Systèmes de dialogue pour l'apprentissage des langues : typologie des systèmes et mesure des effets

Serge Bibauw

CENTAL, UCLouvain
ITEC, imec research group at KU Leuven
Universidad Central del Ecuador
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…)

Hey Siri
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

**PUBLICATIONS DATABASES SEARCH**
- Scopus: 153 records
- Web of Science Core Collection: 75 records
- Inspec: 68 records
- PsycINFO: 38 records
- LLBA: 38 records
- ERIC: 36 records
- ProQuest Central: 13 records
- MLA International Bibliography: 9 records
- LISA: 4 records
- Total identified: 434 records

**MANUAL SEARCH**
- Forward and ancestry search: 193 records
- Additional identified: 193 records

**IDENTIFICATION**
- Duplicate removal: 250 unique publications

**SCREENING FOR AVAILABILITY**
- 419 unique publications
- 27 records unavailable
- 20 records not an automated interlocutor
- 39 records item-based (no multi-turn dialogue)
- 13 records dialogue for scaffolding, not as task
- 136 records excluded

**INCLUSION-EXCLUSION**
- Conceptual eligibility review: 386 publications
- Methodological eligibility review: 250 publications on dialogue-based CALL

**TOTAL**
- 250 papers
- 114 different systems
Co-citation network of systems
(Bibauw, François & Desmet, 2019)

Chatbots
(AI/...)

Intelligent tutoring systems (ICALL)

Spoken dialogue systems (Speech/NLP)

focus on corrective feedback

focus on dialogue management
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

- 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
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)
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

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

Goal-oriented systems

SPELL (Morton, Gunson & Jack, 2012)
Typology of systems (Bibauw, François & Desmet, 2019)

Four types of dialogue-based CALL systems

- **Form-focused systems**
  - Constraints on meaning: Explicit
  - Reactive systems
    - Open-ended dialogue
      - ELIZA (Weizenbaum, 1966)

- **Goal-oriented systems**
  - Constraints on meaning: Implicit
  - Narrative systems
    - Croquelandia (Sykes, 2008)
      - Branching dialogue
        - Pre-set form

- **Reactive systems**
  - Constraints on meaning: None
    - 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 Management

- Action selection based on recognized intent, context, knowledge base, dialogue state...

Natural Language Understanding

- Dialogue act recognition
- Syntactic & semantic parsing

- Recognition
- Preprocessing
- Normalization

User

Natural Language Generation

- Surface selection or full generation

- Synthesis

System Agent
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)*

```xml
<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

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

He: Bien le bonjour! Comment t'appelles-tu?
You: bonjour je m'appelle Marco
He: Enchanté de faire ta connaissance, Rlnc! Rlnc. Rlnc. Rlnc. 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'appelles?
He: He does not seem to have heard you...
You: Tu t'appelles comment?
He: He does not seem to have heard you...
You: Tu t'appelles comment?
He: 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.

Microtasks to guide the conversation
Gamification
Contextualization
Corrective feedback
Scaffolding
Free written input
Dialogue guided by **microtasks/instructions**

→ 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

Don’t look at me for all your answers.

Are you useful?
Meta-analysis of effectiveness studies
Inclusion of individual effect sizes

419 unique publications extracted
SCREENING FOR AVAILABILITY

386 publications undergo full-text review
CONCEPTUAL ELIGIBILITY REVIEW

250 publications on dialogue-based CALL
METHODOLOGICAL ELIGIBILITY REVIEW

39 meta-analysable publications, reporting (per sample and per outcome variable)
EFFECT SIZES ELIGIBILITY REVIEW

138 individual effect sizes

27 full-text unavailable
3 full-text in other languages
3 republications
33 records excluded

64 no application to L2 learning
20 not an automated interlocutor
39 item-based (no multi-turn dialogue)
13 dialogue for scaffolding, not as task
136 records excluded

118 without empirical data
40 with technical evaluation
26 with observational/qualitative data
27 with survey data
211 records excluded

13 no central tendency (M, Mdn) reported
8 no variance (SD) reported nor alternate statistics to compute d (e.g., t)
6 lack of reference data (pretest, control)
11 effects on other outcomes
38 effect sizes excluded

$k = 100$ effect sizes
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:

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

\[
d_{ECPP} = J(df_{ECPP}) \left( \frac{M_{post,E} - M_{pre,E}}{SD_{pre,E}} - \frac{M_{post,C} - M_{pre,C}}{SD_{pre,C}} \right)
\]
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>
<thead>
<tr>
<th>Level</th>
<th>Number of clusters/items</th>
<th>Source of variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Samples</td>
<td>$k = 96$ ($n = 803$)</td>
<td>Random sampling variance</td>
</tr>
<tr>
<td>2 Effects sizes</td>
<td>$k = 96$</td>
<td>Variation within study</td>
</tr>
<tr>
<td>3 Studies</td>
<td>$l = 17$</td>
<td>Variation between studies</td>
</tr>
<tr>
<td>Reference</td>
<td>Features of single effect</td>
<td>n_E</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------------------------------------------------------------------------------------------</td>
<td>-----</td>
</tr>
<tr>
<td>Jia et al 2013</td>
<td>(sample Huiwen JHS)</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>(sample Huojia N1 SHS)</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>(sample Jingxian JHS)</td>
<td>48</td>
</tr>
<tr>
<td>Taguchi et al 2017</td>
<td>... gap–filling test *post</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>... gap–filling test *delayed</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>... multiple choice test *post</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>... multiple choice test *delayed</td>
<td>30</td>
</tr>
<tr>
<td>Kim 2016</td>
<td>(A1 sample)</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>(A2 sample)</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>(B1 sample)</td>
<td>21</td>
</tr>
<tr>
<td>Petersen 2010</td>
<td>... QFT, morphology score</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>... QFT, syntax score</td>
<td>19</td>
</tr>
<tr>
<td>Harless et al 1999</td>
<td>... listening comp.</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>... reading comp.</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>... speaking prof.</td>
<td>9</td>
</tr>
<tr>
<td>Hassani et al 2016</td>
<td>... Grammatical errors/sentence</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>... Nb of proper replies</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>... Phonation time/letter</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>... Automatic prof. score</td>
<td>10</td>
</tr>
<tr>
<td>Lee et al 2011a</td>
<td>(A1) ... listening compr.</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>(A2) ... listening compr.</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>(A1) ... hol. grammar rating</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>(A2) ... hol. grammar rating</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>(A1) ... hol. pronunciation rating</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>(A2) ... hol. pronunciation rating</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>(A1) ... hol. communicative ability rating</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>(A2) ... hol. communicative ability rating</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>(A1) ... hol. vocabulary rating</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>(A2) ... hol. vocabulary rating</td>
<td>11</td>
</tr>
<tr>
<td>Lee et al 2014a</td>
<td>... nb of grammatical errors</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>... nb of words</td>
<td>25</td>
</tr>
<tr>
<td>Noh et al 2012</td>
<td>(Engl. major) ... DCT, comprehensibility</td>
<td>40</td>
</tr>
<tr>
<td>Chiu et al 2007</td>
<td>(Engl. major) ... DCT, comprehensibility</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>(not Engl. major) ... DCT, comprehensibility</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>(Engl. major) ... DCT, use of speech acts</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>(not Engl. major) ... DCT, use of speech acts</td>
<td>20</td>
</tr>
</tbody>
</table>
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:

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

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

$d_{\text{raw}} = 0.627$ *** [0.390, 0.863]
Global effect close to the median of meta-analyses in CALL/SLA (Plonsky & Oswald, 2014)
- $\geq$ game-based learning ($d = .53$, Chiu et al, 2012)
- $\leq$ CALL in general ($d = .84$, Plonsky & Ziegler, 2016)

Consistent with effect of face-to-face interaction (Mackey & Goo, 2007) and SCMC.
- $\leq$ F2F interaction ($d = .75$, Mackey & Goo, 2007)
- $\leq$ 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?
Form-focused and goal-oriented systems confirm their effectiveness. Unclear difference though.
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

System modality

Outcome modality

<- Matching modality ->
Meta-analysis: moderator analyses

Outcome ▶ Dimensions

![Graph showing dimensions of comprehension, knowledge test, and production.](image-url)
More promising effects on **fluency** and possibly vocabulary.
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
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.
Compare:

(A) fully interactive, immediate/synchronous dialogue system

(B) classic, asynchronous dialogue completion task

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

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 texte..  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 texte..  Send your reply
Methods

Population and group assignment

4 schools volunteered to participate, with 2-3 classes each:
\[ N_{\text{clusters}} = 11 \quad N_{\text{participants}} = 215 \quad (208 \text{ complete cases}) \]

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\textsuperscript{nd} 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 \textbf{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

Posttest
- Computer-delivered spoken interview
- Perceptions questionnaire
- Target vocabulary test
### Methods · Instruments

**Perceptions questionnaire (post)**

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

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” 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)
• **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 _ _ _ _ _ _ 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 (Cucchiarini et al, 2002...), unrelated to other fluency measures (Segalowitz et al 2017)

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

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

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)

Perceived authenticity
Perceived interactivity
Perceived ease-of-use
Perceived usefulness

N = 32 (pilot)

p = 0.1551
p = 0.0129
p = 0.0027
p = 0.1600
Experimental results

Differences of learners’ perceptions

- Perceived authenticity: $p = 0.264$
- Perceived interactivity: $p = 0.065$
- Perceived ease-of-use: $p = 0.037$
- Perceived usefulness: $p = 0.677$

N = 159
**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

- **Number of messages**
  - Interactive Dialogue System
  - Dialogue Completion Task

- **Number of words / message**
  - Interactive Dialogue System
  - Dialogue Completion Task
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.

\[ d_{DSystem} = 0.56^{***} \]

\[ d_{DCompletion} = 0.39^{***} \]

\[ d_{DControl} = 0.02 \text{ n.s.} \]

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

\[ d_{DSys \text{ vs DCompl}} = 0.23 \text{ n.s.} \]
Results

Fluency

- Interactive Dialogue System
- Dialogue Completion Task
- Control

(Pruned) Length of runs (in syllables)

Timing

- Pre
- Post

p-values:
- p = 0.017
- p = 0.014
- p = 0.144
Results

Fluency

$d_{DSys} = 0.57$

$d_{DSys \text{ vs Ctrl}} = 0.17$

$d_{Ctrl} = 0.48$

$p = 0.0023 \ *[**]$

$p = 0.0016 \ *[**]$

$p = 0.1637 \ n.s.$

No difference

$DSys \text{ vs } DCompl$

$d_{DSys} = 0.57$

$0.17$

$d_{DSys \text{ vs Ctrl}} = 0.17$

$\approx$

$d_{Ctrl} = 0.48$

Interactive Dialogue System

Dialogue Completion Task

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

but very short treatment (2h) $\rightarrow$ expected (effect on general L2 speaking proficiency by written practice)

No difference between DSys and DCompl $\Rightarrow$ 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
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
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