Supplementary Materials of Semantic Network Development in L2 Spanish and Its Impact on Processing Skills: A Multisession Eye-Tracking Study

M. Gabriela Puscama

Supplementary S1: Description of Category Fluency Task

This task is adapted from Baus et al. (2013) and Linck et al. (2009). During this task, participants were asked to name in thirty seconds as many items as possible within a given category. The task was presented on a YouTube video embedded in a Google site and oral responses were recorded using the participants' computer microphone. The task included eight categories, which were divided in two groups and counterbalanced between the languages. One group of categories included *clothing*, *animals*, *vegetables*, and *musical instruments*, and the other group included *colors*, *fruits*, *body parts*, *and furniture*. Each block was preceded by a practice category, which was not analyzed. To obtain an overall fluency score per participant per language, for each new item mentioned within a given category, participants received a point, and then the points for all categories in each language were averaged. If a participant did not mention any items for a given category, that category was not included in the average (e.g., the total number of items was divided by 3 instead of 4 to obtain the score). This decision was made because on numerous occasions participants commented during debriefing that they were not familiar with the category name, for example *muebles* 'furniture', but when they heard the translation of the category in English, they did know items within that category (e.g., *silla* 'chair', *cama* 'bed'). In these cases, participants would have received an artificially low proficiency score, so it is best to exclude such categories where no exemplars were mentioned, as a general rule. Paired-samples t-tests of participants' scores in each language revealed that the L2 learners produced significantly more items in English (M=12.03, SD= 1.40) than in Spanish (M=6.10, SD= 1.35) t(38) 22.01, $p <$ 0.001. This indicates that the L2 learners were more dominant in their L1.

The Category Fluency and the Digit Span scores (Supplementary S2) were used to match participants when dividing them into training groups (thematic lists and visual scenes), to ensure that differences found across experimental groups were not due to proficiency or working memory. A Welch two-sample t-test of participants' Spanish Category Fluency showed that the two groups did not differ significantly in Spanish proficiency (M=6.04, SD=1.31 and M=6.15, SD=1.41), t(36.98)= -0.25, p=0.8).

Supplementary S2: Description of Digit Span Task

During this task, participants saw a sequence of digits on the screen, and they were asked to remember and recall each sequence in the same order. The task was presented on a YouTube video embedded in the Google site, and oral responses were recorded using the participants' computer microphone. The sequences increased progressively from 2 to 9 digits, and there were two sequences of each length, for a total of 16 trials. Participants received 1 point for each sequence recalled correctly, with the maximum possible score being 16. Participants did not receive any points after having failed to recall both sequences of the same length. The average score for the Digit Span task was 9.62 (SD=1.70).

The Category Fluency (Supplementary S1) and the Digit Span scores were used to match participants when dividing them into training groups (thematic lists and visual scenes), to ensure that differences found across experimental groups were not due to proficiency or working memory. A Welch two-sample t-test of participants' Digit Span scores showed that the two groups did not differ significantly in working memory (M=9.53, SD=1.71 and M=9.70, SD=1.72), t(36.91)= -0.32, p=0.8).

Supplementary S3: Process for calculating semantic relatedness scores within each visual world display

For the visual world eye-tracking task, semantic relatedness scores were calculated for each word pair within the displays, using a model designed to train Lightweight Metrics of Semantic Similarity (LMOSS) (Recchia and Jones 2009). The LMOSS was created as an alternative to other semantic relatedness measures, such as latent semantic analysis. According to the authors, it provides a flexible and scalable way of measuring how close two words are in meaning, based on their frequency of co-occurrence in a corpus (Recchia and Jones 2009). Semantic relatedness is calculated by dividing the number of co-occurrences of two words, by the product of the frequency of each word in the corpus. This yields a simplified point mutual information score (Simplified PMI):

$$Simplified\ PMI = \frac{Co - occurrence(word1, word2)}{Frequency(word1) \times Frequency\ (word2)}$$

For the Spanish tokens, the model was trained using the lemmatized version of the *Corpus del Español* (Davies 2016). The lemmatized version was prefered over the original text, because it is beyond the scope of this paper to argue for a decompositional or holistic storage and access of vocabulary items. The main question is whether two concepts can be interrelated in an L2 learner's mind (e.g., *correr* 'to run' and *deporte* 'sport'), not whether different inflections of the same words have different level of relatedness to another concept. Therefore, using lemmas for the analysis allows us to paint a broader picture of semantic relations in the lexicon, for the purposes of this paper.

The lemmatized version of the *Corpus del Español* (Davies 2016) is originally organized into a table, which includes a lemma column and additional columns for information, such as text identifier and part of speech. There are several texts for each of different Hispanic countries, and the text identifier column shows this information. Additionally, the corpus is divided into separate files, each corresponding to different Spanish-speaking countries, and each containing several texts. Since the LMOSS requires the corpus to be in a single file, and with one line for each text in the corpus, the following steps were followed to reformat the documents and make them readable for LMOSS.

First, the individual files were combined, so that there was one text file for each country. For this step, the Unix shell Git Bash for Windows (https://git-scm.com/) was used. Then, using R (R Core Team 2013), for each country file the lemma and the text identifier column were used to wrangle the data and create a file with a single column, each row containing one text of the corpus (i.e., each row contained a "text" from the given Hispanic country). As part of the wrangling, the rows that had symbols instead of words were eliminated. Also, the spelling of the lemmas were modified as following. All diacritics were eliminated, because LMOSS does not support them. Then, to avoid confusion between minimal pairs where the diacritic is the only reliable cue for differentiation (e.g., *cana-caña*, 'white hair'-'cane'), the grapheme that originally had the diacritic was duplicated (in the previous example, *cana-canna*). To illustrate, a noun like *avión* 'airplane' became *avioon*. Furthermore, words with alternative spelling were standardized by choosing one spelling, e.g., *bizcocho/biscocho* 'sponge cake, biscuit' were transcribed as *bizcocho*. Finally, the individual files for each country were combined into a single file using Git Bash (https://git-scm.com/), to train the model.

With the newly formatted corpus, LMOSS was trained using the Spanish prompts and responses from a word association task (Supplementary S4). The model gives the option of calculating regular point mutual information (PMI) scores, or PMI-Order, which takes into consideration word order (i.e., how many times word1 occurs before word2, or vice-versa). For the purpose of this article, word order was not relevant, so PMI scores between word pairs were calculated without considering their order of appearance. The co-occurrence window size selected was 100 words, based on several pilot runs where window size was progressively increased and decreased. A window size smaller than 100 words yielded mostly null results, as words rarely co-occurred so closely. Inversely, a larger window size provided inflated co-

occurrence scores between words that did not have a strong semantic association. Additionally, 100 words is the approximate length of a paragraph, which is a substantial unit of meaning in which words that co-occur are probably closely related.

Supplementary S4: Procedure for generating stimuli through a word association task

Materials

Spanish words were obtained from the beginner and intermediate Spanish textbooks *Mosaicos 2 and 3* (Olivella de Castells et al. 2013, 2015). To choose the words, I selected three prototypical situations familiar to U.S. undergraduate students, and three thematic lists pertaining to each situation, totaling nine thematic lists:

- Moving to campus: **furniture** (e.g., *cama* 'bed'), **house chores** (e.g., *planchar* 'to iron'), and **shopping** (e.g., *tienda* 'store');
- Spring break: **clothing** (e.g., *vestido* 'dress'), **traveling** (e.g., *avión* 'airplane'), and **weather** (e.g., *lluvia* 'rain');
- Tailgating: **food** (e.g., *postre* 'dessert'), **sports** (e.g., *entrenador* 'coach'), and **health** (e.g., *dolor* 'pain').

These three contexts were selected to render the task more meaningful and relevant to the participants with the goal of increasing their interest in the topic and their motivation for learning the vocabulary. Additionally, the themes within each situation were selected based on the thematic lists in Spanish textbooks.

Finally, six words were selected from each of these nine thematic lists, adding up to a total of 54 Spanish words and their English translation equivalents, that is, 108 words. The rationale for selecting the situations a priori, instead of creating a random list of words from different themes, was to generate cohesive stimuli for the main visual world eye-tracking experiment.

Procedure

250 native Spanish speakers were recruited through the crowdsourcing platform *Prolific* (Prolific 2014), and the data was collected in three waves (100 for Group 1, 100 for Group 2 and 50 for Group 3). Demographic and linguistic information was obtained through a brief language history questionnaire (LHQ) administered in *Qualtrics* (Qualtrics 2005), which was completed immediately after the main task, on the same survey. From the 250 native Spanish speakers, 3 were excluded because they answered "No" to the question "Where you born and raised speaking Spanish at home?", 5 were excluded because they accidentally completed the survey twice (only the first iteration was retained), and 1 was excluded because they did not complete the survey, leaving a total of 241 participants. The final number of participants was 94 for Group 1, 98 for Group 2 and 49 for Group 3. Each group completed a round of word associations, as detailed below.

The participants were L1 Spanish-L2 English bilinguals, ages 18-35 (M=24.29 SD=4.49); 138 participants self-identified as female, 101 as male, and 2 as non-binary gender. At the time of the study, all of them resided in the United States. 65.56% of the participants reportedly had studied English in a formal setting. Additionally, 60.17% of the native Spanish speakers indicated that at least one of their caregivers spoke English[1].

[1] Among the native Spanish speakers recruited through crowdsourcing, there may have been some heritage speakers as well. Based on the questionnaire they completed, we do not have information to identify and exclude heritage speakers. Therefore, this group could be made up of participants who were raised in Spanish-speaking countries and moved to the United States as adults, and others who were raised with at least 1 Spanish-speaking caregiver in the U.S.

The procedure was similar to the snowball technique employed by De Deyne and Storms (2008) and De Deyne et al. (2013). During the first round, each participant completed one English block and one Spanish block, starting with their L1. The task was administered as a *Qualtrics* survey (Qualtrics 2005), and distributed through *Prolific* (Prolific 2014) (native speakers).

The 54 Spanish words and their English translations were divided equally into two lists. The lists were counterbalanced across participants, so that if participant 1 saw *cama* 'bed' in Spanish and *rain* in English, then participant 2 saw *lluvia* 'rain' in Spanish and *bed* in English. This counterbalancing was to avoid practice effects, which may have occurred if participants saw the same words in both blocks, in different languages.

The Spanish and English words were presented as a prompt one at a time to the first 100 Spanish-English bilinguals. Participants were instructed to provide three associated words to each prompt in the same language (e.g., for the Spanish block, they provided Spanish associates). In this first stage, each participant responded to 54 prompts (27 in each language block). An attention check was included in the break between the two blocks, to make sure that participants were engaged in the task, especially for the *Prolific* (Prolific 2014) data collection. The format of the attention check question followed the format of each trial of word association, and it asked participants to type the word *banana* in each of the three text boxes provided.

From the first stage of associations, the most frequent response to each prompt was extracted to create a second list of 108 prompts (54 in each language). Words that were already in the original stimuli list (Stage 1) were excluded, leaving 83 novel responses (41 English words and 42 Spanish words). Finally, some of the stimuli were repeated (e.g., the most frequent associate for both *partido* 'game' and *pelota* 'ball' was *deporte* 'sport'); to avoid redundancies, only a single repetition was included on the new list. Therefore, the final number of prompts for the second stage of associations was 59 words (28 Spanish words and 31 English words).

Following the procedure from Stage 1, the new stimuli were divided into two lists and counterbalanced across language. In some cases, there were items in English and Spanish that were translation equivalents (e.g., *volar* 'to fly'), so these were separated into different lists, to avoid practice effects. The procedure was the same as in Stage 1. The second stage was completed by 100 native Spanish speakers, who had not participated in Stage 1.

Finally, the same cleaning procedure and final round of associations was completed for Stage 3 with 50 L1 Spanish-L2 English bilinguals who had not been part of Stages 1 or 2. From the second stage of associations, the most frequent response to each prompt was extracted to create a third list of 59 prompts (28 Spanish words and 31 English words from Stage 2). Words that were already in the first two stimuli lists were excluded, leaving a total of 29 novel responses (12 Spanish words and 17 English words). On two of the Spanish prompts, there were two responses with equal highest frequency. For the prompt *carro* 'car', the two highest frequency responses were *coche* 'car' and *gasoline* 'gas', and for the prompt *soda* 'soda', the two highest frequency responses were *bebida* 'drink' and *refresco* 'soda'. In the case of the *carro* responses, both responses were included in the final list for round 3, as they were both equally frequent and novel. In the case of *soda*, however, only *bebida* was included in the final list, as *refresco* was not a novel response (it belonged to the list of Stage 1).

In the last stage, there was only one list of stimuli, and the task was completed by 50 participants. Unlike Stages 1 and 2, there was no need to create two counterbalanced lists. There were no English-Spanish translation equivalents in this last round, hence the risk of a practice effect was non-existent. The divergence of responses across the two languages of the same group may be an indicator of the language-specific nature of semantic networks, reflected in the word association data.

Analysis

Each token in the word association data was lemmatized, converted to lower case, and all diacritics were eliminated. To avoid confusion between minimal pairs where the diacritic is the only reliable cue for differentiation (e.g., *cana-caña*, 'white hair'-'cane'), the grapheme that originally had the diacritic was

duplicated (in the previous example, *cana-canna*). To illustrate, a noun like *sueños* 'dreams' became *suenno* 'dream', and an adjective like *botánicas* 'botanical$_{\text{FEM-PL}}$', became *botaanico* 'botanical$_{\text{MASC-SING}}$'. Furthermore, words with alternative spelling were standardized by choosing one spelling, e.g., *bizcocho/biscocho* 'sponge cake, biscuit' were transcribed as *bizcocho*. Finally, following the word association literature (e.g., Dubossarsky et al. 2017, De Deyne et al. 2019), for responses that included multiple words (e.g., *a la plancha* 'grilled') prepositions and determiners were removed, keeping only the main word in the phrase, which was typically a noun (in the example, *plancha*). In other cases, such as noun + modifier combinations, we evaluated whether it was a fixed expression with a specific referent (e.g., *aire acondicionado* 'air conditioner') or merely a noun with a modifier (e.g., *fresas con crema* 'strawberries with cream'). For fixed expressions, both words were kept, and for nouns with modifiers, only the noun was kept.

Supplementary S5: Output of additional linear mixed-effects models for visual world data

Results of LME on pre-test (Session 1) for the pause region, for L2 learners only with median-centered proficiency and condition as predictors.

Reference value: related condition, median proficiency

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.01	0.03	0.49	n.s.
Condition	0.10	0.02	4.16	<0.001
L2 Proficiency	-0.00	0.02	0.02	n.s.

Results of LME on pre-test (Session 1) for the target region, for L2 learners only with median-centered proficiency and condition as predictors.

Reference value: related condition, median proficiency

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.16	0.03	4.71	<0.001
Condition	0.09	0.03	3.19	<0.01
L2 Proficiency	-0.02	0.02	-1.34	n.s.

Results of LME on pre-test (Session 1) for the pause region, comparing native speakers and L2 learners, with potential heritage speakers excluded.

Reference value: related condition, L1 English

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.02	0.03	0.82	n.s.
Condition	0.06	0.02	2.49	<0.05
L1	-0.03	0.05	-0.71	n.s.
Condition : L1	0.20	0.05	4.39	<0.001

Results of LME on pre-test (Session 1) for the target region, comparing native speakers and L2 learners, with potential heritage speakers excluded.

Reference value: related condition, L1 English

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.15	0.03	4.22	<0.001
Condition	0.09	0.03	3.34	<0.01
L1	0.11	0.05	1.96	n.s.
Condition : L1	0.12	0.05	2.41	<0.05

References

(Baus et al. 2013) Baus, Cristina, Albert Costa, and Manuel Carreiras. 2013. On the effects of second language immersion on first language production. *Acta Psychologica* 142(3): 402-409. https://doi.org/10.1016/j.actpsy.2013.01.010.

(Davies 2016) Davies, Mark. 2016-. Corpus del Español: Two billion words, 21 countries. Available online: http://www.corpusdelespanol.org/web-dial/ (accessed on 4 August 2021).

(De Deyne et al. 2019) De Deyne, Simon, Daniel J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" English word association norms for over 12,000 cue words. *Behavior Research Methods* 51: 987-1006. https://doi.org/10.3758/s13428-018-1115-7.

(De Deyne et al. 2013) De Deyne, Simon, Daniel J. Navarro, and Gert Storms. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods* 45: 480-498. https://doi.org/10.3758/s13428-012-0260-7.

(De Deyne et al. 2008) De Deyne, Simon, and Gert Storms. 2008. Word associations: Network and semantic properties. *Behavior Research Methods* 40(1): 213-231. https://doi.org/10.3758/BRM.40.1.213.

(Dubossarsky et al. 2017) Dubossarsky, Haim, Simon De Deyne, and Thomas T. Hills. 2017. Quantifying the structure of free association networks across the life span. *Developmental Psychology* 53(8): 1560-1570. https://doi.org/10.1037/dev0000347.

(Linck et al. 2009) Linck, Jared A., Judith F. Kroll, and Gretchen Sunderman. 2009. Losing access to the native language while immersed in a second language: Evidence for the role of inhibition in second-language learning. *Psychological Science* 20(12): 1507-1515. https://doi.org/10.1111/j.1467-9280.2009.02480.x.

(Olivella de Castells et al. 2013) Olivella de Castells, Matilde, Elizabeth E. Guzmán, Paloma Lapuerta, P., and Judith E. Liskin-Gasparro. 2013. *Mosaicos: Course Materials for Spanish 2* (4th Custom ed.). Boston, MA: Pearson.

(Olivella de Castells et al. 2015) Olivella de Castells, Matilde, Elizabeth E. Guzmán, Paloma Lapuerta, P., and Judith E. Liskin-Gasparro. 2015. *Mosaicos: Course Materials for Spanish 3* (5th Custom ed.). Boston, MA: Pearson.

(Prolific 2014) Prolific. (2014). Available online: https://www.prolific.co (accessed on 18 November 2021).

(Qualtrics 2005) Qualtrics. 2005. Provo, UT. Available online: https://www.qualtrics.com (accessed on 18 November 2021).

(R Core Team 2013) R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online: http://www.R-project.org/ (accessed on 15 November 2023).

(Recchia and Jones 2009) Recchia, Gabriel, and Michael N. Jones. 2009. More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods* 41(3): 647-656. https://doi.org/10.3758/BRM.41.3.647.