
¡Gracias!

Serge Bibauw
serge.bibauw@uclouvain.be

More info: <https://serge.bibauw.be>