Semantic Ambiguity Effects: A Matter of Time?

Joyse Medeiros (j.medeiros@bcbl.eu)
BCBL. Basque Center on Cognition, Brain, and Language
UPV. University of the Basque Country

Blair C. Armstrong (blair.armstrong@utoronto.ca)
Department of Psychology & Centre for French & Linguistics at Scarborough, University of Toronto
BCBL. Basque Center on Cognition, Brain, and Language

Abstract

Are different amounts of semantic processing associated with different semantic ambiguity effects? Could this explain some discrepant ambiguity effects observed between and across tasks? Armstrong and Plaut (2016) provided an initial set of neural network simulations indicating this is indeed the case. However, their empirical findings using a lexical decision task were not clear-cut. Here, we use improved methods and five different experimental manipulations to slow responding---and the presumed amount of semantic processing---to evaluate their account more rigorously. We also expanded the empirical horizon to another language: Spanish. The results are partially consistent with the predictions of the neural network and differ in several important ways from English data. Potential causes of these discrepancies are discussed in relation to theories of ambiguity resolution and cross-linguistic differences.

Keywords: semantic ambiguity; slow vs. fast lexical decision; semantic settling dynamics, neural networks.

Understanding how the meaning of ambiguous words is resolved is critical because the meaning of most words depends on context (e.g., *cricket* can refer either to a game or to an insect). Developing an account of ambiguity resolution has, however, been challenged by two complications: 1) the complex and often apparently contradictory effects of ambiguity observed between and sometimes even within a given experimental task, discussed below, and 2) the often inconsistent effects observed for polysemes with related senses (e.g., *chicken* can refer to an animal or its meat) vs. homonyms with unrelated meanings (e.g., *cricket*) compared to (relatively) unambiguous control words (e.g., *chalk*).

Recently, Armstrong and Plaut (2016) reported neural network simulations suggesting that many apparently inconsistent effects can be reconciled as a function of (a) how the number and relatedness of a word's meanings are activated over time, (b) the amount of processing that takes place before a response can be generated in a given task (see Figure 1). This *semantic settling dynamics* (SSD) account posits that early processing is dominated by excitatory/cooperative neural dynamics that would facilitate the processing of polysemes. In contrast, later processing would be dominated by inhibitory/competitive neural dynamics that would impair the processing of homonyms. Thus, "fast" tasks like typical lexical decision, in which participants must decide whether a letter string forms a word (e.g., *cricket*) or not (e.g., *blicket*), would show a polysemy advantage (e.g., e.g., Armstrong & Plaut, 2016; Beretta, Fiorentino, & Poeppel, 2005; Rodd, Gaskell, & Marslen-Wilson, 2002). In contrast, "slow" tasks like typical semantic categorization, in which participants must determine whether a word refers to a member of a particular category (e.g., does *cricket* refer to a *vegetable?*), would show a homonymy disadvantage (e.g., Hino, Pexman, & Lupker., 2006).

The SSD account offers both a contrasting and a complementary explanation to an account positing that different ambiguity effects are due to task-specific configurations of the decision system (Hino et al., 2006). In contrast to the decision system account, the SSD hypothesis stresses how dynamics within semantics can critically shape the ambiguity effects observed in a given task. The decision system should, however, play an important role in determining "when" sufficient evidence has accumulated to generate a response---and thus which portion of semantics is being tapped (for a broader discussion, see Armstrong & Plaut, 2016).

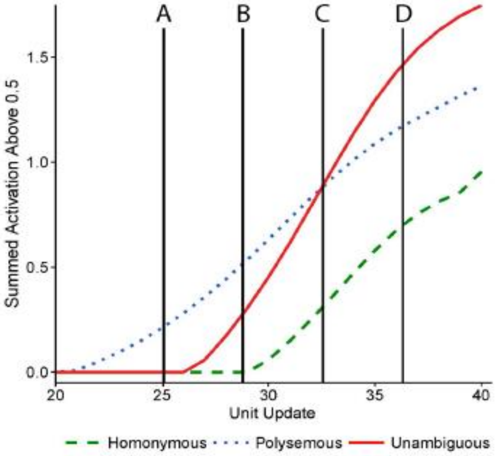

Figure 1. Semantic activity as a function of processing time for homonyms, polysemes, and unambiguous controls in the neural network simulation reported by Armstrong and Plaut (2016). Slices A-D highlight how sampling these trajectories at different time points aligns with different behavioural and neural effects reported in the literature, such as typical lexical decision (Slice A) and semantic categorization (Slice C).

Armstrong and Plaut (2016) also put the SSD account to the test in an empirical setting. In their experiments, the overall task (lexical decision) was held constant. They then manipulated additional properties of the task to slow responses (manipulations of nonword difficulty and/or the brightness of the letters on the screen). Insofar as these slow-downs enabled additional semantic processing to take place, the SSD account predicts this would lead to a shift from a polysemy advantage in the easy/fast conditions (Figure 1, Slice A) and a homonymy disadvantage in the slow/hard conditions (Figure 1, Slice C).

The results were generally---although not perfectly---consistent with these predictions. A polysemy advantage was typically observed in the easy/fast condition, but evidence for this advantage in the harder conditions was more limited. Similarly, there was evidence that a homonymy disadvantage was present in some (but not all) of the hard/slow conditions, but, critically, not in the easy/fast conditions. One possible interpretation of these results is that they are attributable to a slight increase in semantic processing and thus reflect only a small step along the predicted semantic settling dynamics (e.g., Figure 1, Slice A to Slice B, rather than Slice A to Slice C). Additional investigations are needed, however, to better explore this possibility and the validity of the SSD account more broadly.

The present work is a major extension of Armstrong and Plaut's (2016) initial empirical studies. From a theoretical perspective, it follows the abductive reasoning: if a range of different manipulations designed to slow responding all yield the same changes in ambiguity effects, this will provide broad convergent support for the SSD account. Our work also builds upon past work in several important ways: First, for all but one condition, it uses within-participant manipulations to boost statistical power. Second, the experiments were run in Spanish, a language in which it is easier to control for several potential confounding variables (e.g., with few exceptions, each Spanish letter maps to a single sound and vice versa, so matching word lengths in number of letters also matches word lengths in number of phonemes). Doing so also allows for the evaluation of the robustness of particular ambiguity effects and facilitates the development of general as opposed to Anglocentric theories (Share, 2008). Further, recent Spanish homonym meaning frequency norms (Armstrong et al., 2015) allow us to select homonyms with balanced meaning frequencies. This should boost the competitive dynamics assumed to be associated with homonyms during late processing.

Behavioral Studies of Lexical Decision

We evaluated whether slowing participants' lexical decision responses using several different manipulations reproduced the different semantic ambiguity effects predicted by the SSD account. If these different manipulations produce the anticipated effects, this would support the notion that the time-point at which the response was made---and the corresponding amount of semantic settling---is a critical component of any theory of semantic ambiguity resolution. (Without denying that these dynamics interact and are further shaped by other systems; e.g., the response system.) If the results do not produce the predicted effects, this would support claims that qualitative differences in the configuration of the response system, as opposed to semantic settling dynamics, explain many discrepant ambiguity effects.

We applied the following manipulations to a standard visual and/or auditory lexical decision task, which we describe in detail subsequently. The first two manipulations relate closely to those in Armstrong & Plaut (2016) for comparison purposes, whereas the remaining three have never been used in studies of semantic ambiguity.

1. Visual Lexical Decision: Nonword Wordlikeness: "Easy" nonwords with lower bigram frequencies and higher Orthographic Levenshtein distances (OLD; Yarkoni, Balota, & Yap, 2008) than the word stimuli were used in the *baseline*; "Hard" nonwords with higher bigram frequencies and lower OLDs than the words were used in the *slowed* condition. This was the only between-participant manipulation because previous experiments have found carry-over effects when nonword difficulty is blocked within participants (Armstrong, 2012). All other manipulations were within participants and used easy nonwords to avoid potential ceiling effects on how slow lexical decision can be pushed.
2. Visual Lexical Decision: Visual Noise: Standard text was presented in the *baseline*; visual noise (950 3px dots) was superimposed to degrade the text in the *slowed* condition. This condition is similar to the contrast reduction manipulation in Armstrong & Plaut (2016).
3. Intermodal Lexical Decision: Visual lexical decision served as the *baseline*, auditory lexical decision as the *slowed* condition. This experiment was motivated by different ambiguity effects observed in audio vs. visual lexical decision in Rodd et al. (2002).
4. Auditory Lexical Decision: Auditory Noise: Clear sound recordings were presented in the *baseline*; noisy recordings---created by replacing 75% of the auditory signal with signal-correlated noise---were used in the *slowed* condition.
5. Auditory Lexical Decision: Compression/Expansion: Recordings were played 30% faster in the *baseline* and 30% slower in the *slowed* condition. The "similarity" time effect in Goldwave ® (v6.13) was used to preserve pitch and the naturalness of the vocalization.

Participants. Each experiment was completed by 42 Spanish native speakers (avg. age = 24 years, 70% female). All had normal or corrected-to-normal vision and no history of language or psychological disorders. Participants received a monetary payment. Consent was obtained in accordance with the declaration of Helsinki.

Stimuli. *Words.* The stimuli filled a 2x2 factorial design that crossed number of unrelated meanings (NoM: one vs. two) with number of related senses (NoS: few [range: 1-5] vs. many [range: 6-14]), similar to past work (Rodd et al., 2002; Armstrong & Plaut, 2016). NoM and NoS were based on the number of separate entries vs. sub-entries for each word in the Real Academia Española Spanish dictionary (RAE, 2014). For convenience, we will refer to the four conditions as (relatively) unambiguous words (NoM: 1, NoS: few), homonyms (NoM: 2, NoS: few), polysemes (NoM: 1, NoS: many) and hybrids (NoM: 2, NoS: many).

To maximize the potential for competition between the interpretations of words with two unrelated meanings, we only included homonyms and hybrids with dominant relative meaning frequencies below 0.82 in the Spanish eDom norms (Armstrong et al., 2015). Using the EsPal Spanish word database (Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013), the candidate items were also constrained to have no homophones, be between 4 and 10 letters long, have word frequencies between 0.1 and 50, and have only noun or verb meanings (all had at least one noun meaning). This database also provided information regarding the word's summed bigram frequency, length in phonemes, and length in syllables.

The SOS stimulus optimization software (Armstrong, Watson & Plaut, 2012) identified 36 items in each cell of the design that were also matched on a range of psycholinguistic covariates (see Table 1). Finally, we collected separate norms for the imageability and familiarity of the words from two groups of 25 native speakers who did not participate in the main experiments.

Nonwords. Candidate nonwords were generated for each of ~80,000 words sampled from Espal (Duchon et al, 2013) to match the psycholinguistic properties of the experimental words, except for NoM and NoS. Nonwords were generated via the Wuggy nonword generator using the default settings (Keuleers & Brysbaert, 2010). In total, 144 "easy" nonwords were sampled to have lower bigram frequency and higher OLD than the words, whereas 144 "hard" nonwords were selected to have a higher bigram frequency and lower OLD than the words.

Table 1. Properties of the Word Stimuli

	Unambig.	Polyseme	Homonym	Hybrid
Example	tractor	vaina	pinta	pipa
# Meanings	1	1	2.1	2.4
# Senses	3.2	9.8	3.3	9.0
Word Freq.	5.3	5.5	5.0	6.3
OLD	1.9	1.8	1.8	1.5
# Letters	6.6	6.5	6.7	6.0
# Phonemes	6.6	6.3	6.6	5.9
# Syllables	2.8	2.8	2.9	2.6
Familiarity	4.2	4.7	4.0	4.6
Imageability	4.3	5.1	4.5	4.9
Dom. Freq.	-	-	0.5	0.5

Note. Dom. Freq. = Relative Frequency of dominant meaning.

Table 2. Properties of the Word and Nonword Stimuli

	Words	Easy Nonwords	Hard Nonwords
Bigram Freq.	1602	445	2782
OLD	2.0	2.9	1.5

Audio Recordings. Audio recordings were produced by a male native speaker. Volume was normalized to half the dynamic range. Auditory stimuli were pre-processed using Audacity (Mazzoni, 2013).

Procedure. The experiments were run on a desktop computer with a CRT monitor using Psychopy (Peirce, 2007). Auditory stimuli were presented over headphones.

Each experiment began with 4 practice trials. Participants then completed four blocks of 72 experimental trials, each of which began with 4 unanalyzed warm-up trials. An equal number of words from each cell of the design were presented in each block. The order of the stimuli was pseudorandom, with the constraint that no more than three words or nonwords could be presented in a row.

Each trial began with blank screen for 250ms, followed by a fixation cross (+) for 750ms, which was briefly replaced by a blank screen again for 50ms before the presentation of the word or nonword. In the visual conditions, text was presented in the center of the screen. In the auditory conditions, the recording was played, instead. Response latency was measured from stimulus onset, and the next trial began automatically after a response. A message was displayed if no response was made within 2500ms. Participants responded by pressing the left and right control keys with their right and left index fingers. Word responses were always made with the dominant hand. The experiment took about 20 minutes to complete.

Results

Data screening. Participants and items were screened for outliers using the Mahalanobis Distance Statistic and a critical p-value of .001. This eliminated no more than two participants in each experiment and no more than two words of any type. Trials with latencies < 200 ms or > 2000 ms were also discarded (0.66% of trials).

Analytical approach. The analyses reported here focused on the critical effects of homonymy and polysemy relative to unambiguous controls, as well as how these variables were affected by the slowing manipulations. We also report exploratory analyses of the hybrids, which should be affected by both cooperative and competitive dynamics.

All of the word data were analyzed with linear mixed-effect models (Bates, Maechler, Bolker & Walker, 2015) using R (R Core Team, 2016). The models included the key fixed effects of manipulation (with the faster/easier condition used as the baseline) and item type (with separate contrasts between an unambiguous baseline and homonyms, polysemes, and hybrids). To address potential confounds, the models also included fixed effects of imageability, residual familiarity[1], log-transformed word frequency, OLD, length in letters, and bigram frequency. All of the aforementioned fixed effects were allowed to interact with the effect of manipulation. Further, to reduce auto-correlation effects from the previous trials (Baayen, & Milin, 2010), the models included fixed effects of stimulus type repetition, previous trial accuracy, previous trial lexicality, previous trial latency, and trial rank. All continuous variables were centered and normalized. Additionally, the models included random intercepts for

---

[1] Residual familiarity was derived by regressing out NoM, NoS, and NoM vs. NoS from raw familiarity.

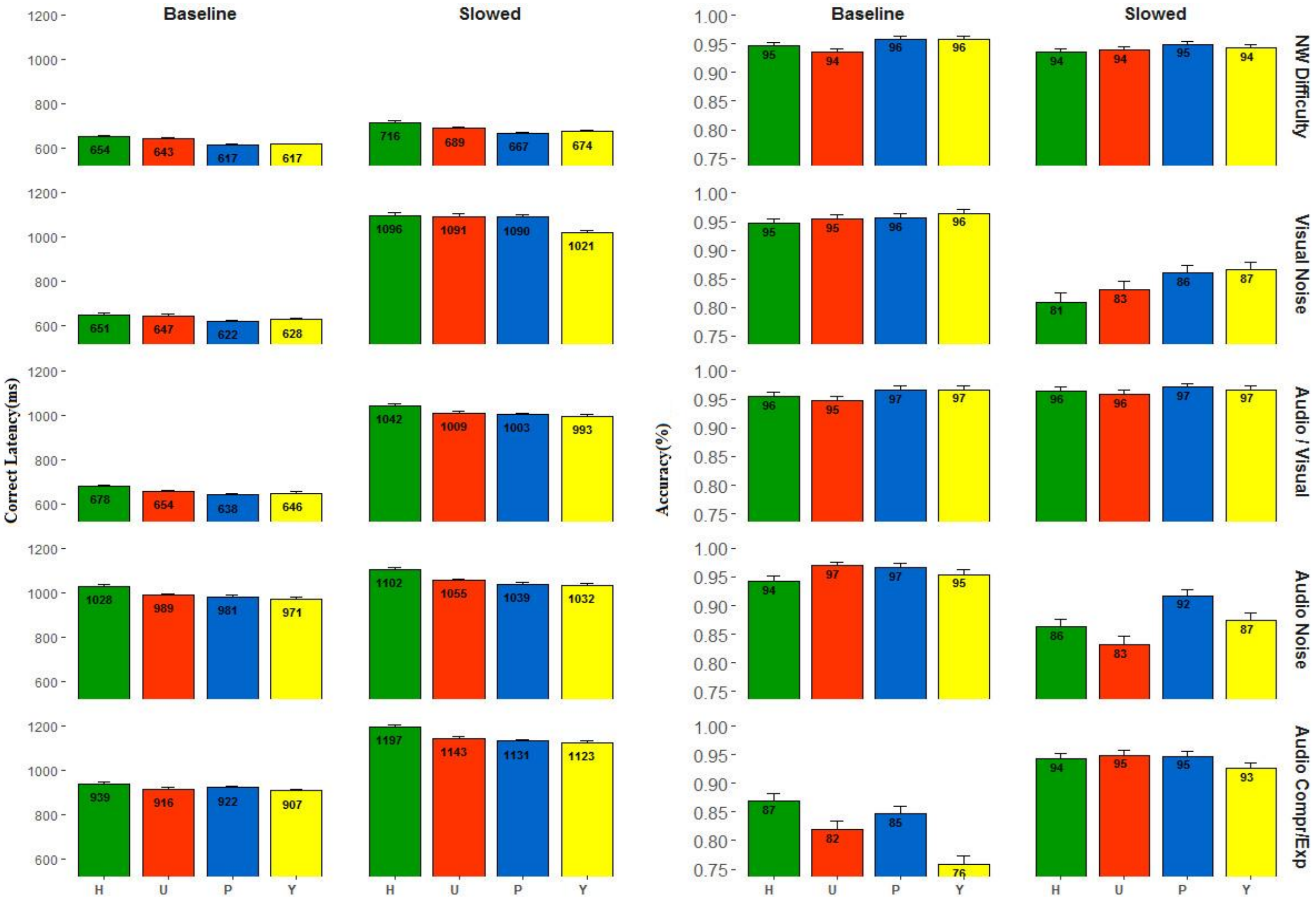

Figure 2. Correct latency [left] and accuracy [right] for the experiments. H=homonym, U=unambiguous, P=Polysemous, Y=Hybrid. Error bars = SEM.

item and participant. Random slopes were omitted because these models did not always converge. Latency was modeled with a Gaussian distribution, whereas accuracy was modeled with a binomial distribution. Effects were considered significant if $p \leq .05$, and trends are considered marginal if $p \leq .15$. All tests were two-tailed.

Correct Latency. The latency data are presented in the left panel of Figure 2. *Slowing Manipulations.* All five manipulations slowed overall response speed (all ps $\leq .02$). *Homonyms.* A main effect indicating a homonymy disadvantage was observed in the intermodal and auditory noise manipulations ($b = 24.0$, SE = 10.3, $t = 2.4$, $p = .02$ and $b = 34.0$, SE = 15.0, $t = 2.3$, $p = .03$, respectively). The homonymy by slowing manipulation interaction produced a significant increase in the homonymy disadvantage in the slower condition of the auditory compression/expansion experiment ($b = 29.1$, SE = 13.3, $t = 2.2$, $p = .03$). A similar marginal trend was observed in the nonword wordlikeness experiment ($b = 13.2$, SE = 8.2, $t = 1.6$, $p = .11$). *Polysemes.* A main effect indicating a polysemy advantage was only detected in the baseline condition of the nonword wordlikeness manipulation ($b = -19.8$, SE = 9.4, $t = -2.1$, $p = .04$). The polysemy by slowing manipulation interaction indicated that the polysemy advantage marginally decreased in the visual noise experiment ($b = 30.7$, SE = 16.3, $t = 1.9$, $p = .06$). *Hybrids.* There were no significant effects involving hybrids in any experiment. *Imageability*. There was always a marginal or significant facilitatory main effect of imageability (all ps $\leq .06$). The imageability by slowing manipulation interaction indicated this effect increased marginally in the slowed conditions of the intermodal ($b = -6.1$, SE = 4.0, $t = -1.5$, $p = .12$), visual noise ($b = -9.3$, SE = 5.9, $t = -1.6$, $p = .12$), and audio compression/expansion experiments ($b = -9.2$, SE = 4.9, $t = -1.9$, $p = .06$).

Accuracy. The accuracy data are presented in the right panel of Figure 2. *Slowing Manipulations.* The slowing manipulation decreased overall accuracy in the visual noise condition ($b = -2.1$, SE = 0.3, $z = -8.2$, $p < .001$), whereas it increased overall accuracy in the audio expansion condition ($b = 2.0$, SE = 0.3, $z = 7.0$, $p < .001$). *Homonyms.* The homonymy by slowing manipulation interaction in the compression/expansion experiment indicated that there was a marginal decrease in homonym accuracy after slowing ($b = -0.6$, SE = 0.4, $z = -1.7$, $p = .10$). *Polysemes.* A marginal main effect indicating a polysemy advantage was observed in the nonword wordlikeness experiment ($b = 0.4$, SE = 0.2, $z = 1.7$, $p < .09$). There were also marginal polysemy by slowing manipulation interactions in the nonword wordlikeness and audio compression/expansion experiments, indicating that there was decrease in polyseme accuracy relative to the unambiguous baseline in the slowed conditions (b = -0.6, SE = 0.4, z = -1.8, p =.09). *Hybrids.* As in the latency data, no significant effects involving the hybrids were observed. *Imageability.* The facilitatory main effect of imageability was always significant (all ps ≤ .02), except for in the case of auditory noise (the model did not converge) and in the audio compression/expansion experiment, where the effect was marginal (p = .15). There was a marginal interaction between imageability and the slowing manipulation in the visual noise experiment indicating differentially decreased facilitation after slowing (b = -0.2, SE =0.1, z = -1.6, p = .11), whereas in the compression/ extension experiment (b = 0.2, SE = 0.1, z = 1.5, p = .14) there was increased facilitation.

Summary. A significant or marginal homonymy disadvantage, or an increased homonymy disadvantage in the slowed condition, was observed in all but the visual noise experiment. A main effect of polysemy was only detected in one experiment and the polysemy advantage marginally decreased in two experiments. Hybrid items were never significantly different from the unambiguous controls, which is likely due, at least in part, to difficulties matching these rare items on other covariates. The facilitatory effect of imageability was significant or marginal in all experiments. The magnitude of these facilitation effects increased marginally in three experiments (intermodal, visual noise, compression /expansion).

Discussion

The aim of our study was to evaluate whether a range of different manipulations designed to slow responses would lead to different ambiguity effects, as predicted by the SSD account. At first glance, except for speed-accuracy trade-offs, virtually all of the effects that were significant or marginal were consistent with the SSD account. Additionally, most of non-significant results showed the predicted trends numerically. Thus, this collective body of work does add some additional support to the notion that processing time---and the presumed amount of semantic settling---plays a role in explaining many ambiguity effects. These results also suggest that some broad ambiguity effects transcend different languages.

Additionally, taking a more critical view of the observed effects promises to reveal additional aspects of how and why discrepant ambiguity effects are observed within and between tasks. To begin, our ideal a priori aim was to reproduce a polysemy advantage only in the easiest/fastest tasks (Figure 1, Slice A) and observe a homonymy disadvantage only in the hardest/fastest tasks (Figure 1, Slice C). The overall pattern of results, however, would appear to be more consistent with the easiest task beginning closer to Slice B, where both a weaker homonymy disadvantage and polysemy advantage are predicted. This result is surprising for several reasons. First, Armstrong and Plaut (2016) went to great lengths to make their lexical task as difficult as possible, and yet their results were consistent with earlier processing dynamics (primarily Figure 1, Slice A-B). Their overall latencies were also approximately 100ms faster than in the analogous conditions in the present work. The present work did use words with slightly lower frequencies, but it also used considerably easier nonwords, so there is no clear explanation for this large discrepancy. Further, we have conducted an additional experiment with "very easy" nonwords (nonwords with extremely low bigram frequencies and neighborhood sizes) and still not been able to increase overall performance by a substantial degree. These results are also inconsistent with Jager, Green, & Cleland's (2016) prediction that a polysemy advantage should be strongest for low frequency words because their meanings overlap more.

Another possibility worth considering is that whereas past research has typically struggled to produce a homonymy disadvantage and had more success in obtaining a polysemy advantage, the present work may have experienced the opposite difficulties. This may be due to having used atypically large set of balanced homonyms. This was accomplished by sampling from a database of subjective meaning frequency norms (Armstrong et al., 2015) and may have differentially boosted the power of the homonymy effects. This more powerful manipulation of homonymy may also have coincided with a less powerful manipulation of polysemy based on the recent results of Fraga, Padrón, Perea, & Comesaña (2016). They found that although the number of senses provided in a subjective meaning norming study and those available in the RAE dictionary (the source of our polysemy counts) correlated highly, only the subjective norms were significant predictors of latencies in lexical decision and naming tasks. Unfortunately, there was insufficient overlap between our items and theirs to corroborate their findings in our own data. However, this recent observation clearly stresses the importance of how polysemy is measured. In English, several studies have used dictionary counts to predict polysemy successfully (e.g., Armstrong et al., 2016; Rodd et al., 2002 both used counts from Wordsmyth; Parks, 1999). Thus, our findings in Spanish suggests that the lexographers administering the RAE dictionary use a different classification scheme for ambiguity, and/or English and Spanish vary in their distributions of polysemes in ways that shape performance to a substantial degree. The latter possibility gains support from the Armstrong et al. (2015) homonym norming study. They observed that despite Spanish and English having similar total numbers of homonyms, Spanish homonyms are much more likely to have a strongly dominant meaning. (This also posed challenges for us finding balanced and well matched hybrids.) Clearly, a more extensive set of polyseme norms with high external validity must be collected in both languages to evaluate these possibilities.

The prior discussion has focused primarily on potential differences in objective or subjective measures of ambiguity. However, is also possible that broader properties of the language and/or of our participants may have contributed to the aforementioned discrepancies. Our use of

Spanish, an orthographically transparent language, may have been advantageous when controlling for orthographic and phonological confounds. However, it may also have allowed for the rapid spreading of activation between orthography and phonology. This could have, in turn, allowed these representations, as opposed to semantics, to be the primary drivers of the response system. Although the significant effects of imageability indicate that semantics did always influence responses, it is possible that semantic effects may have been attenuated such that only the strong effect of homonymy could be detected.

On a related front, the participants tested by Armstrong and Plaut (2016) were all native English speakers in the USA and presumably had limited exposure to other languages. In contrast, the participant population in the Basque Country is bilingual and all participants reported proficiency in one or more other languages that share at least a partially overlapping phonology and/or orthography (e.g., Basque, French, English). Bilingualism in and of itself has been reported to slow responses in some tasks (e.g., Luo, Luk, & Bialystok, 2010). These results have typically been explained by focusing on dynamics at the (sub)lexical level, however (e.g., in the Bilingual Interactive Activation model; Dijkstra & van Heuven, 1998). Our results suggest that some of these differences could also be attributable to processing differences at a semantic level. Consistent with this hypothesis, Taler, Zunini, and Kousaiev (2016) found that monolinguals exhibited greater facilitation as a function of increased NoS than bilinguals in a lexical decision task. This was true both in response latency and in EEG measures of the N400, which is known to index semantic processing. Collectively, these results suggest that semantic settling dynamics and ambiguity resolution could be impacted by knowledge of multiple languages. The field would therefore benefit from additional carefully matched experiments across a broad span of languages.

Returning to the initial question that motived our work, does processing time play a critical role in shaping some ambiguity effects? Our results provide partial support that this is, indeed the case. However, the cases in which such support did not materialize are perhaps just as theoretically relevant. These cases highlight how certain core effects in the semantic ambiguity literature may vary as a function of the language in which the test is conducted, and/or as a function of knowledge of a second language. They also point to important methodological issues that remain to be addressed, such as how to classify and compare polysemy across languages. Taken together, the present work therefore serves not only advances our understanding of the semantic settling dynamics in ambiguity resolution. It also highlights the value of cross linguistic comparisons in developing a general as opposed to a language-specific understanding of semantic ambiguity.

Acknowledgments

This work was funded by the BCBL's Severo Ochoa Center grant SEV-2015-049, CAPES Foundation BEX grant 1692-13-5 to JM, NSERC, and the University of Toronto. We thank Asier Zarraga for assistance with data processing, and Arthur Samuel for his tool for generating auditory noise.

References

Armstrong, B. C. (2012). *The temporal dynamics of word comprehension and response selection: Computational and behavioral studies*. Doctoral dissertation, Psychology Department, Carnegie Mellon University, Pittsburgh.

Armstrong, B. C., & Plaut, D. C. (2016). Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. *Language, Cognition and Neuroscience, 31*(7), 940-966.

Armstrong, B. C., Watson, C. E., & Plaut, D. C. (2012). SOS! An algorithm and software for the stochastic optimization of stimuli. *Behavior Research Methods*, 44(3), 675-705.

Armstrong, B. C., Zugarramurdi, C., Cabana, Á., Lisboa, J. V., & Plaut, D. C. (2015). Relative meaning frequencies for 578 homonyms in two Spanish dialects: A cross-linguistic extension of the English eDom norms. *Behavior Research Methods*, 1-13.

Baayen, R.H., Milin, P. (2010). Analyzing Reaction Times. *International Journal of Psychological Research*, 3(2), 12-28.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: An MEG study. *Cognitive Brain Research*, 24(1), 57-65.

Dijkstra, T., & Van Heuven, W. J. (1998). *The BIA model and bilingual word recognition. Localist connectionist approaches to human cognition*, 189-225., Mahwah, New Jersey: Lawrence Erlbaum Associates.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods*, 45(4), 1246-1258.

Fraga, I., Padrón, I., Perea, M., & Comesaña, M. (2016). I saw this somewhere else: The Spanish Ambiguous Words (SAW) database. *Lingua*. 185, 1-10.

GoldWave (v6.13) [Software]. St. John's, NF: GoldWave ® Inc.

Hino, Y., Pexman, P. M., & Lupker, S. J. (2006). Ambiguity and relatedness effects in semantic tasks: Are they due to semantic coding?. *Journal of Memory and Language*, 55(2), 247-273.

Jager, B., Green, M. J., & Cleland, A. A. (2016). Polysemy in the mental lexicon: relatedness and frequency affect representational overlap. *Language, Cognition and Neuroscience*, 31(3), 425-429.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3), 627-633.

Luo, L., Luk, G., & Bialystok, E. (2010). Effect of language proficiency and executive control on verbal fluency performance in bilinguals. *Cognition*, 114(1), 29-41.

Mazzoni, D. (2013). Audacity (Version 2.0.5) [Software].

Parks, R. (1999). *Wordsmyth English Dictionary-Theasurus*.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1), 8-13.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Real Academia Española. (2014). *Diccionario de la lengua española* (23rd ed.). Madrid, Spain.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, 46(2), 245-266.

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: the perils of overreliance on an" outlier" orthography. *Psychological Bulletin*, 134(4), 584.

Taler, V., Zunini, R. L., & Kousaie, S. (2016). Effects of Semantic Richness on Lexical Processing in Monolinguals and Bilinguals. *Frontiers in Human Neuroscience*, 10.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5), 971-979.