


Download these slides   
<http://bit.do/eurocall>



# Interactivity in dialogue-based CALL practice: effects on learners' perception & production

Serge Bibauw<sup>1,2,3</sup>    Thomas François<sup>2</sup>    Piet Desmet<sup>1</sup>

<sup>1</sup> ITEC, imec research group at **KU Leuven**

<sup>2</sup> CENTAL, **UCLouvain**

<sup>3</sup> **Universidad Central del Ecuador**

# Interactivity in dialogue-based CALL practice: effects on learners' perception and L2 fluency

Interactive practice with a **dialogue system**, in this case, integrated within an educational game.



Randomized controlled experiment with young Flemish learners of French (N=215)

# Effects of interactivity of written practice on incidental vocabulary acquisition



## Dialogue systems for language learning

Definition, summarized typology, research questions

## Experimental evaluation methodology

Conditions, population, instruments, treatment

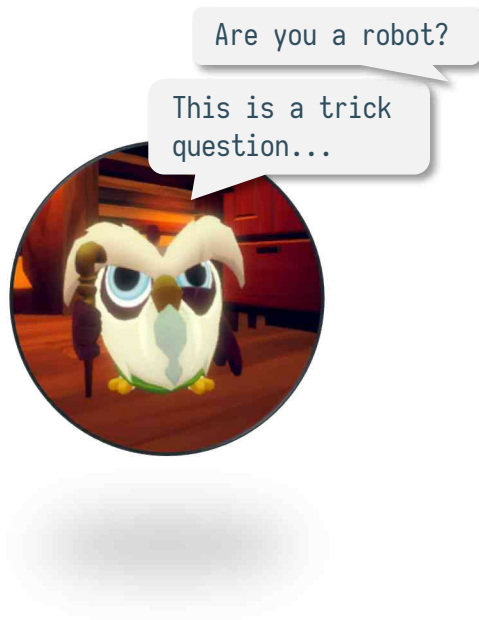
## Results and discussion

Effects on perception

Effects on target vocabulary acquisition

Effects on fluency of L2 production

# Effects of interactivity of written practice on incidental vocabulary acquisition



## Dialogue systems for language learning

Definition, summarized typology, research questions

## Experimental evaluation methodology

Conditions, population, instruments, treatment

## Results and discussion

Effects on perception

Effects on target vocabulary acquisition

Effects on fluency of L2 production

# 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

## Types of systems (Bibauw, François & Desmet, 2019)



### Form-focused systems



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

### Goal-oriented systems



SPELL (Morton, Gunson & Jack, 2012)

# Dialogue systems for language learning

## Types of systems (Bibauw, François & Desmet, 2019)



### 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 L2 research

## Research questions

1. Dialogue-based CALL systems exhibit large variation in terms of **interactivity and freedom** vs. **constraints** of the learner within the dialogue: what does it change?
2. Technologically, it is **considerably easier to “fake” the interaction** by restraining/ignoring the learner: it is worth it pedagogically to implement advanced natural language understanding and dialogue management?



# Effects of interactivity of written practice on incidental vocabulary acquisition



## Dialogue systems for language learning

Definition, summarized typology, research questions

## Experimental evaluation methodology

Conditions, population, instruments, treatment

## Results and discussion

Effects on perception

Effects on target vocabulary acquisition

Effects on fluency of L2 production

Intervention · Dialogue system

*LanguageHero*, dialogue-based game for young learners

15

Codeveloped with Leuven-based start-up *Linguineo*.

(Main) target audience: teenagers (10-14).

Prototype developed for French for Dutch-speaking learners.

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

# 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 image shows two screenshots of a game interface, likely from 'The Legend of Zelda: Breath of the Wild', comparing two dialogue systems. The top screenshot, labeled 'dialogue system', shows a character (an owl) in a room with a dialogue box on the right containing several lines of text. The bottom screenshot, labeled 'dialogue completion', shows the same character and room, but with a dialogue box on the right that has input fields for the player to respond to each line of text. The input fields are labeled with instructions like 'Greet the old owl.', 'Say how it is going.', 'Ask where you are.', 'Ask who he is.', 'Tell your name.', and 'Ask what happened.'. A 'Terminer' button is visible at the bottom right of the dialogue completion interface.

dialogue system

dialogue completion

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 2<sup>nd</sup> 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

( $M_{\text{score}}$  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   | $\alpha$ | 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)

## 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

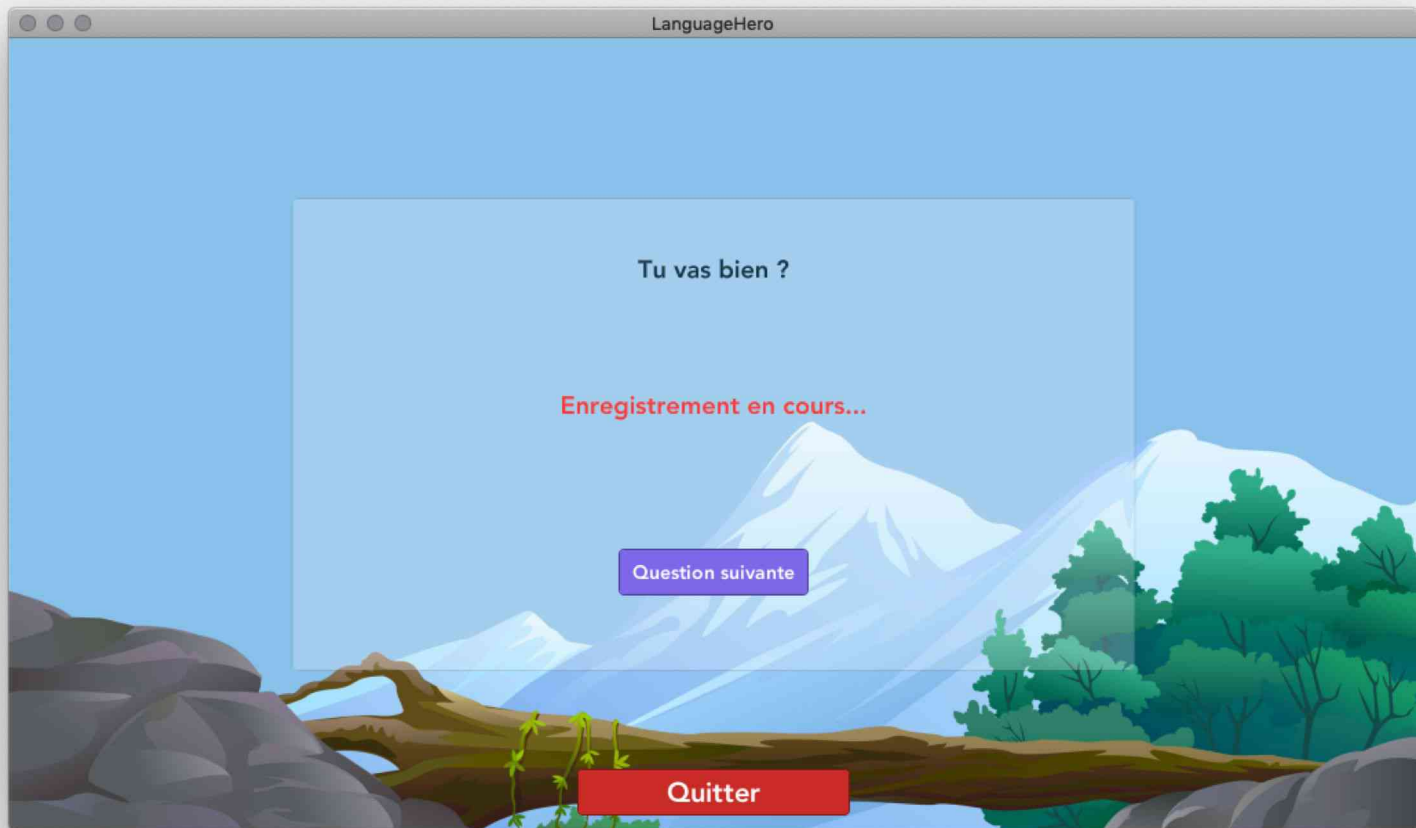
from basic ("How are you?") to questions targeting  
specific communicative functions ("Can you  
describe your French teacher?")

## 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...)

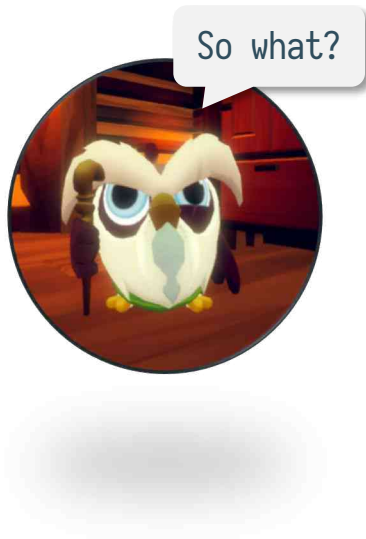
• ~~filled pauses: not good differentiator~~ (Cucchiari et al, 2002...),  
unrelated to other fluency measures (Segalowitz et al 2017)

• ~~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)

# Effects of interactivity of written practice on incidental vocabulary acquisition



## Dialogue systems for language learning

Definition, summarized typology, research questions

## Experimental evaluation methodology

Conditions, population, instruments, treatment

## Results and discussion

Effects on perception

Effects on target vocabulary acquisition

Effects on fluency of L2 production



## 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

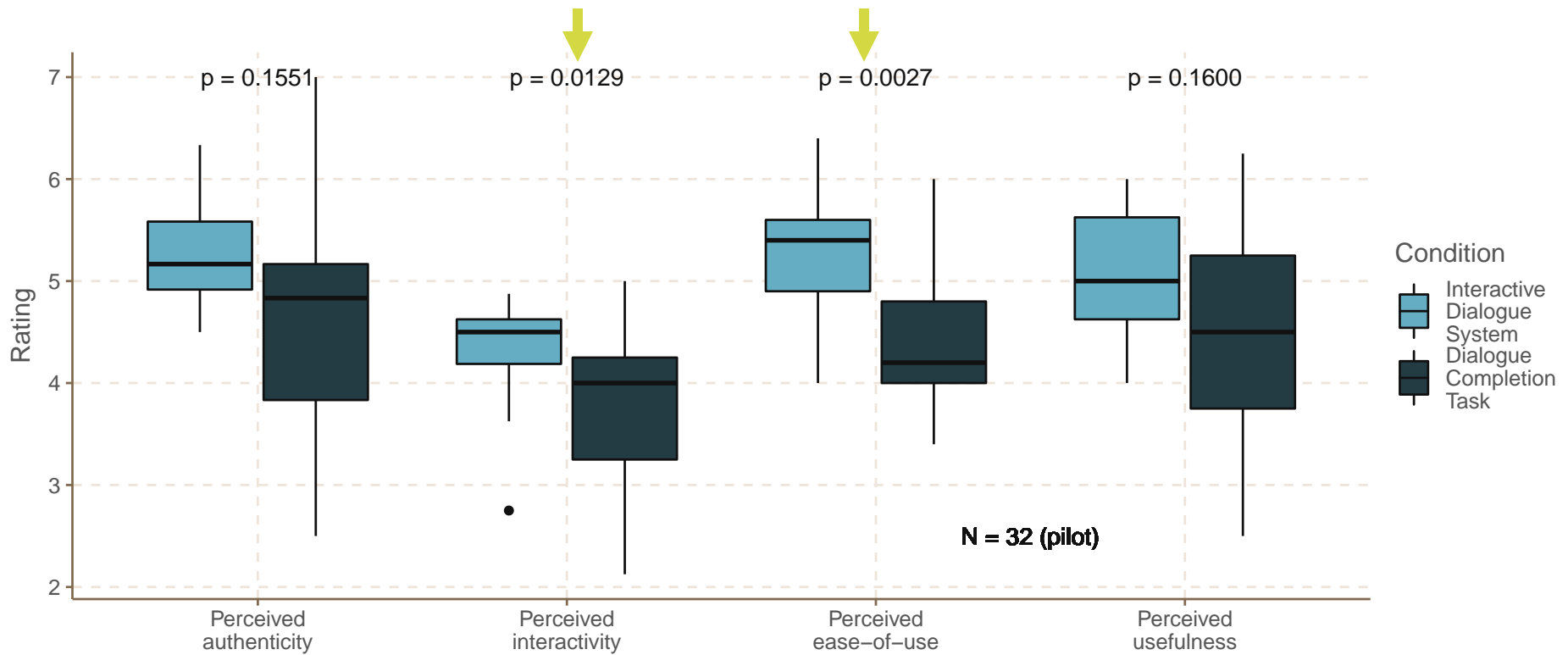
23.7% of messages containing “voluntary noise”

→ Raised ethical issues

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

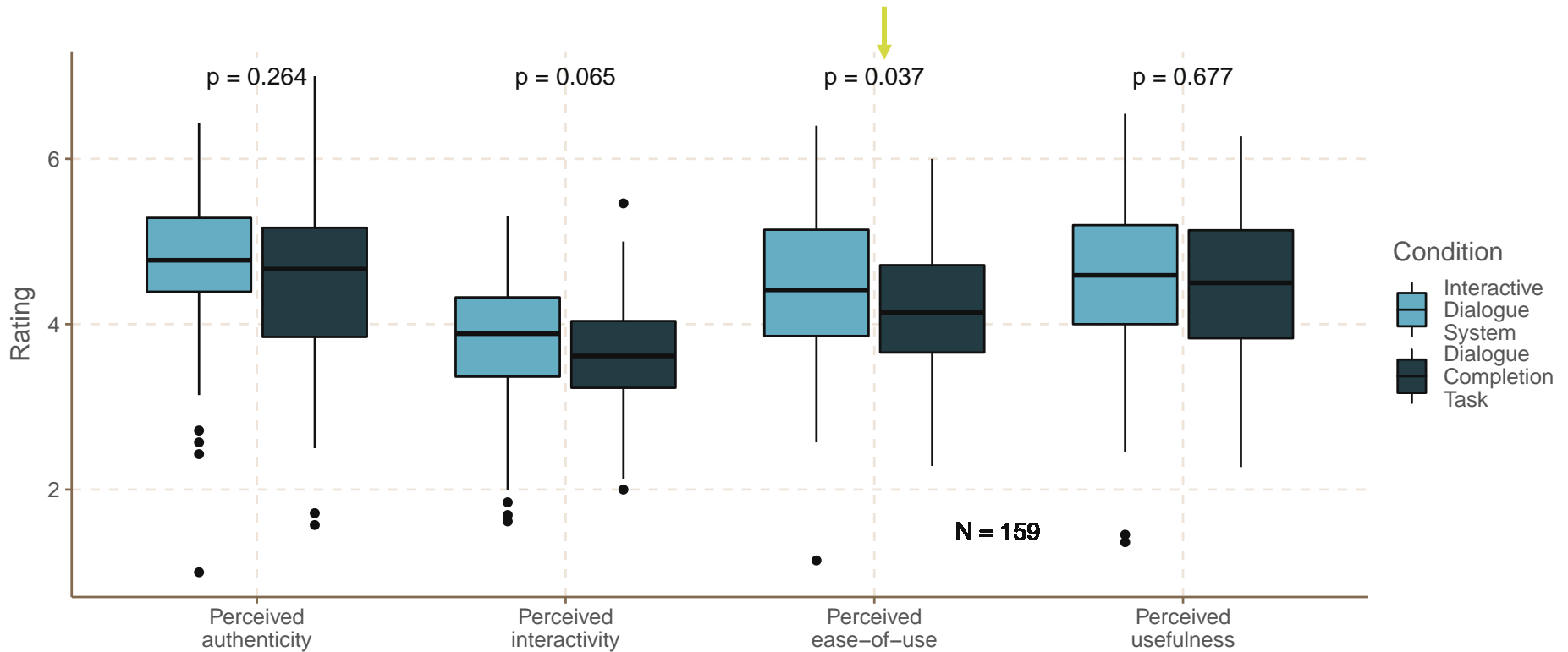
# Results

## Differences of learners' perceptions (pilot only)



# 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**

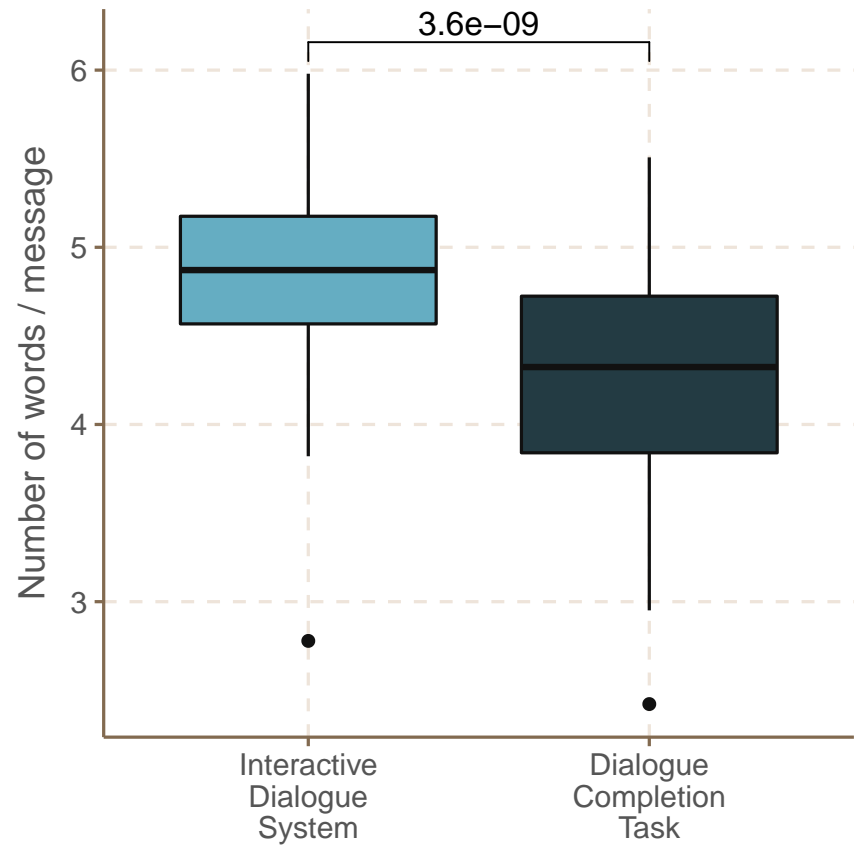
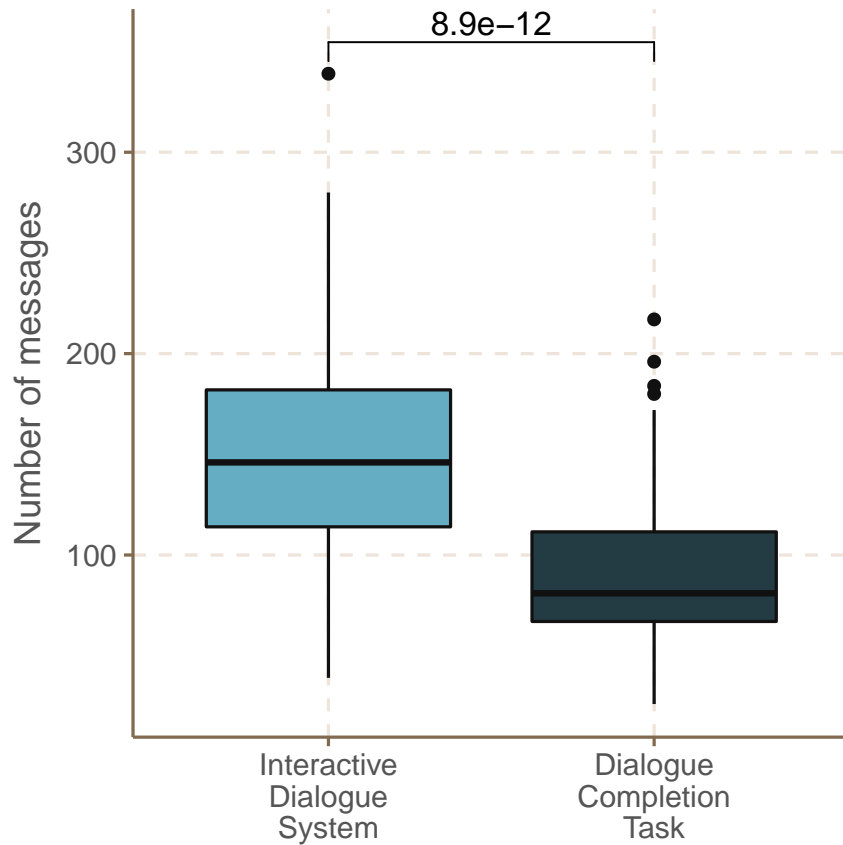
form-orientation behaviour/'exercise mindset'  
among many participants from both conditions:

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

→ lack of perception of task goals as meaningful

# Results

## Quantity of in-task production



# Effects of interactivity of written practice on incidental vocabulary acquisition



## Dialogue systems for language learning

Definition, summarized typology, research questions

## Experimental evaluation methodology

Conditions, population, instruments, treatment

## Results and discussion

Effects on perception

Effects on target vocabulary acquisition

Effects on fluency of L2 production

# Results

## Receptive vocabulary

Very significant increase.

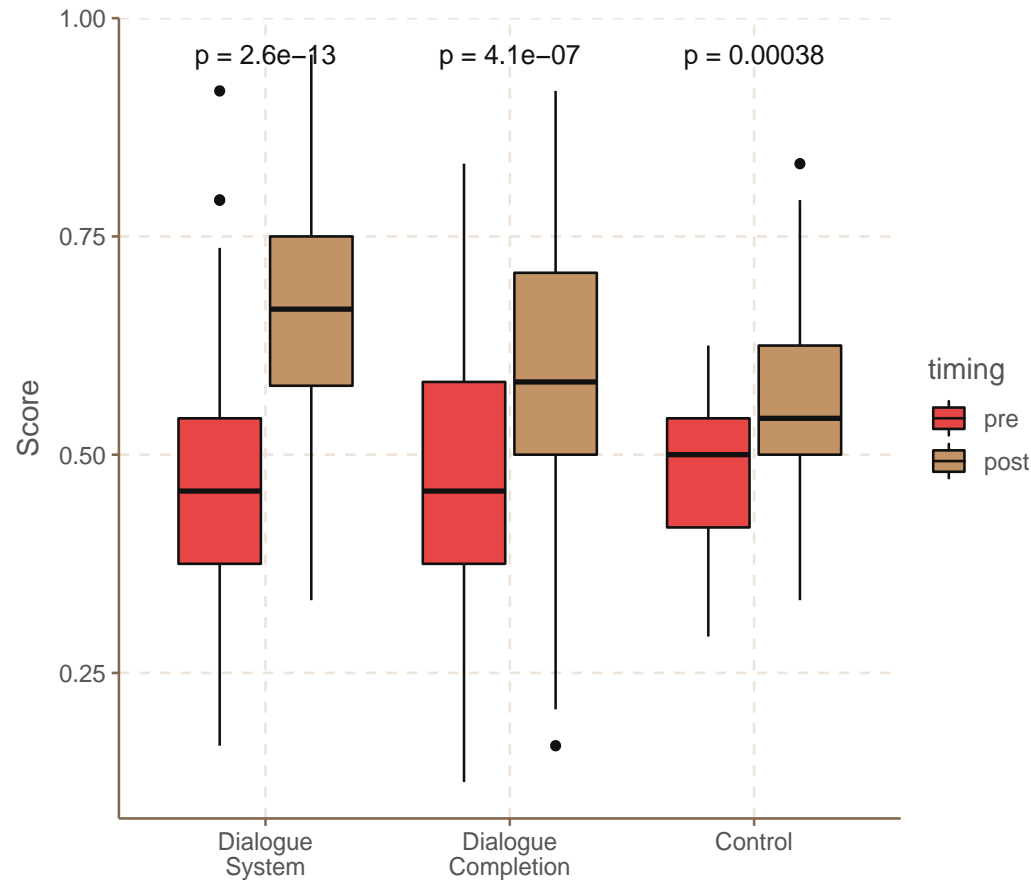
$$d_{\text{DSystem}} = 1.17^{***}$$

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

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

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.

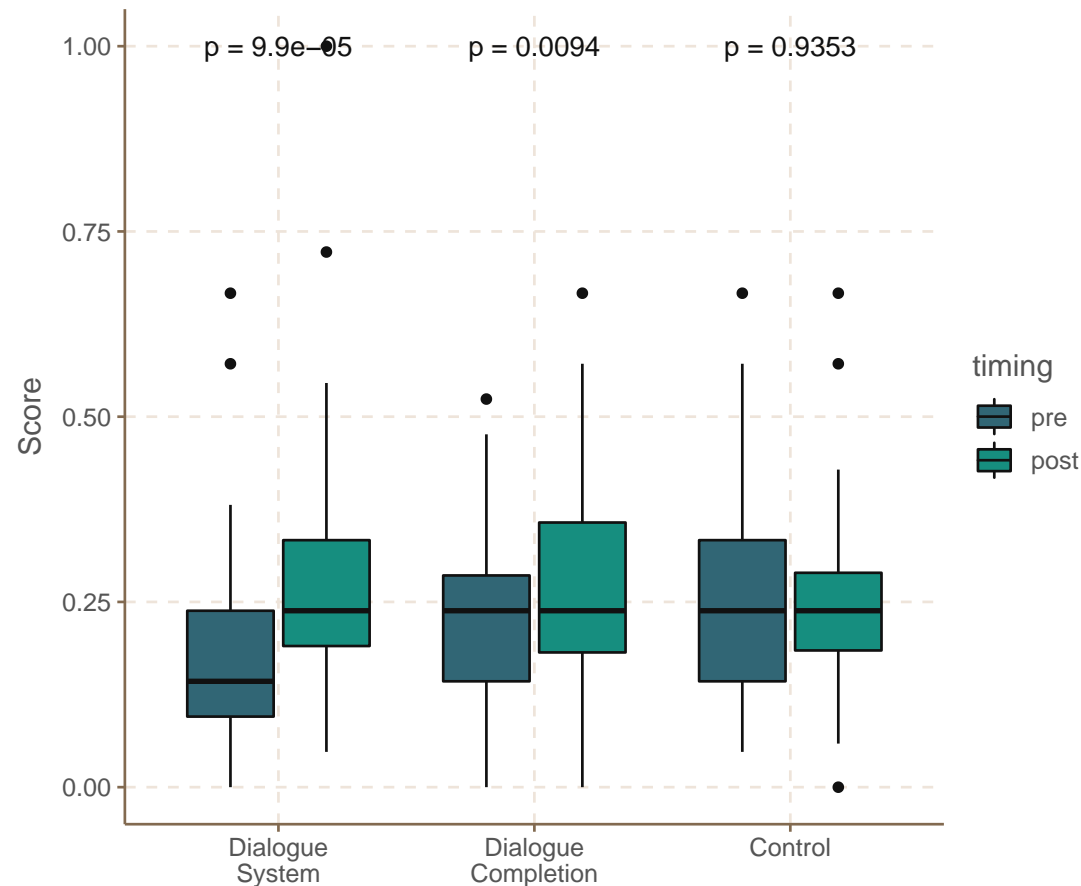
$$d_{\text{DSystem}} = 0.56^{***}$$

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

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

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

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





# Effects of interactivity of written practice on incidental vocabulary acquisition



## Dialogue systems for language learning

Definition, summarized typology, research questions

## Experimental evaluation methodology

Conditions, population, instruments, treatment

## Results and discussion

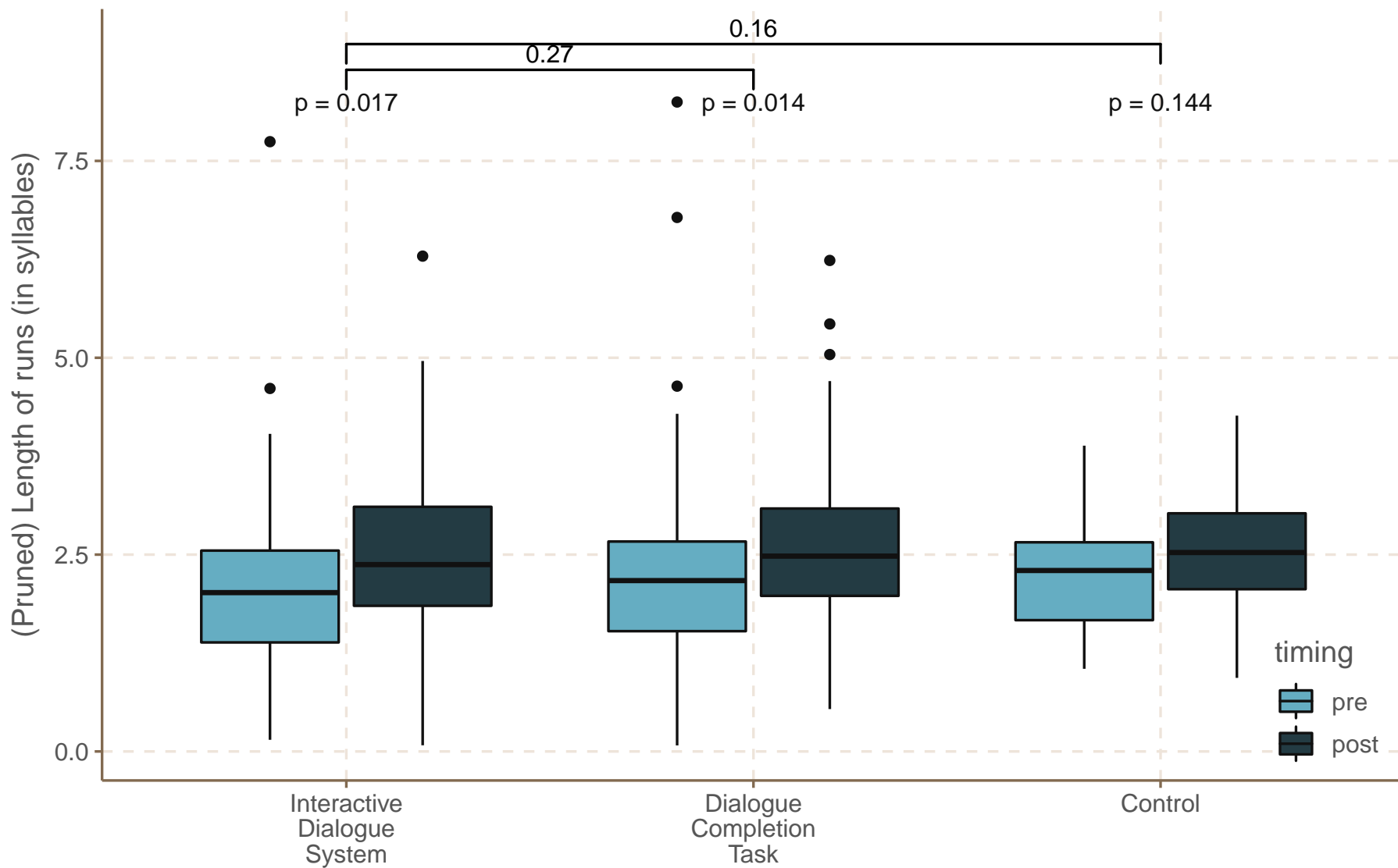
Effects on perception

Effects on target vocabulary acquisition

Effects on fluency of L2 production

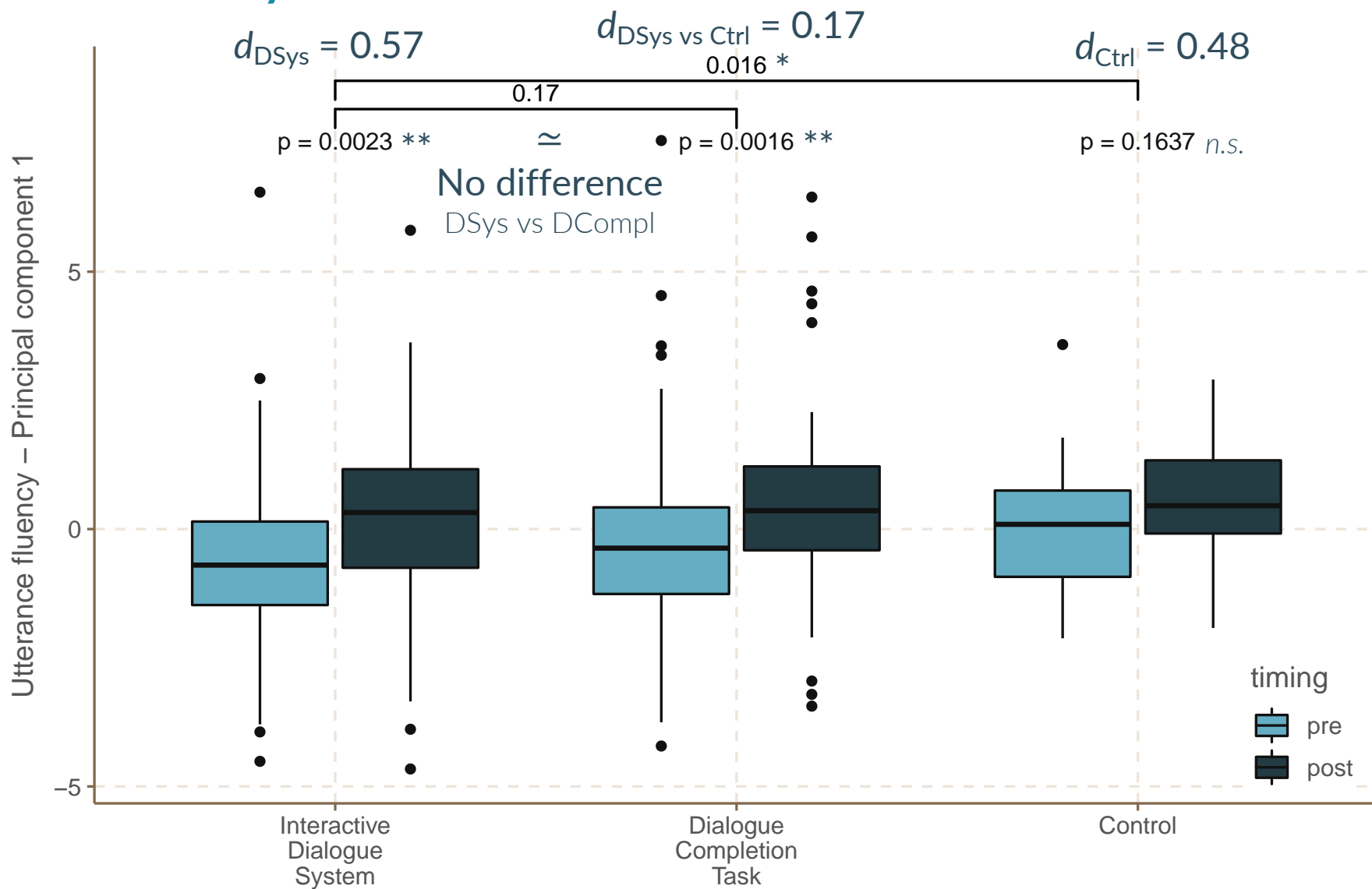
# Results

## Fluency



# Results

## Fluency



# Discussion

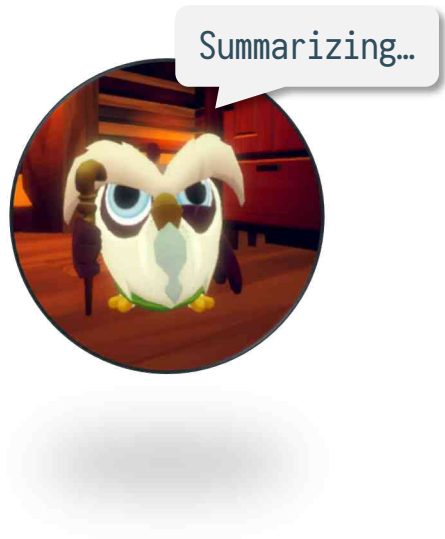
## Fluency

**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

# Effects of interactivity of written practice on incidental vocabulary acquisition



## Dialogue systems for language learning

Definition, summarized typology, research questions

## Experimental evaluation methodology

Conditions, population, instruments, treatment

## Results and discussion

Effects on perception

Effects on target vocabulary acquisition

Effects on fluency of L2 production

# 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, in prep.)



# 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  
Thomas François  
Piet Desmet

Download these slides 

<http://bit.do/eurocall>



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