Language Learning Through Dependency Trees

Alexa Little

Advisor:
Prof. Claire Moore-Cantwell, PhD

Submitted to the faculty of the Department of Linguistics
in partial fulfillment of the requirements for the degree of
Bachelor of Arts.

Yale University

April 20, 2016

Abstract

Alexa Little. Language learning through dependency trees.

With the rise of digital technology, the popularity of computer-assisted language learning (CALL) programs increased. Because these programs allow students to study a language remotely or even independently, CALL is particularly favored for teaching less commonly taught languages (LCTL) such as Japanese. Few methods, however, incorporate explicit grammatical instruction, which was advocated as the most effective method of second language education in the landmark survey by Norris & Ortega (2000).

The purpose of this study was to examine dependency tree construction as a potential means of L2 grammar education. I investigated whether constructing dependency trees in a digital environment caused a reduction in grammatical errors by beginning students of Japanese. I also compared the efficacy of this novel method to existing CALL methods.

The research was conducted online via a web application, and data was collected from 17 beginner-level Japanese students at 6 universities. Each participant translated 7 sentences into Japanese to establish their prior knowledge. Then, they were shown a standardized description of Japanese causative syntax. Participants completed twenty exercises, for which they were randomly assigned to one of three groups. Group 0 completed a digital version of worksheet exercises, group 1 completed phrase-based CALL exercises, and group 2 constructed dependency trees of Japanese sentences. After the exercises, participants again translated 7 sentences to measure their improvement. My hypothesis was that group 2 (tree-based CALL) would show the greatest improvement.

A one-sample t test indicated that the mean improvement across all groups was greater than zero (mean = 6.47, st. dev. = 7.84, 95% CI = (2.44, 10.50), $p = 0.004$). This suggested that participants did, on average, make fewer errors after completing the study. However, one-way ANOVA (d.f. = 2, f-value = 0.24, p-value = 0.790) and the Kruskal-Wallis Test ($H = 0.34$, d.f. = 2, $p = 0.845$) suggested that there were no statistically significant differences in the mean error reduction for each group. In other words, participants' improvement appeared to be consistent across all treatment groups. Further analysis of the data showed that self-reported weakness (chosen from "speaking", "grammar", "script", and "vocabulary") did not statistically correlate with baseline performance nor with error reduction in that area. The only variable that showed a statistically significant effect was years of previous study – by one-way ANOVA, participants with two or more years of Japanese study made fewer initial errors (d.f. = 1, f-value = 5.20, p-value = 0.038) and showed more modest improvement (d.f. = 1, f-value = 5.89, p-value = 0.028).

The results of this study, the first to investigate dependency trees as a means of CALL, suggest that tree-based CALL is in fact an effective method and that it reduces subject errors on par with other methods of computer-assisted language instruction.

Primary Field: second language acquisition
Secondary Field: computer-assisted language learning

Keywords: CALL, dependency trees, second language acquisition, Japanese

Table of Contents

1 Introduction

As digital technology becomes increasingly sophisticated, there is ever-increasing interest in leveraging computerized tools for second language education. Computer-Assisted Language Learning, or CALL, is particularly useful for less-commonly-taught languages (LCTLs), because it offers students the opportunity to learn such languages even when physical, local classes are unrealistic. The goal of CALL as a discipline is to produce programs that are maximally effective at teaching language, yet maximally efficient so as not to bore students or take unrealistic amounts of time.

At present, a number of factors have impeded CALL from reaching this goal. For one, CALL is not well-linked to traditional second language acquisition (SLA) research, and CALL projects are influenced as much by the latest improvements in computer technology as by scientific research. Second, CALL, as an increasingly lucrative and high-profile industry, must maintain a balance between what is most effective at teaching language and what will entertain the users. This has triggered an "anti-grammar" trend, in which many CALL companies and applications have rejected the teaching of grammar altogether, in favor of focusing on words and phrases.

This paper presents a novel approach to CALL by introducing the construction of dependency trees as a method of L2 grammar practice. In this experiment, the tree-based approach was compared to two established methods of CALL, one of which was the worksheet-type drills favored in early-generation CALL, and the other of which was phrase-based CALL similar to today's popular language-learning mobile applications.

The results of the experiment indicated that dependency tree construction was, in fact, an effective form of CALL and that all three methods performed equally well at reducing student errors. Data analysis also revealed that subjects did not accurately detect their own weaknesses in spelling, grammar, et cetera; while not the focus of this research, this finding suggests that subject-directed CALL may not produce the optimal result.

As a whole, this work establishes tree-based CALL as a productive means of computer-assisted language instruction and proposes extensive opportunities for future research of this novel method.

2 Background

2.1 Second Language Acquisition

Second language acquisition (SLA) concerns the way we learn a second, or non-native, language. We seem to acquire our first language easily, yet many people struggle to learn a second language. SLA researchers study the differences between learning a native language (L1) and learning a second language (L2). They also research the way our native language influences the languages we learn later in life. For example, we are familiar with the pronunciation and grammar errors that non-native speakers make – but how and why do those errors actually occur? The goal of many SLA specialists is to use the scientific research of second language acquisition to make learning another language easier, faster, and more enjoyable.

One of the most influential theories in the field of second language acquisition is the Critical Period Hypothesis. First proposed by Penfield and Roberts in 1959, the Critical Period Hypothesis claims that humans are most capable of learning language by a certain age, after which changes in our brains make language acquisition more difficult. This may seem intuitive – babies, after all, learn to speak without ever attending a class – but it is actually a hotly debated topic in SLA. First, children may not actually be better at acquiring language than adults. In a 2000 study, researchers found that English-speaking adults could speak accurately and expansively in a closely related L2 (e.g., Danish, Dutch, Italian, etc.) after just 24 weeks of intensive study (Omaggio-Hadley 1993: 28). In comparison, children take about four years to learn basic L1 grammar (Hudson 2000: 121). This is not to say that adults are better at all aspects of learning language. A 1999 study found that children, as a group, achieve better fluency in grammar and pronunciation than adults do (Flege et al. 1999: 85, 88). Although some adults were able to speak a second language with near-native fluency, their scores ranged wildly. Children, in contrast, consistently achieved 75-90% accuracy on the tasks. The combined results of these studies suggest that, while the "critical period" may exist, it may not be as influential as previously thought (Brown & Larson-Hall 2012: 15). SLA researchers continue to investigate this concept.

Because SLA research is designed to improve language learning in a practical way, many SLA studies take place in the classroom. Researchers may split students according to ability level or language background, then observe them to discover differences between the groups. This information is useful to researchers attempting to find the key differences between proficient speakers – i.e., those who speak a second language well but imperfectly – and near-native speakers – i.e., those who speak a second language as well as their first. In other studies, students are split evenly, and both groups are taught the same concept, each group using a different methodology. The students are then tested on that concept (for example, a grammatical pattern), and the results are compared to determine which method was more effective. Studies like this are particularly common in the "input versus output" debate, where linguists hope to determine whether students can acquire a language simply by listening and reading, or if they also need to practice speaking and writing the language.

Implicit grammar education—the process of acquiring language without explicit grammatical instruction—also remains a controversial topic in SLA research. The most outspoken proponent of implicit learning is Steven Krashen, who proposed the Input Hypothesis in 1985 (later renamed the Comprehension Hypothesis) (Brown & Larson-Hall 2012: 38). Krashen drew a distinction between learning, which is conscious, and acquisition, which is unconscious, and he argued that conscious knowledge cannot contribute to naturalistic speech (Brown & Larson-Hall 2012: 38-39). As a result, Krashen proposed that students will learn best by comprehending input, not by producing output according to explicit rules (Brown & Larson-Hall 2012: 39). He also predicted that, given enough input, students would be able to produce language even without substantial production practice (Brown & Larson-Hall 2012: 46). However, a 1996 study by DeKeyser and Solkalski showed that students do need production experience in order to acquire production skills—in other words, input without output is not enough (Brown & Larson-Hall 2012: 47). Although the input-only element of Krashen's hypothesis has been disproven, many researchers do still support his assertion that language *acquisition* (unconscious knowledge) must be implicitly learned (Brown & Larson-Hall 2012: 85). In essence, these researchers claim that explicit grammatical instruction is ineffective because students will never be able to constructively and unconsciously apply that knowledge to produce natural language (Brown & Larson-Hall 2012: 85).

Proponents of explicit grammar education, in contrast, argue that such instruction *is* useful because students, with practice, eventually assimilate the conscious knowledge into their unconscious language production (Brown & Larson-Hall 2012: 86). Ellis (2005) argued, for example, that explicit instruction enables a student to produce a grammatical structure consciously, and over time that production leads to acquisition in the student's unconscious, or procedural, memory (213). Critics, like Lee and VanPatten (2003), have countered with the assertion that explicit instruction appears effective only because other, implicit systems are also at work (Brown & Larson-Hall 2012: 87).

In 2000, a landmark paper by Norris & Ortega[1] reviewed all the work to date on implicit versus explicit instruction, filtered out any studies with questionable methodology, and proposed conclusions based on the aggregate results of the remaining studies. Norris & Ortega concluded that "the current state of findings within this research domain suggests that treatments involving an explicit focus on the rule-governed nature of L2 structures are more effective than treatments that do not include such a focus" (2000: 483). In other words, the research to date supports the conclusion that explicit instruction is more effective than implicit instruction.

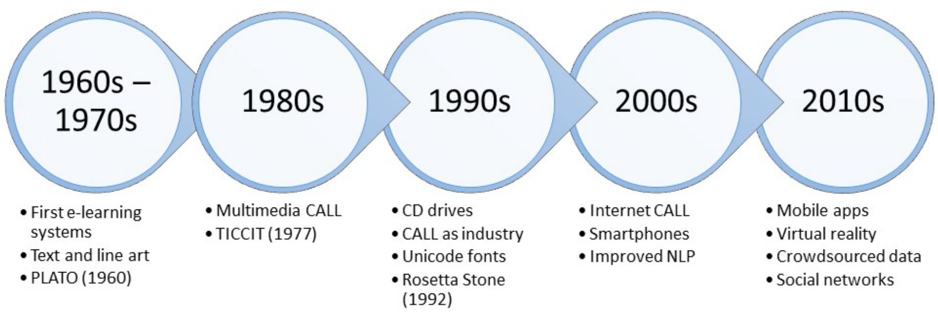
This conclusion, unfortunately, is not borne out in the current range of CALL offerings. The leading services emphasize "immersion" over explicit instruction, which generally means that students produce the surface output without studying the underlying grammar. The Rosetta Stone program, for example, teaches concepts using only images and target language phrases. Duolingo, another popular program, simply prompts users to translate English sentences into the target language. These implicit, phrase-based methods may work well if the user's L1 and L2 are similar enough (i.e., historically related or contact languages), but they require the user to make significant—and perhaps impossible—inferences if the languages are dissimilar. This raises a problem in the case of less-commonly-taught languages, which are simultaneously less likely to resemble the learner's L1 grammar and more likely to be taught using CALL. The CALL industry, however, is hesitant to adapt for LCTLs, which represent only a fraction of their user base, and even

[1] According to scholar.google.com, this work has been cited over 1,300 times since its publication.

more hesitant to include explicit grammar instruction, for fear that teaching too much grammar will cause their users to leave.

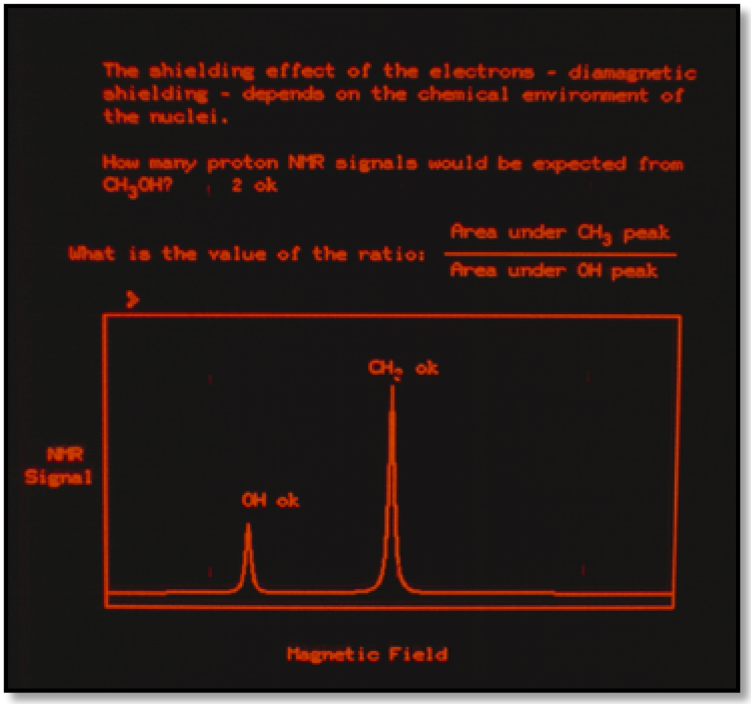
In short, SLA research firmly supports explicit instruction over implicit instruction for effectively teaching L2 grammar, but due to user demographics and perceived user preferences, the CALL industry is reluctant to implement this. This experiment is an attempt to find a compromise: a way of teaching grammar through a short, explicit introduction, then reinforcing it with a novel means of practice—dependency tree construction—that incorporates both the rule-consciousness advocated by SLA researchers and the user-friendly gamification supported by the CALL industry.

2.2 Computer-Assisted Language Learning: A History



Computer-Assisted Language Learning, or CALL, is the use of computers to aid humans in acquiring a language. As a discipline that bridges two fields, linguistics and computer science, CALL has historically gained from advances on both sides. Here, in order to give context to my research, I will give a short overview of the development of CALL over the past fifty years.

The PLATO (Programmed Logic for Automatic Teaching Operations) system, built at the University of Illinois in 1960, is widely regarded as the first major e-learning system as well as the first instance of Computer-Assisted Language Learning (Hubbard 2009: 3). Like many early computer programs, PLATO was originally restricted to a small physical network, with no external or remote access possible (Dooijes n.d.: n.p.). It allowed text and eventually line-art graphics, as shown in Figure 1, and the courses consisted mainly of pre-programmed lessons with limited error feedback (Dooijes n.d.: n.p.).



[Figure 1. A chemistry exercise from a 1970s edition of PLATO (Dooijes n.d.: n.p.).]

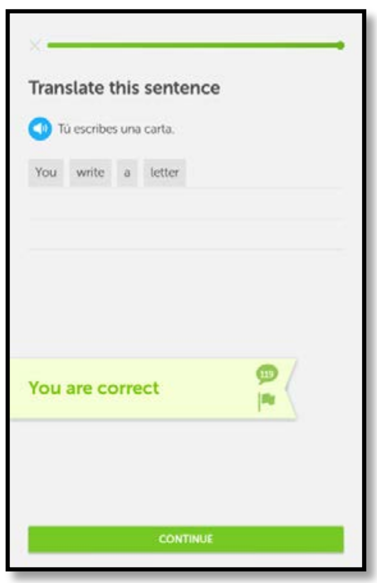
In the 1970s and 1980s with the advent of miniaturized "personal computers" (PCs), the numbers of both course creators and users increased considerably (Daves 2008: n.p.). Universities began designing and distributing CALL programs to the general public (Davies 2008: n.p.). The 1980s in particular saw the rise of "multimedia CALL" as PCs became capable of displaying photos, videos, and audio (Davies 2008: n.p.). Contemporary CALL systems began to include video and audio exercises alongside the textual drills of the past, and line-art graphics were replaced with more sophisticated renderings (Davies 2008: n.p.). The prime example from this period was the Time-Shared Interactive Computer Controlled Information Television System (TICCIT) developed at Brigham Young University (McNeil 2003: n.p.). The TICCIT program, which began in 1977, combined minicomputers and a color TV display to create one of the earliest multimedia CALL systems (McNeil 2003: n.p.). Unlike earlier systems, TICCIT was designed to be a standalone course, rather than supplementary material for a traditional college class (McNeil 2003: n.p.). This demanded, for the first time, that CALL developers consider the educational needs of students beyond simply providing practice exercises, and TICCIT became a major milestone in the field of instructional design (McNeil 2003: n.p.).

CALL was transformed again in the early 1990s with the arrival of two revolutionary technologies: the CD drive and the Internet (Davies 2008: n.p.). Until this point, nearly all CALL programs were developed at universities, but the popularization of the CD drive, and with it the CD- and DVD-ROM, made it possible to distribute multimedia CALL programs to a far wider audience (Delcloque 2000: 33, 53). The first CALL businesses were soon to follow; Transparent Language was founded in 1991, and Rosetta Stone in 1992 (Transparent Language 2015: n.p., Rosetta Stone 2016: n.p.). Figure 2 shows an early edition of Rosetta Stone's CALL software.

[Figure 2. Rosetta Stone version 2.0, released in 2001 (Rosetta Stone 2001: n.p.).]

At this time, developers also began experimenting with Internet-based CALL systems, although contemporary download speeds made web-hosted multimedia CALL difficult to realize. This coincided with the creation of Unicode, which allowed for the first time the representation of non-Latin characters as digital text. Unicode 1.0 Volume 1, released in 1991, offered Cyrillic, Arabic, and many other scripts; Chinese characters were added for Unicode 1.0 Volume 2, released in 1992 (The Unicode Consortium 2015: n.p.).

In the 21st century, CALL has begun a shift toward highly interactive systems. Demand for increasingly sophisticated systems placed new emphasis on ICALL, or intelligent CALL, which draws on cutting-edge Natural Language Processing techniques to produce more accurate and user-specific feedback (Davies 2008: n.p.). The state of computer-generated graphics in CALL has also improved considerably: systems like the DARWARS Tactical Language Training System, pioneered in 2004, now incorporate video game technology to provide realistic simulations for language learners (Johnson et al. 2004: 4). This technology continues to improve, with virtual reality CALL intended for release within the next year (Moss 2016: n.p.). The rise of the smartphone, meanwhile, has triggered an unprecedented boom in mobile CALL applications. The CALL application Duolingo, which also uses gamification to encourage language learning, earned Apple's iPhone App of the Year in 2013 and now boasts over 100 million users (Duolingo n.d.: n.p.); its mobile interface is shown below. Transparent Language and Rosetta Stone have also made the transition to mobile learning and offer users the option to synchronize their progress across multiple computing devices.

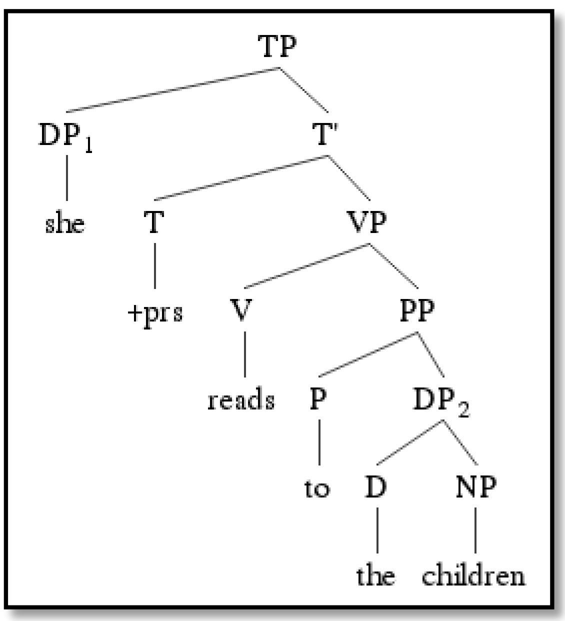


[Figure 3. An exercise on the Duolingo Android application. (Little 2016: n.p.)]

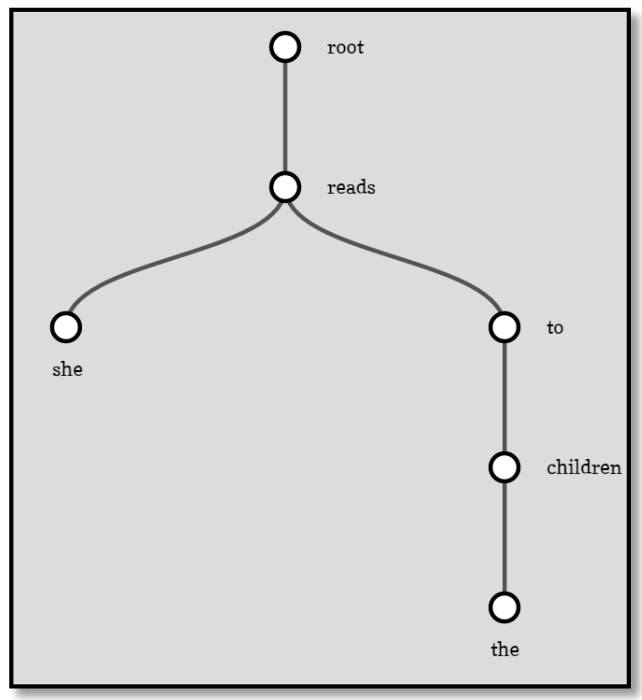
In recent years, computer technology has improved at an exponential rate, but our understanding of CALL and how to maximize its effectiveness is still in its infancy. To complicate matters, the lucrative CALL market attracts a constant influx of language-learning companies, not all of which base their software on scientific SLA research. In the end, as long as the precise components of a successful CALL system remain a matter of research and debate, the CALL field will continue to grow, adapt, and experiment alongside our digital technology.

2.3 Dependency Trees

For decades, linguists have used tree structures to visually represent grammatical concepts. One such structure, called a dependency tree, loosely relates the words of a sentence so that each word is dependent on another word in the sentence. Unlike syntax trees, which are detailed and require understanding of linguistic principles to interpret, dependency trees show only the broad-strokes patterns of grammar and can be interpreted without significant training. See Figures 4 and 5, which contrast a simple version of a syntax tree with its dependency tree equivalent.



She reads to the children.
[Figure 4. A syntax tree]



She reads to the children.
[Figure 5. A dependency tree]

Because of their relative simplicity, dependency trees are frequently used in Natural Language Processing (NLP) parsing tasks, in which a computer will predict the tree for a sentence, check that guess against the correct answer, and repeat until the probability of the correct answer is maximized. This intuition is the same one behind the tree-based CALL exercises; participants will attempt to build a tree, see the correct tree, and repeat for subsequent examples. Unlike the machine, however, the subjects will need to transfer their knowledge of the tree structures to the practical matter of translation in order for the exercises to be truly effective.

3 Experiment

3.1 Purpose

The purpose of this experiment was to investigate whether the construction of dependency trees could be used as a method of L2 grammar acquisition and to determine the effectiveness of this tree-based method relative to typical CALL exercises.

3.2 Hypothesis

I hypothesized that, of the three treatment groups, the tree-based CALL group would show the greatest reduction in errors between the pre-test and the post-test. I also anticipated that the phrase-based CALL group, which approximates state-of-the-art CALL systems, would perform better than the worksheet-based CALL group, which approximates the earliest CALL systems.

3.3 Experimental Preparation

3.3.1 Selecting Japanese as L2

A combination of linguistic and practical concerns led me to choose Japanese as the target language for this experiment. Firstly, Japanese is syntactically very different from English. Japanese is a Japonic language, head-final, and has robust case-marking, while English is an Indo-European language, head-initial, and has limited case-marking. Due to these syntactic differences, there are a wide range of structures found in Japanese (L2) that are not found in English (L1). This provided a full selection of non-L1 grammatical concepts from which to choose the experimental focus.

On the most practical level, I have working proficiency in Japanese and professional contacts with many native speakers. This allowed me to develop and edit a corpus easily, as well as locate native Japanese speakers to proofread my work. My grasp of the Japanese language also assisted me in writing a learner-friendly description of the chosen grammatical concept, and it allowed me to choose vocabulary terms that were appropriate for elementary Japanese learners. Additionally, although Japanese is a less-commonly-taught language (LCTL), well-established Japanese programs now exist at many universities across the United States. Since the experiment is web-based, this allowed me to recruit subjects from a slightly larger population than the average LCTL.

3.3.2 Selecting Causatives as the Grammar Concept

The grammar concept used in this experiment needed to be (a) challenging enough that it would not have been taught in the subjects' elementary-level classes and (b) distinct enough from English to require explicit explanation of its form and structure.

I chose to focus on causatives: a syntactic construction in which a third party causes an agent to perform an action. In English, these sentences are analytic, i.e. they use a multi-verb structure:

(1) She **made** me **write** a letter.

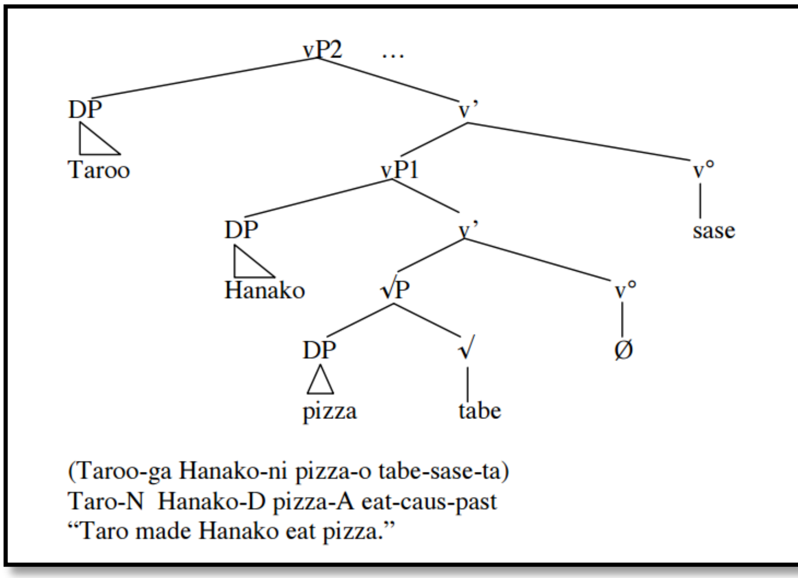
In Japanese, however, the structure is single-verb, or synthetic, and requires the use of particular cases:

(2) *Kanojo-wa watashi-ni tegami-wo kakaseta*
she-TOP I-DAT letter-ACC **write.CAU.PST**
'She made me write a letter.'

Beyond, or perhaps due to, the obvious distinctions in syntax, causatives are well-known among L2 Japanese learners as a challenging grammar topic.

In order to use causatives as the focus of my experiment, I needed to develop a dependency tree structure representing Japanese causative syntax. I started with the formal analysis of causatives under a syntactic framework.

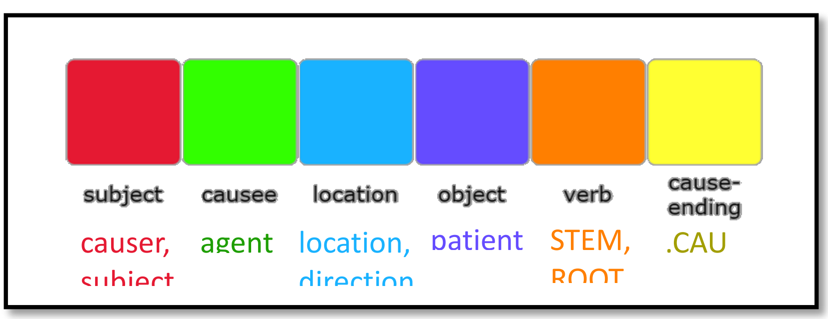
Harley (2008) presents the current accepted syntactic analysis of the productive (i.e., non-lexical) causative in Japanese. According to this analysis, the causative structure is several layers of vPs, each with a filled specifier position that takes on a particular theta role and case (Harley 2008: 30). For example, vP2 contains the causer "Taro" in its specifier v', and "Taro" is marked with nominative case. vP1 contains the agent "Hanako", which takes the dative case, in its specifier, and the lowest specifier position contains the patient "pizza" with accusative case. See Figure 6 for an image of this structure.



[Figure 6. Syntax tree showing a causative sentence, from Harley (2006: 30)]

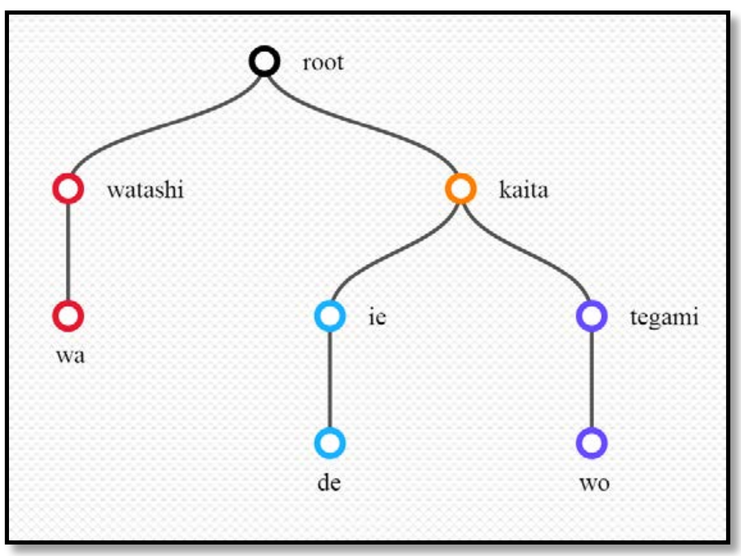
Because the participants in the experiment were unlikely to have a grasp of formal linguistic syntax, I simplified both the structure and the terms used to make them more appropriate for the average L2 learner.

Nodes in the tree were color-coded to what I called "parts of speech", which were actually renamed versions of the theta roles. In order to encourage the subjects to associate the case with the theta role of its head, the case markings were shown as separate nodes and color-coded to match the theta role of their respective heads. Similarly, the stem was called simply "verb", and the causative affix was renamed "cause-ending".



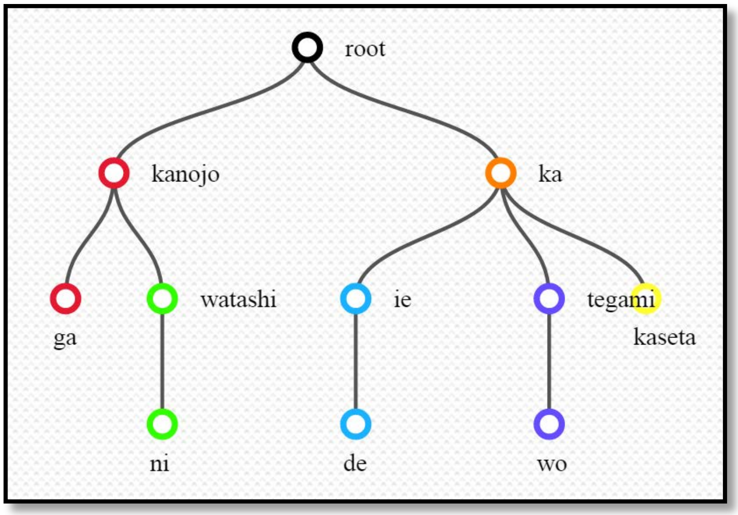
[Figure 7. The "parts of speech" and their corresponding theta roles or glosses]

The trees themselves were restructured also. This required a compromise between the rigid structure of a syntax tree and the very loose structure of a dependency tree. Keeping too much of the syntax tree structure would require a variety of dummy (i.e. non-word) verb nodes, while adopting a true dependency framework would result in too many items simply connecting to the verb stem. To avoid either of these outcomes while maintaining as much of the structure proposed by Harley (2008) as I could, I split the structure into "DP" and "VP" heads. The DP parent, on the left side, would be either the agent or (in a causative sentence) the causer, and if the sentence was causative, the DP parent would dominate the agent. On the right side, the VP parent was the main verb. This was a departure from Harley (2008), in which the causative affix dominates the main verb, but it allowed for a certain degree of parallelism between the simple and the causative structures. Any locative phrases or direct objects (patients) likewise were dominated by the main verb, because they either modify or are selected by their main verb head. Figure 8, below, shows the modified dependency tree for a simple Japanese sentence, while Figure 9 shows the tree for a minimally contrastive causative sentence.



watashi-wa ie-de tegami-wo kaita
I-TOP home-LOC letter-ACC write-PST
"I wrote a letter at home."

[Figure 8. Modified dependency tree for a simple sentence]



kanojo-ga watashi-ni ie-de tegami-wo kakaseta
she-NOM I-DAT home-LOC letter-ACC write-CAU.PST
"She made me write a letter at home."

[Figure 9. Modified dependency tree for a causative sentence]

3.3.3 Corpus Development

The lexicon and character (i.e., *kanji*) set for this experiment was restricted to items found on the beginner level of the Japanese Language Proficiency Test. Although the Japan Foundation no longer publishes a complete list of vocabulary items for its exams, archived versions provide an approximation of the lexicon and characters a beginning student of Japanese can be expected to know.

In order to implicitly model the difference between causative and plain (i.e., non-causative) sentences, the corpus needed to contain instances of both patterns. Using the lexicon described above, I developed a preliminary corpus of 34 sentences, in parallel-text Japanese and English, to be used in the course of the experiment. 20 of these (12 causative, and 8 non-causative) were for use in the exercises. The remaining 14 formed the test sets – each containing 7 sentences (5 causative and 2 non-causative) – which were used for the pre-test and post-test. This preliminary corpus was reviewed and edited by a native speaker of Japanese, and those edits were incorporated into the final corpus.

3.3.4 Software Development

The front end of the experimental website was built using HTML5, CSS, JavaScript, and jQuery. The back end of the website was coded in PHP and MySQL and hosted as a web application. In order to keep identifying information separate from the experimental data, the survey entries were collected via a private Google Form.

Because one group of subjects would be learning via tree-based grammar instruction, a means of digitally constructing dependency trees was necessary. I used an adapted version of EasyTree (Little & Tratz, forthcoming), a program based on the d3 JavaScript library (Bostock et al. 2011) that allows users to construct trees and save in the browser via drag-and-drop. The entire system is shown in detail below, in the Experimental Methodology section.

3.4 Subjects

Participants were recruited from 39 universities with established Japanese language programs. Ultimately, students from the following universities took part: Yale University, University of Wisconsin—Madison, University of Kentucky, Colgate University, Emory University, and University of California, Davis. Only students currently enrolled in an elementary level Japanese class were eligible for participation in the experiment. Students participated online, via an experimental website, and in exchange for their participation they were entered to win an Amazon gift card. Data and metadata from the subjects is presented and discussed in Section 4.

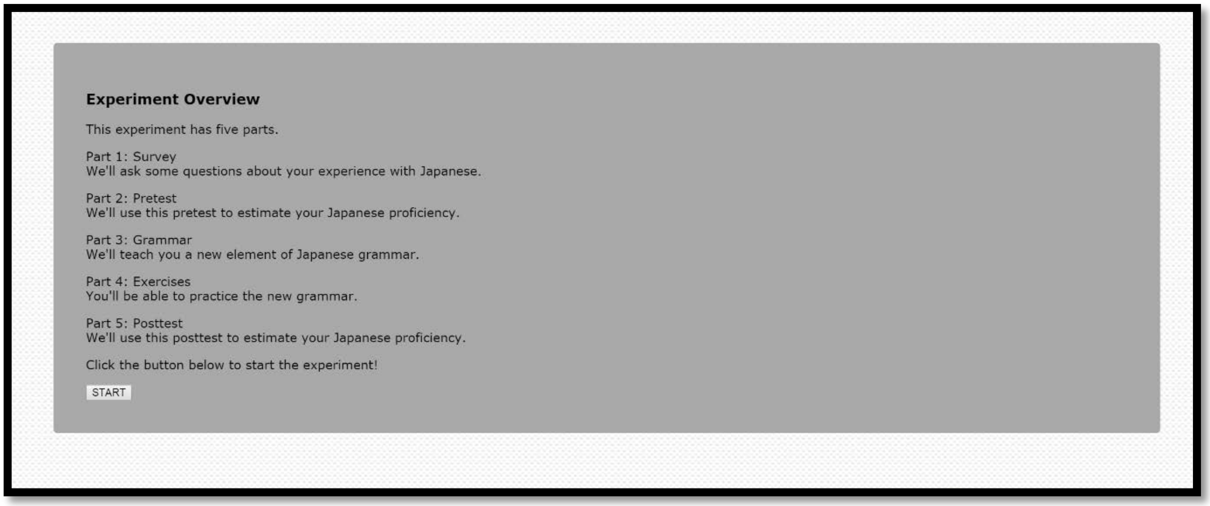
3.5 Experimental Methodology

The experiment consisted of six main stages: onboarding, pre-test, grammar lesson, exercises, post-test, and debriefing. In this section, I describe each stage in detail and include images of the experimental interface.

3.5.1 Onboarding

When participants arrived at the experimental website, they were presented with a description of the project, followed by a consent form. This was controlled with JavaScript so that participants could not proceed with the experiment until they indicated their consent.

From the consent form, participants were directed to an overview of the experiment. This overview indicated the stages of the experiment and their general content, as shown below in Figure 10.



[Figure 10. The overview of the experiment shown to participants.]

Finally, the participants proceeded to a metadata survey, which gathered information about their L1 background, exposure to the Japanese language, and typical study habits. This information was used to form categorical variables for statistical analysis of the data, in order to identify any trends or confounding variables outside the intended treatments. The questions were as follows:

Q1. What is your native language?

Q2. Please list any other languages you speak.

Q3. How many years have you studied Japanese?

- less than 1 year
- 1 year
- 2 years
- more than 2 years

Q4. How old were you when you started learning Japanese?

- age 8 or younger
- age 8-13
- age 13-18
- age 18 or older

Q5. Are you a heritage speaker?
(Do you have native Japanese speakers in your family?)

Q6. Have you ever been to Japan?

Q6b. If yes, for how long?

Q7. Choose your main source of Japanese instruction so far:

- self-study (books)
- self-study (software)
- individual tutoring
- classes
- immersion travel

Q8. Which part of Japanese do you find most challenging?

- kanji
- vocabulary
- particles[2]
- grammar

Q9. Which form(s) of Japanese have you studied?

- plain (だ, する) form
- polite (です, します) form
- both forms

I will address the use of these responses in Section 4.

3.5.2 Pre-test

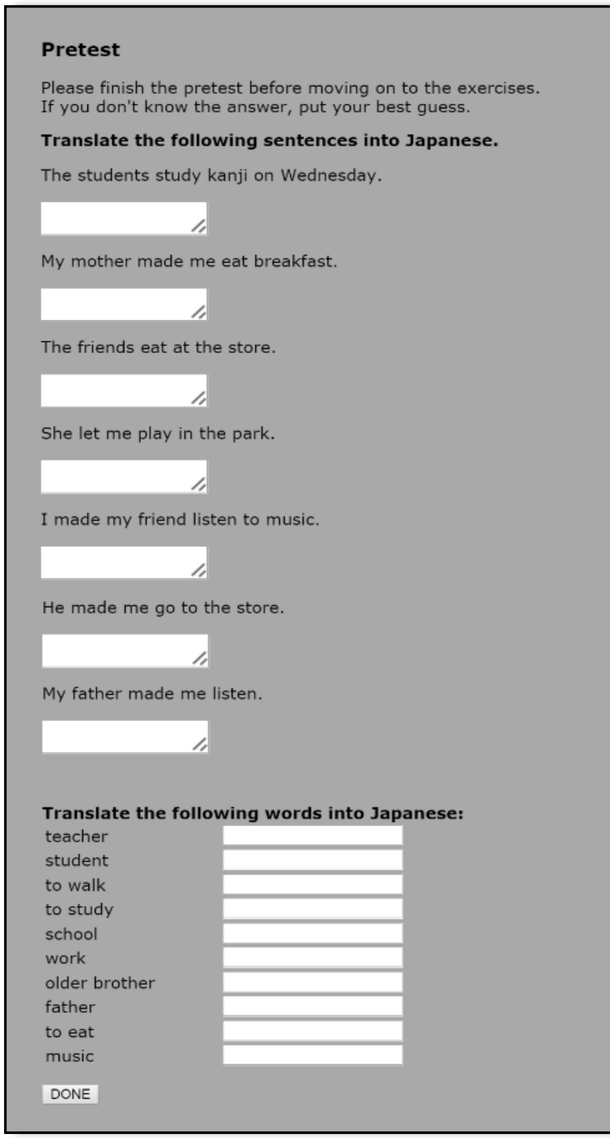
Once their consent and the relevant metadata were recorded, participants were directed to a pre-test designed to measure their initial level of Japanese proficiency. The pre-test consisted of two parts: word-level translation, and sentence-level translation. A sample pre-test is available in Figure 11, below.

In the first section, participants were asked to translate seven English sentences into Japanese. Five of those sentences were causative; the remaining two were non-causative. For this part of the test, I generated two sets of seven sentences. Half of the participants were shown set A in the pre-test, and set B in the post-test; the other half of the participants had set B as a pre-test and set B as a post-test. This randomization helped

[2] *particle* is a common term in L2 Japanese education; it refers to the case-marking suffix

eliminate the chance that a particular sentence or set of sentences had a confounding effect on the results.

In the second section, participants were asked to translate ten words into Japanese. This section was designed to measure each participant's approximate lexical inventory, for use in statistical analysis. The words were chosen manually from the corpus to cover various part-of-speech categories, and the same set of words was used for all participants.



Pretest

Please finish the pretest before moving on to the exercises.
If you don't know the answer, put your best guess.

Translate the following sentences into Japanese.

The students study kanji on Wednesday.

My mother made me eat breakfast.

The friends eat at the store.

She let me play in the park.

I made my friend listen to music.

He made me go to the store.

My father made me listen.

Translate the following words into Japanese:

teacher
student
to walk
to study
school
work
older brother
father
to eat
music

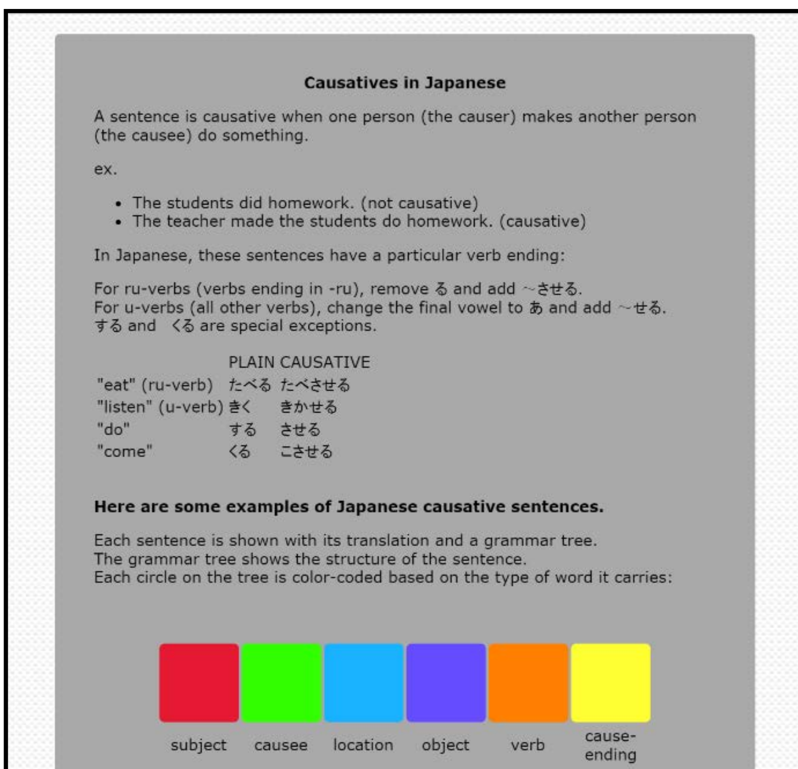
DONE

[Figure 11. One variation of the pre-test.]

3.5.3 Grammar Lesson

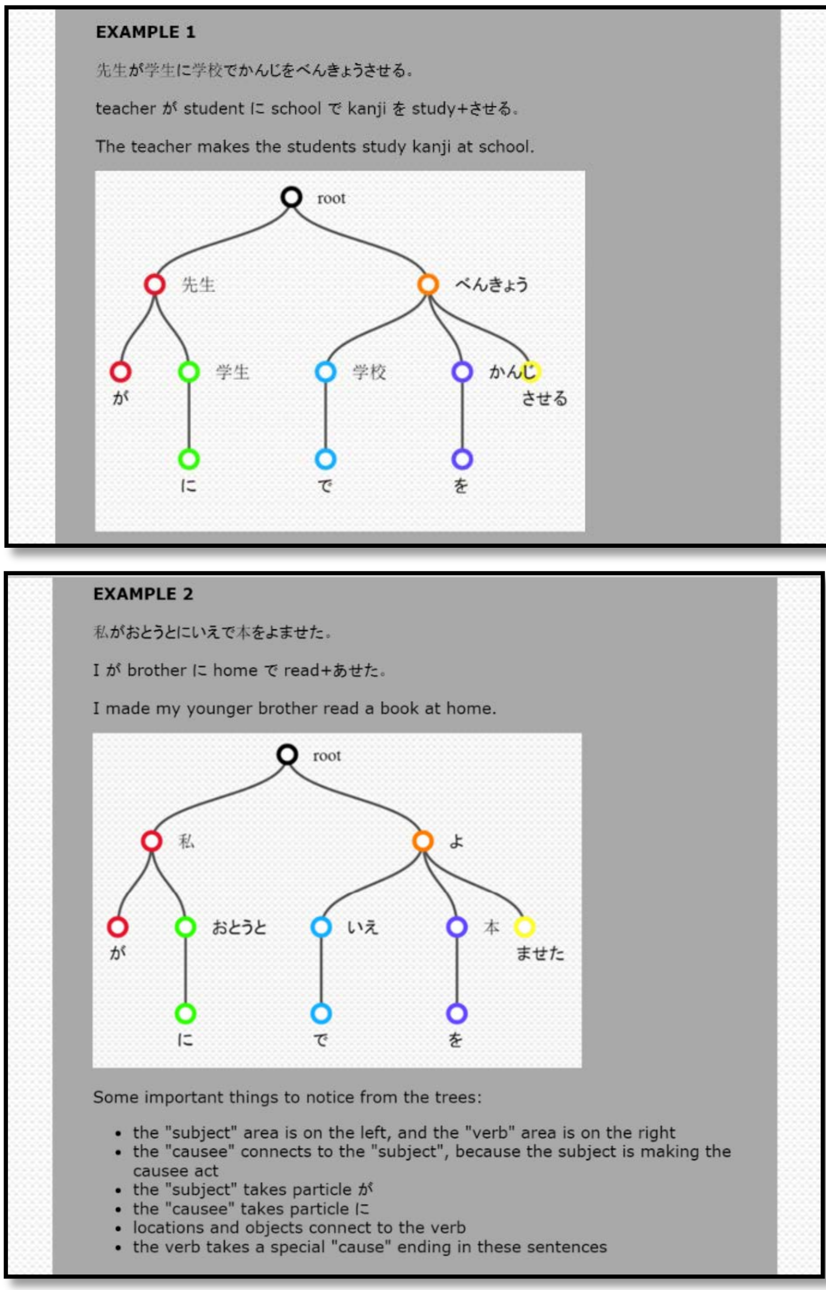
During this portion, all subjects were shown the same explicit lesson on Japanese causatives. This lesson was purposefully kept constant across all treatment groups in order to avoid introducing confounding variables. This required, however, that the grammar notes include material essential to the Tree-Based CALL group, such as a color-coded part of speech key and several tree diagrams. The full content of the grammar lesson, as it appeared to participants, is shown in Figures 12-14.

The grammar lesson began with an overview of causatives and the conjugation paradigm, plus an introduction to the parts of speech involved in the experiment. In order to make the lesson more straightforward for the subjects, some terms were changed and some descriptions were simplified. In the parts of speech, for example, the "subject" category covered both topicalized elements and elements in nominative case. The accusative was renamed "object", and the agent in the causative sentence was renamed "causee". Similarly, the descriptors *u*-verbs and *ru*-verbs, familiar terms for most students of Japanese, were used to explain the paradigms.



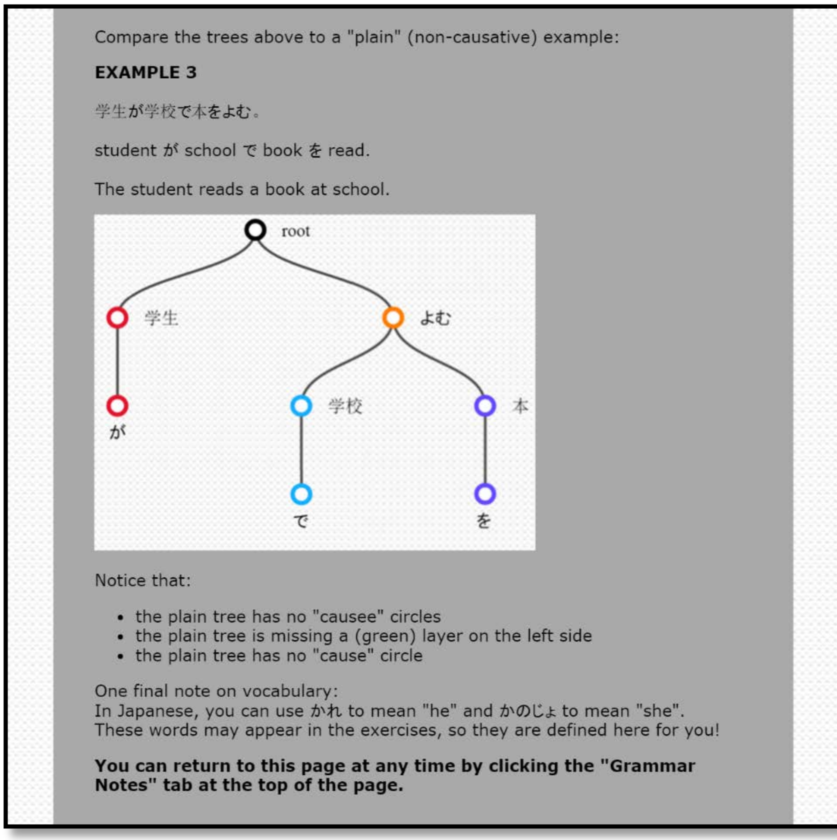
[Figure 12. The grammar lesson, part one.]

The overview then presented two causative tree structures as trees and highlighted the important aspects of causative grammar. Because L2 Japanese learners tend to conceptualize the case-marking affixes, also called "particles", as separate words, the case-marking affixes were shown as separate nodes on the tree structure. I hypothesized that attaching the affix to the word it modifies might help the participants learn to associate particular cases with particular grammatical functions.



[Figure 13. The causative structures and associated descriptions.]

These examples were followed by a tree structure depicting a simple sentence. Notes at the bottom contrasted the simple structure with the causative ones from previous examples.



[Figure 14. The simple sentence and contrastive notes.]

Finally, the grammar overview contained a short explanation on the Japanese terms *kare* 'he' and *kanojo* 'she'. Pronouns are relatively uncommon in Japanese, particularly in comparison to English, and are generally used only when necessary to disambiguate meaning. Because pronoun use in Japanese was not the focus of this experiment, these words were explained overtly for the subjects' benefit.

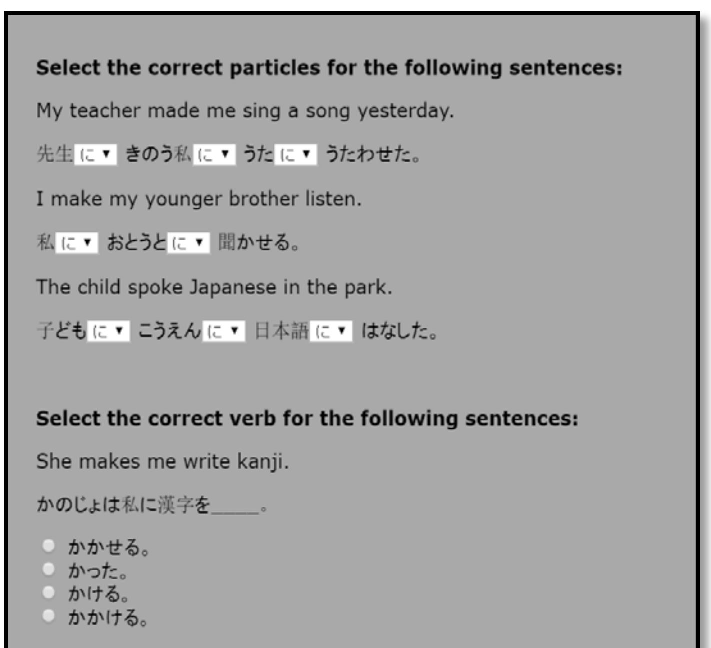
At the bottom of the grammar overview, subjects were reminded that they can access the explanations at any time during the exercises. They were then prompted to click a button to transition to the exercise phase of the experiment.

3.5.4 Exercises

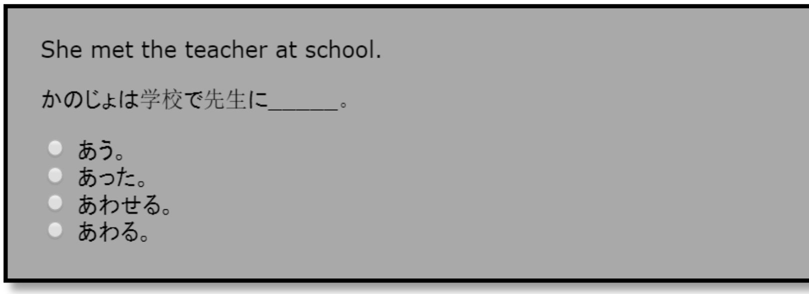
Upon reaching the exercise component of the experiment, subjects were randomly sorted into one of three treatment groups. Each group practiced using the same twenty-sentence corpus, and each group had access to the grammar overview. The key distinction among the three groups was the type of exercises used to practice. One group practiced using worksheet-type CALL drills, the second practiced using translation (phrase-based CALL), and the final group practiced using the novel tree-construction method. Each of these systems is shown and described in this section.

Group One – Worksheet CALL

The first treatment group imitated the naïve drills that dominated early CALL systems until the emergence of multimedia CALL. For this group, the exercises consisted of two ten-question digital worksheets. These worksheets consisted of a variety of exercises, including selecting the correct case-marking suffixes, choosing the correct verb conjugation, and translating sentences (see Figures 15 and 16 for examples). Upon submitting their responses, the subjects could view their responses alongside the correct response, and once two worksheets had been completed and reviewed, the subjects were redirected to the experimental post-test.



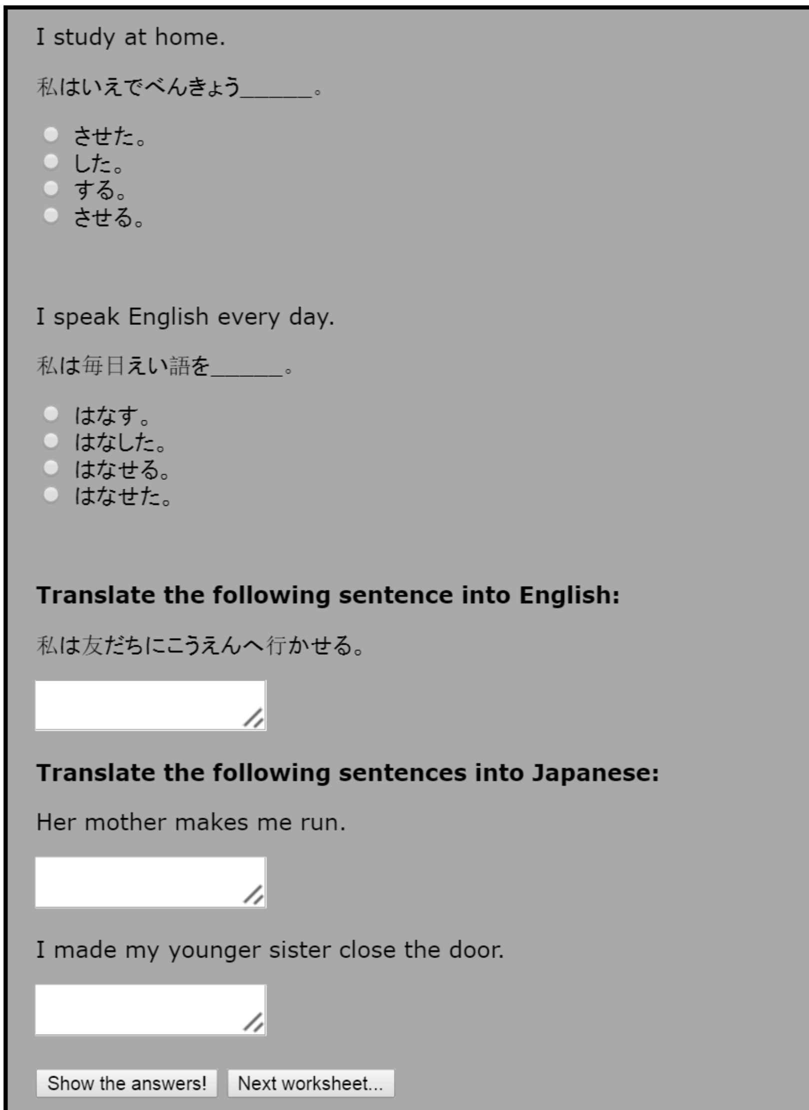
[Figure 15. Worksheet exercises]



She met the teacher at school.

かのじょは学校で先生に______。

- あう。
- あった。
- あわせる。
- あわる。



I study at home.

私はいえでべんきょう______。

- させた。
- した。
- する。
- させる。

I speak English every day.

私は毎日えい語を______。

- はなす。
- はなした。
- はなせる。
- はなせた。

Translate the following sentence into English:

私は友だちにこうえんへ行かせる。

Translate the following sentences into Japanese:

Her mother makes me run.

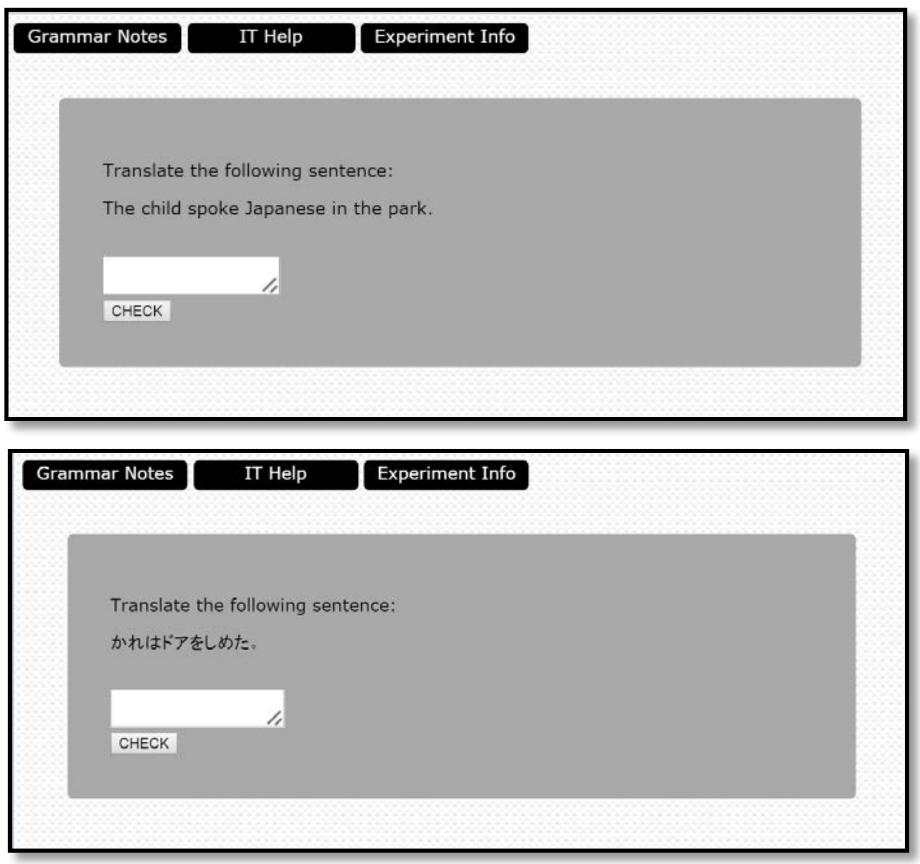
I made my younger sister close the door.

Show the answers! Next worksheet...

[Figure 16. Worksheet exercises, continued.]

Group Two – Phrase-based CALL

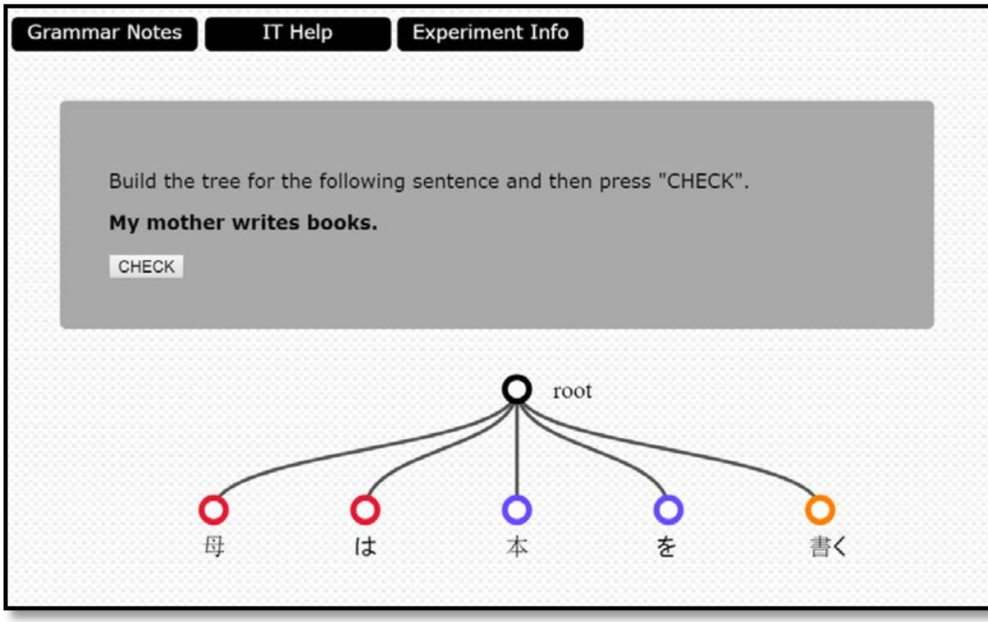
The second treatment was based on popular mobile and internet CALL systems, which focus on translation and surface structure over the grammatical specifications of the language. In this group, subjects completed twenty translation exercises of the type shown in Figure 17. After each exercise, they were shown their response and the desired response. This consistent stream of feedback also distinguished them from the first group; while the first group had only two opportunities to receive feedback on their answers, this group received feedback twenty times – once after every prompt. Upon completing those twenty exercises, the subjects were automatically redirected to the post-test.



[Figure 17. Sample exercises for group two.]

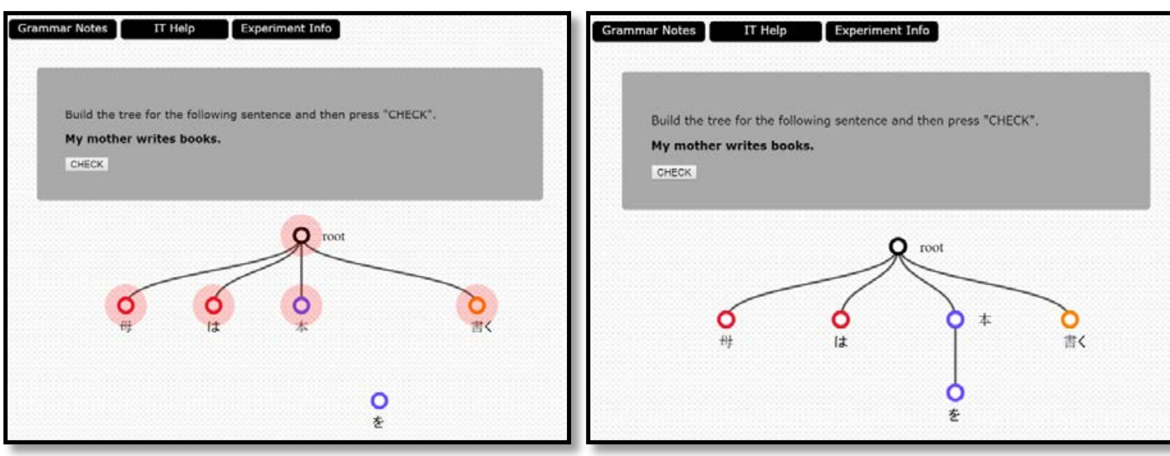
Group Three – Tree-based CALL

The third and final treatment involved tree construction as a method of CALL. For each exercise, the subject was shown an English sentence, plus a flat tree structure in the word order of the Japanese sentence.



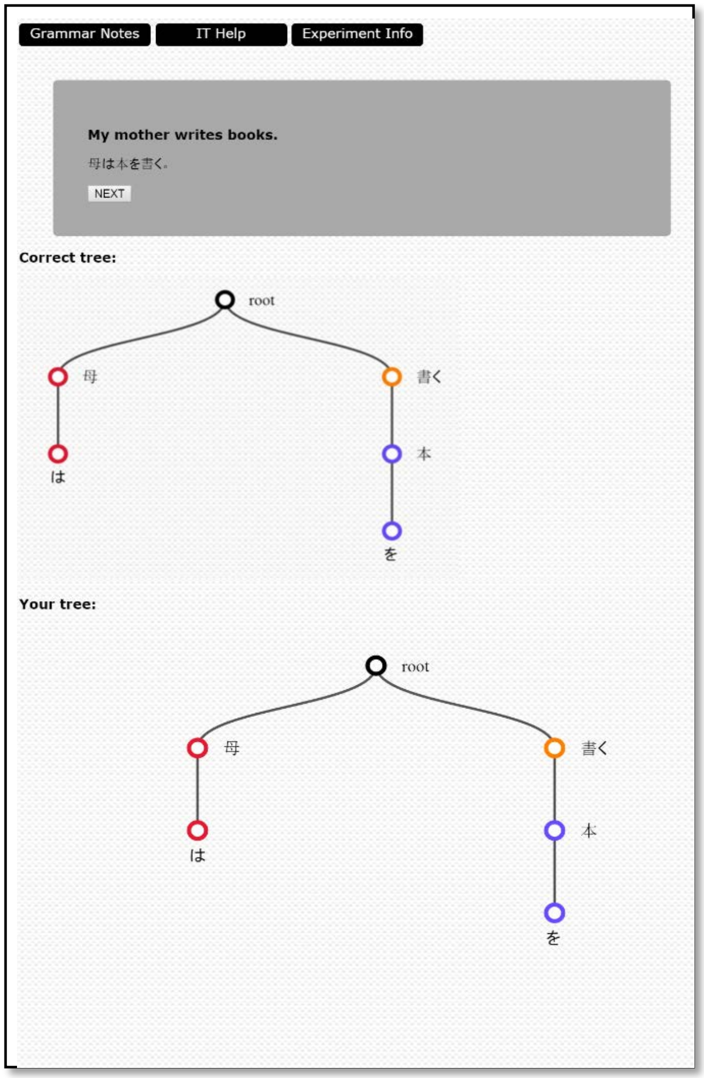
[Figure 18. Initial prompt for a tree-based exercise.]

Subjects were instructed to construct the tree for the sentence. This could be achieved by dragging and dropping the nodes onto the appropriate parent, as shown in Figure 19.



[Figure 19. Dragging and dropping a node to form the tree.]

Once the subject completed the tree and pressed 'check,' the correct answer was shown with the subject's tree underneath (see Figure 20). Although the answer tree was an image, the subject's tree remained editable at this stage, so subjects could opt to revise their tree, using the correct response as a reference, or to move on to the next prompt.



[Figure 20. Displaying the correct answer and the subject's submission.]

Subjects in this group completed twenty exercises, each time constructing a tree and viewing it alongside the desired response. Afterward, they were sent to the post-test.

3.5.5 Post-test

Upon completing the allotted drills, subjects were automatically redirected to the post-test. The post-test had the same structure as part two of the pre-test – that is, participants were asked to translate seven sentences into Japanese, five of which were causative and two of which were not. As described in Section 3.5.2 above, participants were shown the opposite set of sentences from those they translated in the pre-test.

3.5.6 Debriefing

After the post-test, the subjects were thanked for their time and released from the study. They were also given a link to the Google Form for the gift card raffle, which ensured that participant data was kept separate from any identifying information.

4 Results and Analysis

4.1 Scoring the Pre- and Post-tests

To maintain as much objectivity as possible, errors in the pre- and post-tests were tabulated using a detailed, pre-determined rubric. Furthermore, once observations had been collected, the errors were tabulated without reference to the treatment group or other metadata in order to avoid unconscious bias.

Errors were sorted into four main categories: spelling errors, vocabulary errors, case-marking errors, and grammar errors. I describe each of these categories below, and give examples for the individual error types. A full list of possible errors is available as Figure 21.

Error Types
Spelling Errors
1. Incorrect Kanji
2. Misspelled Word
Vocabulary Errors
3. Incorrect Word
4. Incorrect Formality Level
5. Missing Content Word
6. Substituted English Word
Case-Marking Errors
7. Accusative Case Incorrect / Missing
8. Locative Marker Incorrect / Missing
9. Nominative Case Incorrect / Missing (non-causative)
10. Nominative Case Incorrect / Missing (causative)
11. Dative Case for “Causee” Incorrect / Missing
Grammar Errors
12. Incorrect Verb Ending
13. Misspelled Verb Ending (Causative)
14. Incorrect Verb Tense
15. Misspelled Verb Tense

[Figure 21. List of potential errors.]

4.1.1 Spelling Errors

Error 1 - Incorrect Kanji

Kanji (content-word character) errors are common for beginning and intermediate students of Japanese, especially when typing. On most modern operating systems, Japanese is typed phonetically, and Natural Language Processing tools are used to convert the input into the most likely character in a series of homophones. Because this system was statistically trained based on input from native Japanese speakers, it is highly accurate as long as the input is native-like. However, errors elsewhere in the sentence, such as errors in case-marking, can cause the system to produce the incorrect *kanji* character. Users can easily correct this to the proper character using a drop-down menu, but beginning students of Japanese are likely to overlook the error or choose the wrong alternative character. In this scoring system, such errors were classed as error type 1. See example 3 below:

(3)

Correct response:	Error:
私は**漢字**をべんきょうする。 *watashi-wa kanji-wo benkyou-suru* "I study **kanji**."	私は**感じ**をべんきょうする。 *watashi-wa **kanji**-wo benkyou-suru* "I study **feeling**."

Error 2 - Misspelled Word

Beginning students of Japanese may try to avoid *kanji* errors by typing exclusively in *kana*, the phonetic script of Japanese. This is acceptable, but it makes the subjects prone to another type of error: misspellings. Misspellings are most likely to occur on vowel-length distinctions and geminate words. Unlike English, Japanese contrastively distinguishes based on mora length, i.e. heavy (or long) versus light (or short) syllables. This distinction is difficult for beginning students to hear and spell accurately. Similarly, geminate consonants, another contrastive distinction found in Japanese but not in English, can present a spelling challenge for learners of Japanese. Under this rubric, legible misspellings of content words were marked as error type 2. Words are defined as "legible" as long as the intended word can be inferred and the error has not transformed the word into another, inaccurate lexical item. See the following example:

(4)

Correct response:	Error:
私は漢字を**べんきょう**する。 *watashi-wa kanji-wo* ***benkyou****-suru* “I **study** kanji.”	*私は漢字を**べんきょ**する。 **watashi-wa kanji-wo* ***benkyo****-suru* “I **???** kanji.”

4.1.2 Vocabulary Errors

Error 3 - Incorrect Word

This error type is fairly self-explanatory; it covers all word errors not attributable to spelling problems.

Error 4 - Incorrect Formality Level

In the Japanese language, certain words take on either a humble or a polite form based on the social relationships between the speakers. This is particularly true of kinship terms. The humble forms of kinship terms are used to refer to one’s own family; a related set of polite kinship terms are used to refer to others’ family members. In Example 5, the subject has mistakenly used the humble form of “father”, *chichi*, rather than the polite form *otousan*. This implies that she is referring to her own father.

(5)

Correct response:	Error:
父は本をよむ。 ***chichi****-wa hon-wo yomu* “**(My) father** reads books.”	**お父さん**は本をよむ。 ***otousan****-wa hon-wo yomu* “**(Your) father** reads books.”

Error 5 - Missing Content Word

Every content word included in the English prompt must be included in the Japanese translation. Any omitted words counted as an error.

Error 6 - Substituted English Word

Even if Japanese syntax was used, substituting an English word for an unknown or forgotten Japanese word was not permitted and counted as an error. (Note that no errors of this type were observed in the experimental data).

4.1.3 Case-Marking Errors

Error 7 - Accusative Case Incorrect or Missing

In Japanese, the direct object, or patient, of a sentence generally takes the accusative case, indicated by the suffix *wo*. If an element that should be accusative lacks the proper case marking, this is counted as an error.

Error 8 - Locative Markers

Japanese contains two affixes associated with locative or directional phrases. The first of these, the stative locative suffix *ni*, is used to indicate movement toward the head element (as in example 6), or to indicate the location of said element when used with a stative verb (as in example 7) (Ohtani 2013: 382).

(6) *jon-san-wa gakkou-ni iku*
John-HON-TOP school-to go.PRS
"John goes to school."

(7) *jon-san-wa gakkou-ni iru*
John-HON-TOP school-at be.PRS
"John is at school."

The locative suffix *de*, in contrast, is used to indicate the setting of an action (i.e., corresponds with non-stative verbs) (Ohtani 2013: 382):

(8) *jon-san-wa gakkou-de yomu*
John-HON-TOP school-in read.PRS
"John reads at school."

The suffixes *ni* and *de* are not interchangeable, and using the stative suffix *ni* in place of the non-stative suffix *de* results in an ungrammatical sentence:

(9) **jon-san-wa gakkou-ni yomu*
John-HON-TOP school-in read.PRS
*John reads at school. (intended)

Mixing up the stative and non-stative locative affixes, or failing to use them altogether, was marked as an error.

Errors 9 and 10 – Nominative Case Incorrect or Missing

Depending on the context, the subject of a Japanese sentence can take either nominative case *ga* or the topic marker *wa*. Distinguishing between proper use of *ga* versus *wa* is a challenge even for advanced learners of Japanese, so mistakes of this nature were *not* considered an error. However, failing to use any case whatsoever on the subject or using a case marker that is neither *wa* nor *ga* was considered an error of this type. In order to distinguish between mistakes in simple sentences and those in causative sentences, this error type was split into errors 9 (non-causative) and 10 (causative).

Error 11 – Dative Case for Causee Incorrect or Missing

The causee in non-lexical Japanese causatives, i.e. the agent, always takes the dative affix *ni* (see example 10) (Harley 2008: 5).

(10) *watashi-wa imouto-ni doa-wo shimesaseta*
I-TOP younger.sister-DAT door-ACC close.CAU.PST
"I made my younger sister close the door."

The omission of the dative case suffix, or the substitution of another suffix, was considered an error.

4.1.4 Grammar Errors

Error 12 - Incorrect Verb Tense

For the purposes of this experiment, all sentences in the corpus are either in past or present tense, with the occasional infinitive verb. The Japanese translation submitted by the subject should have aligned in tense with the English prompt; if it did not, each instance of incorrect tense was considered an error.

Error 13 - Misspelled Verb Tense

The conjugation of informal register past tense verbs in Japanese is relatively complex (see Figure 22), and students of Japanese who have not yet grasped the full paradigm are likely to make mistakes. An error was counted each time the subject misspelled a tense ending.

hashiru run.PRS	*hashitta* run.PST
nomu drink.PRS	*nonda* drink.PST
taberu eat.PRS	*tabeta* eat.PST
kaku write.PRS	*kaita* write.PST
dasu submit.PRS	*dashita* submit.PST

[Figure 22. Present and past tense of Japanese verbs]

Error 14 - Incorrect Verb Ending

The use of the simple verb ending in a sentence that should be causative, or vice versa, was considered an error.

Error 15 - Misspelled Causative Ending

Japanese verbs are generally split into two classes of regular verbs, *u*-verbs and *ru*-verbs, and the conjugation of the causative ending depends on the verb class (see Figure 23). If the subject attempted to utilize the causative ending, but commits one or more spelling errors, this was marked as a single violation under "Misspelled Causative Ending."

	Simple	Causative
***u*-verb**	*tsukau* use.PRS	*tsukawaseru* use.CAU.PRS
***ru*-verb**	*taberu* eat.PRS	*tabesaseru* eat.CAU.PRS

[Figure 23. Causative paradigm for two classes of Japanese verbs.]

4.2 Presentation of Data

A total of 17 subjects participated in the experiment. Figure 24 contains the essential data for those participants.

ID	Test Set	Group	Pre-test Errors	Post-test Errors
1	2	1	20	4
2	2	3	14	5
3	1	3	12	13
4	1	3	2	1
5	2	2	3	4
6	1	2	6	6
7	2	1	18	3
8	1	1	1	2
9	1	1	15	3
10	2	3	27	14
11	1	3	22	14
12	1	2	9	5
13	1	3	13	13
14	2	2	20	2
15	2	3	10	6
16	2	2	22	3
17	2	1	7	13

[Figure 24. Basic experimental data.]

During the experiment, each participant was randomly assigned a test set and a treatment group. These values were stored as part of the experimental data. In Figure 24, for example, the first participant belonged to test set 2 and treatment group 1. This means that participant was given test B as the pre-test and test A as the post-test—participants in test set 1 were given the inverse—and was a member of group 1, the worksheet-based CALL treatment.

The overall errors in the pre-test and post-test are also included in Figure 24. In order to further analyze the types of errors made by participants, and in order to identify any important statistical correlations with other experimental variables, the overall errors were broken down by spelling, vocabulary, case, and grammar subcategories, as shown in Figure 25.

For the pre-test and post-test, the error tallies for all 15 error categories were recorded for all 15 error categories for later analysis (see Figures 26, 27).

	PRE-TEST					POST-TEST				
ID	Errors	Spelling	Vocab.	Case	Gram.	Errors	Spelling	Vocab.	Case	Gram.
1	20	0	7	5	8	4	0	1	1	2
2	14	0	9	5	0	5	1	1	0	3
3	12	1	1	10	0	13	0	3	9	1
4	2	1	1	0	0	1	0	1	0	0
5	3	0	2	1	0	4	0	2	1	1
6	6	0	2	2	2	6	0	2	2	2
7	18	0	3	6	9	3	1	1	0	1
8	1	0	1	0	0	2	0	2	0	0
9	15	1	3	6	5	3	0	2	1	0
10	27	3	11	9	4	14	3	2	3	6
11	22	0	8	7	7	14	0	5	4	5
12	9	0	1	5	3	5	0	2	0	3
13	13	0	2	5	6	13	3	4	5	1
14	20	0	6	5	9	2	0	1	0	1
15	10	1	4	0	5	6	0	2	1	3
16	22	2	8	7	5	3	0	1	0	2
17	7	0	4	2	1	13	0	4	5	4

[Figure 25. Errors by category]

PRE-TEST															
ID	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15
1	0	0	0	1	6	0	2	0	0	2	1	5	0	1	2
2	0	0	2	0	7	0	0	1	0	2	2	0	0	0	0
3	1	0	1	0	0	0	1	0	0	4	5	0	0	0	0
4	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0
6	0	0	0	0	2	0	0	0	0	0	2	0	1	1	0
7	0	0	0	0	3	0	0	0	0	1	5	5	0	1	3
8	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
9	0	1	1	1	1	0	0	0	0	3	3	5	0	0	0
10	0	3	1	0	10	0	0	0	2	3	4	4	0	0	0
11	0	0	1	1	6	0	1	0	0	1	5	5	0	2	0
12	0	0	0	1	0	0	0	0	0	0	5	0	1	2	0
13	0	0	2	0	0	0	0	0	0	0	5	5	0	1	0
14	0	0	1	0	5	0	0	0	1	2	2	3	2	4	0
15	0	1	0	1	3	0	0	0	0	0	0	0	5	0	0
16	0	2	0	0	8	0	1	0	0	2	4	5	0	0	0
17	0	0	0	0	4	0	0	0	1	0	1	0	1	0	0

[Figure 26. Pre-test errors]

									POST-TEST					
ID	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15
1	0	0	0	0	1	0	0	0	0	1	0	0	1	1	0
2	0	1	0	0	1	0	0	0	0	0	0	1	0	2	0
3	0	0	1	0	2	0	0	0	0	4	5	1	0	0	0
4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5	0	0	0	1	1	0	0	0	0	0	1	0	0	1	0
6	0	0	0	0	2	0	0	0	0	0	2	0	1	1	0
7	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0
8	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	2	0	0	0	1	0	0	0	0	0	0
10	0	3	2	0	0	0	0	2	0	1	0	5	0	1	0
11	0	0	1	0	4	0	0	1	0	1	2	0	1	4	0
12	0	0	0	0	2	0	0	0	0	0	0	0	1	2	0
13	0	3	2	0	2	0	0	0	0	0	5	0	1	0	0
14	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
15	0	0	0	1	1	0	0	1	0	0	0	1	1	1	0
16	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0
17	0	0	0	0	4	0	0	0	0	3	2	0	0	4	0

[Figure 27. Post-test errors]

Although vocabulary scores were calculated based on the second portion of the pretest, these results are not reported here because all participants translated either 9 or 10 of the 10 words correctly.

Metadata regarding each subject's linguistic background and experience studying Japanese were also collected. In Figure 28, the participants' answers to these questions are recorded.

All participants gave the same answer for questions 7 and 9 on the metadata survey. For question 7, participants indicated their main source of Japanese instruction as "classes", and for question 9, participants indicated that they had studied both the plain and polite forms of Japanese. In the interest of simplicity, these have been omitted from the chart above.

ID	University	L1	L2+	Japan visit	Time in Japan (mo.)	Years studied	Age when started	Heritage speaker?	Self-reported weakness
1	Yale	Chinese	English, French	yes	0-2	0	18+	no	speech
2	UC Davis	Chinese	English, Cantonese	yes	0-2	>2	13-18	no	grammar
3	UC Davis	English	Yoruba	yes	0-2	>2	18+	no	vocab
4	UC Davis	English	n/a	yes	2-6	>2	18+	no	speech
5	UC Davis	English	n/a	no	n/a	2	18+	no	speech
6	Colgate	Chinese	Vietnamese	no	n/a	0	18+	no	grammar
7	UW	English	Cantonese	no	n/a	0	18+	no	speech
8	UW	English	n/a	no	n/a	0	13-18	no	speech
9	UW	English	n/a	yes	6-12	0	18+	no	script
10	Kentucky	English	Spanish	no	n/a	0	13-18	no	script
11	Kentucky	English	n/a	no	n/a	1	18+	no	speech
12	UC Davis	English	Vietnamese	yes	0-2	>2	8-13	no	speech
13	Colgate	English	n/a	no	n/a	1	13-18	no	script
14	Colgate	English	n/a	no	n/a	0	18+	no	grammar
15	UC Davis	English	Vietnamese	no	n/a	0	18+	no	script
16	UC Davis	English	n/a	no	0-2	0	18+	yes	grammar
17	Emory	English	Vietnamese	yes	n/a	2	18+	no	grammar

[Figure 28. Metadata]

4.3 Data Analysis and Discussion

Due to the small size of the sample, many of the metadata variables had to be simplified to allow for valid statistical analysis. The results of these changes are shown below.

ID	University	L1	Japan visit	Years studied	Age when started	Self-reported weakness
1	Other	Chinese	yes	<2	18+	speech
2	UC Davis	Chinese	yes	2+	<18	grammar
3	UC Davis	English	yes	2+	18+	vocab
4	UC Davis	English	yes	2+	18+	speech
5	UC Davis	English	no	2+	18+	speech
6	Colgate	Chinese	no	<2	18+	grammar
7	UW	English	no	<2	18+	speech
8	UW	English	no	<2	<18	speech
9	UW	English	yes	<2	18+	script
10	Kentucky	English	no	<2	<18	script
11	Kentucky	English	no	<2	18+	speech
12	UC Davis	English	yes	2+	<18	speech
13	Colgate	English	no	<2	<18	script
14	Colgate	English	no	<2	18+	grammar
15	UC Davis	English	no	<2	18+	script
16	UC Davis	English	no	<2	18+	grammar
17	Other	English	yes	2+	18+	grammar

[Figure 29. Simplified metadata.]

For the university category, the single observations from Yale and Emory were combined into a group titled "Other". The L2+ category was omitted, due to lack of overlapping observations to condense into reasonable groups. The binary variable for visiting Japan remained, but the duration of the visit had to be omitted, again due to lack of overlapping observations. "Years studied" and "Age when started" were each collapsed into binary values—less than or greater than 2, and less than or greater than 18, respectively. Finally, because only one participant was a heritage speaker, that variable was also eliminated.

I began my statistical analysis by developing a measure of improvement, which I titled "totaldifference". This value was equivalent to each participant's pre-test errors minus their post-test errors, in other words the reduction of errors for each participant. Similarly, I calculated the difference for spelling errors, vocabulary errors, case errors, and grammar errors, as well as the difference for each of the 15 error types.

First, I performed a one-sample t test on totaldifference. This would indicate if the mean for totaldifference was greater than zero, which would indicate that the subjects overall had experienced a reduction in errors between the pre-test and the post-test. The output of the t-test was as follows:

Test of $\mu = 0$ vs $\neq 0$

Variable	N	Mean	StDev	SE Mean	95% CI	T	P
totaldifference	17	6.47	7.84	1.90	(2.44, 10.50)	3.40	0.004

[Figure 30. T-test output]

The null hypothesis was that the true mean equals zero. Because the p-value is less than the threshold $\alpha = 0.05$, I rejected the null hypothesis. There was a statistically significant reduction of errors between the pre-test and post-test means. The 95% confidence interval for the t-test puts this true mean value between 2.44 and 10.50 errors eliminated.

After concluding via the t-test that participants had improved between the pre-test and the post-test, I investigated the metadata variables with ANOVA, seeking any statistically significant differences in means among the groups. Figure 31 shows the results of one-way ANOVA. The factors were various metadata variables, given along the y-axis of the table, while the responses investigated were the difference between pre-test and post-test errors overall, as well as the difference between pre-test and post-test errors for the spelling, vocabulary, case, and grammar error categories. Statistically significant findings are indicated with yellow shading.

	totaldifference	spelling	vocabulary	case	grammar
University	d.f. = 4 f = 0.22 p = 0.922	d.f. = 4 f = 1.14 p = 0.383	d.f. = 4 f = 0.788 p = 0.559	d.f. = 4 f = 0.53 p = 0.717	d.f. = 4 f = 1.47 p = 0.271
L1	d.f. = 1 f = 0.20 p = 0.665	d.f. = 1 f = 0.46 p = 0.509	d.f. = 1 f = 1.88 p = 0.190	d.f. = 1 f = 0.08 p = 0.778	d.f. = 1 f = 0.13 p = 0.720
Japan visit	d.f. = 1 f = 0.85 p = 0.372	d.f. = 1 f = 2.97 p = 0.106	d.f. = 1 f = 0.26 p = 0.618	d.f. = 1 f = 2.47 p = 0.137	d.f. = 1 f = 0.07 p = 0.799
Years studied	d.f. = 1 f = 5.89 p = 0.028	d.f. = 1 f = 0.09 p = 0.773	d.f. = 1 f = 1.35 p = 0.263	d.f. = 1 f = 1.47 p = 0.243	d.f. = 1 f = 10.72 p = 0.005
Age when started	d.f. = 1 f = 0.24 p = 0.633	d.f. = 1 f = 5.71 p = 0.030	d.f. = 1 f = 0.09 p = 0.762	d.f. = 1 f = 0.33 p = 0.575	d.f. = 1 f = 1.67 p = 0.216

[Figure 31. One-way ANOVA with metadata factors]

The age at which a participant started studying Japanese (less than 18 versus greater than 18) had a statistically significant effect on spelling errors. Because this effect was *not* similarly observed in the Years Studied variable, I am uncertain why starting age had an effect on spelling error reduction. It is unlikely that this affected the core statistical analyses most relevant to the experiment, so I report it here and suggest it as a possible focus of future investigations.

Most notably, the one-way ANOVA revealed that the Years Studied variable had an extremely statistically significant effect on the reduction of grammar errors. Ultimately, this is sensible—subjects with more experience studying Japanese are more likely to have encountered causative grammar in the past, even if it was not taught to them explicitly. This influence of Years Studied over grammar errors appears to contribute to the statistically significant difference in overall errors (totaldifference), which is a more modest effect.

In order to analyze this effect more thoroughly, I examined the effects of Years Studied on initial grammar errors, the reduction of grammar errors, initial overall errors, and the reduction of overall errors.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
2+yrsstudied	1	89.00	88.998	15.51	0.001
Error	15	86.06	5.737		
Total	16	175.06			

Means

2+yrsstudied	N	Mean	StDev	95% CI
false	11	5.455	2.806	(3.915, 6.994)
true	6	0.667	1.211	(-1.418, 2.751)

Pooled StDev = 2.39528

[Figure 32. One-way ANOVA of pre-test grammar errors by Years Studied]

As shown in Figure 32, subjects who had studied Japanese for two or more years had significantly fewer grammar errors on the pre-test. Those subjects, to a lesser degree, also had significantly fewer errors on the pre-test overall:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
2+yrsstudied	1	247.5	247.53	5.20	0.038
Error	15	714.5	47.63		
Total	16	962.0			

Means

2+yrsstudied	N	Mean	StDev	95% CI
false	11	15.82	7.74	(11.38, 20.25)
true	6	7.83	4.79	(1.83, 13.84)

Pooled StDev = 6.90154

[Figure 33. One-way ANOVA of pre-test errors by Years Studied]

As might be expected, the group with fewer years of Japanese experience showed far greater improvement in the experiment. Because the group with more experience performed better on the pre-test, they had a reduced opportunity to improve their errors relative to the group with less experience. (See Figures 34 and 35 for details).

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
2+yrsstudied	1	85.65	85.651	10.72	0.005
Error	15	119.88	7.992		
Total	16	205.53			

Means

2+yrsstudied	N	Mean	StDev	95% CI
false	11	3.36	3.32	(1.55, 5.18)
true	6	-1.333	1.366	(-3.793, 1.127)

Pooled StDev = 2.82700

[Figure 34. One-way ANOVA of grammar error reduction by Years Studied]

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
2+yrsstudied	1	277.5	277.51	5.89	0.028
Error	15	706.7	47.12		
Total	16	984.2			

Means

2+yrsstudied	N	Mean	StDev	95% CI
false	11	9.45	7.59	(5.04, 13.87)
true	6	1.00	5.10	(-4.97, 6.97)

Pooled StDev = 6.86405

[Figure 35. One-way ANOVA of overall error reduction by Years Studied]

Due to the statistical significance of the Years Studied variable, I tested for interaction effects and even ran statistics with the more experienced users held out. However, I observed no significant statistical effect of Years Studied on the performance of other variables in statistical tests, so I can tentatively claim that its effects are limited to the range of improvement possible for an individual user, and that it did not affect the overall results of this study.

After analyzing the effect of the metadata variables, I analyzed the core focus of this experiment: the effect on treatment group on mean error reduction. This involved one-way ANOVA to measure the difference in mean “totaldifference” (pre-test errors minus post-test errors) for each of the three groups. The results of the ANOVA analysis are reported in Figures 36 and 37 below.

Null hypothesis	All means are equal
Alternative hypothesis	At least one mean is different
Significance level	$\alpha = 0.05$

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
groupid	3	0, 1, 2

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
groupid	2	32.58	16.29	0.24	0.790
Error	14	951.66	67.98		
Total	16	984.24			

Model Summary

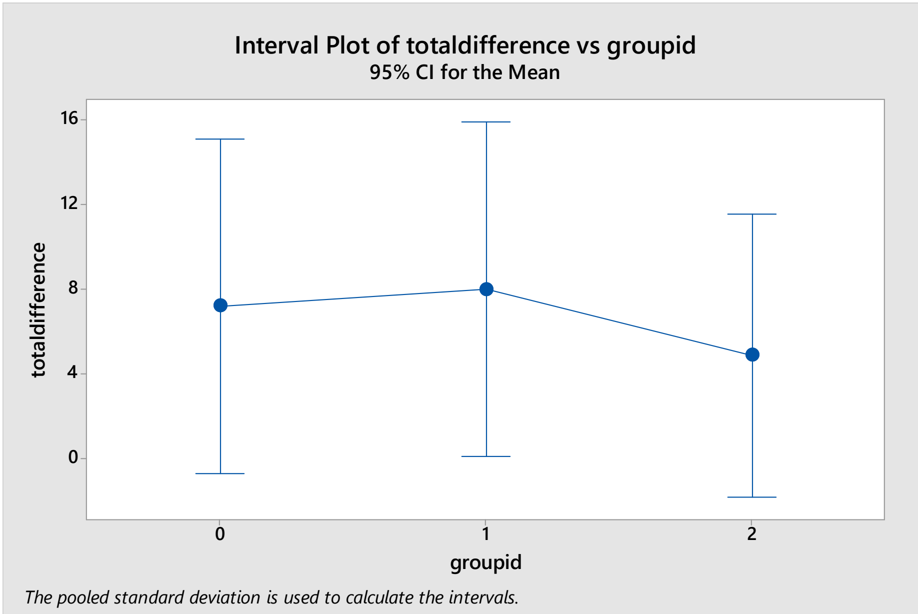
S	R-sq	R-sq(adj)	R-sq(pred)
8.24473	3.31%	0.00%	0.00%

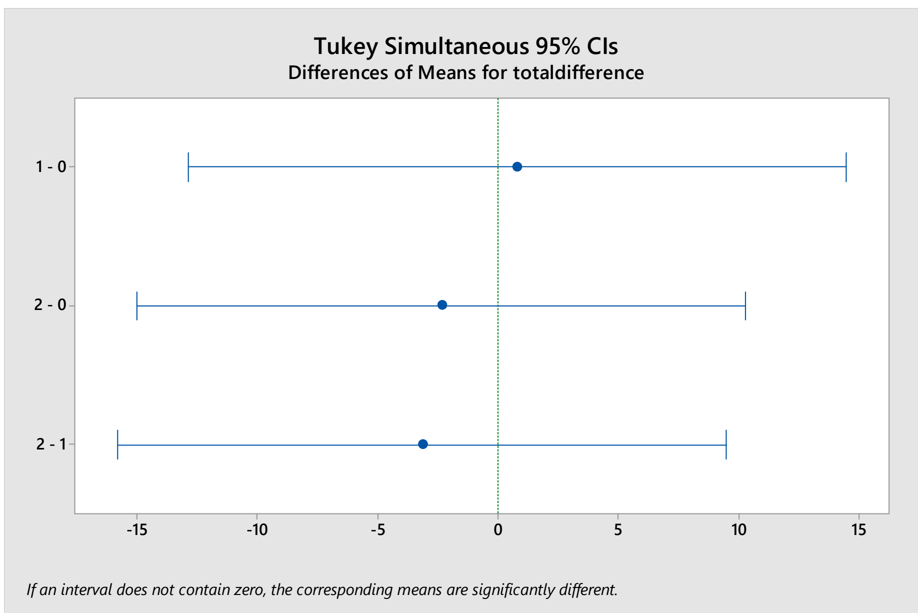
Means

groupid	N	Mean	StDev	95% CI
0	5	7.20	10.03	(-0.71, 15.11)
1	5	8.00	9.77	(0.09, 15.91)
2	7	4.86	5.27	(-1.83, 11.54)

Pooled StDev = 8.24473

[Figure 36. One-way ANOVA: totaldifference by group]





Tukey Simultaneous Tests for Differences of Means

Difference of Levels	Difference of Means	SE of Difference	Adjusted 95% CI	T-Value	P-Value
1 - 0	0.80	5.21	(-12.84, 14.44)	0.15	0.987
2 - 0	-2.34	4.83	(-14.97, 10.29)	-0.49	0.879
2 - 1	-3.14	4.83	(-15.77, 9.49)	-0.65	0.795

Individual confidence level = 97.97%

[Figure 37. Interval plots and Tukey comparisons]

Given the p-value of 0.790, far above the threshold $\alpha = 0.5$, I rejected the null hypothesis. This indicated that there were no statistically significant differences in the mean error reduction among the three groups. In other words, each treatment seemed to be equally effective, or nearly as effective, compared to the others at facilitating subject improvement.

Due to the small sample size, I decided to also perform a nonparametric Kruskal-Wallis test, in case the distribution was not normal. This analysis, as shown in Figure 38, echoed the results of the one-way ANOVA: there were no statistically significant differences in total error reduction among the three treatment groups.

Kruskal-Wallis Test on totaldifference

groupid	N	Median	Ave Rank	Z
0	5	12.000	9.0	0.00
1	5	4.000	10.0	0.53
2	7	4.000	8.3	-0.49
Overall	17	9.0		

H = 0.34 DF = 2 P = 0.845
H = 0.34 DF = 2 P = 0.844 (adjusted for ties)

[Figure 38. Kruskal-Wallis test output]

As a final test of treatment group effects, I fit a generalized linear model to check for interaction between the testset variable (i.e., whether participants saw test A or test B as the pre-test, and vice versa for the post-test) and the treatment group. The outcome, as reported in Figure 39, showed no statistically significant variation in the mean error reduction based on treatment group, test set, or the interaction of the two.

Finally, I analyzed whether the self-reported weaknesses of the subjects actually correlated with their performance both before and after the exercises. I created categorical variables for three of the four weaknesses, eliminating vocabulary because there was only one observation. Then, for each weakness, I ran one-way ANOVA to determine whether subjects who indicated that as their weakness performed significantly worse in pre-test errors overall, pre-test errors by category, error reduction overall, and error reduction by category. The p-values resulting from these analyses are reported in Figures 40 and 41 below.

```
Analysis of Variance

Source            DF  Adj SS   Adj MS   F-Value   P-Value
  groupid          2   10.58    5.290    0.08      0.925
  testset          1  169.00  169.000    2.49      0.143
  groupid*testset  2   30.88   15.440    0.23      0.800
Error             11  745.83   67.803
Total             16  984.24

Model Summary

S          R-sq     R-sq(adj)  R-sq(pred)
8.23426    24.22%   0.00%      0.00%
```

[Figure 39. Output of fitting a generalized linear model.]

PRE-TEST					
	total	spelling	vocabulary	case	grammar
script	0.354	0.056	0.633	0.679	0.411
grammar	0.793	0.707	0.226	0.863	0.780
speech	0.325	0.131	0.297	0.289	0.927

[Figure 40. P-values: ANOVA for pre-test errors by self-reported weakness.]

			OVERALL		
	totaldifference	spelling	vocabulary	case	grammar
script	0.829	0.534	0.843	0.983	0.628
grammar	0.620	0.742	0.180	0.822	0.616
speech	0.844	0.859	0.405	0.964	0.688

[Figure 41. P-values: ANOVA for error reduction by self-reported weakness.]

As shown in the above tables, none of the self-reported weaknesses showed a statistically significant correlation with either pre-test errors or overall error reduction in that area—or any other area, for that matter. This indicates that L2 learners may not be effective at identifying the areas in which they need improvement.

Before transitioning to a discussion of the results, I will once again summarize the results of the statistical analysis. A one-sample t-test indicated that the subjects overall did show a reduction of error after the exercises (mean = 6.47, st. dev. = 7.84, 95% CI = (2.44, 10.50), $t = 3.40$, $p = 0.004$), but one-way ANOVA (d.f. = 2, f-value = 0.24, p-value = 0.790) and the Kruskal-Wallis Test ($H = 0.34$, d.f. = 2, $p = 0.845$) suggested that there were no statistically significant differences in the mean error reduction for each group. Statistical analysis also indicated that participants with two or more years of Japanese study made fewer initial errors (one-way ANOVA, d.f. = 1, f-value = 5.20, p-value = 0.038) and showed more modest error reduction overall (one-way ANOVA, d.f. = 1, f-value = 5.89, p-value = 0.028). Finally, despite numerous ANOVA tests of both pre-test errors and overall error reduction, no statistically significant correlation was found between self-reported weakness and errors of that type.

5 General Discussion

5.1 Significance

Although tree-based CALL did not outperform the other two methods as predicted, it was equally effective at reducing subject errors, indicating it is a worthwhile topic for future research. This experiment showed that tree-based CALL exercises were effective at reducing participant errors on Japanese causative translations when paired with explicit instruction. It remains to be investigated whether tree-based CALL itself can be used as a means of implicit instruction, whether it is similarly effective for other grammar concepts and other languages, and whether it is preferred by subjects over other methods. Because this experiment is the first to investigate dependency tree construction for computer-assisted language instruction, it represents a modest finding compared to the vast amount of potential research.

The secondary findings of this experiment demonstrate the importance of individualization in CALL. As elementary-level students of Japanese, all participants technically had the same L2 proficiency. Those with two or more years of exposure to Japanese study, however, made significantly fewer errors on the pre-test, and improved less than their inexperienced counterparts as a result. On the other hand, there was no significant correlation between perceived weakness and actual pre-test or post-test performance. This indicates that individual users may not be accurate at reporting their own deficiencies; thus, in order to cater to individual needs, CALL systems must find another way of approximating those needs. Adapting to individual deficiencies or levels of experience was certainly outside the range of this study, but this finding does indicate an opportunity for CALL systems to implement better individualized instruction, as well as an opportunity to explore tree-based CALL for that purpose.

5.2 Strengths and Weaknesses of the Experiment

The key strength of this experiment is that it considers a means of CALL never before implemented or investigated: the construction of tree structures. In my background research of second language acquisition, computer-assisted language learning, and dependency trees, I have been unable to find a single paper that utilizes trees structures of any kind for second language instruction. The finding of this study that tree-based practice is equally as effective as worksheet-type drills and phrase-based translations supports adding tree construction exercises to the arsenal of language-learning tools.

Another strength of this experiment is its incorporation of both academic research and a realistic understanding of the CALL industry. While researching the history of CALL, I observed that advances in CALL tend to coincide with advances in technology, not necessarily advances in SLA research. Additionally, because the CALL of the past 20 years has become dominated by industry, rather than academia, feasible changes to CALL must consider industrial concerns as well as the effectiveness of the method. In designing this experiment, I kept in mind the elements supported by SLA research—such as explicit grammar instruction—while attempting to build a system that could be successfully adapted by a CALL company—i.e., one that did not rely too heavily on explicit explanations and did not require a large amount of exercises to produce a result. As shown by the one-sample t test and the one-way ANOVA of error reduction by group, all three treatment groups I created were indeed effective at improving the users' Japanese.

The key weakness in this experiment, as in many studies involving less-commonly-taught languages (LCTLs), is sample size. Although I contacted 39 universities, only 6 universities ultimately participated, and from those universities only 17 subjects completed the study. This resulted in an average 5.5 participants per treatment group, which decreased the chances of observing a statistically significant effect if one existed. Because some of the statistical tests (the one-way ANOVA of Years Studied, as well as the one-sample t test) did show statistically significant results, I posit that there may not have been a statistically significance between the true means for the treatment groups. A larger sample size would, however, allow for more solid statistics. It is worth noting here that the sample size for this experiment was within the norms, if at the lower end, for second language acquisition research. Norris and Ortega report a mode sample size of 34, and average group sizes ranging from 5 to 35 (2000: 456). Thus, while this experiment is small, it deserves consideration among and comparison with other studies in the SLA field.

Finally, as Norris and Ortega (2000) eloquently states, "no single investigation of the effectiveness of L2 instruction can begin to provide trustworthy answers" (423). Only repeated experiments will conclusively reveal whether tree-based CALL is an effective means of reducing grammatical errors. The preliminary results provided by this experiment, however, are promising, and in the next section I propose future work to investigate the potential of tree-based CALL.

6 Conclusions and Future Work

While tree-based CALL is not significantly more effective than existing methods of CALL, it did perform equally well at reducing subject errors. Future work should establish whether these results can be replicated, as well as whether tree-based CALL is effective for other grammar concepts, other LCTLs, and perhaps even for syntactically similar languages.

One area not addressed by this study was subject reactions to the tree-based activity. Since all three methods were shown to produce relatively equivalent improvement in the subjects' abilities, perhaps any real-world implementation of these exercises should allow students to choose, rather than randomly assigning the exercise type. Given such a system, it would be particularly interesting to determine which method the subjects prefer, and whether the ability to choose the method of practice had any impact on subjects' ultimate improvement.

This study also did not establish whether these methods are effective as a means of implicit grammatical instruction. Although SLA research strongly supports explicit instruction as the more powerful method, many popular CALL systems refuse to integrate explicit instruction. For this reason, it may be worthwhile to investigate the effectiveness of the methods as implicit instruction, rather than practice following explicit instruction. If the tree-based method is proven to outperform the others in implicit instruction, it could prove a valuable tool for current CALL programs.

Two metadata variables were of particular interest in this study: the subject's years of Japanese experience, and the subject's self-reported area of weakness. Statistical analysis showed that subjects with two or more years of Japanese experience, even if enrolled in a beginner-level class, had significantly fewer grammatical errors before the experiment and therefore made smaller gains. Although outside the scope of research on tree-based CALL, it would be interesting to determine ways to better gauge such participants' fluency and weak points and to focus exercises on those areas. This would optimize the effectiveness of the exercises for each individual and allow every participant to improve with equal magnitude, regardless of their initial skill level.

Along those same lines of individual optimization, the failure of self-reported weakness to correlate with actual errors indicates that another measure of detecting individual deficiencies may be needed. In future studies, it would be of interest to determine whether particular methods of CALL are more effective for subjects with particular weaknesses. Dependency tree construction, for example, might prove more effective for subjects who demonstrate difficulty with case-marking and principles of causative grammar. If correlations of this type were discovered, it would enable CALL to provide a new level of individualization and offer more efficient instruction to its users.

This study is, to my knowledge, the first exploration of tree-based CALL. Much work, therefore, remains to be done on the subject, and now that the effectiveness of tree-based CALL has been indicated by this experiment, it is my sincere hope that other researchers and the CALL industry will consider the potential of this new and promising form of computer-assisted language instruction.

Acknowledgements

I am deeply grateful to Dr. Claire Moore-Cantwell, my thesis advisor, for agreeing to supervise me on this nontraditional thesis and for her patient help and feedback at every stage.

I would also like to express my sincere thanks to Dr. Raffaella Zanuttini, the Director of Undergraduate Studies in Linguistics, for her guidance over the past four years and particularly during the development and completion of this thesis.

I owe the success of this experiment to the Japanese instructors who encouraged their students to participate: Professor Mari Stever at Yale University, Professor Yukari Hirata at Colgate University, Yoko Horikawa at University of Kentucky, Professor Julia Bullock at Emory University, and Professor Nobuko Koyama at University of California, Davis. Thank you very much!

A special thank you to Maki Sato, who helped revise the Japanese sentences in the corpus, and to Dr. Jonathan Reuning-Scherer, who advised me on experiment design and statistical analysis. All errors are my own.

Finally, I would like to thank my parents Adrienne and Steve Little, and my sister Caroline Little, for their enduring love and support throughout my life.

References

Bostock, Michael, Vadim Ogievetsky, & Jeffrey Heer. 2011. D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions* 17(12). 2301-2309.

Brown, Steven & Jennifer Larson-Hall. 2012. *Second Language Acquisition Myths*. Ann Arbor: University of Michigan Press.

Davies, Graham. 2008. CALL (computer assisted language learning). *Centre for Languages Linguistics & Area Studies*. http://www.llas.ac.uk/resources/gpg/61#toc_1 (7 March 2016).

Delcloque, Philippe (ed.). 2000. The history of computer assisted language learning web exhibition. *Computer Assisted Language Instruction Consortium (CALICO)*. http://www.ict4lt.org/en/History_of_CALL.pdf (17 April 2016).

Dooijes, Edo H. n.d. The PLATO-IV system for computer aided instruction. *Computer Museuem.* Amsterdam: University of Amsterdam. https://ub.fnwi.uva.nl/computermuseum/PLATO.php (10 March 2016).

Duolingo. n.d. About Duolingo. https://www.duolingo.com/press (10 March 2016).

Ellis, Rod. 2005. Principles of instructed language learning. *System* 33(2). 209-224.

Flege, James E., Grace Yeni-Komshian & Serena Liu. 1999. Age constraints on second-language acquisition. *Journal of Memory and Language* 41. 78-104.

Harley, Heidi. 2008. On the causative construction. In Miyagawa, Shigeru and Mamoru Sato (ed.), *Handbook of Japanese Linguistics.* Oxford: OUP. http://babel.ucsc.edu/~hank/mrg.readings/harley_06_On-the-causativ.pdf (18 April 2016).

Hubbard, Philip (ed.). 2009. *Computer Assisted Language Learning*, vol. 1. New York: Routledge.

Hudson, Grover. 2000. *Essential introductory linguistics.* Malden, MA: Blackwell.

Johnson, W. Lewis, Stacy Marsella & Hannes Vilhjálmsson. 2004. The DARWARS tactical language training system. *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).*

Little, Alexa & Stephen Tratz. Forthcoming. EasyTree: a Graphical Tool for Dependency Tree Annotation. *Language Resources and Evaluation Conference (LREC).*

Little, Alexa. 2016. Duolingo content. *Duolingo*, n.d. Author's screenshot. http://www.duolingo.com (8 January 2016).

McNeil, Sara. 2003. A hypertext history of instructional design. http://faculty.coe.uh.edu/smcneil/cuin6373/idhistory/ticcit.html (9 March 2016).

Moss, Richard. 2014. Learn Immersive teaches language in virtual reality. *Gizmag.* http://www.gizmag.com/learn-immersive-language-virtual-reality/35128 (8 March 2016).

Norris, John M. & Lourdes Ortega. 2000. Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language learning* 50(3). 417-528.

Omaggio-Hadley, Alice. 2000. *Teaching language in context.* Boston: Heinle.

Ohtani, Akira. 2013. Locative postpositions and conceptual structure in Japanese. *PACLIC 27*. http://www.aclweb.org/anthology/Y13-1039 (19 April 2016).

Rosetta Stone. 2001. Detail page Rosetta v2 screen 2. http://www.amazon.com/dp/B000TFI08U (8 March 2016).

Rosetta Stone. 2016. History. http://www.rosettastone.com/history (10 March 2016).

The Unicode Consortium. 2015. Unicode 1.0. http://www.unicode.org/versions/Unicode1.0.0 (7 March 2016).

Transparent Language. 2015. About us. http://www.transparent.com/about/company_info.html (10 March 2016).