



Overspecified references: An experiment on lexical acquisition in a virtual environment



Alexandra Luccioni^{a,*}, Luciana Benotti^a, Frédéric Landragin^b

^aLogic, Interaction and Intelligent Systems Group, FAMAf, Universidad Nacional de Córdoba/CONICET, Medina Allende s/n, Ciudad Universitaria, Córdoba CP 5000, Argentina

^bLaTITCe Laboratory, ENS, UMR 8094, 1, Rue Maurice Aroux, F-92120 Montrouge, France

ARTICLE INFO

Keywords:

Referring expression generation
Overspecification
Lexical acquisition
Virtual worlds
Computer Assisted Language Learning

ABSTRACT

This study explores the role of quantity of information on vocabulary acquisition in a virtual world. Previous studies have shown that, although it makes interpretation more lengthy, speakers include more information in their referring expressions than is strictly necessary to identify an object – they *overspecify*. We aim to study the impact of this kind of overspecification on the acquisition of new lexemes in a foreign language. Our hypothesis is that using overspecified expressions during the practice of recently acquired vocabulary will help learners to better remember the lexemes and to exploit them more efficiently later on. In this paper, we describe an experimental study designed to evaluate this hypothesis, comparing two groups of learners who received overspecified and minimally specified referring expressions while practising newly acquired lexemes in the context of a language learning game in Russian. The game is situated in a virtual environment and the interaction is similar to that of a video game. Our results, based on experimental data from participants' performance as well as a post-experiment questionnaire, support the claim that overspecification improves lexical acquisition rates compared to minimal specification. Some pedagogical suggestions are provided for the design of referring expression generation algorithms in Technology-Enhanced Language Learning (TELL) and Computer Assisted Language Learning (CALL) Systems.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

In our increasingly connected and internationalized world, learning a foreign language at an adult age is something that more and more people are facing. However, due to the number of learners worldwide, it is difficult to provide personalized, one-on-one teaching to all of them. Given the potential of technology to enhance and improve education in general, it is to be expected for learners to turn to CALL (Computer Assisted Language Learning) to learn languages (Dickinson & Herring, 2008; Johnson, 2007; Zhao, 2003; Sørensen & Meyer, 2007). This brings new questions as to the characteristics that CALL software should have in order to provide the most efficient learning experience.

Language acquisition is a complex process, with many variables that can affect a learner's progress on his or her unique learning path. In terms of vocabulary acquisition, the quantity of new

lexemes to be provided at once during the language acquisition process is the subject of much debate (Hulstijn & Laufer, 2001; Sagarra & Alba, 2006). We propose that instead of teaching isolated lexemes or minimally specified expressions, a CALL system should provide learners with overspecified referring expressions (REs). This is based on previous research in the field of Referring Expression Generation (REG), which has shown that, although it makes interpretation more costly, redundant information is frequently used in the REs produced by speakers (Engelhardt, Bailey, & Ferreira, 2006; Pechmann, 1989; Wu & Keysar, 2007). For example, Viethen and Dale (2011) found that, when identifying a big red ball sitting next to a big red cube in Fig. 1, most people use the referring expression “*the big red ball*”. They do so even though there are no small red balls in the figure, and “*the red ball*” provides enough information to identify the target object. We say that the RE “*the big red ball*” is *overspecified* because it includes more information than is necessary to univocally distinguish the referent.

In the literature, there are two explanations that have been proposed as contradictory for the overspecification phenomenon. Explanation (1) holds that overspecification impairs communication and is a result of a cognitive limitation of the speakers. Explanation (2) argues that overspecification is a useful part of

* Corresponding author at: Laboratoire de Gestion, Diffusion et Acquisition de Connaissances, Université du Québec à Montréal, PK-4285, 201 Avenue du Président-Kennedy, H2X 3Y7 Montréal, Canada. Tel.: +1 514 436 4865.

E-mail addresses: alexandra.vorobyova@gmail.com (A. Luccioni), benotti@famaf.unc.edu.ar (L. Benotti), frederic.landragin@ens.fr (F. Landragin).

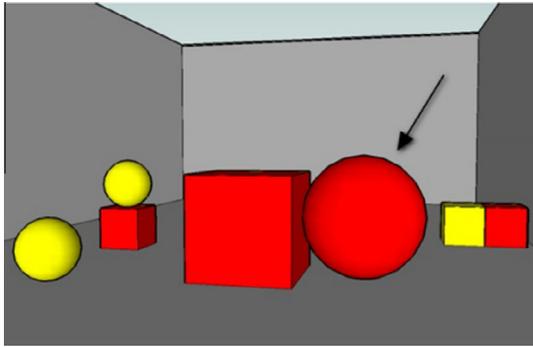


Fig. 1. An example of a visual scene with multiple objects. Source: Viethen and Dale (2011).

communication because it encourages lexical alignment between the speaker and the listener. Below we discuss the cognitive mechanisms for both explanations in turn.

Explanation (1) holds that the cognitive effort of producing non-redundant (i.e., minimal) referring expressions is too high (Pechmann, 1989). Pechmann's *Principle of least effort* suggests that language production is an incremental process and that it is not cognitively efficient for speakers to evaluate the entirety of their surroundings before producing a minimal specification. Waiting for such a complete representation of the context can be too costly for the speaker and can result in postponing the completion of language production processes. On the contrary, including salient and redundant properties in a referential expression as they come to mind is less cognitively demanding than explicitly verifying that they are redundant in order to omit them. Experimental results supporting this theory see overspecification as a result of a cognitive limitation of speakers, which impairs listeners' comprehension (Engelhardt, Demiral, & Ferreira, 2011). Engelhardt et al. found that listeners take longer to process overspecified referring expressions in comparison to minimal ones and concluded that listeners are confused and their comprehension impaired by overspecification. To the best of our knowledge, no empirical data exists showing that it takes longer to produce minimal referring expressions compared to overspecified ones.

Explanation (2) claims that overspecification is a useful part of communication because it allows speaker and listener to align by increasing their *common ground*, that is, the information they share regarding their environment (Nadig & Sedivy, 2002). As a result, future communication becomes more effective (Wu & Keysar, 2007). This hypothesis is also supported by empirical data: Clark and Wilkes-Gibbs (1986) have found that both speaker and listener use redundant properties in early references to an object, and reuse these properties in later references to the same object or to other objects with the same properties. This phenomenon can be perceived as a mechanism aiming to provide the listener with more information than is strictly necessary in order to compensate for potential perceptual difficulties or interference, as well as to coordinate on which lexemes are suitable for describing controversial physical properties. For example, a speaker may refer to a particular jacket as red using the overspecified RE "the red jacket on the sofa" when this is the only jacket on the sofa. Even if his listener thinks that the jacket is pink, she will nonetheless be able to resolve the RE because the property red is redundant. Later on, the listener may begin to refer to it as red herself in order facilitate communication.

Contrary to what has been argued in previous work, we believe that explanations (1) and (2) are not contradictory. They differ in how they study communication. Explanation (1) analyzes only single turns between speaker and listener while explanation

(2) consider multiple turns. In the short term, any delay to the listeners interpretation such as the one found by Engelhardt et al. (2011) can be seen as detrimental for communication. However, in the long run, overspecification leads to more shared information, that is, to more lexical alignment between speaker and listener. This in turn, leads to more effective communication. We believe that communication is a multi-turn process and therefore we adhere to explanation (2).¹

We see second language lexical acquisition as an extreme case of alignment, since learners are obliged to use the same vocabulary and grammatical constructions as their teachers due to the limited choice of linguistic resources that they can use (Atkinson, Churchill, Nishino, & Okada, 2007). A study such as ours, focusing on lexical acquisition, permits us to control the speakers knowledge of a language and therefore to target the 'square one' from where alignment starts.

In this paper, we aim to support explanation (2) by empirically evaluating the effect of overspecification on lexical acquisition in second language (L2) learning. To this end, we created a system that produces minimal and overspecified REs describing objects in a 3-dimensional virtual world. We divided the participants into two groups: one group received minimal REs and the second group received overspecified REs during the practice phase following an initial exposure to new lexemes. We aim to see if, by giving overspecified REs during the practice period of lexical acquisition, the effectiveness of vocabulary learning increases. If this is the case, it has theoretical and practical implications. On the theoretical side, it provides evidence to support explanation (2), which goes to show that overspecification is useful for communication. On the practical side, it has a direct application in the design of referring expression generation algorithms for Computer-Assisted Language Learning (CALL): the descriptions used during lexical acquisition should be overspecified in order to increase lexical acquisition rates.

The rest of the paper proceeds as follows. In Section 2, we describe the various practical aspects of our experiment: the participants, the virtual world used, and the procedure. We present both the objective and subjective results of the experiment in Section 3, specifying the metrics we looked at and their pertinence to evaluating our hypothesis. The discussion of the results follows in Section 4, where we compare the results of the two groups of participants studied. Finally, Section 5 concludes and discusses potential applications to expert systems useful for CALL.

2. Materials and methods

We designed an experiment that uses a virtual environment setting, screenshots of which are shown in Figs. 2 and 3, and the map of the environment can be found in Appendix A. We programmed an instruction-giving system to guide participants in the world and permit them to acquire new vocabulary words. By comparing participants who practised new words with minimal and overspecified referring expressions, we aim to observe the utility of overspecification for L2 lexical acquisition.

2.1. The participants

We recruited 50 participants (15 women and 35 men) by sending out an invitation to test a language learning game. All of the participants were university students from the Universidad Nacional de Cordoba in Argentina. The average age was 28. All of the participants were native Spanish speakers who had no prior

¹ Communication is a multi-turn process both in speech-based interactions and in text-based interactions such, as the chat used in our experiment.



Fig. 2. Screenshot of the 3D virtual environment used by the participants. Navigation instructions were given in Spanish.

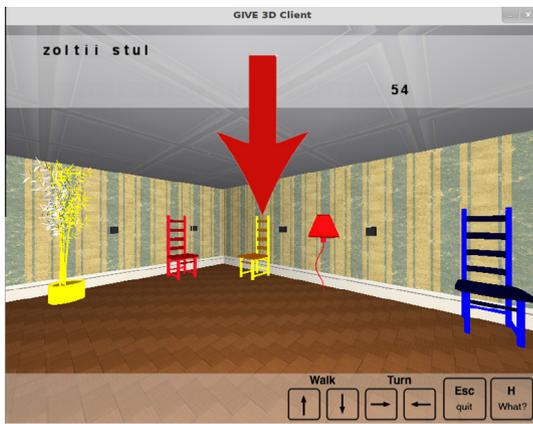


Fig. 3. Screenshot from the Practice Phase – RE received by a participant in the MR condition.

knowledge of Russian, the language taught during the experiment, and they were not given any additional information about the experiment. We collected demographic information such as age, gender, and video game familiarity. We also asked for the participants' subjective opinion regarding the experiment via a questionnaire after they had completed the task.

We randomly assigned the participants to one of two groups: the *MR (Minimal Reference) Group* ($n = 25$), in which the participants received vocabulary exercises with minimal referring expressions during the practice phase, and the *OR (Overspecified Reference) Group* ($n = 25$), who received overspecified referring expressions during the same phase. Minimal referring expressions provide just enough information needed to identify the target object, whereas overspecified referring expressions include redundant properties beyond what is strictly necessary for identification. Both groups received an identical amount of lexemes during the initial learning and testing stages of the experiment.

2.2. The experiment

Our experiment employed an instruction-giving system that guided participants through a series of language-learning and practice tasks in a virtual world. The system gave instructions in Spanish and taught the participants a set of words in the context of a 3D virtual language learning game. At first, the participants were invited to explore the world and to discover the names of objects in Russian at their own pace by pressing a button beside

the object. After an initial exposure to the words, they passed a first test, then were given exercises to practice the words (while receiving feedback from the system), followed by a second test, identical to the first, to measure their improvement.

The lexemes the participants were taught were from 3 categories: Nouns (3 names of objects), Colors (4 colors) and Locations ('Left of' and 'Right of'), for a total of 9 words. They were initially taught the lexemes in isolation, but the lexemes were later assembled to describe complex objects (color + object in the MR condition, and color + object + location for the OR condition). Since our participants could not read the Cyrillic alphabet, we transliterated the REs in Russian using the ISO 9:1995 standard Romanization of Cyrillic. We also replaced recognizable Russian words with Latin roots by non-Latin synonyms, in order to avoid vocabulary bias.

2.2.1. The virtual world

We carried out an experiment in a virtual world specially designed to test our hypothesis. The 3D view of a room in the virtual world is shown in Fig. 2. The virtual world was designed using the GIVE platform (Koller et al., 2008). The GIVE platform includes a set of tools for designing virtual houses that can contain different virtual objects with distinguishing properties (such as color and shape). We chose a virtual environment because we find it to be more immersive for language acquisition than flashcards or images: participants could see and interact with the objects from a variety of angles and viewpoints, and examine them more closely by moving closer to the objects.

The GIVE platform and the design of the experiment were such that each time a button was pressed, we logged the referring expression given, the corresponding referent, and whether the participant pressed the correct button. The participants had 60 s to identify the referents at each time, with a countdown timer appearing on their screen. If this time passed and no button was manipulated, we considered this as an error. Furthermore, we were able to register the participants' visibility area and progress, which enabled us to calculate their speed and track their movement during the experiment. Each phase of the experiment took place in a different room of the virtual world, between which the participants navigated independently.

2.2.2. Phases of the experiment

The experiment had five phases: the Tutorial Phase, the Learning Phase, the First Test Phase, the Practice Phase, and the Second Test Phase. Some phases had a single room, while other (such as the Learning Phase) were composed of a series of rooms through which the participant navigated one by one. A map of the virtual world containing the rooms where the phases occurred is shown in Appendix A. The average completion time for the experiment was 13.4 min.

In the *Tutorial Phase*, participants learned how to navigate the world and interact with the objects in it. In the next phase of the experiment, the *Learning Phase*, we asked the participants to press the buttons beside the objects in a room and gave them the name of the object in Russian. Each word stayed visible for 5 s and could only be received once for each lexeme. No specific order of the objects was enforced and participants were free to move in the room as they wished. Fig. 2 shows the view that a participant perceived at the beginning of the Learning Phase. The door to the First Test Phase only opened when all buttons in the Learning Phase had been pressed.

The *First Test Phase*, which came next, evaluated how well the participants retained the lexemes they were exposed to during the Learning Phase. Participants were given single words in Russian and asked to find the correct object in the room and press its button. This was done for 7 of the 9 lexemes they had been

taught (except locations, since they could not be tested in isolation), in a random order. No feedback was given to the participants in this phase. Up to this point, the phases were exactly the same for both groups.

The following phase, the *Practice Phase*, differed for each group. In this phase, the participants were given exercises to practice identifying complex referential expressions (combinations of colors and types of objects, for a total of 12 complex objects). The quantity of information given differed for the Minimal Reference (MR) and Overspecified Reference (OR) conditions: The MR participants received a minimally specified expression, e.g. *zoltii stul* (“yellow chair” in Russian), and the OR participants received an overspecified expression for the same object, e.g. *zoltii stul sleva ot krasnii tsvet* (“yellow chair to the left of the red light” in Russian). Both groups of participants received feedback as to whether or not they correctly resolved the reference, but they were not given a second chance to resolve it correctly if they failed.

Fig. 3 shows a RE as received by a participant in the MR condition – in this situation, the target referent, the yellow chair, was visible to the participant. It is indicated by a red arrow on the figure (the arrow was not visible during the experiment). In the OR condition the participants received the RE “*zoltii stul sleva ot krasnii svet*”, meaning “yellow chair on the left of the red light”, which is overspecified since there was only one yellow chair in the room.

The REs in both figures refer to the same object in the virtual world; however, “yellow chair” gives only the minimal semantic content needed to find the object, whereas “yellow chair on the left of the red light” gives an overspecified description. We decided to overspecify the REs—such as *yellow chair*—with a relation to a neighboring object—to the left of the red light—since there are case studies that show that this is the property that is most frequently overspecified in corpora (Viethen & Dale, 2008).

The final phase, the *Second Test Phase*, with an identical procedure to the First Test Phase, was taken after the Practice Phase to calculate the participants relative improvement. Once again, no feedback was given during this phase. By comparing the scores of participants from both groups in the two test phases, we aimed to see to what extent the Practice Phase helped learners improve their acquisition of new lexemes.

3. Results

In order to have a more complete view of the participants' experience, we completed the data logged during the experiment with results that were obtained via questionnaires given to the participants after the experiment was completed. We designated the first type of metrics as objective metrics (participant performance and behavior) and the second type as subjective metrics (participants' opinion of the experiment) and we will present the results of both below. We describe the objective metrics in Section 3.1, and the subjective metrics in Section 3.2.

3.1. Objective metrics

In order to test our hypothesis, we extracted information on whether and how much each participant's number of errors decreased between the First Test Phase and the Second Test Phase. That is, we measured whether the Practice Phase (which came between the First Test Phase and the Second Test phase) helped the participants to better remember the words they were taught. These metrics were meant to indicate whether practising new vocabulary with the *Minimal Reference* (MR) or the *Overspecified Reference* (OR) condition was most effective.

Table 1

Percentages of participants that decreased their number of errors comparing the first and second test phase classified by the type of lexeme. The values in bold show that the difference between the two conditions is significant $p < 0.01$ (independent t -tests, $df = 48$).

Lexeme acquisition rate	MR	OR	t -value	p -value
Any lexeme ^a (%)	47	89	3.99	0.0001
Object type lexemes (%)	53	89	5.18	0.00001
Color lexemes (%)	50	82.5	2.78	0.007

^a Notice that these values are not equal to adding the other two since many participants improved on both properties.

Table 1 shows the percentage of participants that decreased their number of errors between the First and Second Test Phases (*lexeme acquisition rate*) and the percentage of improvement on the different kinds of properties included in the referring expressions. The expressions given in the Test Phases included types of objects (e.g. chair), realized as nouns, and colors, (e.g. blue), realized as adjectives. We performed independent t -tests on all the metrics in Table 1. The corresponding t -values and p -values are reported in the table per metric. As it can be observed, the improvement was statistically significant for all the metrics in Table 1, with the OR group significantly outperforming the MR group.

Table 2, below, shows the sum of errors for all the participants in the First Test Phase and in the Second Test Phase, the total percentage of errors overcome in the Second Test Phase (*global error overcoming rate*), and the average number of errors per participant during both test phases. We performed independent t -tests on all the averages reported in the table. The average error overcoming rate per participant is considerably higher for the OR condition compared to the MR condition with $p = 0.05$, $t = 1.89$ and $df = 48$. The differences in the averages error per participant in both tests are not statistically significant showing that our OR and MR experimental groups were balanced.

In traditional reference resolution experiments (e.g. Clark & Wilkes-Gibbs, 1986), participants' reaction time is measured as the time between the moment at which the referring expression is presented and the moment at which an object is selected. Since our participants could move freely in the virtual world, the distance from the participant to the referent had an impact on the resolution time: if the participant was further away, it would take them longer to identify the object. In order to normalize our results and factor out the distance, we calculated *resolution speed* instead of resolution time.

We did this by recording the time it took the participant to select an object, from the moment the expression was given to the moment a button was pressed, and dividing it by the distance at which the participant was from the object when the RE was given. We thereby calculate the average speed at which the participants in each condition resolved the referring expressions in the Practice Phase. We complement this with the average success rate achieved by each group during the Practice Phase (meaning the percentage of correctly resolved REs) – we present these metrics in Table 3. We performed independent t -tests on all the metrics reported in the table and we found that the difference in resolution speed is statistically significant ($t = 7.4$, $p = 0.00001$). The success rate in the practice phase for the OR condition is higher than that of the MR condition but the difference is not statistically significant.

Finally, we also collected demographic data, namely age, gender, number of languages spoken, familiarity with 3D video games and profession via a pre-questionnaire. None of these factors had a significant effect on any of the metrics reported except for one: we found a positive correlation of video game familiarity on speed:

Table 2

Number of errors committed in the first and second test phase and the error overcome rate both in total and in average per participant.

	MR	OR
Global first test errors	55	65
Global second test errors	39	37
Global error overcoming rate (%)	29	43
Average error per participant in first test	2.2	2.6
Average error per participant in second test	1.56	1.48
Average error overcoming rate per participant	0.64	1.12

The values in bold show that the difference between the two conditions is significant.

gamers were faster. Our MR and OR groups were balanced with respect to video game familiarity.

3.2. Subjective metrics

After the experiment, participants were asked to complete a questionnaire collecting subjective ratings of various aspects of the experiment. There were four subjective metrics in the post-questionnaire. Table 4 shows the average and standard deviation of the ratings given by the participants, on a scale from 0 to 100. Only for values in bold are the differences between the MR and OR conditions are significant ($p < 0.05$, independent t -tests).

Metrics Q1 and Q2 verify that there were no technical problems with the experiment that could affect the results. Q1 was meant to ensure that the instructions given by the system were clear and easy to understand by the participants; if the participants did not understand what they were supposed to do, it would have a significant effect on their performance. Equally, if the participants did not have enough time to read the instructions (Q2), this would have affected their behavior. However, the results of both metrics were high, meaning that neither significantly affected results.

Q3 aimed to reproduce previous experiments that show that people do not rate overspecified expressions worse than minimal ones even though it may take them longer to resolve them (Engelhardt et al., 2006). In our case, we asked participants whether or not they considered the descriptions given in the exercise room (Practice phase) as 'good' descriptions, comparing the two conditions to see if overspecified REs were judged equal to minimal ones.

The last question, Q4, was the metric that aimed to support our hypothesis. In this question, we asked the participants to evaluate whether they perceived that the Practice Phase was useful for them to acquire the new Russian vocabulary that they had learned.

4. Discussion

Our initial hypothesis was that giving overspecified REs during the practice of newly acquired lexemes would help participants to better remember the new words they had learned. According to this hypothesis, the participants who received overspecification during the Practice Phase should acquire vocabulary more effectively than those who did not. This was confirmed by our results – we found that the improvement between the First Test and Second Test Phases was greater for participants in the OR condition

Table 3

Average reaction speed and success rate for each condition in the Practice Phase.

Practice phase performance	MR	OR
Success rate (%)	59.33	68.00
Resolution speed (cm/s) ^a	101.1	49.88

The values in bold show that the difference between the two conditions is significant.

^a The metric unit used for speed is an interpretation of perceived size in the virtual world.

Table 4

subjective Metrics collected in the post-questionnaire. Only for values in bold show that the differences between the MR and OR conditions are significant (independent t -tests).

Metric evaluated	MR (%)	OR (%)	t -value	p -value
(Q1) The instructions in Spanish were easy to understand	95.1 (9)	94.2 (11)	0.31	0.76
(Q2) The descriptions given in the practice phase were visible enough time for me to read them	90.2 (13)	89.8 (14)	0.19	0.84
(Q3) The descriptions of objects in Russian were good descriptions	80.9 (15)	79.4 (12)	0.11	0.91
(Q4) The exercise room helped me remember the Russian words better	75.1 (21)	85.8 (17)	1.95	0.05

than in the MR condition. We also found that the participants from the OR condition perceived the Practice Phase as more useful than the MR participants. We discuss the implications of these results in the current section.

4.1. Discussion of objective results

Our results indicate that the error overcoming rate was greater for the OR condition. By the Second Test Phase, the number of errors committed by the OR participants dropped by 43%, compared to 29% for MR participants. Also, the average error overcoming rate per participant indicates that each individual participant improved more in the OR condition. This is, arguably, the most important metric, because it not only measures whether the subjects improved consistently but also how much they improved overall (see Table 2).

Furthermore, the success rate in the Practice Phase for the OR group was 8.7% higher than for the MR group (see Table 3), meaning that, in average, people were better at identifying objects when overspecification was provided; however, this difference is not statistically significant. We believe this may be due to the timeout set to 60 s for each exercise in the practice phase. Indeed, the timeout rate in the OR condition was twice higher (20%) than in the MR condition (9%).

Comparing the resolution speeds of the two conditions, we can see that the average MR resolution speed (101 m/s) is more than twice faster than that of the OR condition (50 m/s); these results are coherent with previous research that found that listeners take longer to resolve overspecified referring expressions compared to minimal ones (Engelhardt et al., 2006; Engelhardt et al., 2011).

However, we believe that the OR group's increased resolution time is not indicative of hesitation or confusion, but rather of time used to interpret and integrate the overspecified properties and for participants to update the semantic content of their mental representation of the objects referred to, coherent with the Explanation 2 described in Section 1.

In Engelhardt et al.'s experiment (Engelhardt et al., 2011), it was found that the processing of overspecified REs resulted in a longer resolution time as well as centroparietal negativity (N400), and the conclusion reached was that "too much information is detrimental to comprehension performance" (p 313). However, the N400 amplitude has also been associated with retrieving and storing conceptual knowledge associated with words, as well as semantic integration (Kutas & Federmeier, 2000). We believe that the ERP observed by Engelhardt and colleagues is not indicative of a lexical-access problem, but rather of a semantic integration of the overspecified lexemes into the existing mental representation of the object. In the case of vocabulary acquisition, this corresponds to the anchoring of the

additional properties provided via overspecification, which results in an improvement of the lexeme acquisition rate.

4.2. Discussion of subjective results

For the subjective metrics gathered from the post-experiment questionnaire, there are two questions in particular that are of interest to us: whether the participants thought that the minimal and overspecified descriptions given in the Practice Phase were good descriptions (Q3), and to what extent they thought that the Practice Phase helped them remember the words learned (Q4), in Table 4. For Q3, we found no statistically significant difference between the two conditions. That is, even though OR descriptions may be considered more “cognitively demanding” in accordance with results of previous experiments, they were not judged more difficult to understand by the participants. Experimental evidence that OR descriptions are, in fact, more cognitively demanding is that the resolution speed was slower for OR participants than for MR participants (50 m/s vs. 101 m/s). However, the questionnaire results indicate that although it took OR participants in average twice as much time to resolve referential expressions, they still did not consider them to be ‘worse’. This is coherent with Engelhardt’s results (Engelhardt et al., 2006) that overspecified descriptions are judged equally good as minimal ones, despite the extra effort needed to process them.

For Q4, we found that the participant’s evaluations of the utility of the Practice Phase to be significantly higher in the OR condition (86%) than in the MR condition (75%), which shows that the participants considered that the overspecified training exercises were more useful to them than the minimal specification exercises. This suggests that the participants found that practising the vocabulary with overspecification helped them memorize new words better than just receiving minimal specification: we see this as further support of our hypothesis.

4.3. Discussion of implications for language learning systems

Since the early 1990s, computer programs and on-line instructional sites have become popular tools in the language classroom. From the early years of these tools, researchers from the new field of CALL (Computer-Assisted Language Learning) have been studying the impact, effectiveness and potential of these new tools on the language learning process (Bush, 2008; Felix, 2002; Hampel & Pleines, 2013; Nagata, 1998). The development of technology in the last decades has pushed the boundaries of CALL, allowing the usage of new tools such as blogs (Miceli, Murray, & Kennedy, 2010), wikis (Mak & Coniam, 2008) and immersive virtual reality environments (O’Brien, Levy, & Orich, 2013), in the language-learning classroom.

While results regarding the effectiveness of various tools and designs have been variable, many studies have concluded that it is necessary to put more emphasis on the appropriate design of language-learning tasks, both to enhance motivation, improve ease of usage, and augment student learning (Doughty & Long, 2003; Felix, 2002; Hampel & Pleines, 2013). It is often argued that in order to be effective, the design of CALL material must be “grounded in theories of second language acquisition (SLA)” (Brandl, 2005, p. 22). Vocabulary acquisition is one of the key applications for CALL; however, after several decades of SLA research, it is now commonly agreed that the vocabulary acquisition process is not as straightforward as was previously thought, and that rote memorization exercises may not be the optimal solution (Cobb, 2007; Laufer, pp. 286–303). However, learning new vocabulary remains an important part of the language learning process (Barcroft, 2007), and second language learners often need to learn

a large quantity of words in a relatively short time, which raises the bar in terms of task design (Cobb & Horst, 2011).

We propose that instead of teaching isolated lexemes or minimally specified expressions, a CALL system should provide learners with overspecified referring expressions (REs) regarding the objects they see before them. This is based on the results presented in this article, inspired by previous research in referring expression generation, which has shown that native speakers spontaneously give too much information regarding an object when referring to it (Engelhardt et al., 2006; Pechmann, 1989; Wu & Keysar, 2007).

5. Conclusion

In this paper, we argued that overspecification in referring expressions enhances lexical acquisition by giving learners more opportunities to anchor new lexemes and exploit them when needed. We have shown, via a 3D virtual-world experiment, that overspecification during the practice phase of lexical acquisition enhances performance on subsequent testing of the lexemes learned. Our experiment consisted in comparing two groups receiving different degrees of specification (minimal and overspecified). Overspecified training exercises yielded the most successful Test Phase performance for all metrics studied. Also, our participants evaluated overspecified REs to be of equal comprehensibility to minimal REs and overspecified vocabulary practice exercises as more useful than minimal ones.

Our results support previous work regarding the utility of overspecification in reference when reference is considered as a multi-turn collaborative process between speaker and listener (Arts, Maes, Noordman, & Jansen, 2008; Lane, Groisman, & Ferreira, 2006). They are consistent with previous work regarding the higher processing time of overspecified references (Arts, Maes, Noordman, & Jansen, 2011; Engelhardt et al., 2006) and provide new evidence that overspecification has a positive impact on the acquisition of lexemes in a foreign language. This, in turn, has implications for the design of referring expression generation algorithms used by Technology-Enhanced Language Learning (TELL) and Computer Assisted Language Learning (CALL) Systems (Bush, 2008; Dickinson & Herring, 2008; Yang & Chen, 2007). If overspecification results from limited cognitive resources on the speaker’s part and causes unjustified extra effort for the listener, we could argue that CALL exercises should only have minimal semantic content and therefore be minimally specified. However, if it is true that overspecified REs play a role in enhancing lexical acquisition and serve a purpose in subsequent communication, we should consider providing learners with overspecified REs during vocabulary practice exercises in order to aid them in learning new lexemes.

Appendix A. Map of the virtual world

Fig. A.4 shows a top view map of the virtual world designed for our experiment. The virtual world was designed using the GIVE platform Koller et al. (2008) and has 6 rooms: (1) the Tutorial Room, the 3 Learning Phase rooms: (2) the Object Room, (3) the Color Room, and (4) the Spatial Relation Rooms, (5) the Test Room (where both the First and Second Test Phases took place) and the (6) the Exercise Room.²

² The tutorial room is not part of any phase of the experiment, its goal is to teach the participant how to move forward, backwards, how to rotate and how to press buttons in the virtual world.

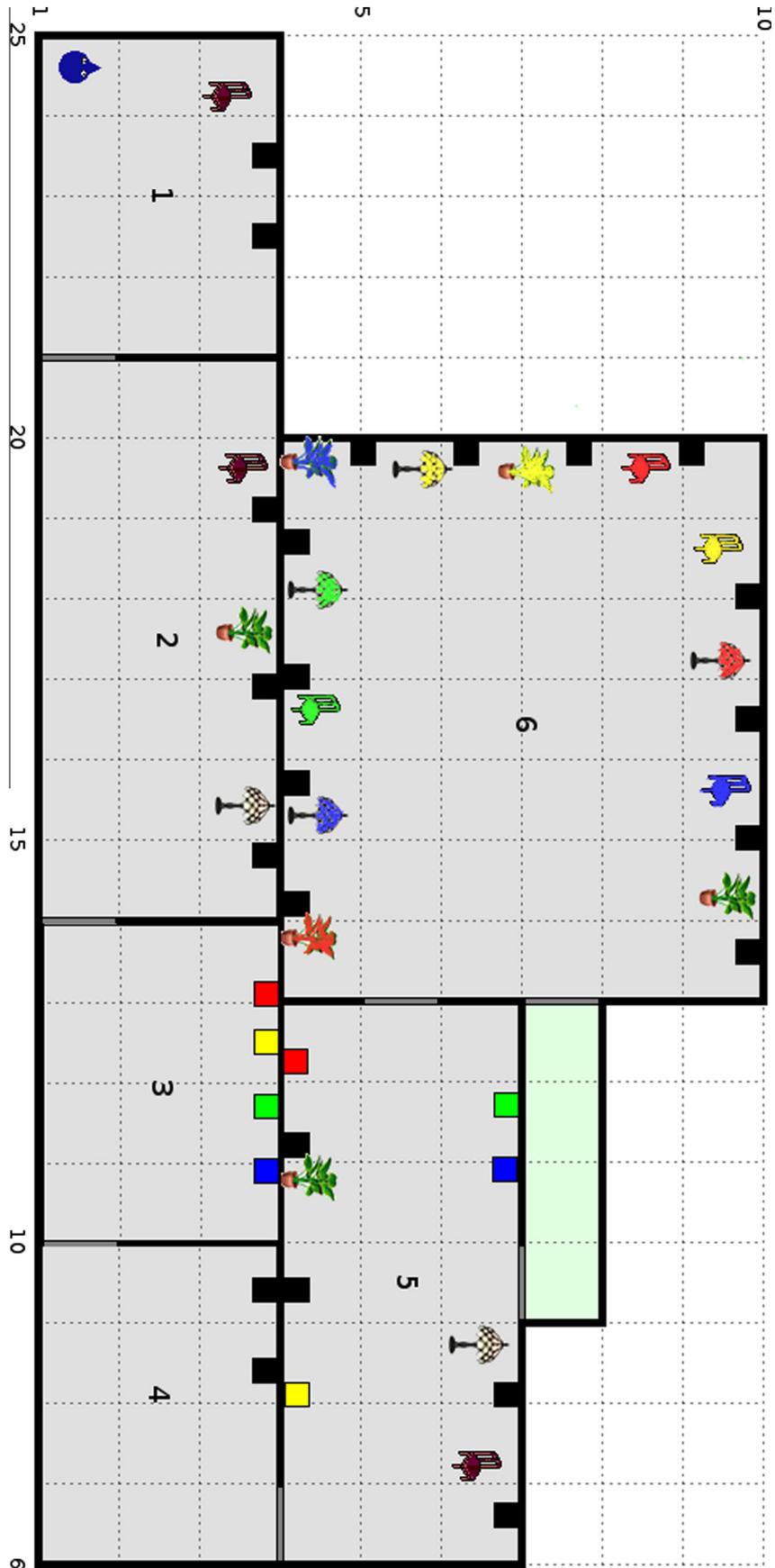


Fig. A.4. A map of the GIVE virtual world used in the experiment.

References

- Arts, A., Maes, A., Noordman, L., & Jansen, C. (2008). Overspecification in action-oriented discourse: Task importance affects the production of overspecifications and overspecifications increase identification efficiency in perception. *Discourse Processes*, 43, 25–27.
- Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, 43, 361–374.
- Atkinson, D., Churchill, E., Nishino, T., & Okada, H. (2007). Alignment and interaction in a sociocognitive approach to second language acquisition. *The Modern Language Journal*, 91, 169–188.
- Barcroft, J. (2007). When knowing grammar depends on knowing vocabulary: Native-speaker grammaticality judgements of sentences with real and unreal words. *The Canadian Modern Language Review*, 63, 313–343.
- Brandl, K. (2005). Are you ready to 'Moodle'? *Language Learning and Technology*, 9, 16–23.
- Bush, M. D. (2008). Computer-assisted language learning: From vision to reality? *CALICO Journal*, 25, 443–470.
- Clark, H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, 1–39.
- Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning and Technology*, 11, 38–64.
- Cobb, T., & Horst, M. (2011). Does word coach coach words? *CALICO Journal*, 28, 639–661.
- Dickinson, M., & Herring, J. (2008). Developing online icall exercises for russian. In *The 3rd workshop on innovative use of NLP for building educational applications (ACL08-NLP-Education)* (pp. 1–9). Columbus, OH.
- Doughty, C. J., & Long, M. H. (2003). Optimal psycholinguistic environments for distance foreign language learning. *Language Learning and Technology*, 7, 50–80.
- Engelhardt, P., Bailey, K., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean maxim of quantity? *Journal of Memory and Language*, 54, 554–573.
- Engelhardt, P. E., Demiral, B., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, 77, 304–314.
- Felix, U. (2002). The web as a vehicle for constructivist approaches in language teaching. *ReCALL*, 14, 2–15.
- Hampel, R., & Pleines, C. (2013). Fostering student interaction and engagement in a virtual learning environment: An investigation into activity design and implementation. *The CALICO Journal*, 30, 342–370.
- Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, 51, 539–558.
- Johnson, W. L. (2007). Serious use of a serious game for language learning. In *Proceedings of the 2007 conference on artificial intelligence in education: building technology rich learning contexts that work* (pp. 67–74). Amsterdam, The Netherlands: IOS Press.
- Koller, A., Striegnitz, K., Gargett, A., Byron, D., Cassell, J., Dale, R. et al. (2008). Report on the second NLG challenge on generating instructions in virtual environments (GIVE-2). In *International conference in natural language generation* (pp. 243–250). Association for Computational Linguistics.
- Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, 4, 463–470.
- Lane, L. W., Groisman, M., & Ferreira, V. S. (2006). Don't talk about pink elephants! *Psychological Science*, 17, 273–277.
- Laufer, B. (2005). Instructed second language vocabulary learning: The fault in the default hypothesis. *Investigations in Instructed Second Language Acquisition*, 286–303.
- Mak, B., & Coniam, D. (2008). Using wikis to enhance and develop writing skills among secondary school students in Hong Kong. *System*, 36, 437–455.
- Miceli, T., Murray, S. V., & Kennedy, C. (2010). Using an L2 blog to enhance learners; participation and sense of community. *Computer Assisted Language Learning*, 23, 321–341.
- Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, 13, 329–336.
- Nagata, N. (1998). Input vs. output practice in educational software for second language acquisition. *Language Learning and Technology*, 1, 23–40.
- O'Brien, M., Levy, R., & Orich, A. (2013). Virtual immersion: The role of cave and pc technology. *CALICO Journal*, 26, 337–362.
- Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27, 89–110.
- Sagarra, N., & Alba, M. (2006). The key is in the keyword: L2 vocabulary learning methods with beginning learners of spanish. *The Modern Language Journal*, 90, 228–243.
- Sørensen, B. H., & Meyer, B. (2007). Serious games in language learning and teaching – A theoretical perspective. In *Proceedings of the 3rd international conference of the digital games research association* (pp. 559–566).
- Viethen, J., & Dale, R. (2008). The use of spatial relations in referring expression generation. In *Proceedings of the fifth international natural language generation conference INLG '08* (pp. 59–67). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Viethen, J., & Dale, R. (2011). Gre3d7: A corpus of distinguishing descriptions for objects in visual scenes. In *Proceedings of the UCNLG + Eval: Language generation and evaluation workshop* (pp. 12–22). Edinburgh, Scotland: Association for Computational Linguistics.
- Wu, S., & Keysar, B. (2007). The effect of information overlap on communication effectiveness. *Cognitive Science: A Multidisciplinary Journal*, 31, 169–181.
- Yang, S. C., & Chen, Y.-J. (2007). Technology-enhanced language learning: A case study. *Computers in Human Behavior*, 23, 860–879.
- Zhao, Y. (2003). Recent developments in technology and language learning: A literature review and meta-analysis. *CALICO Journal*, 21, 7–27.