Effects of **dialogue-based CALL** practice on foreign language learning

Let's look for tickets!

Ok

From New York
To Seatle
Departing on Sept 15
Returning on Sept 19
for 2 Adults

Is this info correct?

Yes

I have found 17 results

See all results

Type your message here...

Serge **Bibauw**

ITEC, KU Leuven · imec
CENTAL, UCLouvain
Universidad Central del Ecuador

ITEC Research Seminar
January 10, 2019

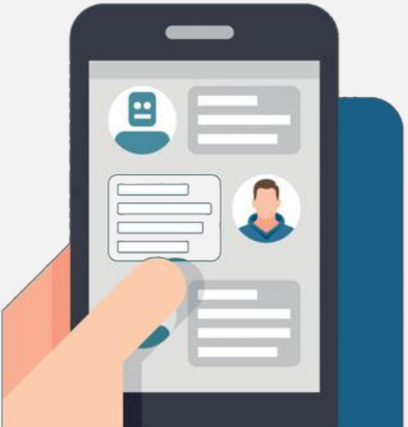
Dialogue-based CALL

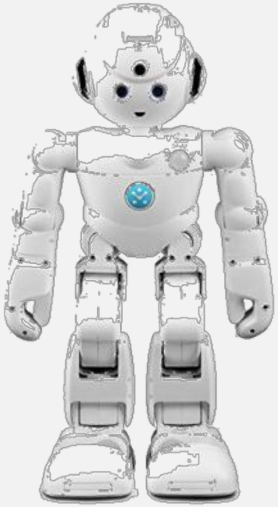
Computer-assisted language learning (CALL)

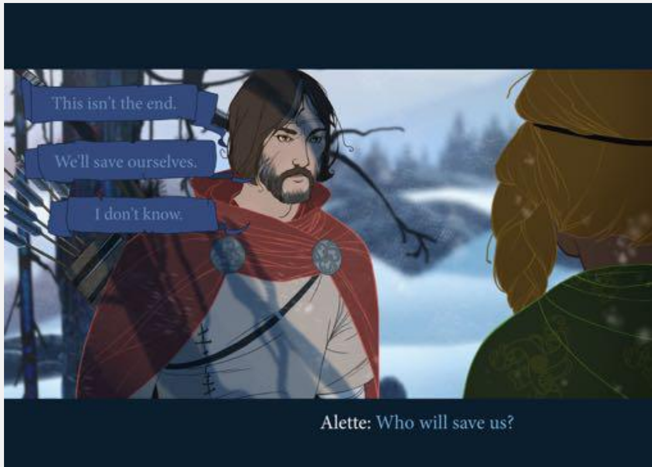
through **dialogues**

with **automated** agents

(chatbot, robot, automated personal assistant, conversational agent, non-player character…)







This isn't the end.

We'll save ourselves.

I don't know.

Alette: Who will save us?

Effects of dialogue-based CALL practice on foreign language learning

Existing systems: state of the art
Research synthesis and conceptual framework

Existing research: effectiveness
Meta-analysis of dialogue-based CALL

LanguageHero: development and evaluation
A randomized controlled evaluation study

Use, perception, and effectiveness results
Preliminary results from the experimental study

Effects of dialogue-based CALL practice on foreign language learning

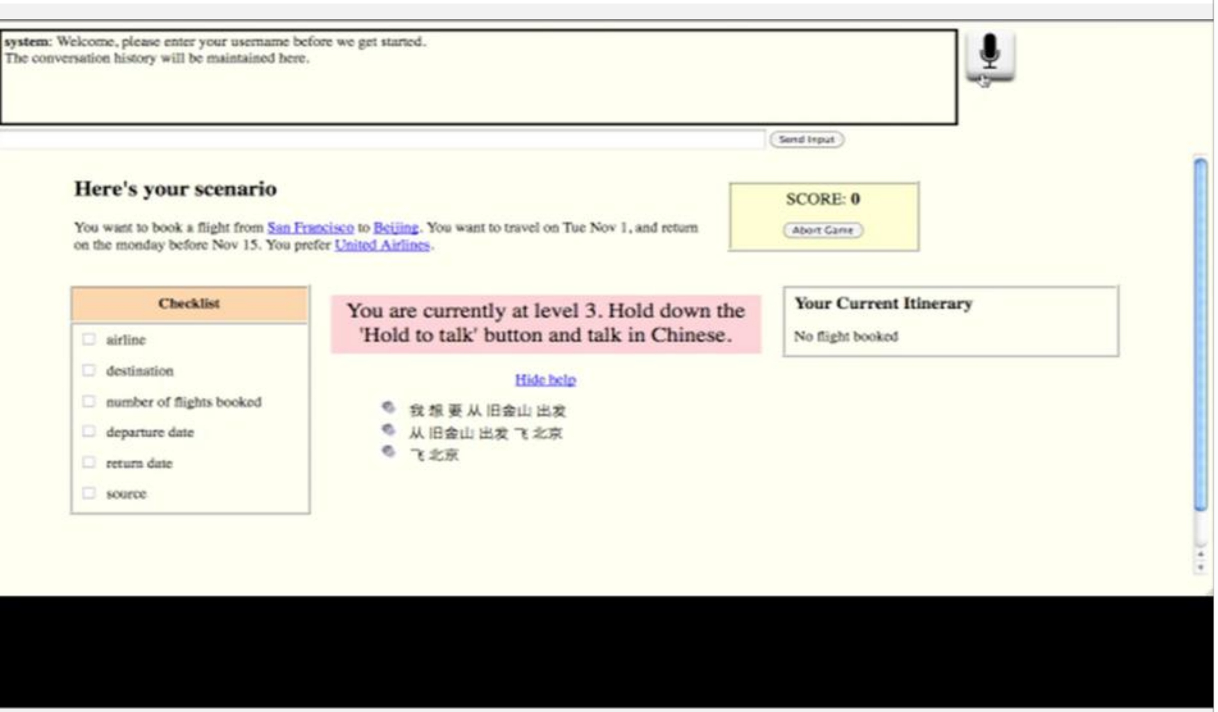
Existing systems: state of the art
Research synthesis and conceptual framework

Existing research: effectiveness
Meta-analysis of dialogue-based CALL

LanguageHero: development and evaluation
A randomized controlled evaluation study

Use, perception, and effectiveness results
Preliminary results from the experimental study

Dialogue-based CALL



system: Welcome, please enter your username before we get started.
The conversation history will be maintained here.

Here's your scenario

You want to book a flight from San Francisco to Beijing. You want to travel on Tue Nov 1, and return on the monday before Nov 15. You prefer United Airlines.

SCORE: **0**

Checklist

- □ airline
- □ destination
- □ number of flights booked
- □ departure date
- □ return date
- □ source

You are currently at level 3. Hold down the 'Hold to talk' button and talk in Chinese.

Hide help

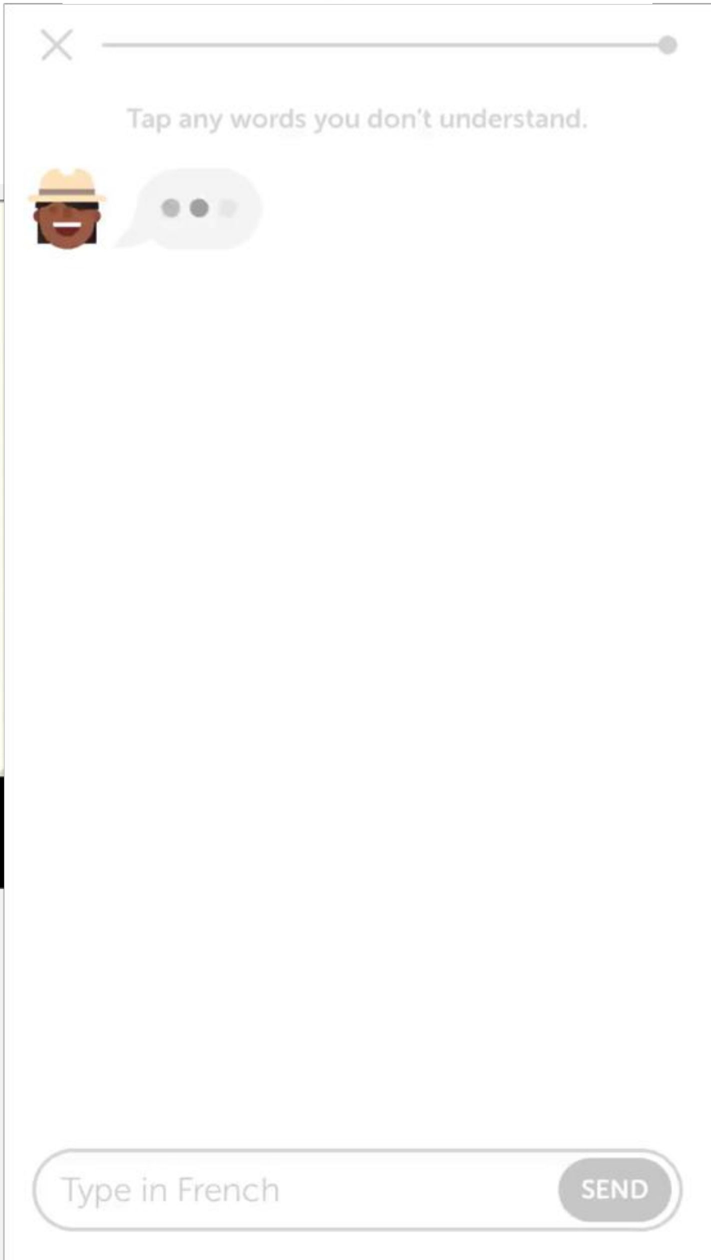
- 我 想 要 从 旧金山 出发
- 从 旧金山 出发 飞 北京
- 飞 北京

Your Current Itinerary

No flight booked

SCILL (Seneff et al, 2007)

duolingo bots



Tap any words you don't understand.

Dialogue-based CALL

Dialogue-based CALL refers to any application or system allowing,

to maintain a dialogue
[immediate, synchronous interaction]
[written or spoken]

with an automated agent
[tutorial CALL (≠ CMC)]

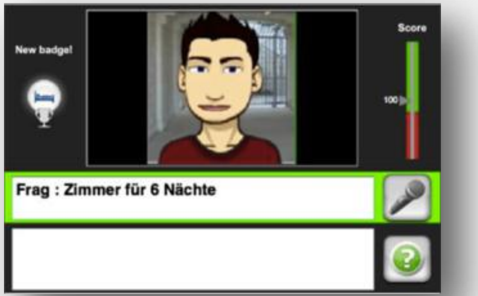
for language learning purposes.

Bibauw, François & Desmet, 2015 (EUROCALL Proceedings); Bibauw, François & Desmet, 2019

Dialogue-based CALL

Typology of systems

(Bibauw et al, *2019*)



Frag : Zimmer für 6 Nächte

Form-focused dialogue systems

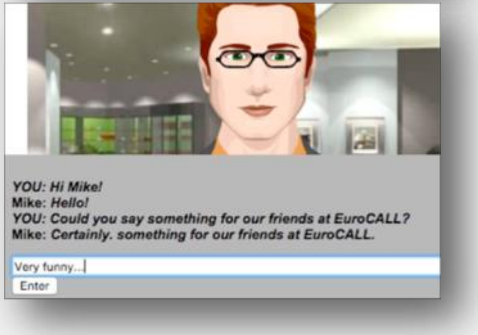
Explicit constraints on meaning,
focus on form/forms

e.g., ICALL intelligent language tutors, and Computer-assisted pronunciation training (CAPT) systems

Goal-oriented dialogue systems

Contextual constraints (task, situated conversation…),
mostly focus on meaning and interaction

e.g., Conversational agents in virtual worlds



YOU: Hi Mike!
Mike: Hello!
YOU: Could you say something for our friends at EuroCALL?
Mike: Certainly. something for our friends at EuroCALL.

Very funny...

Reactive dialogue systems

Free, user-initiated, open-ended dialogue

e.g., Chatbots, and personal assistants

Here, simplified typology (left out *Narrative systems*)

Effects of dialogue-based CALL practice on foreign language learning

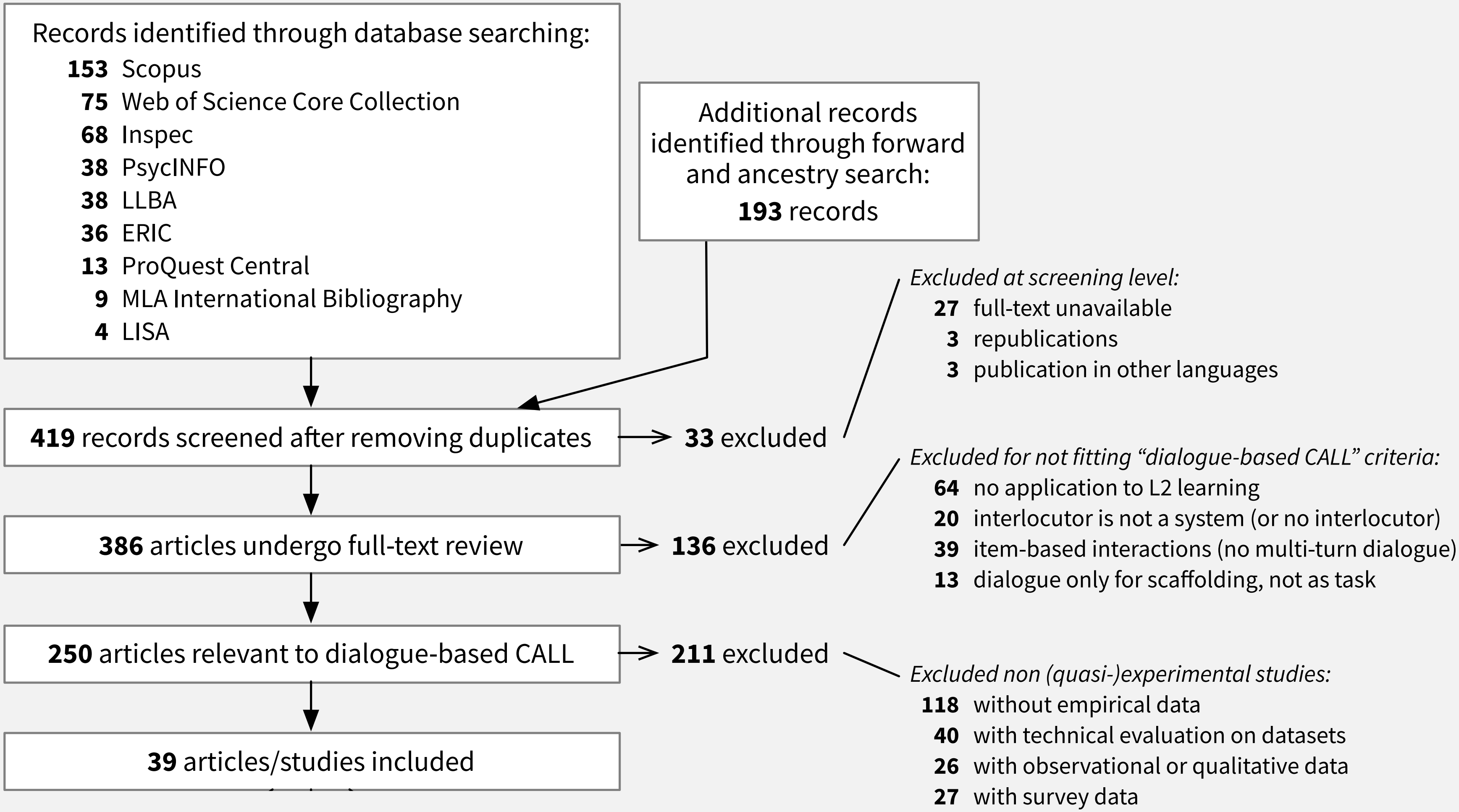
Existing systems: state of the art
Research synthesis and conceptual framework

Existing research: effectiveness
Meta-analysis of dialogue-based CALL

LanguageHero: development and evaluation
A randomized controlled evaluation study

Use, perception, and effectiveness results
Preliminary results from our experimental study

Meta-analysis
Inclusion/exclusion process



Records identified through database searching:

- **153** Scopus
- **75** Web of Science Core Collection
- **68** Inspec
- **38** PsycINFO
- **38** LLBA
- **36** ERIC
- **13** ProQuest Central
- **9** MLA International Bibliography
- **4** LISA

Additional records identified through forward and ancestry search:
193 records

419 records screened after removing duplicates → **33** excluded

Excluded at screening level:

- **27** full-text unavailable
- **3** republications
- **3** publication in other languages

386 articles undergo full-text review → **136** excluded

Excluded for not fitting "dialogue-based CALL" criteria:

- **64** no application to L2 learning
- **20** interlocutor is not a system (or no interlocutor)
- **39** item-based interactions (no multi-turn dialogue)
- **13** dialogue only for scaffolding, not as task

250 articles relevant to dialogue-based CALL → **211** excluded

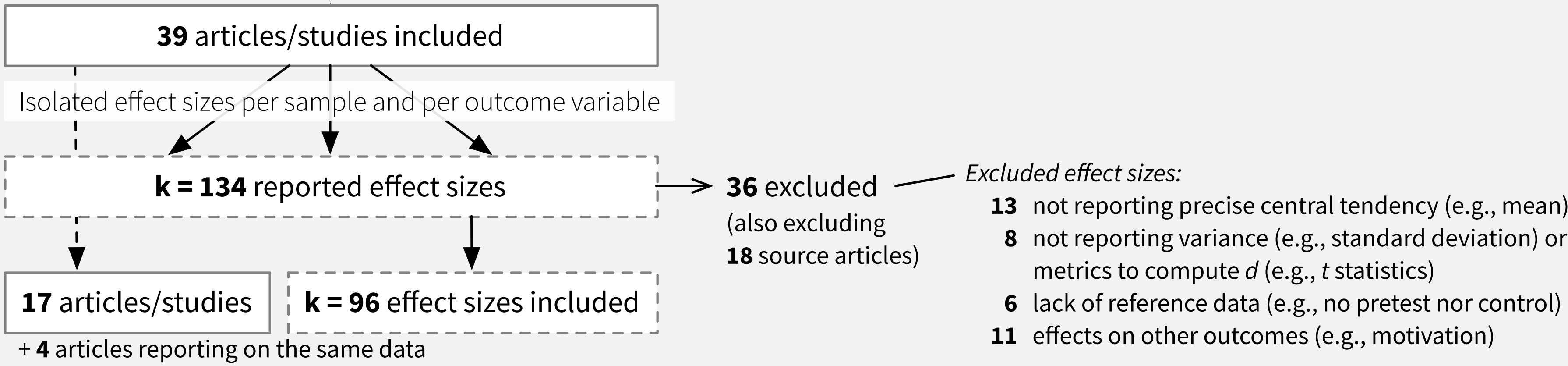
Excluded non (quasi-)experimental studies:

- **118** without empirical data
- **40** with technical evaluation on datasets
- **26** with observational or qualitative data
- **27** with survey data

39 articles/studies included

Meta-analysis

Inclusion of individual effect sizes



39 articles/studies included

Isolated effect sizes per sample and per outcome variable

k = 134 reported effect sizes

36 excluded (also excluding **18** source articles)

Excluded effect sizes:

13 not reporting precise central tendency (e.g., mean)

8 not reporting variance (e.g., standard deviation) or metrics to compute *d* (e.g., *t* statistics)

6 lack of reference data (e.g., no pretest nor control)

11 effects on other outcomes (e.g., motivation)

17 articles/studies

+ **4** articles reporting on the same data

k = 96 effect sizes included

k = 96 effect sizes

Meta-analysis
Effect size calculation

Effect size: standardized measure of the observed (here, learning) effect

Usually, in SLA/CALL:

Standardized Mean Difference

Cohen's d $(M_{\text{post}} - M_{\text{pre}} / \text{SD}_{\text{pooled}})$

Hedge's g

	Exp. Grp	Control
	M (*sd*)	*M* (*sd*)
Post	61 (6.2)	57 (7.4)

EC

	M (*sd*)
Pre	56 (4.3)
Post	61 (6.2)

PP

	Exp. Grp	Control
	M (*sd*)	*M* (*sd*)
Pre	56 (4.3)	54 (5.6)
Post	61 (6.2)	57 (7.4)

ECPP

Standardized Mean Change

Meta-analysis
A comparable effect size metrics

Morris & DeShon (2002) offer a solution: 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:

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

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

Meta-analysis
Multilevel modeling

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 loosing thus all the information on distinct implementations

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

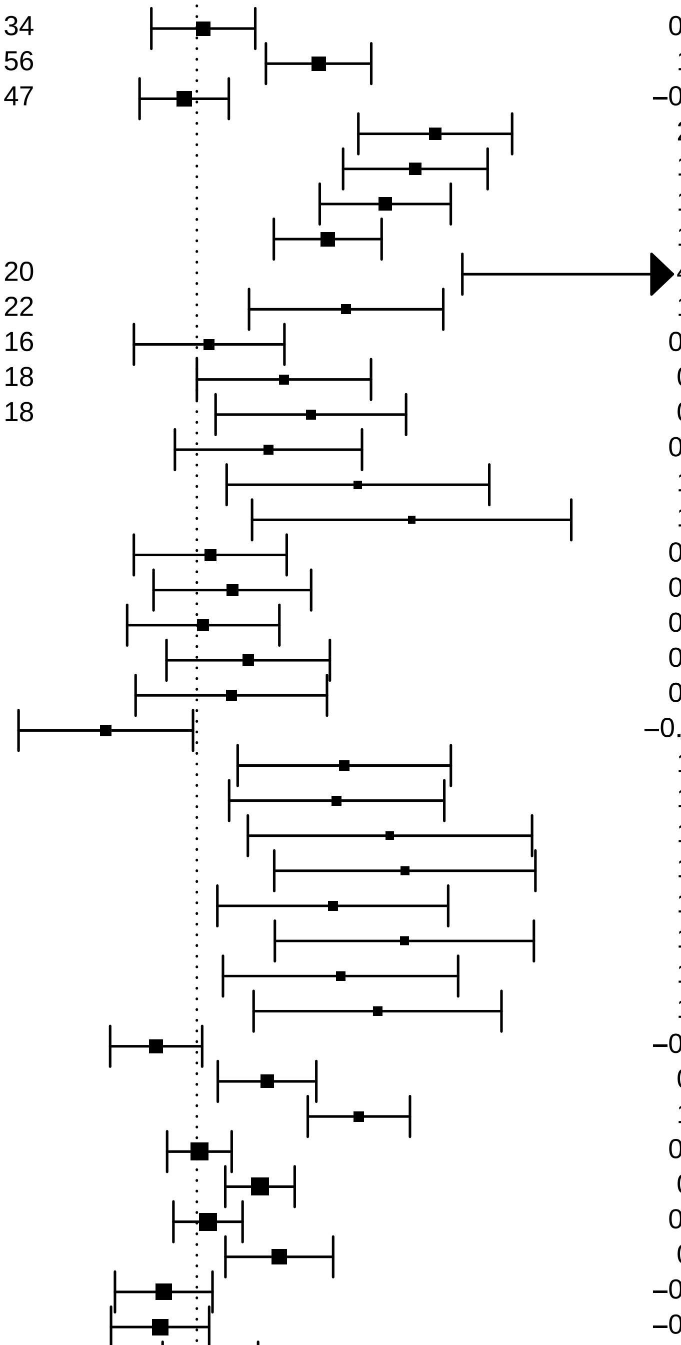
Multilevel modelling

> 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

Reference	Features of single effect	n_E	n_C	d [95% CI]
Jia et al 2013	(sample Huiwen JHS)	37	34	0.05 [–0.38, 0.49]
	(sample Huojia N1 SHS)	56	56	1.02 [0.58, 1.47]
	(sample Jingxian JHS)	48	47	–0.11 [–0.48, 0.27]
Taguchi et al 2017	... gap–filling test *post	30		2.00 [1.36, 2.65]
	... gap–filling test *delayed	30		1.84 [1.23, 2.44]
	... multiple choice test *post	30		1.58 [1.03, 2.13]
	... multiple choice test *delayed	30		1.10 [0.65, 1.55]
Kim 2016	(A1 sample)	20	20	4.60 [2.23, 6.98]
	(A2 sample)	22	22	1.25 [0.44, 2.07]
	(B1 sample)	21	16	0.10 [–0.53, 0.74]
Petersen 2010	... QFT, morphology score	19	18	0.73 [0.00, 1.46]
	... QFT, syntax score	19	18	0.96 [0.16, 1.76]
Harless et al 1999	... listening comp.	9		0.60 [–0.18, 1.39]
	... reading comp.	9		1.35 [0.25, 2.46]
	... speaking prof.	9		1.81 [0.46, 3.15]
Hassani et al 2016	... Grammatical errors/sentence	10		0.11 [–0.53, 0.76]
	... Nb of proper replies	10		0.30 [–0.36, 0.96]
	... Phonation time/letter	10		0.05 [–0.59, 0.69]
	... Automatic prof. score	10		0.43 [–0.26, 1.12]
Lee et al 2011a	(A1) ... listening compr.	10		0.29 [–0.51, 1.09]
	(A2) ... listening compr.	11		–0.77 [–1.50, –0.03]
	(A1) ... hol. grammar rating	10		1.24 [0.34, 2.13]
	(A2) ... hol. grammar rating	11		1.18 [0.27, 2.08]
	(A1) ... hol. pronunciation rating	10		1.62 [0.43, 2.82]
	(A2) ... hol. pronunciation rating	11		1.75 [0.65, 2.85]
	(A1) ... hol. communicative ability rating	10		1.14 [0.17, 2.11]
	(A2) ... hol. communicative ability rating	11		1.74 [0.66, 2.83]
	(A1) ... hol. vocabulary rating	10		1.21 [0.22, 2.20]
	(A2) ... hol. vocabulary rating	11		1.52 [0.48, 2.56]
Lee et al 2014a	... nb of grammatical errors	25		–0.34 [–0.73, 0.04]
	... nb of words	25		0.59 [0.18, 1.00]
Noh et al 2012		40		1.36 [0.93, 1.79]
Chiu et al 2007	(Engl. major) ... DCT, comprehensibility	29		0.02 [–0.25, 0.29]
	(not Engl. major) ... DCT, comprehensibility	20		0.53 [0.24, 0.82]
	(Engl. major) ... DCT, use of speech acts	29		0.09 [–0.20, 0.38]
	(not Engl. major) ... DCT, use of speech acts	20		0.69 [0.24, 1.15]
Rosenthal... et al 2016	Virtual agent, prerecorded voice	22		–0.28 [–0.69, 0.13]
	Virtual agent, TTS voice	22		–0.31 [–0.72, 0.10]



Results
Summary effect

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

d = 0.602 ***
95% CI = [0.373, 0.831]

= Medium effect (Plonsky & Oswald, 2014)

FYI, if converted/computed as *change* metrics:
d_{change} = 0.658 *** [0.414, 0.901]

Immediate effect only (no delayed posttests, k = 73):
d_{raw} = 0.627 *** [0.390, 0.863]

Results & discussion

Summary effect compared to CALL/SLA

Global effect close to the median of meta-analyses in CALL/SLA (Plonsky & Oswald, 2014)

- ≳ game-based learning (d = .53, Chiu et al, 2012)
- ≲ CALL in general (d = .84, Plonsky & Ziegler, 2016)

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

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

Slightly inferior, but logical:

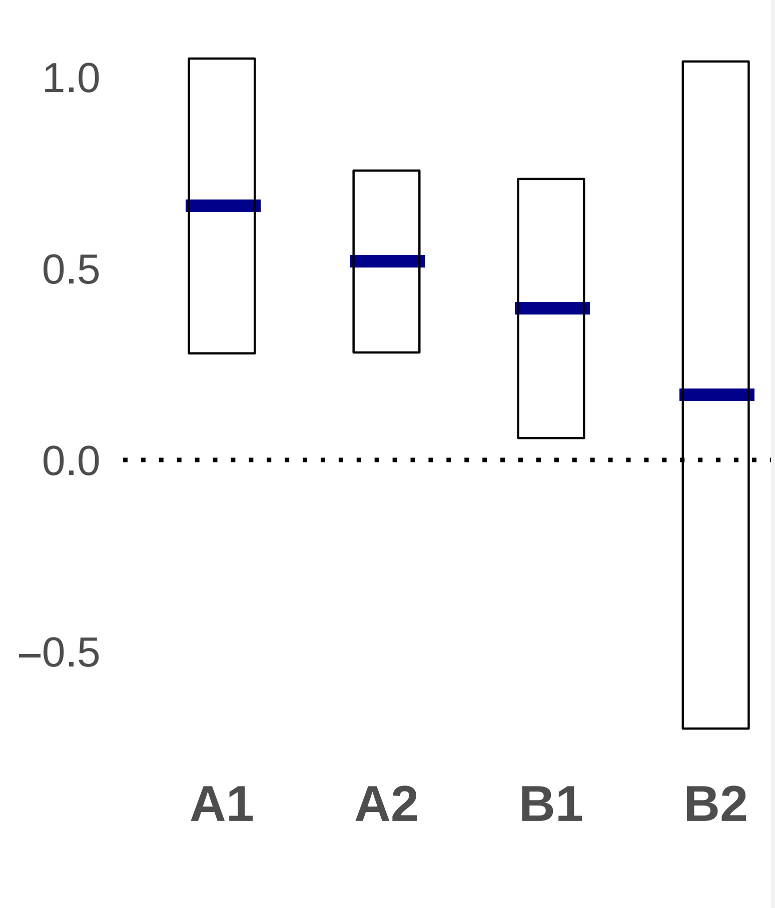
- Human interlocutors remain the gold standard!
- Outcome variables often very ambitious (holistic proficiency…) and treatment duration often very reduced (≤ 3h)

Moderator analysis

Participants: L2 proficiency

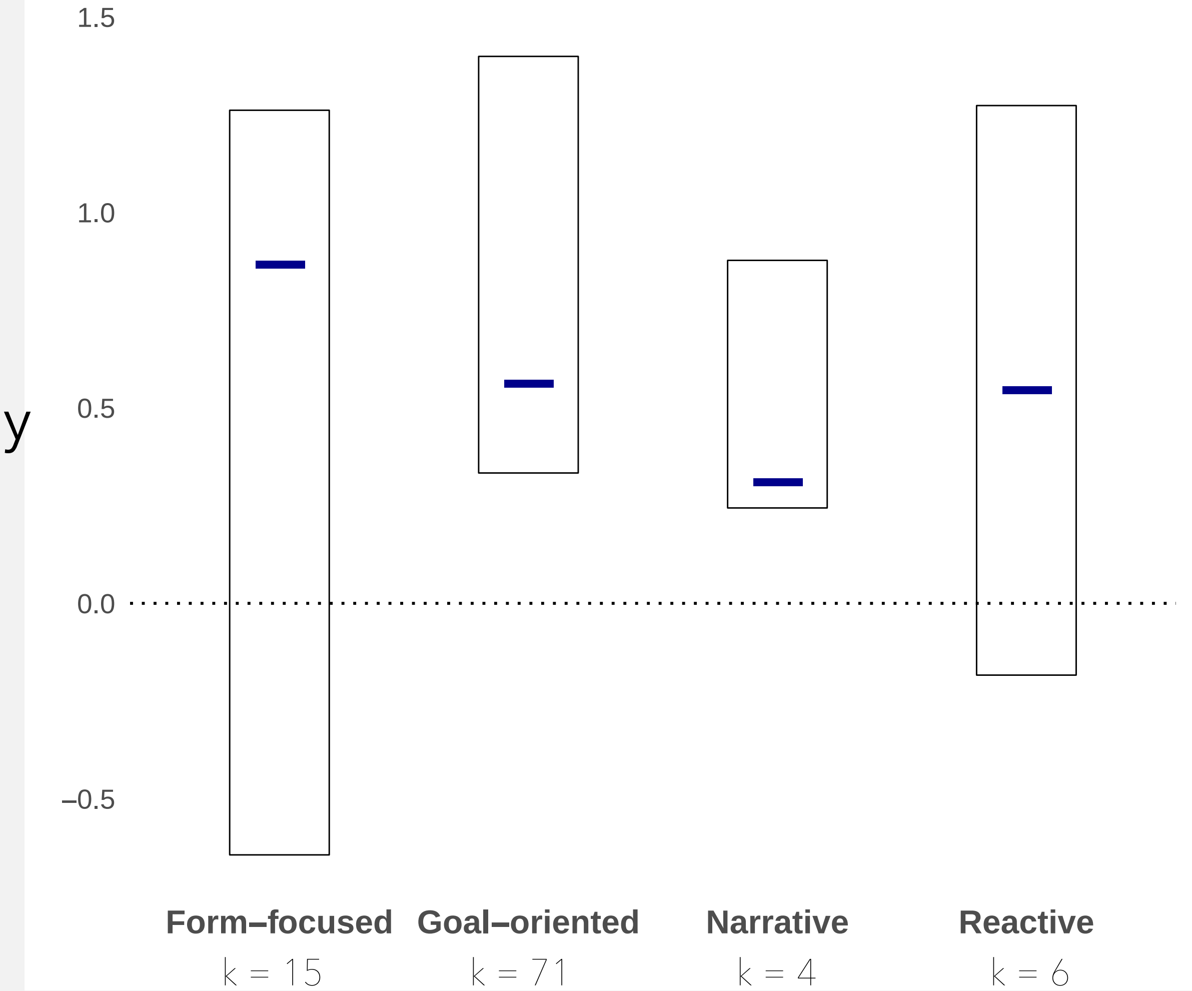
Beginners benefit more from these systems than advanced learners

Very significant difference and predictor
(Q(df=3) = 10.8, p < .001)



Moderator analysis
System: DBCALL type

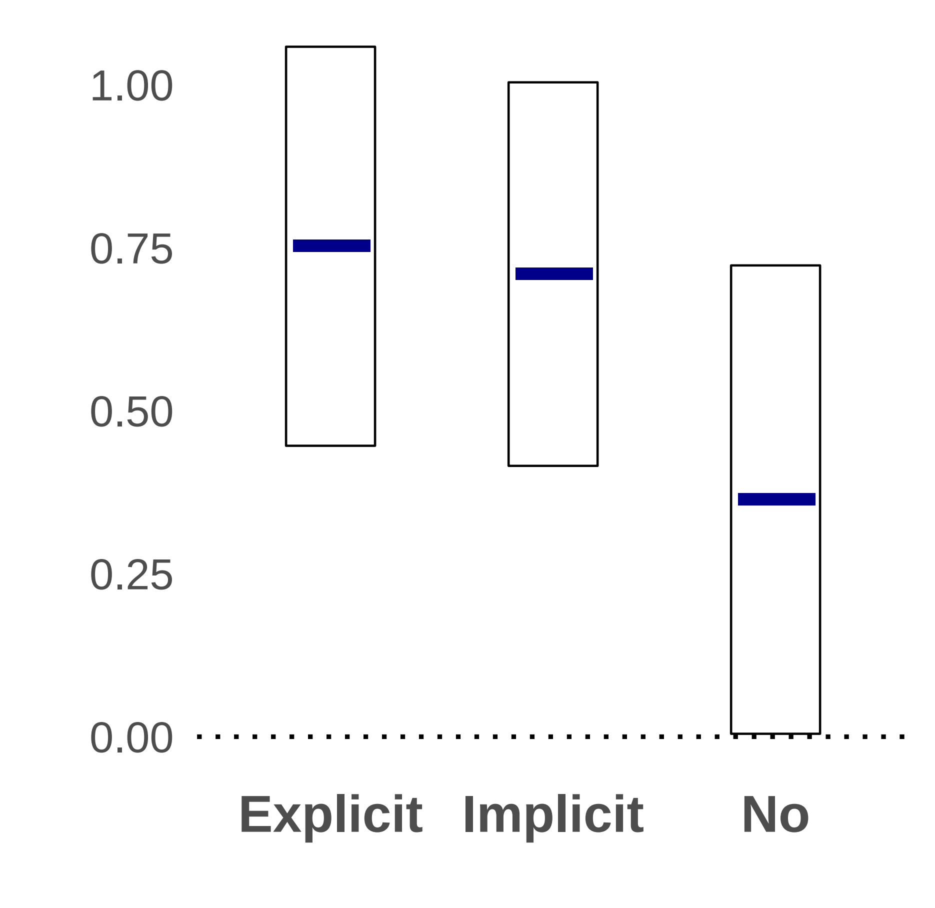
Goal-oriented systems
seem to be globally more effective.



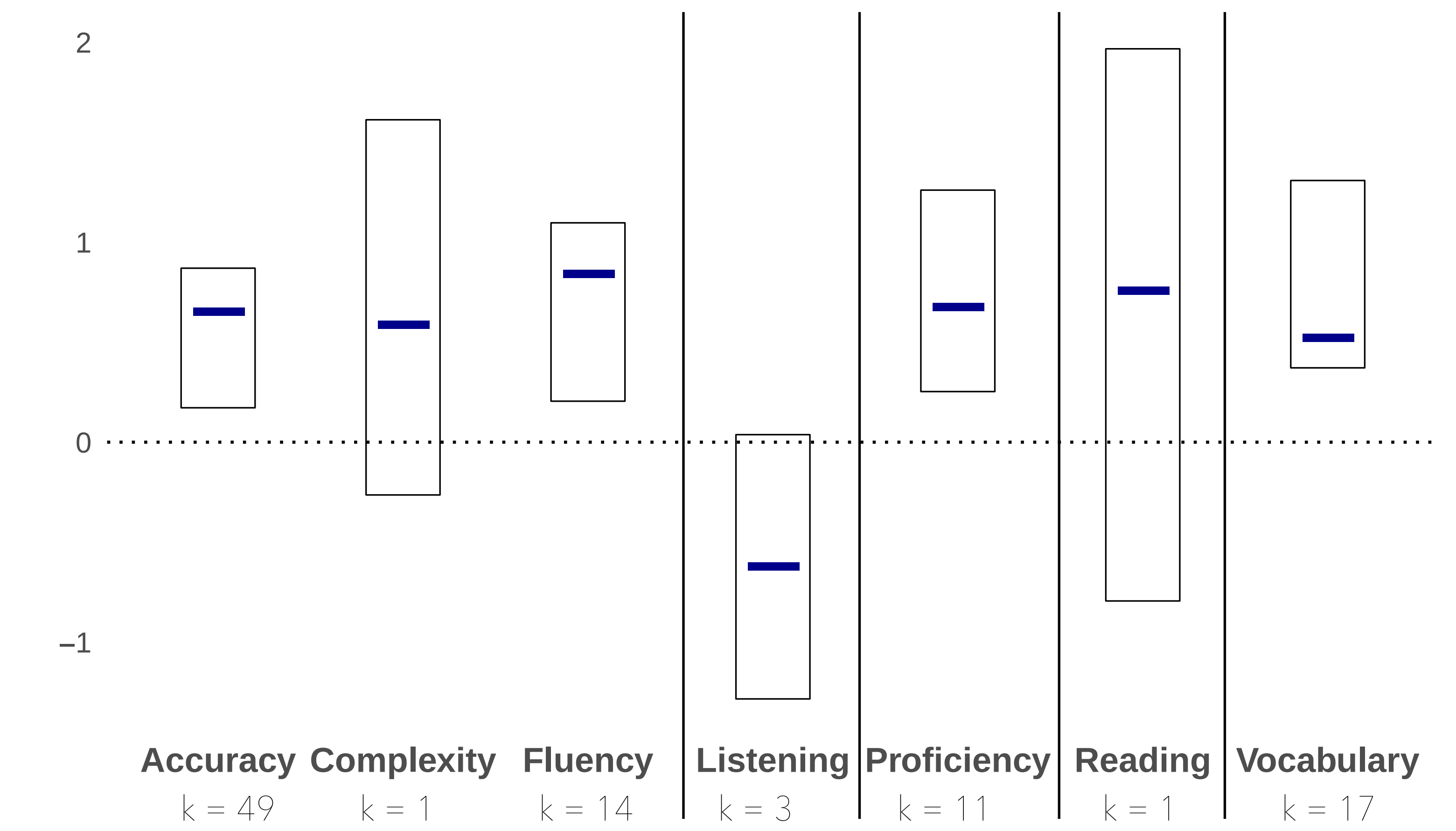
Moderator analysis
System: Corrective feedback

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

If binary (Y/N CF):
QM(df = 1) = 2.53, p-val = 0.111

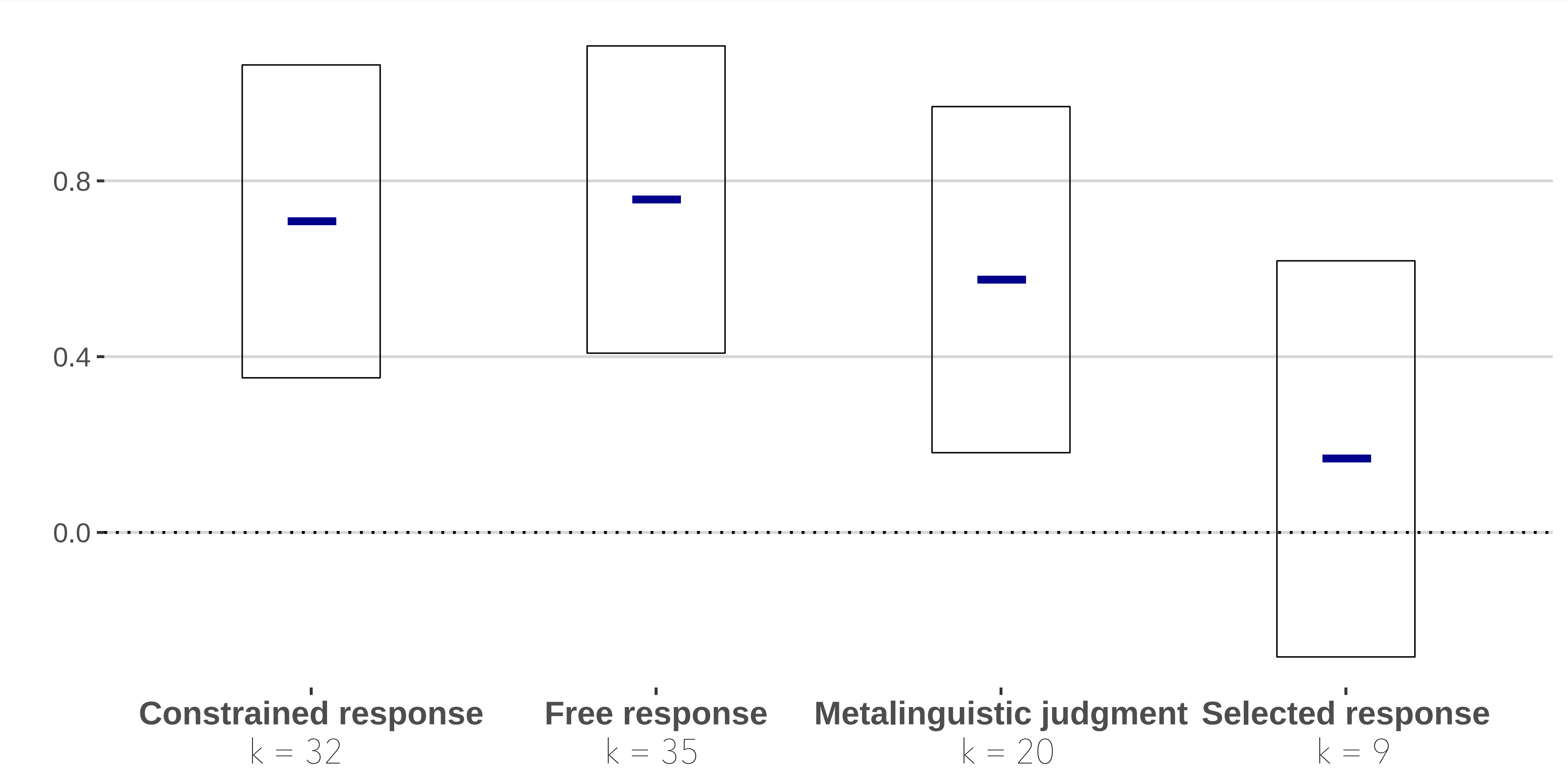


Moderator analysis
Outcome: Dimensions



More promising effects on **fluency**

Moderator analysis
Outcome: Instrument (Norris & Ortega, 2000; Spada & Tomita, 2010)



$Q_{M\ (df = 3)} = 7.77,\ p = 0.051$

Effects of dialogue-based CALL practice on foreign language learning

Existing systems: state of the art
Research synthesis and conceptual framework

Existing research: effectiveness
Meta-analysis of dialogue-based CALL

LanguageHero: development and evaluation
A randomized controlled evaluation study

Use, perception, and effectiveness results
Preliminary results from the experimental study

LanguageHero: Task-based dialogue-based CALL game

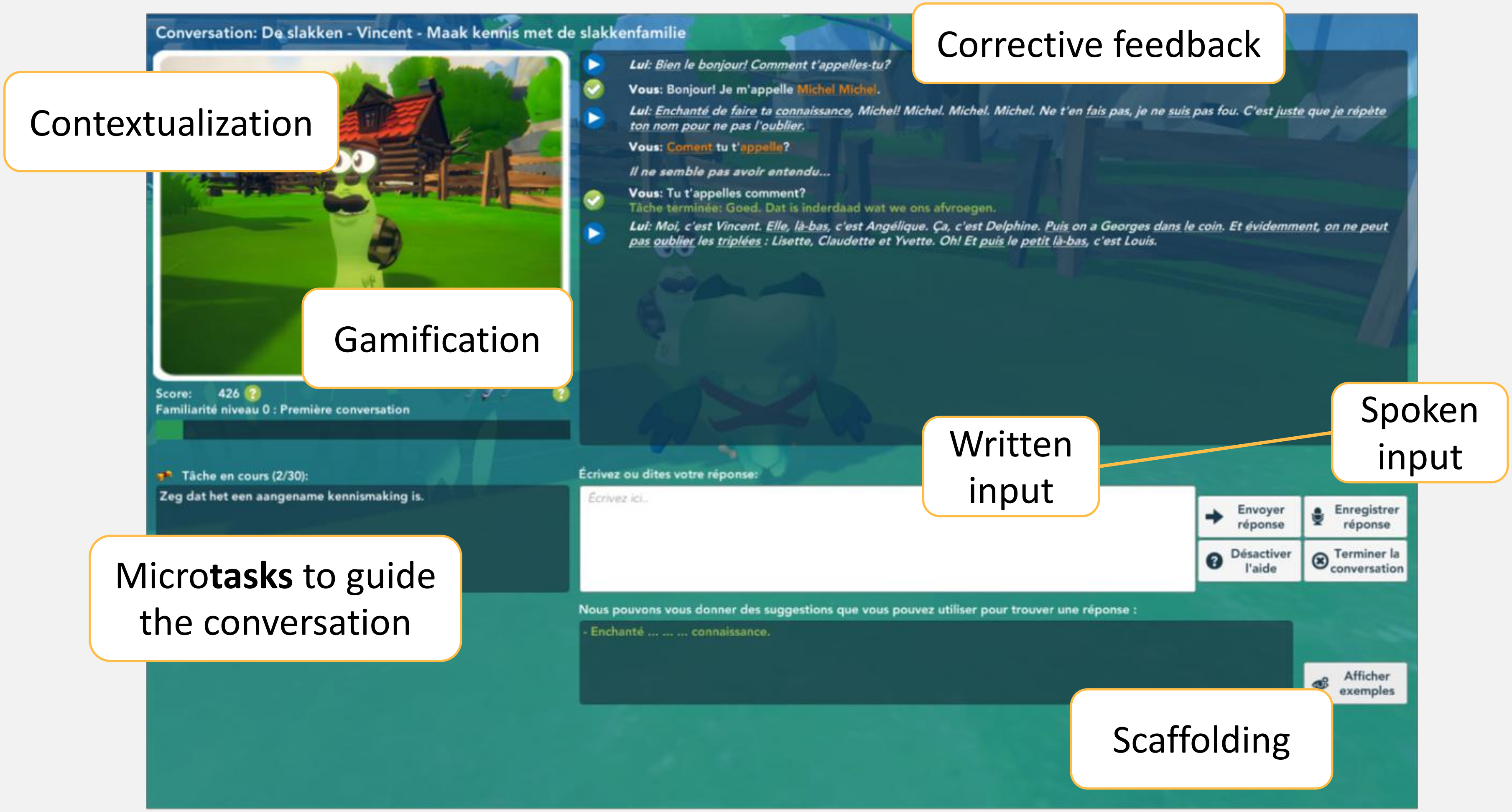
Developed by Leuven-based startup Linguineo

Target audience: teenagers (10-14)

Prototype developed for French for Dutch-speaking learners (other target and origin languages in the future)

3D “walking-around” game with task-based free dialogic written/spoken interaction

Task-based dialogue-based CALL game
LanguageHero



Conversation: De slakken - Vincent - Maak kennis met de slakkenfamilie

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

Vous: Bonjour! Je m'appelle Michel Michel.

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

Vous: Coment tu t'appelle?

Il ne semble pas avoir entendu...

Vous: Tu t'appelles comment?
Tâche terminée: Goed. Dat is inderdaad wat we ons afvroegen.

Lui: 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.

Score: 426
Familiarité niveau 0 : Première conversation

Tâche en cours (2/30):
Zeg dat het een aangename kennismaking is.

Écrivez ou dites votre réponse:

Écrivez ici..

Nous pouvons vous donner des suggestions que vous pouvez utiliser pour trouver une réponse :

- Enchanté ... ... connaissance.

Corrective feedback

Contextualization

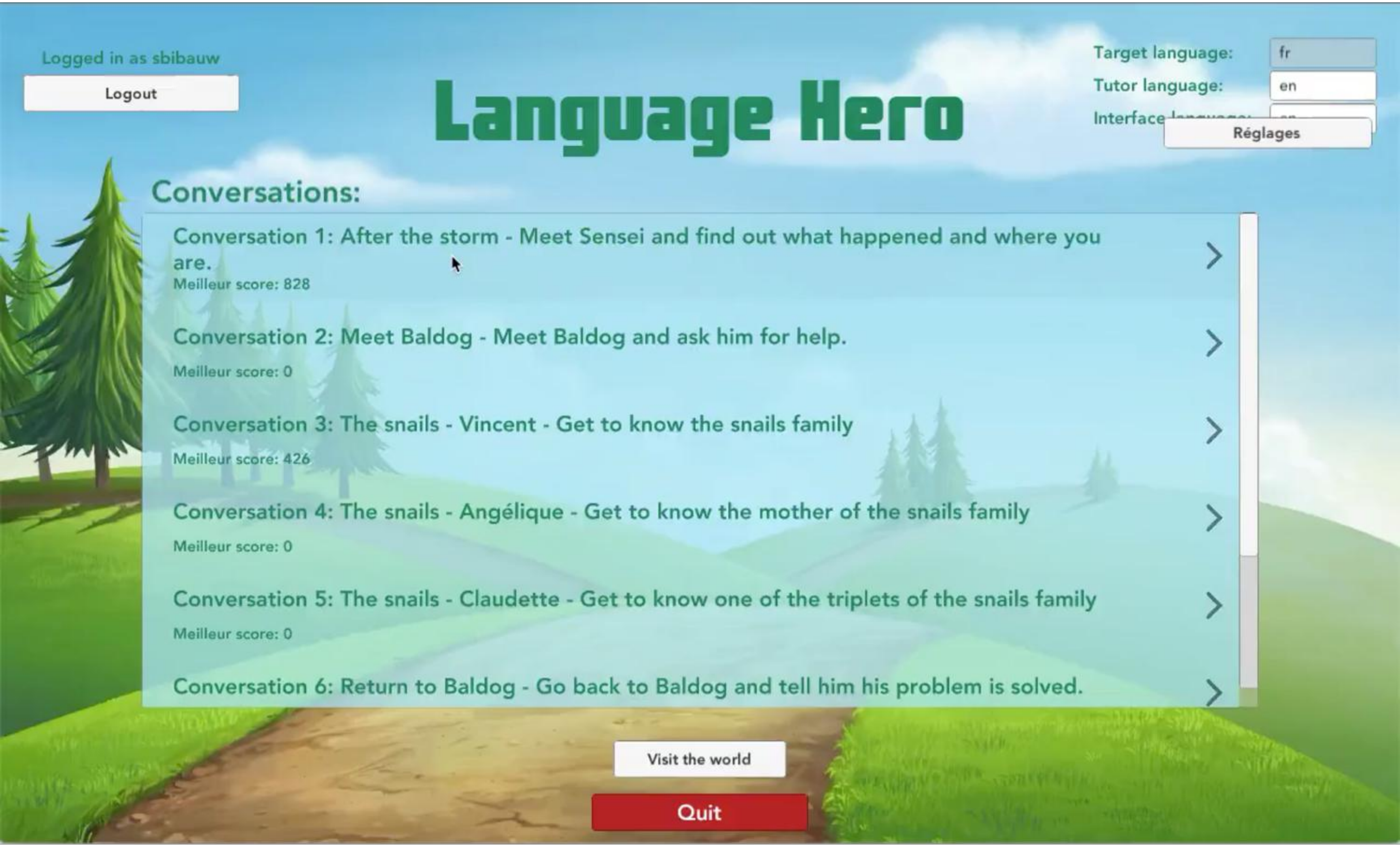
Gamification

Spoken input

Written input

Microtasks to guide the conversation

Scaffolding



Logged in as sbibauw

Language Hero

Target language: fr
Tutor language: en
Interface language: en

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.

Experimental study
Research questions (1)

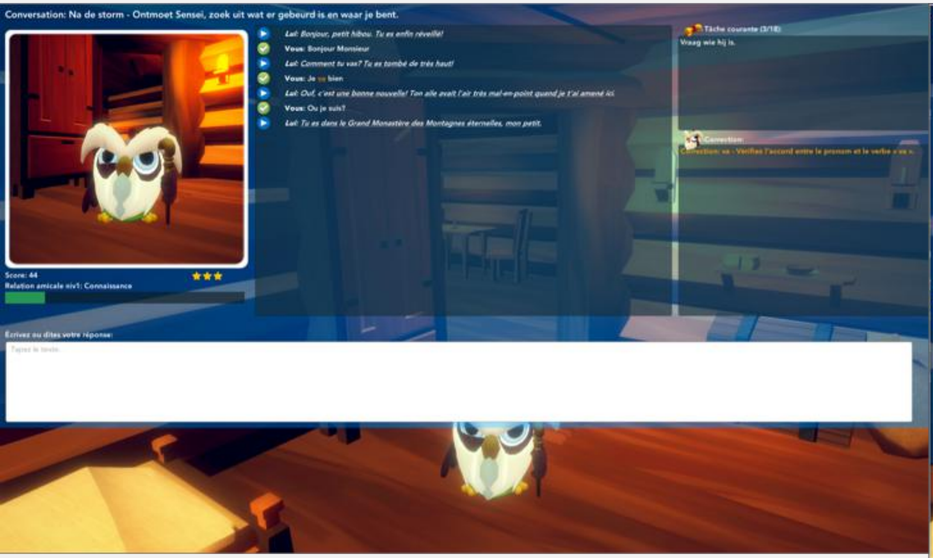
Use, perception and effectiveness of task-based written dialogue-based CALL system/game for L2 development

- Effect of task-based DBCALL on L2 fluency?
  - Are fluency gains even possible in such a short-term intervention (3 sessions of ±40 min)?
- Effect on acquisition of specific lexical items (given in-task exposure)?

Experimental study
Research questions (2)

Compare fully interactive, immediate/synchronous DBCALL condition with a more classic « **dialogue completion task** » with identical input/tasks/environmt

- Perceived difference of interactivity, authenticity?
- Effect of interactivity?



Conversation: Na de storm - Ontmoet Sensei, zoek uit wat er gebeurd is en waar je bent.

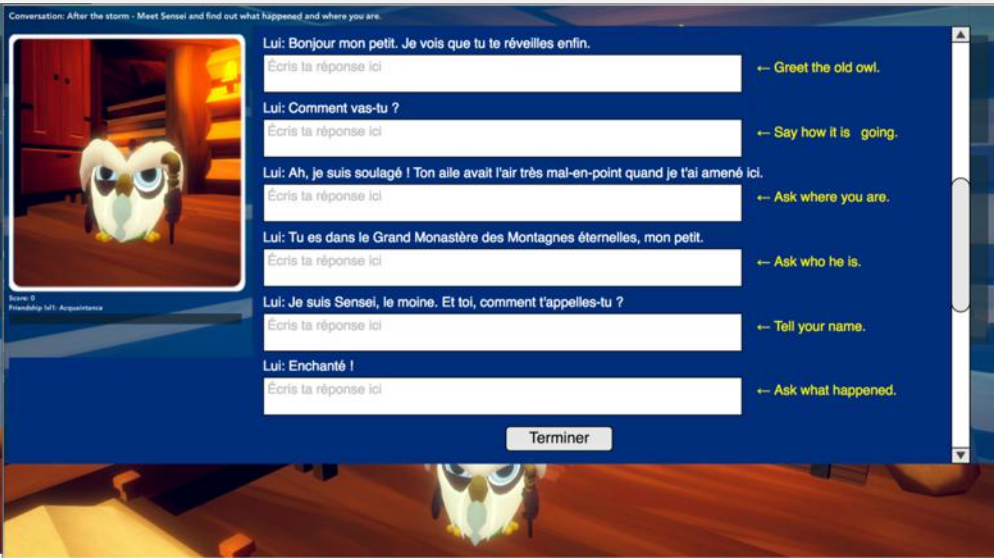
Lui: Bonjour, petit hibou. Tu es enfin réveillé!
Vous: Bonjour Monsieur
Lui: Comment tu vas? Tu es tombé de très haut!
Vous: Je va bien
Lui: Ouf, c'est une bonne nouvelle! Ton aile avait l'air très mal-en-point quand je t'ai amené ici.
Vous: Ou je suis?
Lui: Tu es dans le Grand Monastère des Montagnes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

Tâche courante (3/18)
Vraag wie hij is.

Correction:
Correction: va - Vérifiez l'accord entre le pronom et le verbe « va ».

Score: 44
Relation amicale niv1: Connaissance

Ecrivez ou dites votre réponse:
Tapez le texte



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

Lui: Bonjour mon petit. Je vois que tu te réveilles enfin.
Écris ta réponse ici ← Greet the old owl.

Lui: Comment vas-tu ?
Écris ta réponse ici ← 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.
Écris ta réponse ici ← Ask where you are.

Lui: Tu es dans le Grand Monastère des Montagnes éternelles, mon petit.
Écris ta réponse ici ← Ask who he is.

Lui: Je suis Sensei, le moine. Et toi, comment t'appelles-tu ?
Écris ta réponse ici ← Tell your name.

Lui: Enchanté !
Écris ta réponse ici ← Ask what happened.

Score: 0
Friendship lvl1: Acquaintance

Experimental study
Design and assignment

4 schools volunteered to participate, with 2-3 class groups each: **11 groups** in total

Random assignment (inside the school cluster) to 3 conditions:

- **Experimental** (interactive dialogue system)
- **Alternate** (‘dialogue completion task’)
- **Control** (no-intervention, business-as-usual)

Experimental study
Participants

4 schools (Ingelmunster, Harelbeke, Heule & Heverlee), 11 groups

Initially **N = 218 participants**

After exclusion of abstentees (no pre or post test), **N = 206** participants

- $n_{\text{Experimental}} = 79$
- $n_{\text{Alternate}} = 78$
- $n_{\text{Control}} = 49$

(For learning effects, N = 196 after exclusion of 10 native/near-native French locutors)

Experimental study

Intervention

Pretest (survey + interview) (1h)

3 in-app sessions (max 50min)

Dialogue System | Dialogue Completion Task

Posttest (survey + interview) (<1h)

(1-4 weeks in total, depending on schools)

Experimental study
Instruments: questionnaire test

- Productive **Vocabulary Size test** (Peters et al)
  - Made adaptative
    (30 1K items + 30 2K items if >= 50%)
  - Proxy of proficiency (at pretest only)
- **Target Vocabulary test**
  - Receptive: Translation (NL->FR) multiple choice (25 items)
  - Productive: Gap-filling on formulaic sequences (25 items)
  - At pre and posttest (identical, randomized order)

Experimental study
Instruments: survey

- Attitudes and practices towards L2 (pre)
- **Perceived effectiveness & ease-of-use** (post)
  - adapted from TAM & validated instruments
- **Perceived interactivity & authenticity** (post)
  - adapted from PAW scale (Behizadeh & Engelhard 2014)

Experimental study
Instruments: interview

- Computer-delivered spoken interview
- 28 items/questions
- Automatic recording

Experimental study
In-treatment

Full logging of all messages read and written in the system

+ **Keystroke logging** for writing fluency evaluation

Effects of dialogue-based CALL practice on foreign language learning

Existing systems: state of the art
Research synthesis and conceptual framework

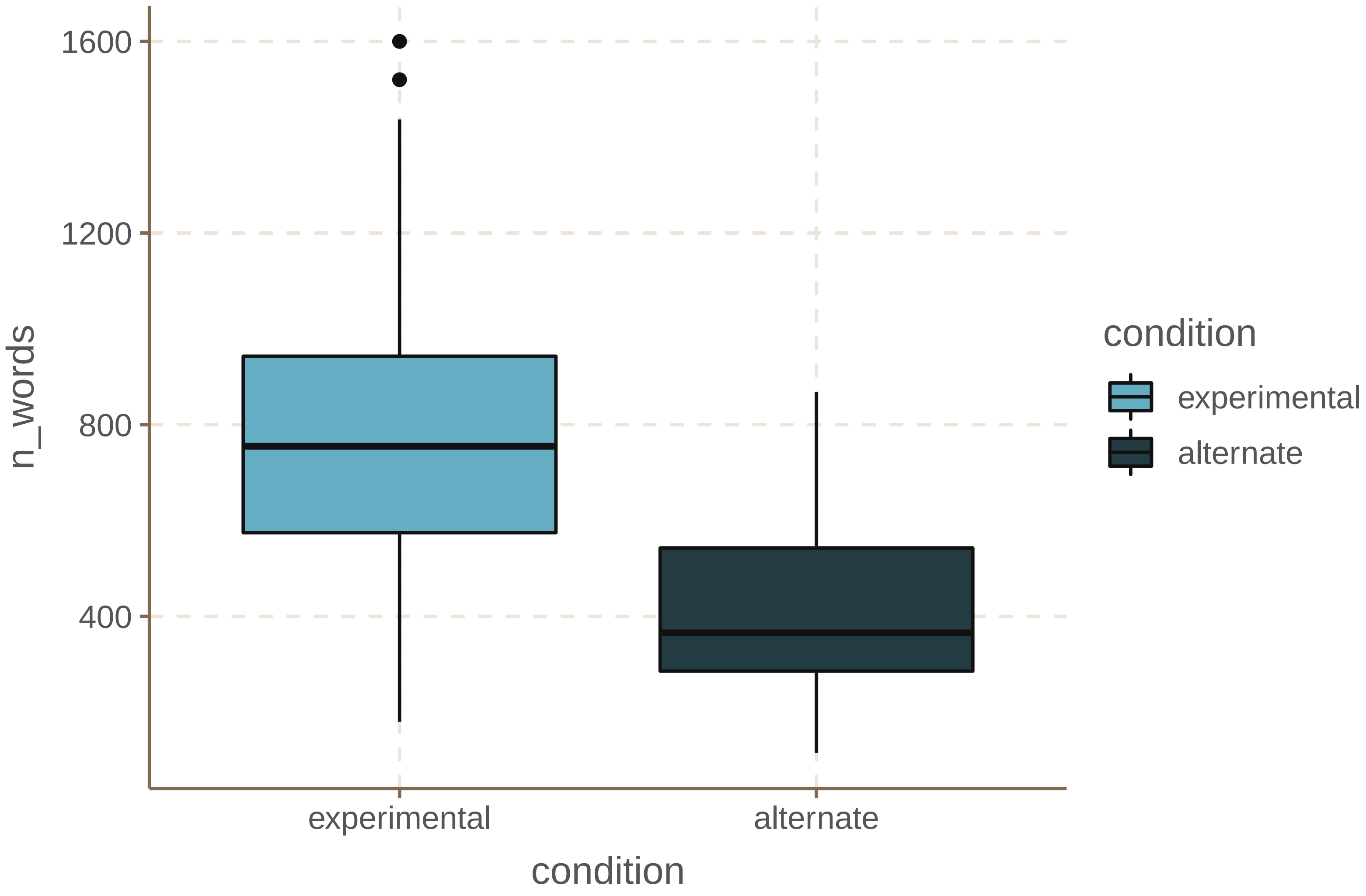
Existing research: effectiveness
Meta-analysis of dialogue-based CALL

LanguageHero: development and evaluation
A randomized controlled evaluation study

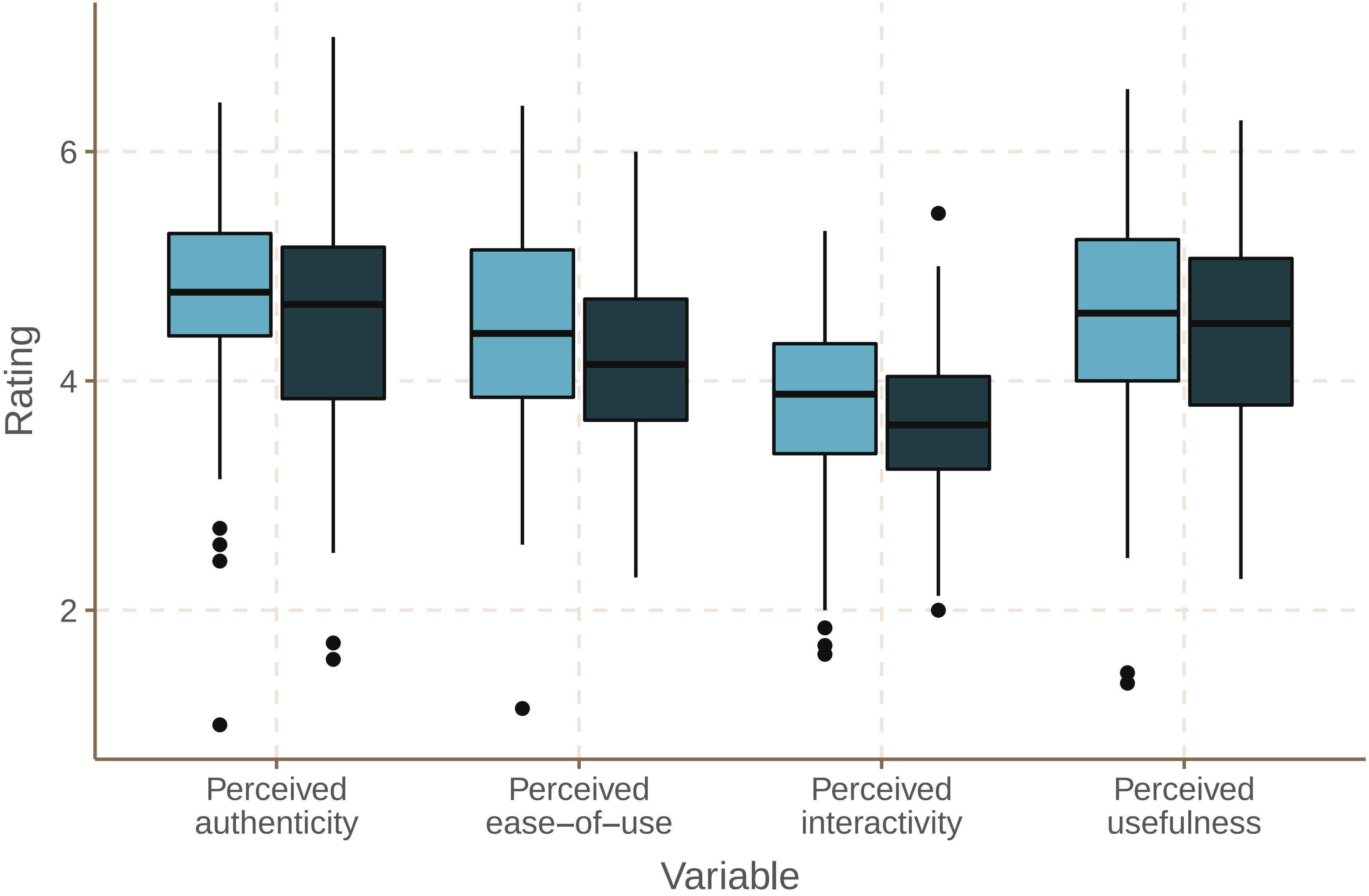
Use, perception, and effectiveness results
Preliminary results from the experimental study

Effective use

Amount of in-task production

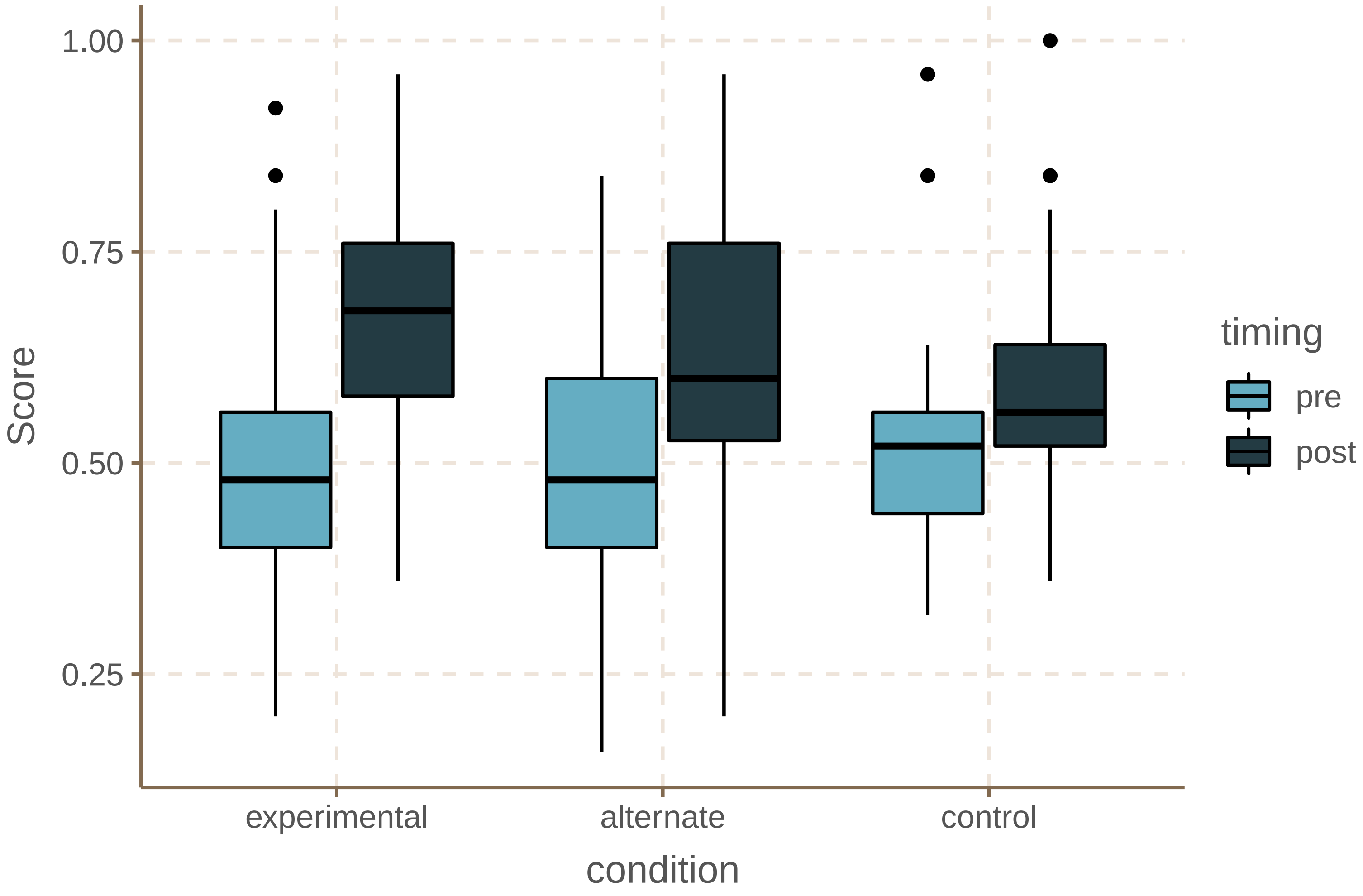


Perceived usefulness, ease-of-use (TAM) and perception of interactivity and authenticity



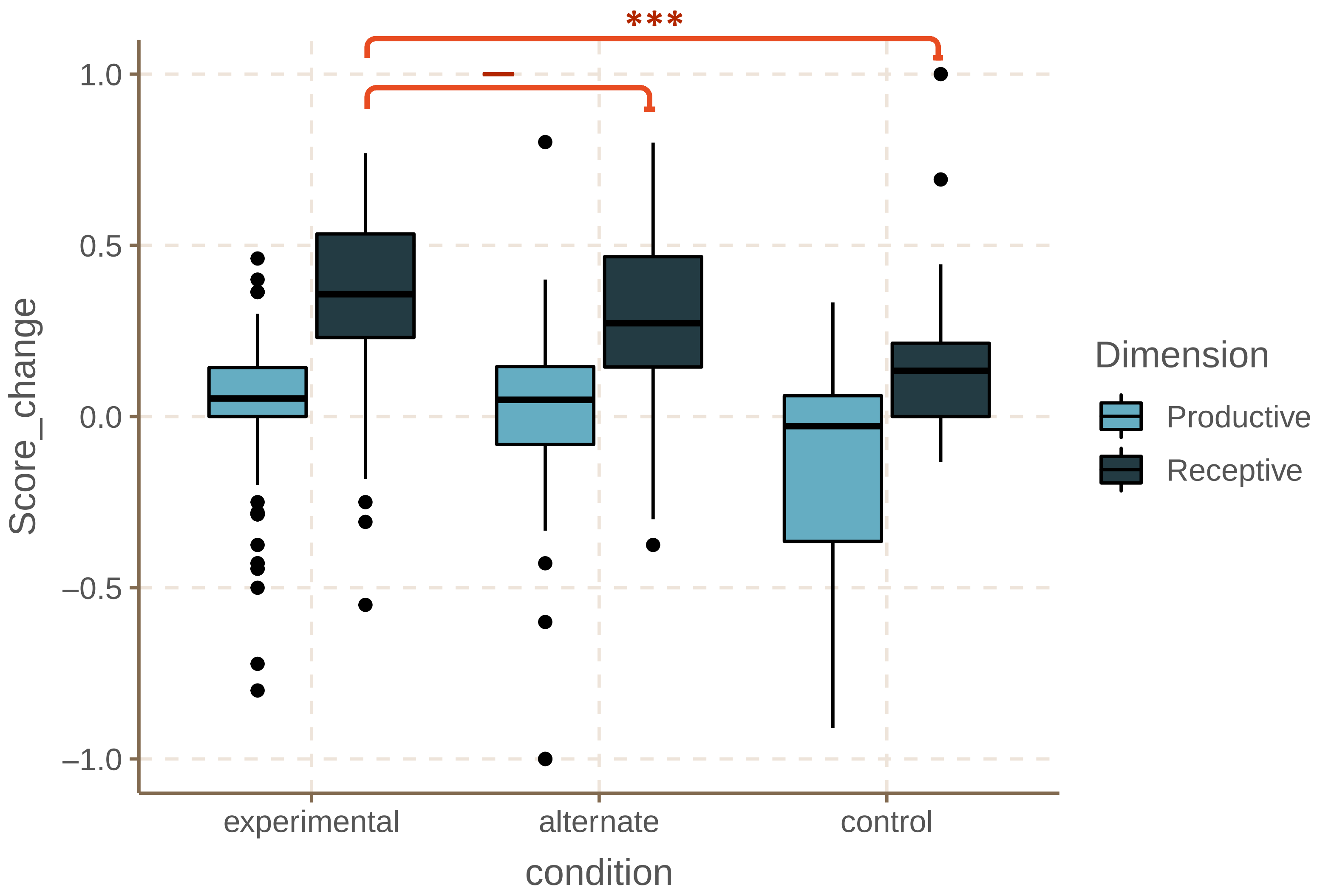
Vocabulary acquisition

Learning effect



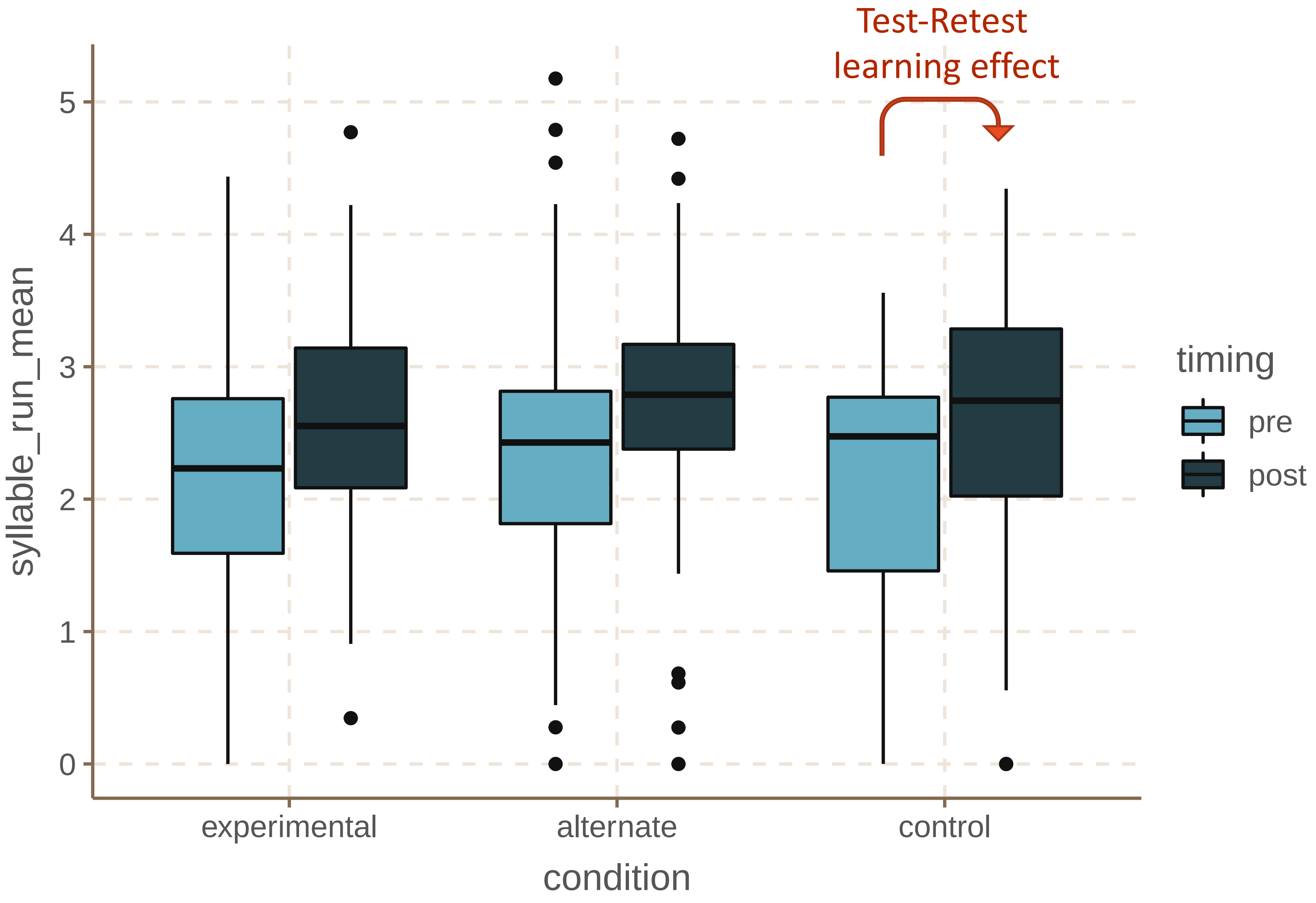
Vocabulary acquisition

Reception vs. production across groups



Effect on fluency development

Mean length of syllable runs



Effectiveness study
Perspectives

Visible effect on fluency development?

Trade-off with complexity development?

Measures of turn fluency/dialogue fluency

Perspectives

Effectiveness of DBCALL systems

Almost all previous systems remained internal, research-only prototypes, never made accessible to the public.
→ No comparability, no replicability

But, recently, **major advances towards publicly available tools** (Duolingo Bots, Alelo Enskill, ETS HALEF) and **joint efforts between industry and researchers** to compare the systems and establish common ground (Sydorenko et al, 2018)

Opportunities to build and compare, in a standardized way, use, perception and effectiveness of dialogue-based CALL environments.

Perspectives
DBCALL as an SLA research environment

Relationship between in-task exposure (both input and output opportunities, taken or not) and acquisition of lexical items.

Relationship between in-task written fluency and spoken fluency?

Serge Bibauw
serge.bibauw@kuleuven.be
https://serge.bibauw.be

Thank you! Dank u! Merci! ¡Gracias!

KU LEUVEN itec imec UCLouvain

CENTAL
centre de traitement
automatique du langage

UNIVERSIDAD CENTRAL DEL ECUADOR
QUITO