Effectiveness of dialogue-based CALL on L2 proficiency development: a meta-analysis

Serge Bibauw
Thomas François
Piet Desmet

CALICO Conference
May 18, 2017
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.
Dialogue-based CALL
A recent example

Duolingo Bots
(Oct. 2016)
Dialogue-based CALL
Three main types of systems

**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

**Reactive** dialogue systems
Free, user-initiated, open-ended dialogue
see Chatbots
Meta-analysis of effectiveness studies

Aggregate results from multiple experimental studies
Treat each study as a subject
Get a more powerful, generalizable, stable and precise idea of the effectiveness of dialogue-based CALL on language learning
Analyzing certain moderator variables to identify tendencies inside the data

Bibauw, François & Desmet, 2015 (EUROCALL Proceedings); Bibauw, François & Desmet, in prep.
Effectiveness of dialogue-based CALL on L2 proficiency development: a meta-analysis

Methodology of meta-analysis
Corpus collection, inclusion/exclusion, effect sizes calculation and multilevel modeling

Main results
A random effects multilevel model to summarize the effect of dialogue-based CALL

Moderator variables
Moderators analysis: effects of specific design choices and experimental context
Effectiveness of dialogue-based CALL on L2 proficiency development: a meta-analysis

Methodology of meta-analysis
Corpus collection, inclusion/exclusion, effect sizes calculation and multilevel modeling

Main results
A random effects multilevel model to summarize the effect of dialogue-based CALL

Moderator variables
Moderators analysis: effects of specific design choices and experimental context
Corpus collection

Search methodology

1. **Database** search
   in Web of Science, Scopus, ProQuest

   Search syntax:
   (chatbot / chat bot / chatterbot /
   conversational agent / conversational companion /
   conversational system / dialog* system /
   dialog* agent / dialog* game / pedagogical agent /
   human-computer dialog* / dialog*-based) +
   ((language / English) (learning / teaching /
   acquisition) / (second / foreign) language / L2 /
   EFL / ESL / ICALL)

2. **Ancestry** search
   Older publications cited by ref

3. **Forward** citations
   New publications citing ref

Note on journal search: 32/183 publications from the 4 major CALL journals (13 CALL, 12 CALICO J., 4 ReCALL, 3 LL&T)
Corpus of studies

256 papers
114 different systems
Corpus collection
Domain definition and inclusion criteria

Based on our operational definition of dialogue-based CALL (dialogue, as a task/main activity, with a system/computer agent, for language learning purposes)

Peer-reviewed publications (journal articles, conference papers, book chapters, dissertations) only

⇒ 183 papers
Coding scheme

Study identification
author, year, team_id, sample_id, study_type...

Sample and context
context, age, L1, L2, proficiency_level

System (treatment) variables
system, system_type, dialogue_type, primary_modality, corrective_feedback, initiative, embodied_agent, gamified...
treatment_duration (in weeks), time_on_task (in hours)

Instruments/outcome variables
proficiency/complexity/accuracy/fluency/vocabulary, speaking/writing, specific test

Quantitative results
n, mean, sd (pre/post, experimental/control)
Studies selection
Effectiveness studies

Effectiveness studies (36 papers)

- Experimental (or quasi-exp.) design
- At least two measurements (pre-post, experimental-control...)
- Motivational or cognitive effects measured
- Few experiments with a real control group
Studies selection

Computable effect sizes

**Effect size**: quantitative measure of the difference between two groups

Typically needs

- **mean**
- **standard deviation**
- **n** (subjects)

for each group/measurement point

Not available for all studies (especially older studies) (6 studies excluded)

Asked the authors for raw data
(worked for some - thanks to them!)
Studies selection

Instruments

Language learning tests

- **Speaking/writing tests**, measuring holistic proficiency or some of its components (complexity, accuracy, fluency)
- **Accuracy tests** (grammar/syntax/morphology)
- **Vocabulary tests**

Excluding **motivational effects** (1 study)

Excluding **non-related** cognitive effects (1 study)

Excluding **uncontrolled** teacher-made exams (1 study)
Meta-analysis

Effect sizes computation

**Standardized Mean Difference (SMD)**

\[ \text{Mean}_{\text{post}} - \text{Mean}_{\text{pre}} / \text{SD}_{\text{pooled}} \]

Cohen’s \( d \)

Hedges’ \( g \)

\[ \hat{g} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)SD_1^2 + (n_2-1)SD_2^2}{n_{\text{total}}-2}}} \times 1 - \left( \frac{3}{4(n_1-n_2)-9} \right) \]

When control group:

**Controlled Standardized Mean Difference**

\[ \text{SMD}_{\text{experimental}} - \text{SMD}_{\text{control}} \]

\textbf{Z-scores} ⇒ Allows to compare results across various measurements scales
Meta-analysis
Summary effect size

Model computes a **summary effect** by aggregating all the single study effect sizes

**Weighting** according to sample size and precision

→ More powerful, more stable, more precise and generalizable than the individual study effect sizes
Meta-analysis
Random effects modeling

**Fixed effects (FE) vs. Random effects (RE)**

**FE** assumes a single common effect size among the studies (All variance is due to random and sampling errors)

**RE** assumes an inherent variance between studies

Considering the variability among systems and outcome measures studied:

**Random effects model**

Confirmed by heterogeneity test

\( Q = 246, p < .0001 \)
Meta-analysis
Multilevel modeling

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

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

**Multilevel modelling**

Here, 3 levels: team / sample / study

\[ K = 11 \quad 21 \quad 40 \]

Weights accordingly, as dependent measures
Effectiveness of dialogue-based CALL on L2 proficiency development: a meta-analysis

Methodology of meta-analysis
Corpus collection, inclusion/exclusion, effect sizes calculation and multilevel modeling

Main results
A random effects multilevel model to summarize the effect of dialogue-based CALL

Moderator variables
Moderators analysis: effects of specific design choices and experimental context
Results

Summary effect

Within-subjects (pre-post) ($k = 40$):

\[ d = 0.904 \text{ *** (within-subjects)} \]

95% CI = [0.511, 1.298]

= Large effect (Cohen’s “rule of thumb”)
= Medium effect (Plonsky & Oswald, 2014, AL/SLA field-specific scale)

Between-subjects (pre-post - control) ($k = 12$):

\[ d = 0.618 \text{ **} \]

95% CI = [0.243, 0.995]

= Medium effect (Plonsky & Oswald, 2014)
### Reference System Outcome measure Prof. N Effect size (g) [95% CI]

**Lee et al 2014**
- POMY
  - Accuracy
    - (~Grammatical errors/words)
    - mixed
    - 25
    - 0.19 [-0.37, 0.74]

**Lee et al 2012**
- POMY
  - Accuracy
    - (Holistic rater judgement)
    - A1
    - 21
    - 1.37 [0.70, 2.04]

**Hassani et al 2016**
- IVELL
  - Accuracy
    - (~Grammatical errors/sentence)
    - A2
    - 10
    - 0.13 [-0.74, 1.01]

**Rayner & Tsourakis 2013**
- CALL-SLT
  - Accuracy
    - (In-app response)
    - A1
    - 12
    - 1.50 [0.60, 2.41]

**Petersen 2010**
- Sasha
  - Accuracy
    - (Question–formation Test, morphology...)
    - mixed
    - 19
    - 0.62 [-0.03, 1.27]

**Petersen 2010**
- Sasha
  - Accuracy
    - (Question–formation Test, syntax score)
    - mixed
    - 19
    - 0.55 [-0.10, 1.19]

**Bouillon et al 2011**
- CALL-SLT
  - Accuracy
    - (Grammar/syntax test)
    - A2
    - 10
    - 1.02 [0.09, 1.95]

**Wolska & Wilske 2010b**
- [Wilske2]
  - Accuracy
    - (Grammaticity Judgement Test)
    - mixed
    - 7
    - 0.50 [-0.56, 1.57]

**Wilska & Wilske 2010a**
- [Wilske2]
  - Accuracy
    - (Grammaticity Judgement Test)
    - A2
    - 6
    - 0.95 [-0.24, 2.15]

**Wilske & Wilska 2011**
- [Wilske2]
  - Accuracy
    - (Grammaticity Judgement Test)
    - mixed
    - 20
    - 0.71 [0.07, 1.35]

**Wilska & Wilske 2010b**
- [Wilske2]
  - Accuracy
    - (Sentence Construction Test)
    - mixed
    - 7
    - 0.33 [-0.72, 1.39]

**Wilska & Wilske 2010a**
- [Wilske2]
  - Accuracy
    - (Sentence Construction Test)
    - A2
    - 6
    - 0.53 [-0.62, 1.69]

**Wilske & Wilska 2011**
- [Wilske2]
  - Accuracy
    - (Sentence Construction Test)
    - mixed
    - 21
    - 0.48 [-0.13, 1.10]

**Hassani et al 2016**
- IVELL
  - Complexity
    - (Nb of proper replies)
    - A2
    - 10
    - 0.36 [-0.53, 1.24]

**Lee et al 2012**
- POMY
  - Comprehension
    - (undisclosed test)
    - mixed
    - 21
    - -0.10 [-0.71, 0.50]

**Harless et al 1999**
- Conversim
  - Comprehension
    - (Defense Language Proficiency Test)
    - mixed
    - 9
    - 0.54 [-0.40, 1.48]

**Lee et al 2012**
- POMY
  - Fluency
    - (Holistic rater judgement)
    - A1
    - 21
    - 1.87 [1.15, 2.60]

**Lee et al 2014**
- POMY
  - Fluency
    - (Nb of words)
    - mixed
    - 25
    - 0.48 [-0.08, 1.04]

**Hassani et al 2016**
- IVELL
  - Fluency
    - (~Phonation time/letter)
    - A2
    - 10
    - 0.05 [-0.82, 0.93]

**Wolska & Wilske 2011**
- [Wilske2]
  - Fluency
    - (Phonation time ratio)
    - mixed
    - 6
    - 1.28 [0.04, 2.53]

**Wilske 2014**
- [Wilske2]
  - Fluency
    - (Phonation time ratio)
    - mixed
    - 7
    - 0.62 [-0.46, 1.69]

**Wilska & Wilske 2011**
- [Wilske2]
  - Fluency
    - (Speech rate)
    - mixed
    - 6
    - 0.31 [-0.83, 1.45]

**Wilske 2014**
- [Wilske2]
  - Fluency
    - (Speech rate)
    - mixed
    - 7
    - 0.11 [-0.94, 1.16]

**Kim 2016**
- Indigo
  - Proficiency
    - (TOEIC Speaking Test)
    - A1
    - 20
    - 4.15 [3.05, 5.26]

**Kim 2016**
- Indigo
  - Proficiency
    - (TOEIC Speaking Test)
    - A2
    - 22
    - 1.95 [1.23, 2.67]

**Kim 2016**
- Indigo
  - Proficiency
    - (TOEIC Speaking Test)
    - B1
    - 21
    - 0.77 [0.14, 1.39]

**Harless et al 1999**
- Conversim
  - Proficiency
    - (Oral Proficiency Interview)
    - mixed
    - 9
    - 1.90 [0.79, 3.02]

**Chiu et al 2007**
- CandleTalk
  - Proficiency
    - (Discourse Completion Test (holistic...)
    - A2
    - 49
    - 0.71 [0.50, 0.92]

**Lee et al 2012**
- POMY
  - Pronunciation
    - (Holistic rater judgement)
    - A1
    - 21
    - 1.90 [1.17, 2.63]

**Lee et al 2012**
- POMY
  - Vocabulary
    - (Holistic rater judgement)
    - A1
    - 21
    - 1.34 [0.67, 2.01]

**Noh et al 2012**
- POMY
  - Vocabulary
    - (undisclosed test)
    - NA
    - 40
    - 1.35 [0.86, 1.83]

**Arispe 2014**
- Langbot
  - Vocabulary
    - (Word Association Depth Test)
    - A2
    - 20
    - -0.41 [-1.04, 0.21]

**Arispe 2014**
- Langbot
  - Vocabulary
    - (Word Association Depth Test)
    - B1
    - 25
    - 0.81 [0.24, 1.39]

**Arispe 2014**
- Langbot
  - Vocabulary
    - (Word Frequency Breadth Test)
    - A2
    - 20
    - 1.20 [0.52, 1.87]

**Arispe 2014**
- Langbot
  - Vocabulary
    - (Word Frequency Breadth Test)
    - B1
    - 25
    - 0.60 [0.03, 1.17]

**Jia et al 2013**
- CSIEC
  - Vocabulary
    - (undisclosed test)
    - mixed
    - 37
    - 1.78 [1.25, 2.32]

**Jia et al 2014**
- CSIEC
  - Vocabulary
    - (undisclosed test)
    - mixed
    - 56
    - 2.04 [1.58, 2.50]

**Jia et al 2013**
- CSIEC
  - Vocabulary
    - (undisclosed test)
    - mixed
    - 48
    - 1.10 [0.67, 1.52]

**Bouillon et al 2011**
- CALL-SLT
  - Vocabulary
    - (Translation test)
    - A2
    - 10
    - 2.80 [1.56, 4.03]

**Rosenthal-von der Putten et al 2016**
- Vocabulary
  - (Cloze test)
    - mixed
    - 130
    - -0.21 [-0.45, 0.03]

---

**Multilevel RE Model for all studies**
- 0.90 [0.51, 1.30]
### Between-subjects Forest plot

<table>
<thead>
<tr>
<th>Reference</th>
<th>Outcome</th>
<th>Prof.</th>
<th>Nexp</th>
<th>Nctl</th>
<th>Effect size (g) [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Petersen 2010</td>
<td>Accuracy</td>
<td>mixed</td>
<td>19</td>
<td>18</td>
<td>0.77 [-0.15, 1.69]</td>
</tr>
<tr>
<td>Petersen 2010</td>
<td>Accuracy</td>
<td>mixed</td>
<td>19</td>
<td>18</td>
<td>0.96 [ 0.04, 1.89]</td>
</tr>
<tr>
<td>Kim 2016</td>
<td>Proficiency</td>
<td>A1</td>
<td>20</td>
<td>20</td>
<td>3.12 [ 1.84, 4.40]</td>
</tr>
<tr>
<td>Kim 2016</td>
<td>Proficiency</td>
<td>A2</td>
<td>22</td>
<td>22</td>
<td>0.91 [-0.04, 1.86]</td>
</tr>
<tr>
<td>Kim 2016</td>
<td>Proficiency</td>
<td>B1</td>
<td>21</td>
<td>16</td>
<td>0.03 [-0.92, 0.99]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>Vocabulary</td>
<td>A2</td>
<td>20</td>
<td>22</td>
<td>0.26 [-0.61, 1.13]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>Vocabulary</td>
<td>B1</td>
<td>25</td>
<td>18</td>
<td>0.55 [-0.33, 1.42]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>Vocabulary</td>
<td>A2</td>
<td>20</td>
<td>22</td>
<td>0.53 [-0.38, 1.43]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>Vocabulary</td>
<td>B1</td>
<td>25</td>
<td>18</td>
<td>0.52 [-0.34, 1.39]</td>
</tr>
<tr>
<td>Jia et al 2013</td>
<td>Vocabulary</td>
<td>mixed</td>
<td>37</td>
<td>34</td>
<td>-0.10 [-0.88, 0.69]</td>
</tr>
<tr>
<td>Jia et al 2013</td>
<td>Vocabulary</td>
<td>mixed</td>
<td>56</td>
<td>56</td>
<td>0.95 [ 0.34, 1.55]</td>
</tr>
<tr>
<td>Jia et al 2013</td>
<td>Vocabulary</td>
<td>mixed</td>
<td>48</td>
<td>47</td>
<td>-0.05 [-0.66, 0.56]</td>
</tr>
</tbody>
</table>

**RE Model**

0.62 [ 0.24, 0.99]
Results
Evolution across time

Maturation of the field?
Discussion

Limitations

- High heterogeneity
- Few studies with strong results
- Publication bias and self-evaluation bias
Effectiveness of dialogue-based CALL on L2 proficiency development: a meta-analysis

Methodology of meta-analysis
Corpus collection, inclusion/exclusion, effect sizes calculation and multilevel modeling

Main results
A random effects multilevel model to summarize the effect of dialogue-based CALL

Moderator variables
Moderators analysis: effects of specific design choices and experimental context
Moderators analysis

Insights about the influence of some covariates/moderators

Sample and context
context, age, L1, L2, proficiency level

System (treatment) variables
system, system type, dialogue type, primary modality, corrective feedback, initiative, embodied agent, gamified... treatment duration (in weeks), time on task (in hours)

Instruments/outcome variables
proficiency/complexity/accuracy/fluency/vocabulary, speaking/writing, specific test
Time on task

$\text{d} = +0.007/hour$ (non significant)

<table>
<thead>
<tr>
<th>Study</th>
<th>Time on task (hours)</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rayner &amp; Tsourakis 2013</td>
<td>1</td>
<td>1.50 [0.60, 2.41]</td>
</tr>
<tr>
<td>Wolska &amp; Wilcke 2010b</td>
<td>1</td>
<td>0.50 [-0.56, 1.57]</td>
</tr>
<tr>
<td>Wilska &amp; Wilske 2010a</td>
<td>1</td>
<td>0.95 [-0.24, 2.15]</td>
</tr>
<tr>
<td>Wilske &amp; Wilske 2011</td>
<td>1</td>
<td>0.71 [0.07, 1.35]</td>
</tr>
<tr>
<td>Wolska &amp; Wilske 2010b</td>
<td>1</td>
<td>0.33 [-0.72, 1.39]</td>
</tr>
<tr>
<td>Wilska &amp; Wilske 2010a</td>
<td>1</td>
<td>0.53 [-0.62, 1.69]</td>
</tr>
<tr>
<td>Wilska &amp; Wilske 2011</td>
<td>1</td>
<td>0.48 [-0.13, 1.10]</td>
</tr>
<tr>
<td>Wilske &amp; Wilske 2011</td>
<td>1</td>
<td>1.28 [0.04, 2.53]</td>
</tr>
<tr>
<td>Wilska &amp; Wilske 2014</td>
<td>1</td>
<td>0.62 [-0.46, 1.69]</td>
</tr>
<tr>
<td>Wilska &amp; Wilske 2014</td>
<td>1</td>
<td>0.31 [-0.83, 1.45]</td>
</tr>
<tr>
<td>Wilska &amp; Wilske 2014</td>
<td>1</td>
<td>0.11 [-0.94, 1.16]</td>
</tr>
<tr>
<td>Rosenthal–von der Putten et al 2016</td>
<td>1</td>
<td>-0.21 [-0.45, 0.03]</td>
</tr>
<tr>
<td>Petersen 2010</td>
<td>1.5</td>
<td>0.62 [-0.03, 1.27]</td>
</tr>
<tr>
<td>Petersen 2010</td>
<td>1.5</td>
<td>0.55 [-0.10, 1.19]</td>
</tr>
<tr>
<td>Hassani et al 2016</td>
<td>2</td>
<td>0.13 [-0.74, 1.01]</td>
</tr>
<tr>
<td>Hassani et al 2016</td>
<td>2</td>
<td>0.36 [-0.53, 1.24]</td>
</tr>
<tr>
<td>Hassani et al 2016</td>
<td>2</td>
<td>0.05 [-0.82, 0.93]</td>
</tr>
<tr>
<td>Bouillon et al 2011</td>
<td>3</td>
<td>1.02 [0.09, 1.95]</td>
</tr>
<tr>
<td>Chiu et al 2007</td>
<td>3</td>
<td>4.15 [3.05, 5.26]</td>
</tr>
<tr>
<td>Kim 2016</td>
<td>1.5</td>
<td>1.95 [1.23, 2.67]</td>
</tr>
<tr>
<td>Kim 2016</td>
<td>1.5</td>
<td>0.77 [0.14, 1.39]</td>
</tr>
<tr>
<td>Kim 2016</td>
<td>1.5</td>
<td>2.80 [1.56, 4.03]</td>
</tr>
<tr>
<td>Bouillon et al 2011</td>
<td>3</td>
<td>1.78 [1.25, 2.32]</td>
</tr>
<tr>
<td>Jia et al 2013</td>
<td>6</td>
<td>2.04 [1.58, 2.50]</td>
</tr>
<tr>
<td>Jia et al 2013</td>
<td>6</td>
<td>1.10 [0.67, 1.52]</td>
</tr>
<tr>
<td>Jia et al 2013</td>
<td>6</td>
<td>1.35 [0.86, 1.83]</td>
</tr>
<tr>
<td>Noh et al 2012</td>
<td>8</td>
<td>0.19 [-0.37, 0.74]</td>
</tr>
<tr>
<td>Lee et al 2014</td>
<td>9</td>
<td>0.48 [-0.08, 1.04]</td>
</tr>
<tr>
<td>Lee et al 2014</td>
<td>9</td>
<td>1.37 [0.70, 2.04]</td>
</tr>
<tr>
<td>Lee et al 2012</td>
<td>10</td>
<td>-0.10 [-0.71, 0.50]</td>
</tr>
<tr>
<td>Lee et al 2012</td>
<td>10</td>
<td>1.87 [1.15, 2.60]</td>
</tr>
<tr>
<td>Lee et al 2012</td>
<td>10</td>
<td>1.90 [1.26, 2.63]</td>
</tr>
<tr>
<td>Lee et al 2012</td>
<td>10</td>
<td>1.34 [0.67, 2.01]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>20</td>
<td>-0.41 [-1.04, 0.21]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>20</td>
<td>0.81 [0.24, 1.39]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>20</td>
<td>1.20 [0.52, 1.87]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>20</td>
<td>0.60 [-0.03, 1.17]</td>
</tr>
<tr>
<td>Noh et al 2012</td>
<td>8</td>
<td>0.54 [-0.40, 1.48]</td>
</tr>
<tr>
<td>Noh et al 2012</td>
<td>8</td>
<td>1.90 [0.79, 3.02]</td>
</tr>
</tbody>
</table>
Standardized Mean Difference

$d = +0.077/\text{week} \times (p = 0.026)$
Learners’ proficiency level

Subjects’ proficiency level

Rayner & Tsourakis 2013
Lee et al 2012
Lee et al 2012
Lee et al 2012
Kim 2016
Lee et al 2012
Lee et al 2012
Hassani et al 2016
Wolska & Wilske 2010a
Wolska & Wilske 2010a
Bouillon et al 2011
Hassani et al 2016
Hassani et al 2016
Chiu et al 2007
Kim 2016
Bouillon et al 2011
Arispe 2014
Arispe 2014
Noh et al 2012
Kim 2016
Arispe 2014
Arispe 2014
Lee et al 2014
Wilska & Wilske 2011
Wolska & Wilske 2010b
Wilska & Wilske 2011
Wolska & Wilske 2010b
Petersen 2010
Petersen 2010
Harless et al 1999
Wolska & Wilske 2011
Wilske 2014
Wolska & Wilske 2011
Lee et al 2014
Wilske 2014
Lee et al 2014
Harless et al 1999
Jia et al 2013
Jia et al 2013
Jia et al 2013

<table>
<thead>
<tr>
<th>Subjects’ proficiency level</th>
<th>A1: d = 1.67 ***</th>
<th>A2: d = 0.92 °</th>
<th>B1: d = 0.61 (n.s.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rayner &amp; Tsourakis 2013</td>
<td>1.50 [0.60, 2.41]</td>
<td>1.37 [0.70, 2.04]</td>
<td>1.87 [1.15, 2.60]</td>
</tr>
<tr>
<td>Lee et al 2012</td>
<td>-0.10 [-0.71, 0.50]</td>
<td>1.90 [1.17, 2.63]</td>
<td>1.34 [0.67, 2.01]</td>
</tr>
<tr>
<td>Lee et al 2012</td>
<td>4.15 [3.05, 5.26]</td>
<td>0.13 [-0.74, 1.01]</td>
<td>0.95 [-0.24, 2.15]</td>
</tr>
<tr>
<td>Lee et al 2012</td>
<td>0.53 [-0.62, 1.69]</td>
<td>0.53 [-0.62, 1.69]</td>
<td>0.36 [-0.53, 1.24]</td>
</tr>
<tr>
<td>Lee et al 2012</td>
<td>1.02 [0.09, 1.95]</td>
<td>0.05 [-0.82, 0.93]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Hassani et al 2016</td>
<td>1.95 [1.23, 2.67]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.36 [-0.53, 1.24]</td>
</tr>
<tr>
<td>Wolska &amp; Wilske 2010a</td>
<td>2.80 [1.56, 4.03]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Wolska &amp; Wilske 2010a</td>
<td>0.41 [-1.04, 0.21]</td>
<td>0.41 [-1.04, 0.21]</td>
<td>0.36 [-0.53, 1.24]</td>
</tr>
<tr>
<td>Bouillon et al 2011</td>
<td>1.20 [0.52, 1.87]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Hassani et al 2016</td>
<td>1.35 [0.86, 1.83]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Hassani et al 2016</td>
<td>0.77 [0.14, 1.39]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Hassani et al 2016</td>
<td>0.81 [0.24, 1.39]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Chiu et al 2007</td>
<td>0.60 [0.03, 1.17]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Bouillon et al 2011</td>
<td>0.19 [-0.37, 0.74]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Kim 2016</td>
<td>0.19 [-0.37, 0.74]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Bouillon et al 2011</td>
<td>0.71 [0.07, 1.35]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>0.71 [0.07, 1.35]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>0.50 [-0.56, 1.57]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Noh et al 2012</td>
<td>0.48 [-0.13, 1.10]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Chiu et al 2007</td>
<td>0.33 [-0.72, 1.39]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Bouillon et al 2011</td>
<td>0.62 [-0.03, 1.27]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>0.55 [-0.10, 1.19]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>0.54 [-0.40, 1.48]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Lee et al 2014</td>
<td>1.28 [0.04, 2.53]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Wilska &amp; Wilske 2011</td>
<td>1.28 [0.04, 2.53]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Wilska &amp; Wilske 2010b</td>
<td>0.62 [-0.46, 1.69]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Wilska &amp; Wilske 2011</td>
<td>0.31 [-0.83, 1.45]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Wilska &amp; Wilske 2010b</td>
<td>0.11 [-0.94, 1.16]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Petersen 2010</td>
<td>0.48 [-0.08, 1.04]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Petersen 2010</td>
<td>1.90 [0.79, 3.02]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Harless et al 1999</td>
<td>-0.21 [-0.45, 0.03]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Harless et al 1999</td>
<td>1.78 [1.25, 2.32]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Rosenthal–von der Putten et al. 2016</td>
<td>2.04 [1.58, 2.50]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Jia et al 2013</td>
<td>1.10 [0.67, 1.52]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
<tr>
<td>Jia et al 2013</td>
<td>1.10 [0.67, 1.52]</td>
<td>0.29 [-0.10, 0.69]</td>
<td>0.29 [-0.10, 0.69]</td>
</tr>
</tbody>
</table>

Test of Moderators:
QM(df=3) = 7.098, p = 0.069
<table>
<thead>
<tr>
<th>Study</th>
<th>System Modality</th>
<th>Test Modality</th>
<th>Standardized Mean Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harless et al 1999</td>
<td>spoken</td>
<td>spoken</td>
<td>0.54 [−0.40, 1.48]</td>
</tr>
<tr>
<td>Lee et al 2012</td>
<td>spoken</td>
<td>spoken</td>
<td>−0.10 [−0.71, 0.50]</td>
</tr>
<tr>
<td>Rayner &amp; Tsourakis 2013</td>
<td>spoken</td>
<td>spoken</td>
<td>1.50 [0.60, 2.41]</td>
</tr>
<tr>
<td>Chiu et al 2007</td>
<td>spoken</td>
<td>spoken</td>
<td>0.29 [−0.10, 0.69]</td>
</tr>
<tr>
<td>Harless et al 1999</td>
<td>spoken</td>
<td>spoken</td>
<td>1.90 [0.79, 3.02]</td>
</tr>
<tr>
<td>Kim 2016</td>
<td>spoken</td>
<td>spoken</td>
<td>4.15 [3.05, 5.26]</td>
</tr>
<tr>
<td>Kim 2016</td>
<td>spoken</td>
<td>spoken</td>
<td>1.95 [1.23, 2.67]</td>
</tr>
<tr>
<td>Hassani et al 2016</td>
<td>spoken</td>
<td>spoken</td>
<td>0.77 [0.14, 1.39]</td>
</tr>
<tr>
<td>Hassani et al 2016</td>
<td>spoken</td>
<td>spoken</td>
<td>0.13 [−0.74, 1.01]</td>
</tr>
<tr>
<td>Hassani et al 2016</td>
<td>spoken</td>
<td>spoken</td>
<td>0.05 [−0.82, 0.93]</td>
</tr>
<tr>
<td>Hassani et al 2016</td>
<td>spoken</td>
<td>spoken</td>
<td>0.36 [−0.53, 1.24]</td>
</tr>
<tr>
<td>Lee et al 2014</td>
<td>spoken</td>
<td>spoken</td>
<td>0.19 [−0.37, 0.74]</td>
</tr>
<tr>
<td>Lee et al 2012</td>
<td>spoken</td>
<td>spoken</td>
<td>1.37 [0.70, 2.04]</td>
</tr>
<tr>
<td>Lee et al 2012</td>
<td>spoken</td>
<td>spoken</td>
<td>1.87 [1.15, 2.60]</td>
</tr>
<tr>
<td>Lee et al 2012</td>
<td>spoken</td>
<td>spoken</td>
<td>1.90 [1.17, 2.63]</td>
</tr>
<tr>
<td>Lee et al 2012</td>
<td>spoken</td>
<td>spoken</td>
<td>1.34 [0.67, 2.01]</td>
</tr>
<tr>
<td>Lee et al 2014</td>
<td>spoken</td>
<td>spoken</td>
<td>0.48 [−0.08, 1.04]</td>
</tr>
<tr>
<td>Rosenthal–von der Putten et al 2016</td>
<td>spoken</td>
<td>written</td>
<td>−0.21 [−0.45, 0.03]</td>
</tr>
<tr>
<td>Bouillon et al 2011</td>
<td>spoken</td>
<td>written</td>
<td>2.80 [1.56, 4.03]</td>
</tr>
<tr>
<td>Noh et al 2012</td>
<td>spoken</td>
<td>written</td>
<td>1.35 [0.86, 1.83]</td>
</tr>
<tr>
<td>Bouillon et al 2011</td>
<td>spoken</td>
<td>written</td>
<td>1.02 [0.09, 1.95]</td>
</tr>
<tr>
<td>Wolska &amp; Wilske 2011</td>
<td>written</td>
<td>spoken</td>
<td>1.28 [0.04, 2.53]</td>
</tr>
<tr>
<td>Wilske 2014</td>
<td>spoken</td>
<td>spoken</td>
<td>0.62 [−0.46, 1.69]</td>
</tr>
<tr>
<td>Wolska &amp; Wilske 2011</td>
<td>written</td>
<td>spoken</td>
<td>0.31 [−0.83, 1.45]</td>
</tr>
<tr>
<td>Wilske 2014</td>
<td>spoken</td>
<td>spoken</td>
<td>0.11 [−0.94, 1.16]</td>
</tr>
<tr>
<td>Jia et al 2013</td>
<td>written</td>
<td>written</td>
<td>1.78 [1.25, 2.32]</td>
</tr>
<tr>
<td>Jia et al 2013</td>
<td>written</td>
<td>written</td>
<td>2.04 [1.58, 2.50]</td>
</tr>
<tr>
<td>Jia et al 2013</td>
<td>written</td>
<td>written</td>
<td>1.10 [0.67, 1.52]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>written</td>
<td>written</td>
<td>−0.41 [−1.04, 0.21]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>written</td>
<td>written</td>
<td>0.81 [0.24, 1.39]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>written</td>
<td>written</td>
<td>1.20 [0.52, 1.87]</td>
</tr>
<tr>
<td>Arispe 2014</td>
<td>written</td>
<td>written</td>
<td>0.60 [0.03, 1.17]</td>
</tr>
<tr>
<td>Wolska &amp; Wilske 2010b</td>
<td>written</td>
<td>written</td>
<td>0.50 [−0.56, 1.57]</td>
</tr>
<tr>
<td>Wolska &amp; Wilske 2010a</td>
<td>written</td>
<td>written</td>
<td>0.95 [−0.24, 2.15]</td>
</tr>
<tr>
<td>Wilske &amp; Wilska 2011</td>
<td>written</td>
<td>written</td>
<td>0.71 [0.07, 1.35]</td>
</tr>
<tr>
<td>Wolska &amp; Wilske 2010b</td>
<td>written</td>
<td>written</td>
<td>0.33 [−0.72, 1.39]</td>
</tr>
<tr>
<td>Wolska &amp; Wilske 2010a</td>
<td>written</td>
<td>written</td>
<td>0.53 [−0.62, 1.69]</td>
</tr>
<tr>
<td>Wilske &amp; Wilska 2011</td>
<td>written</td>
<td>written</td>
<td>0.48 [−0.13, 1.10]</td>
</tr>
<tr>
<td>Petersen 2010</td>
<td>written</td>
<td>written</td>
<td>0.62 [−0.03, 1.27]</td>
</tr>
<tr>
<td>Petersen 2010</td>
<td>written</td>
<td>written</td>
<td>0.55 [−0.10, 1.19]</td>
</tr>
</tbody>
</table>

**Test of Moderators (coefficient(s) 2):**

$QM(df=1) = 0.086, p = 0.769$

**Spoken interaction in system:** $d = 0.95 ***$
Moderators
Other moderators/covariates

Learners variables: L1, age, context...
⇒ non significant

Instruments/outcomes: outcome measure group (accuracy/complexity/fluency/vocabulary)...
⇒ non significant

System characteristics: target language, dialogue type, system type, initiative, embodied agent, gamified...
⇒ non significant

Most likely due to multiple cases/coefficients and too few studies to establish any significance
Dialogue-based CALL: meta-analysis

Summary

Medium effect of dialogue-based CALL on L2 proficiency development
\[d = 0.90 \text{ (within)} / d = 0.62 \text{ (between)}\]

Possibly differentiated effect depending on proficiency level, system modality & test modality
But these observations still need to be confirmed by other studies

Need for more comparable designs, big enough samples and precise instruments
Future research should inscribe itself in this emerging field and compare its results within the field
Thank you! Merci! Dank u! ¡Gracias!

Serge Bibauw
serge.bibauw@kuleuven.be

Thomas François
thomas.francois@uclouvain.be

Piet Desmet
piet.desmet@kuleuven.be

Download this presentation and the full dataset at
http://serge.bibauw.be/calico