LANGUAGE LEARNING
*A Journal of Research in Language Studies*
**Language Learning ISSN 0023-8333**

## EMPIRICAL STUDY

# Interdisciplinary Research at the Intersection of CALL, NLP, and SLA: Methodological Implications From an Input Enhancement Project

Nicole Ziegler,[a] Detmar Meurers,[b] Patrick Rebuschat,[b,c] Simón Ruiz,[b] José L. Moreno-Vega,[c] Maria Chinkina,[b] Wenjing Li,[d] and Sarah Grey[e]

[a]University of Hawai'i at Mānoa, [b]University of Tübingen, [c]Lancaster University, [d]Michigan State University, and [e]Fordham University

Despite the promise of research conducted at the intersection of computer-assisted language learning (CALL), natural language processing, and second language acquisition, few studies have explored the potential benefits of using intelligent CALL systems to deepen our understanding of the process and products of second language (L2) learning. The strategic use of technology offers researchers novel methodological opportunities to examine how incremental changes in L2 development occur during treatment as well as how the longitudinal impacts of experimental interventions on L2 learning outcomes occur on a case-by-case basis. Drawing on the pilot results from a project examining the effects of automatic input enhancement on L2 learners' development, this article explores how the use of technology offers additional methodological and analytical choices for the investigation of the process and outcomes of L2 development, illustrating the opportunities to study what learners do *during* visually enhanced instructional activities.

**Keywords** second language acquisition; intelligent computer-assisted language learning; natural language processing; research methods; input enhancement

---

We would like to thank the anonymous *Language Learning* reviewers for their insightful comments and valuable suggestions on earlier versions of this manuscript. Any remaining errors are of course our own. This research was supported as part of the LEAD Graduate School & Research Network [GSC1028], a project of the Excellence Initiative of the German federal and state governments.

Correspondence concerning this article should be addressed to Nicole Ziegler, Department of Second Language Studies, University of Hawai'i at Manoa, Honolulu, HI 96822. E-mail: nziegler@hawaii.edu



DOI: 10.1111/lang.12227

## Introduction

During the last few decades, computer-assisted language learning (CALL) has rapidly evolved to include a diverse range of computer-mediated language learning activities, tools, materials, and technology-supported learning environments. This development of a diverse and constantly evolving set of technology-mediated tools and environments has provided researchers with opportunities to more closely investigate what we think we know about well-known processes of second language (L2) learning. Although the field of second language acquisition (SLA) has made substantial progress since its establishment (e.g., Ortega, 2014; Pica, 1997), as Plonsky (2013) points out, little research in the field has attempted to describe the how of L2 acquisition, perhaps due to the logistical difficulty of measuring learners' longitudinal, as well as individual, performance during treatment. Seeking to address how researchers might begin to more thoroughly investigate this important aspect of L2 learning, this article examines how the strategic use of natural language processing (NLP) technology integrated into intelligent CALL (ICALL) may offer insight into the processes and products of L2 acquisition by providing information that might otherwise be unavailable via traditional methods of classroom and laboratory data collection, thus potentially illuminating how L2 development may happen incrementally over time. In addition, the current exploratory study also provides an example of how ICALL may facilitate the investigation of instructed L2 acquisition beyond the classroom to real-life language learning in the wild (MacWhinney, 2017), an underexplored context in need of further investigation.

The goal of the current research is to demonstrate the potential of employing an ICALL system, in which learners' actions and engagement with the input are automatically tracked and collected, for mainstream SLA research. In order to deepen our understanding of classic empirical issues in SLA, such as the impact of instructional interventions, we draw on the results of a pilot study using an ICALL system to provide enhanced, authentic texts to learners outside of the classroom. By highlighting the unique methodological affordances available within an ICALL system to examine the effects of input enhancement on L2 development, we aim to illustrate the potential contributions of research at the intersection of ICALL and SLA.

## Input Enhancement and L2 Development

Designed to increase learners' selective attention to input, input enhancement uses a range of cues to increase the saliency and noticing of the target form, including underlining, bolding, italicization, capitalization, color coding, and different font sizes or types (Lee & Huang, 2008; Leow, 2007; Sharwood Smith,

1993). Because noticing is thought to facilitate the acquisition of features in the target language (Schmidt, 1990, 2001), learning contexts that promote the noticing of forms and gaps between learners' interlanguage and the target language, such as visually enhanced input, may optimally facilitate L2 development (Izumi, 2002). Previous research, however, has yielded mixed findings (e.g. Lee, 2007). For example, Polio's (2007) review suggests that visually enhanced input, such as coloring and bolding, had minimal positive effects on L2 development. More recent meta-analytic work, however, has indicated that learners exposed to visually enhanced learning materials performed better on measurements of L2 learning outcomes than learners that did not use enhanced texts (Lee & Huang, 2008).

Although the results of Lee and Huang's (2008) meta-analysis seem to suggest overall positive effects, the authors also pointed out a number of limitations to the body of research in general (see also Han, Park, & Combs, 2008 for a review). For example, findings show a lack of sustained and continued treatment periods, with two-thirds of the meta-analyzed sample employing less than three treatments totaling 2 hours or less of exposure to enhanced materials. Relatedly, Lee and Huang (2008) highlight the relatively small number of word types and tokens that learners are exposed to, with the primary studies included in the sample frequently exposing learners to one text and number of token per target item most commonly under 50. Finally, many of the studies examined in Lee and Huang (2008) used only one measure to assess learners' grammatical development, while the majority also focused on production over receptive skills. By approaching assessment from a range of perspectives, multiple measures may increase the likelihood that a learning effect, which may have been missed by a single measure, might then be detected. Such findings highlight the need for longer and more varied treatments lasting across several sessions, as well as broader methods of measurement, to better our understanding of the role of input enhancement on L2 development.

## Computerized Input Enhancement

Within a CALL context, the effects of input enhancement may be improved, as enhancement options can be more versatile and target items can be underlined, highlighted, repeated, or expanded upon in a dynamic fashion (Chapelle, 1998, 2003). Input enhancement may be particularly natural in Web-based language instruction when compared to static or paper-based texts because designers have many options to enhance the saliency of target forms and structures, including color, graphics, animations, acoustic enhancements, and keyboard transcription (Collentine, 1998; Gascoigne, 2006; Labrie, 2000; Russell, 2012,

2014). In addition to the increased textual enhancement options offered by computerized techniques, computer-mediated input enhancement also expands the opportunities for learners to engage with authentic, contextualized examples of target language features in dynamic ways. For example, they provide the chance for learners to go beyond receptive practice by interacting with the text via clicking, filling in the blank, or selecting multiple-choice options in addition to the automatic colorization of the target item. These options, available in ICALL approaches such as the Working with English Real Texts interactively (WERTi) system (Meurers et al., 2010) that integrate NLP to allow a learner to select any text of interest for enhancement, may further benefit L2 development by supporting learners' processing, comprehension, and retention of input (Chapelle, 2003; Hulstijn, 2001; Loschky, 1994). Indeed, the potential benefits of enhancing the "vividness" of input for recall and retention were noted decades ago (e.g., Baddeley, 1976; Jerisld, 1962). More recently, scholars have highlighted the many ways that computer-mediated contexts might make input more vivid through color, font, sound, and heightened interactivity (Cho & Reinders, 2013; Gascoigne, 2006), thereby increasing opportunities for noticing and subsequent development.

Discussions of the potential benefits of technology for language learning (e.g., Chapelle, 2005; Golonka, Bowles, Frank, Richardson, & Freynik (2014); Grgurović, Chapelle & Shelley, 2013; Heift & Schulze, 2007; Plonsky & Ziegler, 2016) have highlighted the potential of ICALL to support the development of a range of L2 skills (Schulze & Heift, 2012) and emphasized the importance of integrating opportunities for both form- and meaning-focused instruction (Amaral & Meurers, 2011). In Authentic Text ICALL (Meurers, 2012), the input also remains fully integrated in an authentic context. For example, in the WERTi system the salience of specific target language forms (determiners, prepositions, phrasal verbs, and wh-questions) is enhanced in learner-selected Web pages by automatically producing visual enhancement of these pages on the fly using NLP. Learners maintain control in terms of their content selection, supporting autonomy and motivation (e.g., Dörnyei & Ushioda, 2011; Ryan & Deci, 2000) while nonetheless drawing on teacher knowledge about challenging grammatical features, automatically identified through the NLP, in order to raise learners' awareness of these forms.

## The Current Study

ICALL research traditionally has shown relatively limited awareness of SLA theories and constructs, such as focus on form and noticing, and there is very little direct research on the potential methodological contributions of ICALL

research to SLA (for exceptions, see Heift, 2010; Meurers, 2012; Schulze, 2008). The goal of this study is to explore the promise for research conducted at the intersection of ICALL and instructed SLA and to illustrate this by drawing on the results of a small-scale pilot study examining the efficacy of computerized input enhancement on the development of L2 grammar. We ask the following research questions:

1. What are the methodological implications of using an ICALL system to administer automatic computerized input enhancement?
2. What are the effects of different types of automatic computerized input enhancement on the development of learners' implicit and explicit knowledge?

To address the first research question, we examine the possible affordances of research conducted at the intersection of ICALL and SLA, including implications for data collection, analysis, and interpretation, for exploring how this provides new opportunities for deepening our understanding of the processes and products of instructed L2 acquisition. Based on the multifaceted perspective on the potential benefits of using NLP technology for examining and supporting L2 development, we then characterize possible research agendas for future ICALL and SLA investigations.

## Method

### Participants

Fifty postgraduate students at Lancaster University (43 women, 7 men), with a mean age of 25.41 ($SD = 4.01$), volunteered to take part in the experiment and were randomly assigned to one of five conditions (each $n = 10$): Unenhanced (control), Color, Click, Multiple Choice (MC), and Fill-in-the-blank (FIB). All participants were native speakers of East Asian languages, including Mandarin (38), Thai (7), Cantonese (2), Japanese (2), and Hokkien (1). English language proficiency was assessed using learners' most recent International English Language Testing System or Test of English as a Foreign Language scores, with learners falling between intermediate and upper-intermediate. All participants reported owning a computer and being generally comfortable using one, with average use for a typical day at 6.69 hours ($SD = 2.33$). Participants reported spending an average of 2.72 hours per day ($SD = 1.33$) reading news Web sites, blogs, research articles, and social network entries. Participants were eligible to win Amazon vouchers in return for participating.

**Linguistic Target**

English articles were selected as the linguistic target feature for the current study given the well-documented difficulties in acquisition by learners whose first language (L1) does not have articles (Ellis, Sheen, Murakami, & Takashima, 2008; Huong, 2005; Muranoi, 2000; Robertson, 2000). This may stem from English articles having low salience and their use not following a simple rule. To our knowledge, little research on input enhancement has focused on the learning of articles by English learners (although see Ha, 2005), and specifically no other study has tested Chinese, Japanese, and Thai L1 speakers, who represent a substantial portion of worldwide English learners.

**Materials**

The treatment was administered via the WERTi system (Meurers et al., 2010), a Web-based tool providing automatic visual input enhancement of authentic texts on the Web to support L2 learning. The system integrates standard NLP tools to identify the targeted language patterns. While the WERTi tool focuses on English, the general Visual Input Enhancement of the Web approach (http://purl.org/icall/view) supports multiple languages and target patterns. For the experiment, a customized version of WERTi was created and made available online. This version exclusively targets articles and integrates explicit logging of all system interaction based on a login that automatically assigns every subject to one of the four treatment groups or the control group.[1]

The study used a pretest/posttest design consisting of a battery of four tests adapted from Akakura (2009). There were two versions of each test (A and B); these versions were counterbalanced within participants and across the pretest/posttest sessions.

*Elicited Production Task (EPT)*

Following Akakura (2009), participants were shown a PowerPoint document with two stories taken from children's books with no text, namely, *Frog, Where Are You?* (test A; Mayer, 1969) and *Frog Goes to Dinner* (test B; Mayer, 1974), both of which have been used extensively in previous SLA research (e.g., Akakura, 2009, 2012; Housen et al., 2011; Negueruela & Lantolf, 2006; Treffers-Daller, 2011). In addition, this type of task has been shown to elicit a range of linguistic features during spontaneous production (for a more detailed discussion, see Sanchez & Jarvis, 2008). In each case, there were 14 sentences that formed the story and participants had to produce various target articles within the story context. In contrast to traditional elicited imitation tasks, which require participants to repeat decontextualized sentences, this type of

story-based EPT allowed us to test the use of a specific target feature while participants focus on meaning.

*Oral Narration Task (ONT)*
This task was also delivered via PowerPoint and used the same 14 pictures as the EPT. Learners were asked to explain the story in their own words as if they were telling the story to a child. In order to make sure that this test measured implicit knowledge, participants had restricted time (5 minutes) to complete this task (Ellis, 2007).

*Grammaticality Judgment Task (GJT)*
An untimed GJT was used as a measure of explicit knowledge (see Loewen, 2009). This task consisted of 10 items and was delivered via Eprime. For each trial, participants read a sentence on the screen and then judged whether it was grammatical or not. Completion of this task took approximately 10 minutes. Confidence ratings and source attributions were also collected (see Rebuschat, 2013, for discussion). However, the results of the subjective measures of awareness are beyond the scope of this article and are therefore not reported below.

*Metalinguistic Knowledge Task (MKT)*
The untimed MKT test also measured explicit knowledge. This test was delivered on PowerPoint and participants wrote their responses by hand on an answer sheet. Participants were shown five sentences that contained an underlined error related to article usage and were then instructed to correct the underlined error without changing any other part of the sentence. Because this task also assessed explicit knowledge, it was untimed, although completion took approximately 10 minutes.

**Procedure**
Participants were tested individually in a quiet laboratory setting. After providing informed consent, participants were introduced to the WERTi system and completed the four pretests (EPT, ONT, GJT, and MKT). They were asked to read 30 news articles over a 2-week period via the WERTi interface and to complete a short online questionnaire for each article. During the 2-week period, participants' reading and interaction with the system was tracked. In order to promote sustained exposure to the target structures, learners were sent reminders if they were not logging into the system on a regular basis. After two weeks, participants returned to complete the posttests, a debriefing questionnaire, and a background questionnaire.

*Intervention*

Each participant logged into the WERTi system with a unique username that corresponded to one of the exposure conditions. After logging in, participants were presented with a Web page that allowed them to search for current Reuters news items. Learners searched for articles by entering a topic that interested them in the search box. The system returned a list of results, consisting of titles and short abstracts. Once one of the search results was selected, the system showed the corresponding Reuters news Web page, which has been automatically enhanced according to one of four experimental conditions (Color, Click, MC, FIB). For the unenhanced (control) group, the system returned a Web page without any enhancement, except for a clickable box inserted into all news pages to allow the learners to go to the reading questionnaire after reading the text.

The bottom of the Reuters news search page also contained the only instructions that were specific to the group to which the participants were assigned. In the colorizing condition, participants were instructed to pay attention to the target words in a different color font (blue). In the Click condition, they were asked to click on any instance of *a*, *an*, and *the* in the text. Additionally, an example of how to click on the targeted words was provided. After clicking on a word, which could not be undone, the system provided immediate automatic color feedback, bold underlined green for correctly identified targets and red for incorrectly identified ones. For the MC condition, participants were required to select the correct targeted form from dropdown menus in order to complete this spot of the text. As in the Click condition, instant automatic color feedback was provided, but in this condition learners were able to select a different option until the correct target was selected. In addition, clicking on a smiley face next to the blank spaces made the system provide the correct answer (in blue). Subjects were instructed to only use the smiley face as a last resort. In the FIB condition, learners were asked to type in targeted forms to complete blank spaces in the text. Upon completing a blank by pressing enter, instantaneous automatic color feedback appeared just as in the MC case, and the smiley face option was equally available. At the bottom of the Reuters news search page, learners of all groups were instructed to complete a questionnaire after reading the selected text.

*Reading Questionnaire*

After each article, participants reported perceived text difficulty (ranging from 1 = *very easy* to 5 = *very difficult*), enjoyment of the text, and familiarity with text topic (ranging from 1 = *very much* to 5 = *very little*; Pino-Silva, 2006). This was done to ensure that the texts were relatively comparable across groups. Learners were also required to write a two-sentence summary

of the article, as a measure of general comprehension and engagement, and could optionally leave comments. At the time the reading questionnaire was completed, the article was no longer visible.

*Learner Log Files*

The WERTi system kept a log file for each learner, recording information about the news article and the targeted forms that appeared in it; the learner's interaction with these targeted forms in the Click, MC, and FIB conditions; and the completed reading questionnaire. It recorded the total number of words in the news article and the total numbers of targeted items (*a, an, the*) in each text; the total number of *attempted* and *unattempted* items, *incomplete* items (only for the MC and the FIB groups), *correct* and *incorrect* items, and *cheat* instances (clicking on the smiley face in MC or FIB). The *unattempted* items are those that the learner did not interact with in any way, so were mutually exclusive with *attempted*.

## Results and Discussion

### Pretest and Posttest Scores

The first research question examines the efficacy of computerized visual input enhancement on the development of learners' implicit and explicit knowledge. In order to test whether the treatments resulted in learning, we compared pretest and posttest performance across conditions. Results from a paired-samples *t* test indicated just one significant difference between groups. In the MC group, the posttest scores of the GJT were significantly greater than the pretest scores, $t(9) = -5.511$, $p < .001$ suggesting that the MC group was the only group to demonstrate a learning effect and that learning effect was restricted to explicit knowledge. No other pretest–posttest contrasts were significant. Table 1 summarizes participants' performance on the four tests.

The fact that the pretest–posttest comparisons only revealed a positive benefit for the MC group lends support to Polio's (2007) finding that narrowly defined enhancement techniques, such as colorization, are minimally effective at facilitating L2 development, while more broadly operationalized enhancement, such as might be found in the MC or Click conditions where learners were required to interact with the materials, may lead to relatively greater development. In addition, the only significant development was found in a measure of explicit knowledge, suggesting that there may not have been sufficient time for the development of learners' implicit knowledge.

While these results suggest that there may be positive benefits for computerized input enhancement, particularly when using types that require learners

**Table 1** Descriptive statistics for pretest and posttest comparison of percentage scores

| Test | | Control | | Color | | Click | | MC | | FIB | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Pre | Post | Pre | Post | Pre | Post | Pre | Post | Pre | Post |
| Elicited production | *M* | 67.3 | 68.5 | 70.2 | 64.9 | 72.9 | 69.8 | 71.2 | 78.1 | 68.1 | 65.8 |
| | *SD* | 10.9 | 15.8 | 18.3 | 13.4 | 12.4 | 18.0 | 13.2 | 10.7 | 14.5 | 16.2 |
| Oral narration | *M* | 89.9 | 94.6 | 89.4 | 92.5 | 90.5 | 92.5 | 93.4 | 93.9 | 91.4 | 93.3 |
| | *SD* | 5.43 | 4.81 | 7.53 | 5.93 | 7.62 | 7.71 | 5.64 | 7.39 | 6.19 | 7.82 |
| Grammaticality judgment | *M* | 7.4 | 7.9 | 7.2 | 6.7 | 8 | 8.1 | 6.5* | 8.30* | 8.1 | 8.6 |
| | *SD* | 2.07 | 1.66 | 1.87 | 1.7 | 1.56 | 1.91 | 0.71 | 0.95 | 1.1 | 0.97 |
| Metalinguistic knowledge | *M* | 1.82 | 1.88 | 1.92 | 1.92 | 1.9 | 1.98 | 1.88 | 1.96 | 1.78 | 1.86 |
| | *SD* | 0.22 | 0.32 | 0.14 | 0.1 | 0.14 | 0.06 | 0.25 | 0.08 | 0.22 | 0.21 |

*Note.* Significance from chance: $^{*}p < .001$, $t = -5.511$. MC = Multiple Choice; FIB = Fill-in-the-blank.

to actively engage with the input, the results should be interpreted with caution in light of the small number of participants and high number of conditions compared. This points to the more general conclusion that automating the intervention should be accompanied by Web-based administration of all pretests and posttests to support scaling up to sufficiently large sample sizes. In addition, these results highlight the possibility that by relying solely on pretest–posttest comparisons, important results, such as those of the Click condition discussed below, may be obscured by the lack of multiple measures of assessment.

**Treatment**

On average, participants read 27.18 texts ($SD = 6.05$) as part of the treatment. The average text was 414.96 words ($SD = 91.26$) in length and took 7:13 minutes ($SD = 3{:}45$) to read. Participants encountered 33.64 ($SD = 7.67$) target forms per text, with the definite article accounting for 67% of the target forms and the indefinite article (in both spelling variants) accounting for 33%. There were no significant differences between groups in terms of treatment, with all $p > .05$. Table 2 summarizes the reading behavior of the five groups during the treatment period.

Table 3 summarizes the results of the reading questionnaire, which prompted participants to rate each text on a scale from 1 to 5 in terms of difficulty, enjoyment, and familiarity. The mean difficulty rating was 2.25 ($SD = .71$), while the mean enjoyment rating was 2.43 ($SD = .69$) and the mean

**Table 2** Descriptive statistics for reading behavior in each group during the treatment period

| | Control | | Color | | Click | | MC | | FIB | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Article number | 27.70 | 4.90 | 26.80 | 7.97 | 26.60 | 5.99 | 30.20 | 0.63 | 24.60 | 7.56 |
| Article length | 452.85 | 101.62 | 399.04 | 65.10 | 429.14 | 101.83 | 416.36 | 105.41 | 374.90 | 73.56 |
| Time spent | 4:37 | 2:05 | 5:35 | 3:10 | 8:38 | 4:15 | 8:53 | 4:00 | 8:22 | 3:18 |
| Total number of linguistic targets | 1015.20 | 315.78 | 856.20 | 301.78 | 922.10 | 257.29 | 1000.40 | 248.71 | 743.50 | 211.24 |
| Average number of linguistic forms | 36.64 | 8.83 | 31.71 | 5.26 | 35.72 | 9.38 | 33.09 | 8.04 | 31.04 | 5.97 |
| Target *a* | 10.63 | 2.29 | 9.09 | 1.90 | 9.74 | 3.01 | 9.26 | 2.20 | 8.93 | 2.43 |
| Target *an* | 1.79 | 0.60 | 1.68 | 0.51 | 1.58 | 0.45 | 1.44 | 0.51 | 1.46 | 0.54 |
| Target *the* | 24.22 | 6.2 | 21 | 3.6 | 24.4 | 6.5 | 22.4 | 5.8 | 20.7 | 3.47 |

*Note.* MC = Multiple Choice; FIB = Fill-in-the-blank.

**Table 3** Descriptive statistics for Likert-type scale items in the reading questionnaire

| | Control | | Color | | Click | | MC | | FIB | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Difficulty | 2.25 | 0.64 | 2.52 | 0.77 | 2.08 | 0.45 | 1.78 | 0.56 | 2.64 | 0.83 |
| Enjoyment | 2.41 | 0.6 | 2.68 | 0.73 | 2.62 | 0.5 | 1.78 | 0.6 | 2.68 | 0.68 |
| Familiarity | 2.55 | 0.56 | 2.8 | 0.76 | 2.68 | 0.7 | 2.11 | 0.83 | 2.89 | 0.70 |

*Note.* MC = Multiple Choice; FIB = Fill-in-the-blank.

familiarity rating was 2.61 ($SD = .74$). There were no significant differences between groups in terms of text difficulty, enjoyment, or familiarity ($p > .05$).

**Development of Accuracy During Treatment**

Given the purpose of this pilot study to illustrate the options for analyzing the process of learning in addition to the standard pretest/posttest analysis, we next turn to an analysis of the logs providing a more fine-grained, incremental assessment of learner development.

The scatterplot in Figure 1 gives an overall impression of the development of the precision with which individual learners completed the Click, FIB, and MC activities. Each mark corresponds to a document read by a learner, with the *x*-axis showing the cumulative number of items a learner had interacted with up to that point. The linear regression lines for each of the three conditions, with 95% confidence intervals indicated in grey, show that the accuracy of learners in the Click condition was higher than in the other two conditions, often reaching ceiling, and that learners interacted more with the text in the Click condition, clicking on up to 1,250 items in the 30 texts they read.

Focusing on the individual learners for the Click condition, the individual regression lines in Figure 2 show that many individuals perform close to perfect from the start, with half a dozen showing improvement in the first half of their interactions, and only one learner apparently dropping in precision.

Turning from the visual overview to a more precise, statistical analysis, we used a linear mixed-effects regression model to explore the incremental process of learning when participants interacted with the target items in the Click, FIB, and MC conditions. Specifically, we used R (R Core Team, 2015) and *lme4* (Bates, Maechler, & Bolker, 2015) to perform a linear mixed-effects analysis of the detailed system interaction log data. Linear mixed-effects regression allows researchers to account for individual variation among participants and test items. It supports dealing with both fixed and random effects at once (Baayen, 2008; Jaeger, 2008). A fixed effect is a difference between observations with

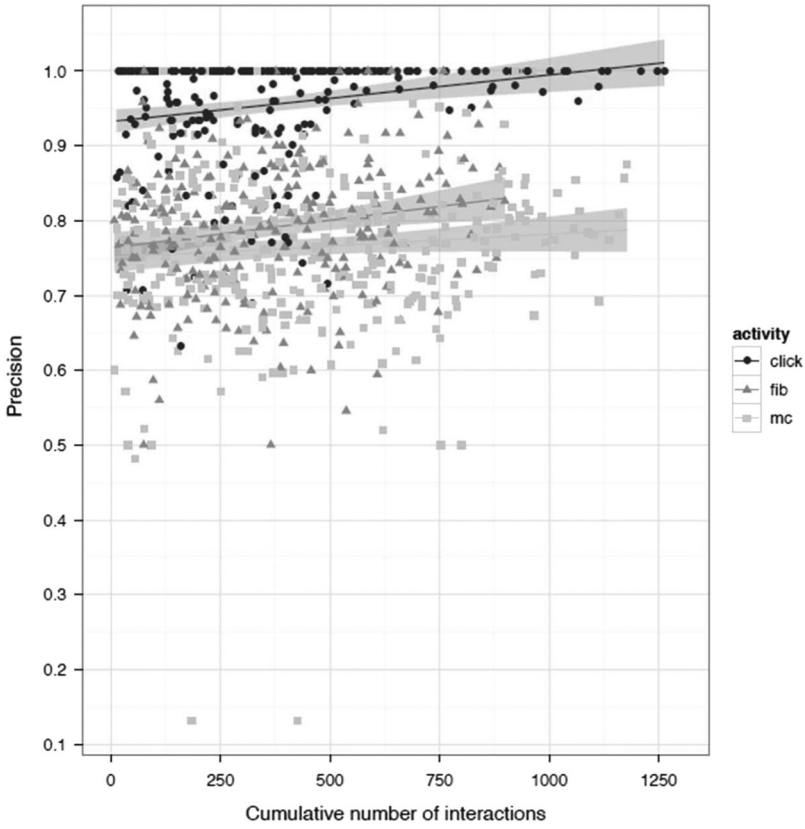


**Figure 1** Scatterplot of precision by interacted items with group regression lines.

different values on an independent variable, with different levels that convey more information than just being different identities (e.g., treatment vs. control groups). A random effect, on the other hand, is the variance of effects as a function of, in principle, randomly drawn observations. Simply put, fixed effects are employed to account for variation due to experimental factors, whereas random effects account for variation due to different observations that cannot be distinguished beyond their being different observations. *Activity type* (three levels: Click, FIB, MC) and the interaction of its levels with *time* (i.e., every time the learner logged into the system, to read the first text, the next text, up to the 30 texts) were entered into the model as fixed effects. As random effect, random intercepts were estimated for *items* (i.e., number of target items interacted with by each learner) and *learners* (i.e., random

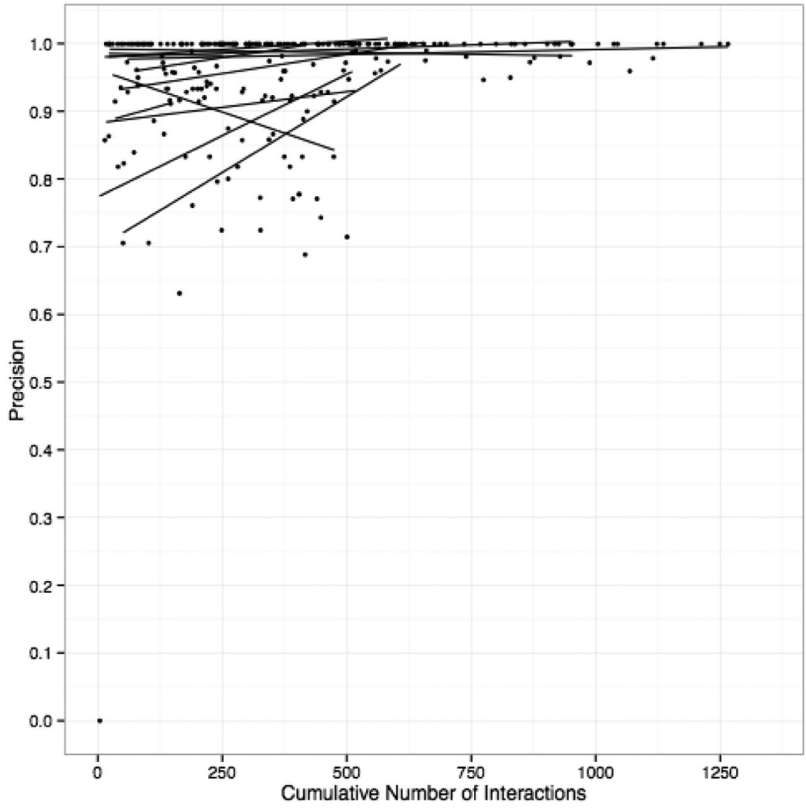


**Figure 2** Scatterplot of precision for Click group with individual regression lines.

differences of difficulty between items and random differences in accuracy between individuals). The regression model included all fixed effects, as well as two two-way interaction terms (i.e., two dummy-coded activity type variables by time). We report a summary of the model in Table 4.

The interaction of activity type and time was significant, $\chi^2(2) = 21.776$, $p < .001$, meaning that the effect of time on accuracy is different across activity type groups. In order to assess the effects of time within the different activity type groups, we calculated simple slopes. For the Click condition, there was a positive effect of time, and it was significant (coefficient $= 0.020$, $z = 2.639$ $p < .001$). For the FIB condition, there was a positive effect of time, but it was nonsignificant (coefficient $= 0.007$, $z = 1.823$, $p > .006839$). For the MC condition, the time slope was significant and negative (coefficient $= 0.006$, $z = –2.166$, $p < .05$).

**Table 4** Linear mixed-effects regression for incremental tracking of performance

| Fixed effect | Estimate | *SE* | *z* value |
|---|---|---|---|
| Intercept | 2.953 | 0.204 | 14.503*** |
| ActivityFIB | −1.728 | 0.259 | −6.679*** |
| ActivityMC | −1.736 | 0.260 | −6.678*** |
| Time | 0.020 | 0.007 | 2.639** |
| ActivityFIB:Time | −0.013 | 0.008 | −1.597 |
| ActivityMC:Time | −0.025 | 0.008 | −3.215** |
| Random effects | Variance | *SD* | |
| Item | 0.017 | 0.130 | |
| Learner | 0.343 | 0.586 | |
| Residual | 1 | | |

*Note.* *SE* = standard error; FIB = Fill-in-the-blank; MC = Multiple Choice. $^{***}p < .0001$, $^{**}p < .001$.

Overall, the strongest time effect was in the Click condition, indicating that the longer learners interacted with the target items, the more accurate they became at identifying articles in the texts over the course of the treatment. Of course, the MC and FIB tasks essentially were different in nature in that the system selected the article in the text and the learner had to choose the right one (*a, an, the*). It should also be noted that for the current research the pilot experiment was not designed to isolate the impact of input enhancement from the effect of learners' selecting one of several options in the MC condition. Future research should consider isolating these variables in order to obtain a better understanding of their empirical effects. However, because the goal of the current study was to explore the possibilities of research conducted at the intersection of ICALL and SLA, this is not addressed further in this article.

Regarding the random effects, *items* explain relatively little variance, given that the random intercept of the items do not vary greatly around the overall intercept. However, *learners* show considerable variance around the intercept, which is about 25% of the entire variance in the design, as explained by individual differences between learners. This confirms the relevance of taking into account explicit measures of individual differences in order to further analyze which specific aspects are potent here.

## Research at the Intersection of CALL, NLP, and SLA: Opportunities and Challenges

Surfacing from the specific pilot study discussed for illustration, we want to argue that the current lack of a stronger connection between SLA and the NLP

research underlying ICALL is indeed a missed opportunity. The strategic use of technology offers not only the potential to provide insight into the processes and products of L2 acquisition, but also opportunities for methodological advancement that are not readily available using traditional experimental methods. While research methods in SLA have improved during the past few decades, there remains room for improvement, for example, in terms of testing practices and sample sizes (Plonsky, 2013). Drawing on the results of the pilot study, we turn to a discussion of how the use of technology in a quantitative research design provides additional perspectives.

The current study examined whether automatic input enhancement of authentic language learning materials contributed to the development of L2 learners' implicit and explicit knowledge of English articles. Initial results, which relied on a traditional approach using learners' pretest and posttest scores, demonstrated a positive effect for the MC enhancement conditions on the measures of explicit knowledge, indicating the benefits of dynamic input enhancement on learners' L2 knowledge. Though these results provide important information regarding the efficacy of input enhancement in computer-mediated contexts, they offer a relatively narrow perspective on learners' possible development. For example, by relying on correlations and gains from pretest to posttest, our understanding of the impact of input enhancement is restricted in two ways: (1) It is only possible to observe trends in terms of the group as a whole population, rather than as a number of individuals and (2) it is only possible to view development in terms of relatively few measurements over time.

Using an ICALL system, however, in which learners' actions and engagement with the system are automatically tracked and collected, provides researchers with opportunities to obtain fine-grained logs of both learners' activity and development over time. For example, the WERTi system automatically collects information on a diverse number of factors directly related to learner interaction with the enhancement conditions. These log files provide researchers with a range of options for analyzing and measuring learner development beyond the perspective of a traditional pretest/posttest design. By tracking learners *during* treatment, it is possible to obtain a more fine-grained and nuanced understanding using a time-series design, providing insight into learners' individual trajectories of development. Time-series designs are particularly suited to examining individual variation and the process of change, a relevant goal given that research has suggested that L2 acquisition is nonlinear and incremental with small changes occurring over long periods of time (e.g. Huebner, 1983; Klein & Purdue, 1997; Lantolf &

Aljaafreh, 1995; Stauble, 1978). Furthermore, traditional group designs are unable to apply findings from group means to individual cases, providing only a general perspective on the effects of the treatment on the population of interest.

Tracking the individual interaction of learners using Web-based CALL systems makes it possible to obtain detailed individual logs of the learning process. In contrast to lab-based methods such as eye tracking, providing fine-grained online data over relatively short periods of time, learners can use Web-based CALL tools for months in their ordinary learning contexts, at home or in an institutional learning environment, leaving an incremental record of their interaction during real-life learning. The individual learner logs thereby can supply the data necessary to examine each learner's path of acquisition (cf. Murakami, 2013). With ICALL systems such as WERTi providing automatic enhancement of any learner-selected content, the individual learner log files record fine-grained data of each of the learners' activities, including their exposure in terms of number of words, number of target items, and amount of engagement with the text. Given that over 60% of the studies examining the effects of input enhancement used designs with less than three treatments totaling 2 hours or less of exposure to enhanced materials (Lee & Huang, 2008), longer studies providing more detailed information regarding the intensity and amount of exposure to input and enhanced materials would be crucial to understand the role of input enhancement in L2 development, as well as the intensity and durability of input enhancement treatment in practice. Based on individual learner logs from ecologically valid activities such as reading news articles on the Web in a Web-based interface such as WERTi, it is possible to control for exposure to input as well as the instances of enhanced texts as potentially mediating variables on the benefits associated with input enhancement. By asking learners to choose the texts they want to read using a linguistically aware search engine such as Form-Focused Language-Aware Information Retrieval (Chinkina & Meurers, 2016), it also is possible to influence learners to select texts that richly represent targeted language patterns. In addition to issues related to such input enrichment and input enhancement, being able to track how learners engage with the texts they read may also deepen our understanding of whether learners' interaction with texts facilitates L2 development by supporting learners' processing, comprehension, and retention of input (Chapelle, 2003; Hulstijn, 2001; Loschky, 1994). Although the sample size of the current study is small, the findings highlight the potential for this type of interdisciplinary work to deepen our understanding of how incremental changes in L2 development might occur during treatment, as well as how the longitudinal impacts of

experimental interventions on L2 learning outcomes occur on a case-by-case basis.

In addition to the analytical advantages that ICALL systems might provide for examining the process and outcomes of L2 acquisition, such systems offer opportunities for large-scale, systematic administrations of testing and treatments. Plonsky's (2013) recent analysis demonstrated that the median group size in SLA research was 19 participants. Previous research has highlighted the potentially negative effects of the small sample sizes typical in L2 research, particularly in terms of reduced statistical power (e.g., Flahive & Ehlers-Zavala, 2010; Norris & Ortega, 2006; Oswald & Plonsky, 2010). Although a number of scholars have called for greater awareness of the implications of statistical power (e.g., Larson-Hall, 2010; Lazaraton, 1991; Plonsky, 2013; Plonsky & Gass, 2011; Ziegler, 2016a, 2016b), due to logistical and situational challenges of recruiting and sustaining large numbers of participants, SLA researchers may be unable to obtain more participants despite understanding the importance of doing so. Though not applicable for every research question and design, the use of Web-based ICALL systems might provide options for increasing sample size by reducing logistical issues, such as the need for learners to attend in-person lab or classroom sessions and returning multiple times for testing administration—at least when the relevant tests are fully integrated into the Web-based setup, which requires more collaboration of SLA and ICALL researchers.

## Conclusion

During the last few decades, CALL and NLP-supported ICALL has rapidly evolved to include a diverse range of computer-mediated language learning activities, tools, materials, and Web-based learning environments. These developments are providing researchers with opportunities to more closely investigate well-known processes and constructs of L2 acquisition and learning. The case study presented in this article illustrates the promise of research conducted at the intersection of CALL, NLP, and instructed SLA by not only providing further empirical evidence supporting the efficacy of input enhancement for L2 development, but by also exploring how innovative technologies might deepen our understanding of L2 acquisition. While ICALL systems employing NLP analysis are far from a cure-all, they offer researchers new or enhanced opportunities to obtain potentially well-scalable, multifaceted perspectives on the impact of instructional treatments, such as input enhancement, on the process and products of L2 learning.

Final revised version accepted 24 November 2016

## Note

1 The system can be accessed at http://purl.org/icall/werti-art by researchers using the logins test.1 (color), test.2 (click), test.3 (multiple choice), test.4 (fill-in-the-blank), test.5 (control) with the password Beatles.

## References

Akakura, M. (2009). Effect of explicit instruction on implicit and explicit second language knowledge: An empirical study on English article acquisition. Unpublished Ph.D. dissertation, The University of Auckland, New Zealand.

Akakura, M. (2012). Evaluating the effectiveness of explicit instruction on implicit and explicit L2 knowledge. *Language Teaching Research*, *16*, 1–29.

Amaral, L., & Meurers, D. (2011). On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, *23*, 4–24.

Baayen, R. H. (2008). *Analyzing linguistics data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baddeley, A. (1976). *The psychology of memory*. London: Basic Books.

Bates, D., Maechler, M., & Bolker, B. (2015). Package "lme4" (Version 1.1-12). Retrieved from http://lme4.r-forge.r-project.org

Chapelle, C. (1998). Multimedia CALL: Lessons to be learned from research on instructed SLA. *Language Learning & Technology*, *2*, 21–39.

Chapelle, C. A. (2003). Linking SLA task theory to technology based tasks. Paper presented at the Conference on Technology for SLL, Memorial Union, Iowa State University.

Chapelle, C. A. (2005). Interactionist SLA theory in CALL research. In J. L. Egbert & G. M. Petrie (Eds.), *CALL research perspectives* (pp. 53–64). Mahwah, NJ: Erlbaum.

Chinkina, M., & Meurers, D. (2016). Linguistically aware information retrieval: Providing input enrichment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 188–198). San Diego, CA: Association for Computational Linguistics.

Cho, M., & Reinders, H. (2013). The effects of aural input enhancement on L2 acquisition. In J. M. Bergsleithner, S. N. Frota, & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 123–138). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Collentine, J. (1998). Cognitive principles and CALL grammar instruction: A mind-centered, input approach. *CALICO*, *15*, 1–18.

Dörnyei, Z., & Ushioda, E. (2011). *Teaching and researching motivation* (2nd ed.). Harlow, UK: Longman.

Ellis, N. C. (2007). Implicit and explicit knowledge about language. In J. Cenoz & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed. Vol. 6): *Knowledge about language* (pp. 119–132). Berlin, Germany: Springer.

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System*, *36*, 353–371.

Flahive, D., & Ehlers-Zavala, F. (2010, March). Power analysis in applied linguistics research. Paper presented at the Annual Conference of the American Association for Applied Linguistics, Atlanta, GA.

Gascoigne, C. (2006). Explicit input enhancement: Effects on target and non-target aspects of second language acquisition. *Foreign Language Annals*, *39*, 551–564.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *1*, 70–105.

Grgurović , M., Chapelle, C., & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, *25*, 165–198.

Ha, J. (2005). *Developing English determiners through Internet chat: An experiment with Korean EFL students*. Unpublished doctoral dissertation, University of Florida, Gainesville.

Han, Z., Park, E. S., & Combs, C. (2008). Textual enhancement: Issues and possibilities. *Applied Linguistics*, *29*, 597–618.

Heift, T. (2010). Bridging computational and applied linguistics: Implementation challenges and benefits. *Language Teaching*, *43*, 102–105.

Heift, T., & Schulze, M. (2007). *Errors and intelligence in computer-assisted language learning: Parsers and pedagogues*. New York: Routledge.

Housen, A., Schoonjans, E., Janssens, S., Welcomme, A., Schoonheere, E., & Pierrard, M. (2011). Conceptualizing and measuring the impact of contextual factors in instructed SLA: The role of language prominence. *IRAL-International Review of Applied Linguistics in Language Teaching*, *49*, 83–112.

Huebner, T. (1983). Linguistic systems and linguistic change in an interlanguage. *Studies in Second Language Acquisition*, *6*(01), 33–53.

Hulstijn, J. H. (2001). Intentional and incidental second-language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge, UK: Cambridge University Press.

Huong, N. T. (2005). *Vietnamese learners mastering English articles* (Unpublished Ph.D. dissertation). University of Groningen, Groningen, Netherlands.

Izumi, S. (2002). Output, input enhancement and the noticing hypothesis: An experimental study on ESL relativization. *Studies in Second Language Acquisition*, *24*, 541–577.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Jerisld, A. (1962). Primacy, recency, frequency, and vividness. *Journal of Experimental Psychology*, *64*, 123–125.

Klein, W., & Perdue, C. (1997). The Basic Variety (or: Couldn"t natural languages be much simpler?). *Second Language Research*, *13*, 301–347.

Labrie, G. (2000). A French vocabulary tutor for the Web. *CALICO Journal*, *17*, 475–500.

Lantolf, J. P., & Aljaafreh, A. (1995). Second language learning in the zone of proximal development: A revolutionary experience. *International Journal of Educational Research*, *23*(7), 619–632.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R*. New York: Routledge.

Lazaraton, A. (1991). Power, effect size, and second language research: A researcher comments. *TESOL Quarterly*, *25*, 759–762.

Lee, S. K. (2007). Effects of textual enhancement and topic familiarity on Korean EFL students' reading comprehension and learning of passive form. *Language Learning*, *57*, 87–118.

Lee, S. K., & Huang, H. T. (2008). Visual input enhancement and grammar learning: A metaanalytic review. *Studies in Second Language Acquisition*, *30*, 307–331.

Leow, R. P. (2007). Input enhancement in classroom-based SLA research: An attentional perspective. In C. Gascoigne (Ed.), *Assessing the impact of input enhancement in second language education: Evolution in theory, research and practice* (pp. 37–52). Stillwater, MN: New Forums.

Loewen, S. (2009). Grammaticality judgment tests and the measurement of implicit and explicit L2 knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 65–93). Bristol, UK: Multilingual Matters.

Loschky, L. C. (1994). Comprehensible input and second language acquisition: What is the relationship? *Studies in Second Language Acquisition*, *16*, 303–323.

MacWhinney, B. (2017). A shared platform for studying second language acquisition. *Language Learning*.

Mayer, M. (1969). *Frog, where are you*? New York: Dial Press.

Mayer, M. (1974). *Frog goes to dinner*. New York: Dial Press.

Meurers, D. (2012). Natural language processing and language learning. In C. A. Chapelle (Ed.), *Encyclopedia of applied linguistics* (pp. 4193–4205). London: Wiley-Blackwell.

Meurers, D., Ramon Z., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V., et al. (2010). Enhancing authentic web pages for language learners. *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 10–18). Los Angeles, CA: Association for Computational Linguistics.

Murakami, A. (2013). Cross-linguistic influence on the accuracy order of L2 English grammatical morphemes. In S. Granger, S. Gaëtanelle, & F. Meunier (Eds.), Twenty years of learner corpus research. Looking back, moving ahead: Corpora and language in use - Proceedings 1 (pp. 325–334).

Muranoi, H. (2000). Focus on form through interaction enhancement: Integrating formal instruction into a communicative task in EFL classrooms. *Language Learning*, *50*, 617–673.

Negueruela, E., & Lantolf, J. P. (2006). Concept-based instruction and the acquisition of L2 Spanish. In R. Salaberry & B. Lafford (Eds.), *The art of teaching Spanish: Second language acquisition from research to praxis*, (pp. 79–102). Washington, D.C.: Georgetown University Press.

Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J.M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 3–50). Amsterdam: John Benjamins.

Ortega, L. (2014). Second language learning explained? SLA across 10 contemporary theories. In B. VanPatten & J. Williams (Eds.), Theories in second language acquisition: An introduction, (pp. 245–272). London: Routledge.

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110.

Pica, T. (1997). Second language teaching and research relationships: A North American view. *Language Teaching Research*, *1*(1), 48–72.

Pino-Silva, J. (2006). Extensive reading through the internet: Is it worth the while? *The Reading Matrix*, *6*(1), 85–96.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366.

Plonsky, L., & Ziegler, N. (2016). The CALL-SLA interface: Insights from a second order synthesis. *Language Learning and Technology*, *20*(2), 17–37.

Polio, C. (2007). A history of input enhancement: Defining an evolving concept. In C. Gascoigne (Ed.), *Assessing the impact of input enhancement in second language education: Evolution in theory, research and practice* (pp. 1–17). Stillwater, MN: New Forums.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Rebuschat, P. (2013). Implicit learning. In P. Robinson (Ed.), *The Routledge encyclopedia of second language acquisition* (pp. 298–302). London: Routledge.

Robertson, D. (2000). Variability in the use of English article system by Chinese learners of English. *Second Language Research*, *16*, 135–172.

Russell, V. (2012). Learning complex grammar in the virtual classroom: A comparison of processing instruction, structured input, computerized visual input enhancement, and traditional instruction. *Foreign Language Annals*, *45*, 42–71.

Russell, V. (2014). A closer look at the Output Hypothesis: The effect of pushed output on noticing and inductive learning of the Spanish future tense. *Foreign Language Annals*, *47*, 25–47.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*, 54–67.

Sanchez, L., & Jarvis, S. (2008). The use of picture stories in the investigation of crosslinguistic influence. *TESOL Quarterly*, *42*, 329–333.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*, 206–226.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge, UK: Cambridge University Press.

Schulze, M. (2008). AI in CALL—artificially inflated or almost imminent? *CALICO Journal*, *25*, 510–527.

Schulze, M. & Heift, T. (2012). Intelligent CALL. In M. Thomas, H. Reinders & M. Warschauer (Eds.), Contemporary computer-assisted language learning (pp. 249–265). London: Continuum.

Sharwood Smith, M. (1981). Consciousness-raising and the second language learner. *Applied Linguistics*, *2*, 159–168.

Sharwood Smith, M. (1993). Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition*, *15*, 165–179.

Stauble, A. M. E. (1978). The process of decreolization: A model for second language development. *Language Learning*, *28*(1), 29–54.

Treffers-Daller, J. (2011). Operationalizing and measuring language dominance. *International Journal of Bilingualism*, *15*(2), 147–163.

Ziegler, N. (2016a). Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*, *38*, 553–586.

Ziegler, N. (2016b). Methodological practices in interaction in synchronous computer mediated communication: A synthetic approach. In A. Mackey & E. Marsden (Eds.), *Instruments for research into second languages: Empirical studies advancing methodology* (pp. 197–223). New York: Routledge.