Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences

Blair C. Armstrong*

Basque Center on Cognition, Brain, and Language

Paseo Mikeletegi 69, Floor 2, San Sebastian, Spain, 20009

blair.c.armstrong@gmail.com or b.armstrong@bcbl.eu; Phone: +34 943 309 300, Ext. 202

David C. Plaut

Department of Psychology and Center for the Neural Basis of Cognition

Carnegie Mellon University

5000 Forbes Avenue, Pittsburgh, PA, 15213

plaut@cmu.edu

Abstract

A core challenge in the semantic ambiguity literature is understanding why the number and relatedness among a word's interpretations are associated with different effects in different tasks. An influential account (Hino, Pexman, & Lupker, 2006, *JML*) attributes these effects to qualitative differences in the response system. We propose instead that these effects reflect changes over time in settling dynamics within semantics. We evaluated the accounts using a single task, lexical decision, thus holding the overall configuration of the response system constant, and manipulated task difficulty—and the presumed amount of semantic processing—by varying nonword wordlikeness and stimulus contrast. We observed that as latencies increased, the effects generally (but not universally) shifted from those observed in standard lexical decision to those typically observed in different tasks with longer latencies. These results highlight the importance of settling dynamics in explaining many ambiguity effects, and of integrating theories of semantic dynamics and response systems.

Keywords: semantic ambiguity; settling dynamics; connectionist models; decision making / response selection

Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences

Most words are semantically ambiguous, in that their interpretation depends on the context in which they occur (Klein & Murphy, 2001). Developing a theory of ambiguous word comprehension is therefore essential for a complete understanding of lexical and discourse processing more generally, and it is no surprise that a vast literature has focused on semantic ambiguity over the past several decades (e.g., Armstrong, Tokowicz, & Plaut, 2012; Armstrong, Zugarramurdi, Cabana, Valle Lisboa, & Plaut, 2015; Azuma & Van Orden, 1997; Beretta, Fiorentino, & Poeppel, 2005; Borowsky & Masson, 1996; Frazier & Rayner, 1990; Gernsbacher, 1984; Hino, Kusunose, & Lupker, 2010; Hino, Pexman, & Lupker, 2006; Jastrzembski, 1981; Joordens & Besner, 1994; Klein & Murphy, 2001; Klepousniotou & Baum, 2007; Klepousniotou, Pike, Steinhauer, & Gracco, 2012; Mirman, Strauss, Dixon, & Magnuson, 2010; Pexman & Lupker, 1999; Rodd, Gaskell, & Marslen-Wilson, 2002; Rubenstein, Garfield, & Millikan, 1970).

Recently, theoretical accounts of semantic ambiguity resolution have faced two main challenges. First, the relatedness of an ambiguous word's interpretations, as well as the word's total number of interpretations, have been found to modulate semantic ambiguity effects. Thus, relative to unambiguous controls, studies often report different patterns of effects for polysemes which have multiple related interpretations (e.g., the <printer> and <academic> interpretations of PAPER; hereafter denoted as <printer>/<academic> PAPER) than for homonyms which have multiple unrelated interpretations (e.g., <river>/<money> BANK). Second, different effects of number-of-interpretations and relatedness-of-interpretations have been reported between—and sometimes within—tasks. For example, a number of visual lexical decision studies have found that polysemous words are associated with a major processing advantage compared to unambiguous controls, whereas homonyms are typically no different from unambiguous controls or are associated with a weak disadvantage (e.g., Beretta et al., 2005; Klepousniotou & Baum, 2007; Klepousniotou, Titone, & Romero, 2008; Locker, Simpson, & Yates, 2003; Rodd et al., 2002). In contrast, semantic categorizations that involve broad categories (e.g., does the presented word denote a LIVING THING? Hino et al., 2006) have found that homonymous words are associated with a processing disadvantage relative to unambiguous controls, whereas there is no difference between polysemes and unambiguous controls. Taken together, these complex and apparently contradictory effects of number of interpretations and relatedness of interpretation across tasks represent a major obstacle for advancing a theory of ambiguous word comprehension.

To address these challenges, Hino and colleagues (2006; see also Hino et al., 2010; Pexman, Hino, & Lupker, 2004) proposed an account of a diverse set of semantic ambiguity effects based on qualitative task differences and their consequences for the configuration of the response system. On their account, a similar semantic representation is activated in all tasks. Consequently, the inconsistent results obtained across tasks must be due to how the response system is configured to select responses based on the same semantic representation. This account is appealing for a number of reasons. First, in some restricted sets of circumstances at least, it is clear that the locus of some inconsistent semantic ambiguity effects is in how the semantic representations map onto the response system. This was well illustrated by Pexman et al. (2004), who showed that a homonymy disadvantage in a semantic relatedness task was strongly modulated by whether a homonym’s meanings all supported a particular response (e.g., all meanings of PUPIL support a "no" response for “are HAWK-PUPIL related?”) or provided partial support for two different responses (e.g., one interpretation of TIE supports a "yes" response for “are SHIRT-TIE related?”). Second, this account essentially permits the explanation of any semantic ambiguity effect observed in any task as resulting from a qualitative difference in how the response system is configured. For instance, Hino et al. (2006) used this account to explain why, they observed an overall advantage for homonymous words [1]and polysemous words relative to unambiguous controls in lexical decision, whereas in semantic categorization (is BANK a LIVING THING?) only a homonymy disadvantage (or, in our classification scheme, a hybrid disadvantage) was observed. Critically, in contrast to Pexman et al. (2004), the Hino et al. experiments used a new set of word stimuli and eliminated response competition per se as an explanation for their semantic categorization results by ensuring that all interpretations of the ambiguous words supported a single response (e.g., all interpretations referred to nonliving things, such as for BANK).

It might seem, then, that an account based on the configuration of the decision system can reconcile all of the apparently contradictory semantic ambiguity effects. However, a closer consideration reveals that this account, at least in its current formulation, is unsatisfactory in a number of respects. First, in its current abstract verbal form, it is unclear exactly how the response system is configured to generate responses in a given task. Consequently, it is impossible to evaluate whether the configuration process can, in fact, generate different configurations of the response system that can fit the results observed in different tasks. It is also unclear how this account can be used to generate predictions regarding effects that should emerge in new tasks. Thus, the account is too flexible and post hoc, in that new experiments will always differ qualitatively to some degree and therefore any new result can, in principle, be accommodated.

Second, the emphasis on a post-semantic process as an explanatory basis for semantic ambiguity effects is inconsistent with the large body of neuroimaging data that point to a semantic source for many of the observed effects—a source that includes, as a hub, the anterior temporal lobes, although other regions may contribute to the semantic network, as well (e.g., Beretta et al., 2005; Carreiras, Armstrong, Perea, & Frost, 2014; Kutas & Hillyard, 1980; Pylkkänen, Llinás, & Murphy, 2006; Van Petten & Kutas, 1987). For example, Beretta et al. (2005) reported the results of a magnetoencephalography (MEG) study that found a neural correlate of semantic processing, the M350 component (the MEG analog to the N400 in the electroencephalography literature) recapitulated the behavioral semantic ambiguity effects reported by Rodd et al. (2002). More specifically, the neural correlates of semantic activity that they recorded suggested that relative to unambiguous controls, homonym representations were activated more slowly; in contrast, polyseme representations were activated more quickly. A simple and parsimonious account of the behavioural results reported by Rodd et al. (2002), who reported similar effects in their behavioural data, is therefore that the different patterns of semantic activity measured via MEG carry forward to influence the speed of behavioural responses. The theoretical locus for explaining all ambiguity effects therefore need not be in the response system, but could, alternatively, be a carry-over effect from semantic processing dynamics in and of itself.

Third, a core assumption of the account is that the response system plays a central role in generating a range of ambiguity effects, and alternative accounts based on temporal processing dynamics in semantics models may have been dismissed prematurely in favor of the decision system account. Specifically, Hino and colleagues (2006) state that "[their] data, as well as those reported by Pexman et al. (2004) and Forster (1999), suggest that ambiguity disadvantages in semantic categorization and relatedness judgment tasks are likely not due to the semantic-coding process as that process is conceptualized within PDP [parallel distributed processing] models (p. 266)." In contrast, our proposed account suggests that the processing dynamics within such models do, in fact, explain the different patterns of semantic ambiguity effects reported in a number of tasks, including many lexical decision and semantic categorization studies, as being the direct consequence of such semantic dynamics. According to this account, the decision system may require more or less evidence—and, by proxy, time—before generating responses in different tasks, but the main reason why the ambiguity effects change as a result of increased evidence/time is how semantic activity changes over time for different types of ambiguous words. Without an explicit description of how their account could be extended to arbitrary tasks and also be sensitive to temporal processing dynamics, the most straightforward predictions for that account are, in our view, that similar semantic representations are tapped across tasks and different patterns emerge only when the response system changes. However, this ignores extensive computational and theoretical literatures showing that temporal processing dynamics generate different amounts of semantic activation at different points in time[2](Armstrong & Plaut, 2008; Joordens, Piercey, & Azarbehi, 2009; Plaut & Shallice, 1993; Simpson, 1994; see also Kawamoto, 1993). Given that the various tasks that are associated with apparently inconsistent ambiguity effects typically differ in overall latency (e.g., Hino et al., 2006; Rodd et al., 2002), often by several hundred milliseconds, this opens the possibility that processing time, rather than qualitative differences in the response system, may be the core principle that underlies the diverse set of reported ambiguity effects.

Building upon this theme, Armstrong and Plaut (2008) proposed the *semantic settling dynamics* (SSD) account as an alternative to an account based on the response system (see also Armstrong & Plaut, 2011, and the related accounts of Piercey & Joordens, 2000, and Rodd, Gaskell, & Marslen-Wilson, 2004). We begin by sketching the main predictions and general processing principles underlying the account, and elaborate on additional mechanistic detail later. Additional supporting evidence is also provided in a simulation reported in Appendix A. The SSD account predicts that different semantic ambiguity effects emerge at different points in time during processing because of the temporal dynamics associated with cooperation and competition between the semantic features of polysemous, homonymous, and unambiguous words. These dynamics are further shaped by contextual constraint which allows for the selection of a contextually appropriate semantic representation of an ambiguous word. However, consistent with past findings (but not with explicit theories of staged contextual integration) the role of context is assumed to be weak during early processing, so the early temporal dynamics are relatively insensitive to context and the influence of context gradually builds up over time (Swinney, 1979; Taler, Klepousniotou, & Phillips, 2009; Van Petten & Kutas, 1987).

Furthermore, to most clearly distinguish between an account based on semantic settling dynamics from other accounts, such as one based on the configuration of the response system, we make an initial simplifying assumption that this semantic activity is able to drive the response system directly regardless of the task in question (e.g., lexical decision, semantic categorization). This allows us to focus our predictions based on the activity in semantics only. This is not to say that we claim that other representations do not also drive the response system to some degree as well (e.g., orthography, phonology), that this semantic activity cannot feed back and interact with other representations[3]or that the response system is not configured in different ways that could also shape performance, as we noted previously.

Indeed, it may very well be the case that these other representations are the major sources of evidence that drive the response system in some cases (e.g., lexical decision) and semantics makes only a minor contribution to this process. For simplicity, however, we are assuming that ambiguous and unambiguous words do not differ in terms of these other representations, so that we can test whether semantic ambiguity effects per se could emerge from semantic representations directly. Related to the last point, this claim is intended to emphasize that, provided that a task is carefully set up to avoid confounds involving these other systems (e.g., ambiguous words that partially support both possible responses in a task), the temporal dynamics in semantic processing—that is, how much semantic processing has taken place when a response is generated—are the primary source of many different semantic ambiguity effects. As a concrete example, consider the case of lexical decision, which is the focus of the empirical work reported here. Orthographic representations may provide a strong basis for making word versus nonword decisions, but for simplicity we assume that orthographic representations are equally informative for all types of ambiguous versus unambiguous words. Semantic activity, however, might differ for polysemes, homonyms, and unambiguous controls and contribute to different patterns of responses for these different word types (as shown in Figure 4 by vertical slices of the x-axis). Such direct contributions from semantics are self-evident in the case of semantic tasks (e.g., semantic categorization), and prior modeling work has shown that semantics can be a reliable basis for making lexical decisions, as well (e.g., Armstrong, Joordens, & Plaut, 2009; Plaut, 1997). After presenting our results, we discuss a more refined overall picture of how performance is driven in lexical decision and how these semantic settling dynamics could be integrated with more refined conceptualizations of processing in other systems (e.g., the response system).

With these general claims laid out, we now describe the representations and processes assumed to underlie the different types of word stimuli. Polysemous words are assumed to share a core set of semantic features across interpretations and to have some idiosyncratic features that do not overlap across interpretations (Klepousniotou et al., 2008; Rodd et al., 2002; Williams, 1992), homonyms are assumed to have multiple distinct and non-overlapping semantic representations, and unambiguous words are assumed to have only a single semantic representation. For simplicity, we assume, both in our verbal description here and in the simulations reported in

Appendix A, that each semantic representation is equally semantically rich (i.e., it has the same number of semantic features), in line with past semantic ambiguity research that counterbalances items or controls for this confound during the subsequent analyses[4](e.g., by including a measure such as imageability or concreteness as a covariate as in the present work and Rodd et al., 2002). In the case of polysemous words, all of the different semantic features can therefore cooperate to activate the core semantic features shared across all interpretations of the word, which gives rise to an initial processing advantage for polysemes relative to unambiguous controls. This advantage is further exacerbated by the strong, numerous, and dense excitatory connectivity within a brain region, as well as the fact that most between-region connections are excitatory-only, which cause overall excitatory/cooperative dynamics to dominate during early processing. The polysemy advantage would fade once the core features are strongly active and the idiosyncratic features associated with each interpretation compete with one another (ultimately leading to a small processing disadvantage for polysemes when extensive semantic processing has taken place).

Given that the polysemy advantage is being driven by an interaction between core features shared across interpretations and idiosyncratic features that are distinct between interpretations, it is also worth noting that the exact magnitude of the polysemy advantage and how long it persists is also assumed to vary as a function of feature overlap. In more detail, we assume that relatedness of interpretation is a continuous measure, ranging from no overlap—as is generally assumed to be the case for homonyms, which activate non-overlapping features in different contexts—to complete overlap—as is basically the case for unambiguous words, which activate the same features in different contexts. Polysemes are a broad category that spans the area in between. Insofar as a particular polyseme is extremely similar to an unambiguous word—that is, there is almost perfect featural overlap across interpretations and almost no idiosyncratic features—performance associated with that item should be similar to unambiguous controls. This is because it has fewer total features to support the core interpretation, when feature counts are summed across interpretations. At the opposite extreme, insofar as a particular polyseme shares only a minimal number of features across interpretations and most features are uniquely associated with a particular interpretation, that polyseme should generate very similar performance to homonyms. This is because even though many features can support the core features, the small number of core features limits the total semantic activity that can build up as a result of this processing advantage.

The optimal polysemy advantage described here will therefore occur for items with some intermediate ratio of core and idiosyncratic features. The literature is in general agreement with these claims, although more detailed empirical and computational work is needed to provide a characterization of how this dynamic operates across the entire range of relatedness of interpretations. For instance, Klepousniotou et al. (2008) have shown that polysemes with senses that have high overlap but are nevertheless somewhat distinct and therefore not completely unambiguous (e.g., metonymic polysemes such as <meat>/<animal> CHICKEN) show a stronger advantage than polysemes with less featural overlap (e.g., metaphoric polysemes such as <movie>/<celestial> STAR). In the modeling work we have conducted to date we have observed that a significant polysemy advantage is detectable for polysemes with between 60-80% feature overlap (as reported in Armstrong & Plaut, 2008, and in the Appendix). More work will be needed, however, to determine exactly when the overlap is too high or too low to generate a sufficiently large ambiguity advantage to be detectable in experimental settings, and relate that overlap to empirical measures of relatedness. In particular, future work could help determine whether there an as-yet unidentified subset of ambiguous words that are still classified as (extremely low) overlap polysemes for which competition between the large number of inconsistent features dominates cooperation among the consistent features—that is, are there polysemes with extremely low featural overlap that behave like typical homonyms? Or, is the amount of featural overlap that determines whether a lexographer classifies an item as polysemous or homonymous in line with the amount of overlap needed to generate processing advantages and disadvantages in our model? Before moving onto those more detailed points, however, we have focused the present work on evaluating the basic validity of the central tenets of our account. [5]

In contrast to polysemes, in the case of a homonym both interpretations would receive partial support during early processing from the surface form (i.e., orthography and/or phonology). For instance, if the homonym had two interpretations, the semantic representation of each interpretation would receive half as much support from the surface form as would be sent to the semantic representation of an unambiguous word. Except for this consideration, the total semantic activation of the homonym would initially build up in much the same way as the activation of the single interpretation of an unambiguous word. More specifically, twice as many semantic features should start to become active but because each interpretation receives half the bottom up support, these two factors would basically cancel each other out, leading to no difference between homonyms and unambiguous controls.[6] As each of these interpretations becomes increasingly active, competition between these inconsistent interpretations increases, giving rise to a homonymy disadvantage relative to unambiguous controls. The initial relatively weak competitive dynamic, which builds in strength over time, is attributable primarily to the relatively sparse/weak inhibitory connectivity, as well as the absence of inhibitory connectivity between brain regions, which result in relatively weak and late inhibitory effects.

Collectively then, the account predicts an early polysemy advantage, consistent with typical lexical decision results, and a later homonymy disadvantage, consistent with semantic categorization findings with broad categories (for a broader review of the empirical effects consistent with the SSD account, seeArmstrong & Plaut, under review). It therefore has the potential to provide a unifying framework for understanding a range of ambiguity effects. To substantiate the SSD account, Armstrong and Plaut (2008) developed a connectionist model that implemented the core assumptions outlined above. The simulation showed that, as predicted, a polysemy advantage emerged during early processing, whereas a homonymy disadvantage emerged during later processing. Moreover, it made the novel prediction that at an intermediate time-point, both a polysemy advantage and a homonymy disadvantage should emerge, albeit to weaker extents than during early and later processing, respectively.

In contrast to an account based on the configuration of the response system, the SSD account makes clear predictions regarding the role that processing time, in and of itself, should play in modulating semantic ambiguity effects, holding task and the overall contributions of the response system constant. In so doing, it presents temporal processing dynamics in semantics per se, particularly as they are processed in parallel distributed processing models, as the mechanistic basis for explaining a range of semantic ambiguity effects (cf. Hino et al., 2006, as discussed earlier in the introduction). Specifically, the SSD account predicts that distinct patterns of semantic ambiguity effects should be observed if responses within a single task are sped up or slowed down. Although the predictions of the decision-system configuration account will remain somewhat unclear until it is specified more formally, presumably it would predict that the same semantic ambiguity effects should hold as long as the task itself remains constant; we will elaborate on a more subtle, mechanistically-oriented view in the discussion. These contrasting predictions are the motivation for the empirical study reported next.

Effects of Task Difficulty in Visual Lexical Decision

The current study was designed to test a key prediction of the SSD account: will variations in the degree of semantic precision required to generate a response determine the pattern of semantic ambiguity effects that are observed? We focused on early processing dynamics because the main controversies in the literature deal with tasks that do not require contextual integration and because changes in semantic effects as a function of processing time are well-established during later context-sensitive processing (e.g., Seidenberg, Waters, Sanders, & Langer, 1984; Swinney, 1979; Tabossi, 1988; for a review, see Simpson, 1994). Visual lexical decision was selected as the target task because of its ubiquity in the study of semantic ambiguity. The overarching logic of the study was to take a “standard” polysemy advantage in visual lexical decision and, by increasing task difficulty and overall processing time, observe a weakened polysemy advantage and the emergence of a homonymy disadvantage.

We elected to extend the experimental design employed by Rodd and colleagues (2002, Experiment 2; see also Beretta et al., 2005) that yielded a polysemy advantage and no homonymy disadvantage, with several improvements to the basic design and methods. These improvements were aimed at minimizing the likelihood of spurious confounds driving the results, and at maximizing statistical power, given that the simulation predicted that at an intermediate time-point, both the polysemy advantage and homonymy disadvantage would be weaker than at either an earlier or a later time point, respectively. To do so, meaning frequency norms measuring the relative meaning frequencies of the dominant and subordinate meanings of homonyms were used to account for the different degrees of competition that should be generated depending on whether a homonym has multiple interpretations that are encountered with roughly equal frequencies or has a single strongly dominant interpretation (Armstrong, Tokowicz, & Plaut, 2012; for related computational principles, see McClelland & Rumelhart, 1981). In addition, a large set experimental items were carefully matched across conditions using a stimulus optimization algorithm (Armstrong, Watson, & Plaut, 2012). Finally, mixed-effects analyses were employed to provide a powerful and unbiased assessment of generalization across both participants and items, in contrast to running separate by participants and by items analysis and not requiring that both tests are significant, and/or not requiring that the min-F′ statistic is also significant (Barr, Levy, Scheepers, & Tily, 2013; Clark, 1973; Raaijmakers, Schrijnemakers, & Gremmen, 1999).

The easiest condition in the experiment, which was nevertheless designed to be at least moderately difficult, was expected to exhibit the standard polysemy advantage and lack of homonymy disadvantage reported in many visual lexical decision studies. Relative to this, the experiment investigated whether a homonymy disadvantage would be observed when difficulty was increased by crossing two between-participant manipulations: three levels of nonword difficulty (orthographically hard nonwords vs. orthographically very hard nonwords vs. pseudohomophones) and two levels of stimulus contrast (“full” white-on-black text vs. “degraded” gray-on-black text). In lexical decision, nonword wordlikeness is known to modulate ambiguity effects (Azuma & Van Orden, 1997; Rodd et al., 2002), and reduced stimulus contrast can slow responses by at least 100-150 ms (e.g., Borowsky & Besner, 1993).

To reduce overall speed-accuracy trade-offs between conditions, we provided feedback at the end of each block to encourage participants to either go faster even at the expense of more mistakes or to try to be more accurate even if it meant responding more slowly, so as maintain an overall accuracy of 90%. We predicted that this would facilitate the interpretation of the data by encouraging differences to emerge in latency. However, because we provided feedback to maintain overall accuracy levels below ceiling, it is also possible that between item-type differences could emerge within a condition. We therefore tested to see if the predicted semantic activity facilitated or impaired efficient responding, as could be reflected in accuracy and/or latency.

Methods.

Participants. Students from the undergraduate participant pool at the University of Pittsburgh participated in the experiment for course credit. Between 72 and 76 participants (447 participants in total) completed each of the six difficulty levels of the lexical decision task, described in detail later, such that the manipulation of task difficulty occurred between participants. All had normal or corrected to normal vision, were native English speakers, and gave informed consent to participate. Students participated in only a single condition or associated norming study. This between-participant manipulation of difficulty was employed to avoid order effects that were observed in preliminary studies that varied difficulty within participants (Armstrong, 2012). The study was approved by the Institutional Review Board of the University of Pittsburgh.

Apparatus. The experiment was run in a dimly lit room on computers running E-prime 2.0.10.182 (Schneider, Eschman, & Zuccolotto, 2010). Participants responded on a standard keyboard. Full-contrast items were presented as white text (162.9 cd/m^2) on a black background (0 cd/m^2), whereas degraded stimuli were presented as dark gray text (1.9 cd/m^2) on black background. Contrast levels were selected in a pilot study so as to cause at least a 100 ms slow-down in the hard nonword condition.

Stimuli and design. Word stimuli were selected to fill a 2 (meanings: one vs. many) x 2 (senses: few vs. many) factorial design analogous to that used by Rodd et al. (2002), although as noted before, we made the a priori decision to take meaning frequency into consideration during the analyses.[7] Meaning and sense counts were derived from the Wordsmyth dictionary, consistent with past studies in this vein (e.g., Azuma & Van Orden, 1997; Beretta et al., 2005; Rodd et al., 2002). For convenience, the one-meaning few-senses cell is referred to as the (relatively) *unambiguous* condition, the many-meanings few-senses cell as the *homonymous* condition, the one-meaning many-senses cell as the *polysemous* condition, and the many-meanings many-senses cell as the *hybrid* condition. The SOS software package (Armstrong, Watson, & Plaut, 2012) was used to find 100 quadruplets of experimental items (400 total), as well as 100 unambiguous filler items for use during familiarization, practice and at the beginning of each block, which were not analyzed. The quadruplets were minimally different from one another on a pair-wise level across a number of factors that influence word recognition (see Table 1). This minimized the presence of both group- and item-level confounds in the later analyses. The homonyms and hybrid items were sampled from approximately 400 homonyms and 130 hybrid items that were identified as suitable for standard psycholinguistic experiments and that were included in the eDom norms (Armstrong et al., 2012). Given that the critical predictions concern homonym-unambiguous-polyseme comparisons, and given the relatively small degrees of freedom available for selecting hybrid items—approximately 75% of the population of such items in English were included in the study—the constraints of the optimization were designed to prioritize matches between the three conditions of primary interest.

Data on familiarity, imageability, and meaning frequency were not available for a sufficiently large set of ambiguous items prior to stimulus selection, so these properties were normed in a separate set of control experiments, with the intent of subsequently addressing any issues with those metrics in the analyses. Note that this approach is feasible only if a large number of items is available—for instance, this allows for approximately 80% of the items in the study to be discarded and still have as many items left as in many studies (e.g., Hino et al., 2010; Mirman et al., 2010), or for these measures to be controlled for when they are treated as continuous variables in a regression, as was done in the analyses we report later. Normative data for familiarity and frequency were collected from approximately 40 different participants for each measure. Standard norming methods similar to those reported by Hino et al. (2006) were employed, in which participants rated these measures on a 7-point Likert scale. Relative meaning frequency was normed by 50 participants and a measure of the frequency of the dominant meaning of the homonym was used in the later analysis (for details, see Armstrong, Tokowicz, & Plaut, 2012).

– – – – – – –

Insert Table 1 Here

– – – – – – –

Three different groups of 500 nonwords were generated (broken down into 400 "experimental" nonwords and 100 "filler" nonwords used during familiarization, practice, and warm-up), each of which matched the distribution of lengths of the word stimuli. Two of these groups were created by sampling from a pool of nonwords created by replacing one consonant in a word sampled from the the SUBTL database (Brysbaert & New, 2009) with another consonant. The hard nonword group consisted of nonwords with positional bigram frequencies that were roughly matched to those of the word stimuli. The very hard nonword condition was created by selecting the nonwords with the highest positional bigram frequencies in the pool. The third group of nonwords consisted of pseudohomophones, which were included because of a report that lexical decisions made in the context of this type of foil may be particularly likely to show semantic ambiguity effects (Rodd et al., 2002), although there are theoretical reasons why our other nonword foils could produce stronger semantic effects—a point we return to in the discussion. The pseudohomophones were sampled from the ARC nonword database (Rastle, Harrington, & Coltheart, 2002). This sample was restricted to contain only pseudohomophones with orthographically existing onsets and bodies and that only contained legal bigrams. These items were rank ordered based on 1) orthographic Levenshtein distance, 2) orthographic neighborhood size, and 3) positional bigram frequency. The most wordlike nonwords in this list were selected, with the constraint that pseudo-plurals and pseudo-past tenses were largely avoided to prevent participants from basing their responses on the last letter of the stimulus. Properties of the nonword and word stimuli are presented in Table 2.

– – – – – – –

Insert Table 2 Here

– – – – – – –

Procedure. Participants were instructed to press the ‘z’ or ‘/’ key to indicate whether a word or nonword was presented, and were provided with examples of each type of trial. Word responses were always made with their dominant hand. To increase the sensitivity of the latency data, avoid speed-accuracy trade-offs, and avoid ceiling effects in accuracy, participants were instructed to respond as quickly as possible and were told that it was acceptable to make incorrect responses up to 10% of the time. After each block, they were also presented with their latencies and accuracies for that block and for the preceding one. If they made less than 10% errors, they were instructed to “try to go faster even if it means making a few more mistakes”, otherwise they were instructed to “try to be more accurate, even if it means slowing down a little.”

An initial block of 20 trials familiarized participants with the task, and was followed by a 100 trial warm-up block. Filler trials were used in both of these cases. Participants then completed 8 experimental blocks of 110 trials each, which were seamlessly divided into 10 warm-up filler trials followed by 100 experimental trials. All blocks contained equal numbers of words and nonwords and the order of the stimuli was randomized, with the constraint that there could be no more than 3 consecutive words or nonwords.

Each trial began with a 250 ms blank screen and a fixation stimulus (####+####) presented for a random duration between 750 and 950 ms. This was followed by a 50 ms blank screen, after which a word or nonword stimulus was presented for 4000 ms or until the participant responded.

Results. Data were screened as follows prior to analysis: All words that at least 10% of participants in the norming studies indicated they did not know, and all items with accuracies below 50%, were dropped. This eliminated 11 words and 17 nonwords, distributed approximately equally across the item types. One participant was also dropped for having a mean accuracy below 50%. We view this portion of data screening as ensuring that only items that were correctly classified by participants were included in subsequent analyses (i.e., the words were typically known to participants, and the nonwords did not inadvertently contain words or colloquialisms that were perceived as words), and that all participants followed the instructions and engaged in the task to a minimum degree. This first phase of data screening eliminated 2.3% of the total data (across all participants, all words, all nonwords). Next, participants' overall performance and item performance for each item type were separately screened for outliers in speed-accuracy space using a Mahalanobis distance statistic and a p-value cut-off of .01 (Mahalanobis, 1936). This eliminated no more than 3 participants per condition (14 total). No more than 4 words were dropped from each of the word types (12 total), and no more than 18 nonwords were dropped for each nonword type (46 total). This eliminated 6.3% of the total data, and minimizes the likelihood that our effects are driven by particular items or participants (which we viewed as particularly important if we were to attempt between-condition/between-participant contrasts). Following this screening, trials with latencies lower than 200 ms or higher than 2000 ms, and trial outliers within each block that exceeded the z-score associated with a two tailed p-value of .005 were removed for each participant (2.3% of the total data), leaving 89.1% of the total data.

Preliminary analyses on the full set of items that did not control for the effects of relative meaning frequency (i.e., analyses that included only a two-level few vs. many meanings predictor) failed to find any significant homonymy effects. This was an expected outcome given that, after norming, the bulk of our homonyms were found to have a strongly dominant interpretation, and was one of the motivations for including a large number of items, so as to address this issue in the analyses. Here, we report one set of analyses wherein the relative frequency of the dominant interpretation of the homonym was used as the predictor for assessing homonymy effects, instead of simply treating number of meanings as a two-level one vs.many factor that is insensitive to meaning frequency. In these analyses, polysemes and unambiguous words were treated as words with a single completely dominant interpretation (dominant meaning frequency = 100). That is, the effects of multiple unrelated meanings were not simply analyzed as whether a word had one meaning or two or more meanings as a factorial manipulation. Rather, we treated the effects of homonymy as varying from maximally strong in the case of items with balanced meaning frequencies (dominant meaning frequency = 50 for items with two meanings) and then gradually transitioning to being completely unambiguous (i.e., dominant meaning frequencies of 100 for homonyms; recall that unambiguous words and polysemes were also assigned dominant meaning frequencies of 100 given that they only have one meaning). Thus, a significant effect of dominant meaning frequency also implies a significant effect of number of unrelated meanings (i.e., homonymy) which is affected by meaning frequency.

Strictly for simplicity, the accuracy and correct latency plots (Figures 1 and 2) were generated from the descriptive statistics from all of the unambiguous and polysemous items and for the homonyms and hybrid items with balanced meaning frequencies. A word with multiple meanings was considered to have balanced relative meaning frequencies if the dominant meaning frequency was less than 65%. A total of 14 homonyms and 22 hybrid items satisfied this constraint. However, note that the statistical analyses reported below involved the data from all of the homonyms and hybrid items, with dominant meaning frequency factor included as a continuous predictor.[8] As quantified in the analysis, if the dominant meaning frequency increased, the homonymy disadvantage decreased and homonyms and unambiguous words showed more similar response patterns. This type of effect also implies an effect of number of unrelated meanings, which is modulated by meaning frequency. Significance tests and the summary of the significant effects related to the homonyms and polysemes that are presented in the figures are based on the regression models described below, which treat dominant meaning frequency as a continuous variable and which include all of the data from all of the homonyms. In additional analyses reported elsewhere, we observed similar results if, instead of using this regression approach to control for relative meaning frequency and including the data from all of the hybrids and homonyms, we dropped all of the unbalanced items and only analyzed a subset that contained relatively balanced meaning frequencies and using a qualitative ‘one vsṁany‘ meanings factor (for details, see Armstrong, 2012; Armstrong & Plaut, 2011; Armstrong, Tokowicz, & Plaut, 2012). Thus, the results we report are not contingent upon only this analytical approach.

All of the analyses were conducted using linear mixed-effect models in R (Baayen, Davidson, & Bates, 2008; Bates, Maechler, Bolker, & Walker, 2014). Due to criticisms regarding the lack of reporting of critical model details and the complexities that can often arise from this type of modeling (Barr et al., 2013), a brief report of the most critical model details are described here.[9] To allow us to evaluate the core predictions of the SSD account, the models included crossed random effects of participant and item and, as applicable given the subset of the data being analyzed, fixed effects of dominant meaning frequency (i.e., a measure of homonymy that is also sensitive to how competition could be modulated by meaning frequency), number of senses (few vs. many), nonword difficulty, and contrast. Strictly to control for possible confounds, the models also contained the psycholinguistic properties listed in Table 1, or some transformation thereof.[10] Additionally, the models included the trial rank, lexicality, accuracy, and latency of the previous trial to eliminate auto-correlations that violate model assumptions (for discussion, see Baayen & Milin, 2010; Barr et al., 2013; Bolker et al., 2009).

Because pilot analyses indicated that, as is often the case for complex models, specifying the maximal random effects structure was not possible due to convergence issues (Barr et al., 2013), we simplified the model so that only the variables related to the critical manipulations—the nonword difficulty and contrast manipulations—were allowed to interact with the effects of meaning and sense, and only meaning and sense were included as nested random slopes for each participant to maximize the accuracy of our inferences within a difficulty manipulation. The correlation term between slope and intercept was not estimated to avoid convergence issues. Accuracy was modeled with a binomial distribution (Quené & Van den Bergh, 2008) and latency with a Gaussian distribution. In pilot omnibus analyses, positional bigram frequency and imageability did not consistently predict significant amounts of variance and so were dropped to facilitate convergence.

Because the most critical test of the SSD account relates to the presence, or lack thereof, of a homonymy disadvantage (particularly for homonyms with relatively balanced meaning frequencies) and of a polysemy advantage relative to unambiguous controls in visual lexical decision, the bulk of the analyses reported here target those comparisons specifically. Additionally, we report exploratory analyses of the hybrid items relative to the unambiguous controls. These analyses are considered exploratory because whether the hybrid items will show a processing advantage, disadvantage, or no difference relative to unambiguous controls will depend on the degree to which cooperative or competitive dynamics dominate processing. For instance, in the easier conditions, hybrids could show a greater processing advantage than polysemes because they have more related senses; however, strong activation of these related senses could be counteracted because these senses would compete across meanings more strongly. Similarly, in harder conditions hybrid items could be associated with performance similar to homonyms because of strong competition between meanings; or, the strong advantage accumulated through cooperation may persist longer than in the case of polysemes and they could still show a processing advantage even after the polysemy advantage has dissipated. The analysis of the hybrid items may therefore provide constraints with respect to the relative strength of each of these dynamics. Recall, however, that the original population of hybrid items precluded controlling for the other psycholinguistic properties to the same degree as the other word types. All significant effects have a p-value less than .05 and all marginal effects have a p-value less than 0.1, as calculated by appropriate functions in R (Kuznetsova, Brockhoff, & Christensen, 2014).

Because the model specifically predicts the direction of the polysemy and homonymy effects, those tests were one-tailed.

Accuracy. A summary of the accuracy data and the results of pairwise tests between the unambiguous words and the homonyms and polysemes are presented in Figure 1. Overall, this figure shows that average accuracy levels remained relatively constant across conditions and were near the 90% accuracy threshold that was specified in the instructions and reinforced via feedback. All of the conditions also showed that the hybrid words exhibited performance that was quite similar to the polysemes, consistent with the results reported by Rodd et al. (2002). This effect diminishes the informativeness of omnibus meaning × sense interactions because they do not precisely test our predictions. For this reason, we focused our analyses on the critical pairwise comparisons.

– – – – – – –

Insert Figure 1 Here

– – – – – – –

With respect to the homonyms, the SSD account predicts that, relative to standard lexical decision performance in which homonyms tend to pattern like unambiguous words, more difficult conditions associated with slower responses would generate a homonymy disadvantage. We evaluated this hypothesis in two main sets of analyses. The first set consisted of within-participants tests for a homonymy disadvantage within each difficulty condition separately using relative meaning frequency as the predictor to make the test of homonymy sensitive to how competition could be modulated by frequency. Thus, when we report a significant homonymy advantage, it is implied that this homonymy disadvantage is shaded my relative meaning frequency. The second set tested for an interaction between the magnitude of the homonymy disadvantage (again, shaded by meaning frequency) and difficulty in a mixed within-and-between participants design spanning two difficulty levels.

We begin by reporting the results of the separate analyses of each condition. In the full contrast–hard nonword condition, only a marginal homonymy disadvantage was observed relative to the unambiguous words ($b = 0.0046$, $SE = 0.0029$, $p = .055$). This is consistent with the standard finding of a weak or absent homonymy disadvantage in lexical decision. However, a significant homonymy disadvantage was detected in all of the other conditions[11] (very hard-full: $b = 0.014$, $SE = 0.004$, $p < .001$; hard-degraded: $b = 0.011$, $SE = 0.004$, $p = .002$; very hard-degraded: $b = 0.014$, $SE = 0.004$, $p = .001$; pseudohomophone-degraded: $b = 0.012$, $SE = 0.004$, $p = .005$), except for the full contrast pseudohomophone condition, in which the effect was marginal ($b = 0.009$, $SE = 0.006$, $p = .07$).

Next, we tested to see if the magnitude of the homonymy disadvantage increased as a function of difficulty, as would be reflected in an interaction between homonymy and condition. These tests were broken down into two subsets. In the first subset, the data were partitioned based on stimulus contrast and within each partition the hard nonword condition was compared to the very hard nonword and the pseudohomophone conditions. In the second subset, the data were partitioned based on nonword difficulty and the full and degraded conditions were compared separately for each level of nonword difficulty. All of these tests were non-significant (p's $\geq .19$), and so are not reported in detail, with the exception of a significant increase in the homonymy disadvantage between the hard to very hard conditions in the full contrast condition ($b = 0.0070$, $SE = 0.0038$, $p = .03$). Note, that weaker effects were expected in these comparisons because they involved a between-participant comparison.

We repeated a similar analytical protocol for the polysemes. According to the SSD account, ideally the polysemes should show less or no advantage relative to unambiguous words, depending on exactly how difficult the task is. We first tested for the presence of a polysemy advantage within each condition. The polysemy advantage was significant in all of the full contrast conditions (hard-full: $b = 0.21$, $SE = 0.11$, $p = .024$; very hard-full: $b = 0.26$, *SE* $= 0.11, p < .01$; pseudohomophone-full: $b = 0.57$, *SE* $= 0.14$, $p < .001$) but only for very hard nonwords in the degraded condition ($b = 0.37$, *SE* $= 0.11$, $p = .001$; all other *p*'s $\geq .15$).

Next, we tested to see if the magnitude of the polysemy advantage decreased as a function of difficulty, as would be reflected in an interaction between number of senses and condition, again, using the same partitioning approach described for the homonyms. The results showed that the polysemy advantage decreased between the pseudohomophone-full contrast and pseudohomophone-degraded contrast conditions ($b = -0.30$, *SE* $= 0.12$, $p = .007$). However, The polysemy advantage increased between the hard-full and pseudohomophone-full contrast conditions ($b = 0.23$, *SE* $= 0.12$, $p = .02$) and increased between the hard-degraded and very-hard degraded conditions ($b = 0.20$, *SE* $= 0.10$, $p = .03$). All other comparisons were non-significant (all *p*'s $\geq .13$).

Finally, we explored how performance for the hybrid items compared to the unambiguous controls. In this case, we did not make specific a priori predictions regarding whether an advantage or a disadvantage would be observed, because either pattern could emerge depending on the relative strength of the cooperative and competitive dynamics. The results of this comparison can therefore serve to gain insight into the relative strength of each of these dynamics at different points during processing. An examination of Figure 1 suggests that the hybrid items tended to pattern more like the polysemes than with the homonyms. This is consistent with the notion that the strong cooperative dynamics related to the shared features of the hybrid items are dominating the overall processing dynamics. To evaluate this possibility quantitatively, we contrasted the unambiguous controls and the hybrid items in regression analyses that included main effects of homonymy (as assessed using relative meaning frequency as a predictor) and number of senses. As before, one set of analyses was run separately for each condition and another set of analyses compared performance across difficulty conditions. In the separate analyses of each condition, the analyses never showed a significant effect of number of senses (all *p*'s $> .13$). These analysis also only showed a significant effect of meaning frequency in the very hard-full contrast condition ($b = -0.012$, *SE* $= 0.006$, $p = .05$) and in the pseudohomophone-full contrast condition ($b = -0.017$, $SE = 0.007$, $p = .02$). In both of those cases, however, the relationship between meaning frequency and accuracy was the opposite of that observed for homonyms—accuracy increased as the meaning frequencies became less balanced for the hybrid items, whereas for the homonyms, accuracy decreased as the meaning frequencies became less balanced. In the analyses testing for an interaction between difficulty and meaning frequency or number of senses, only two of the interactions reached significance (all other p's > 0.1). Both of these exceptions were observed when comparing the hard nonword condition to the pseudohomophone condition under full contrast. The first of these interactions indicated that there was an increased disadvantage for hybrid items with more balanced meaning frequencies as difficulty increased ($b = -0.014$, $SE = 0.005$, $p = .01$). The second of these interactions indicated that there was an increased disadvantage for words with many senses as difficulty increased ($b = -0.35$, $SE = 0.14$, $p = .02$).

Correct latency. Mean latency data from correct trials, as well as the results of pairwise tests between unambiguous words and the homonyms and polysemes, are presented in Figure 2. In general, these results show small increases in overall latencies as a function of nonword difficulty, and substantial overall latency increases as a function of contrast. These effects are more pronounced for the nonwords than for the words.[12]

– – – – – – –

Insert Figure 2 Here

– – – – – – –

The same analytical protocol used for the accuracy data was followed in analyzing the latency data. With respect to the detection of homonymy effects within each condition, as was the case for the accuracy data, there was only a marginal homonymy disadvantage in the full contrast–hard nonword condition ($b = -0.35$, $SE = 0.25$, $p = .09$). However, there was a significant homonymy disadvantage in all of the other conditions (very hard-full: $b = -0.33$, $SE = 0.17$, $p = .03$; pseudohomophone-full: $b = -0.56$, $SE = 1.9$, $p < .001$; hard-degraded: $b = -0.49$, $SE = 0.17$, $p < .001$; very hard-degraded: $b = -0.45$, $SE = 0.20$, $p = .01$; pseudohomophone-degraded: $b = -0.37$, $SE = 0.21, p = .04$). In the between-condition tests for interactions between homonomy and difficulty, no significant effects were detected, although there was a marginal trend for a larger homonymy disadvantage in the degraded condition relative to the full contrast condition in the context of hard nonwords ($b = -0.21$, $SE = 0.16$, $p = .09$), as well as in the pseudohomophone condition relative to the hard nonword condition under full contrast ($b = -0.23$, $SE = 1.59$, $p = .08$; all other p's $> .23$).

With respect to polysemy, a significant advantage was detected in all of the within-condition analyses (hard-full: $b = -9.8$, $SE = 3.5$, $p = .002$; very hard-full: $b = -11.5, SE = 3.5$, $p < .001$; pseudohomophone-full: $b = -12.2$, $SE = 4.4$, $p = .003$; very hard-degraded: $b = -11.7$, $SE = 3.8$, $p < .001$; pseudohomophone-degraded: $b = -10.4$, $SE = 4.4$, $p = .009$), except the hard-degraded condition ($p = .16$). The tests for interactions between the magnitude of the polysemy advantage and difficulty indicated that the polysemy advantage marginally decreased between the full contrast and degraded contrast condition in the context of hard nonwords ($b = 4.9$, $SE = 2.9$, $p = .05$), and significantly increased between the hard and very hard nonwords under degraded contrast ($b = -7.6$, $SE = 3.3$, $p = .01$). All of the other comparisons were non-significant (p's $\geq .13$).

Finally, we explored how the hybrid items patterned relative to the unambiguous controls. As in the accuracy data, an examination of Figure 2 showed that latencies for the hybrid items tended to group more with the polysemes than with the homonyms. In identical statistical analyses to those conducted on the accuracy data, which compared the hybrid items and the unambiguous controls, the within-condition analyses indicated that in no case was the effect of meaning frequency significant (all p's $> .35$), nor was the effect of number of senses (all p's $> .20$). The tests for interactions between conditions revealed a marginal effect of meaning frequency, such that hybrid items with more balanced meaning frequencies were responded to marginally more slowly in the pseudohomophone condition than in the hard nonword condition when stimuli were presented at full contrast ($b = 0.35$, $SE = 0.18$, $p = .053$; all other p's ≥ 0.14).

Summary of results. The within-condition analyses were largely consistent with the predictions of the SSD account. With respect to the homonyms, in the easiest condition (hard-full) which serves as our baseline, no homonymy disadvantage was detected in either accuracy or latency, although there were marginal trends in that direction. However, a significant homonymy disadvantage was detected in the harder difficulty conditions. This disadvantage was always significant in the latency analyses, and significant in all but the pseudohomophone-full condition in the accuracy analyses. With respect to the polysemes, there was a significant polysemy advantage in both accuracy and latency. This advantage remained significant in all of the full contrast conditions, which were also associated with small increases in overall latency relative to the baseline condition. In the degraded contrast conditions, in which latencies were considerably longer, however, the latency analyses only showed evidence for a polysemy advantage in the case of pseudohomophones and very hard nonwords, and the accuracy data only showed evidence for a polysemy advantage in the case of very hard nonwords. Thus, the polysemy advantage appears to have dissipated somewhat as difficulty increased.

The between-condition analyses rarely reached significance, which is at least partially due to the between-participants nature of the comparisons. Nevertheless, with respect to the homonyms, when effects were detected they showed that the homonymy disadvantage increased as difficulty increased. This increase was significant in between hard and very hard nonwords under full contrast in accuracy, and was marginal in the latency data between the degraded contrast and full contrast conditions involving hard nonwords, as well as between the pseudohomophone and hard nonword conditions under full contrast. In the case of the polysemes, the effects were also rarely significant and the effects that were detected were somewhat mixed. The polysemy advantage significantly decreased between the pseudohomophone-full and pseudohomophone-degraded conditions in accuracy, and marginally decreased between the hard-full and hard-degraded conditions in latency. However, the polysemy advantage significantly increased between the hard-full and pseudohomophone-full conditions and the hard-degraded and very hard-degraded conditions in accuracy, as well as between the hard-degraded and very hard-degraded conditions in latency. Finally, in general, the hybrid items performed more similarly to the polysemes than to the homonyms.

Discussion

The SSD account predicts that varying processing time—and the amount of semantic settling that has taken place—should give rise to different ambiguity effects, even when the task and the basic configuration of the response system are held constant. The experiment provided an empirical platform for evaluating the specific predictions of the account regarding the timecourse of semantic settling via two manipulations of task difficulty. The results of the experiment showed that stimulus degradation substantially increased overall latencies and nonword difficulty slightly increased overall latencies. These latency increases were also generally associated with the emergence of a homonymy disadvantage and a reduced sense advantage, particularly in the case the contrast manipulation, as shown in the within-condition analyses. The results of the experiment are thus broadly consistent with the specific predictions of the SSD account with respect to the detection of a weak or absent homonymy disadvantage in standard (relatively) easy/fast lexical decision, which becomes stronger in harder/slower difficulty conditions; and with a strong polysemy advantage in standard lexical decision which becomes weaker or disappears in harder/slower difficulty conditions.

The results also call into question the need to attribute the different ambiguity effects that have been observed in different tasks to qualitative differences in the configuration of the response system (Hino et al., 2006) without denying that some adaptation of the response system could occur. In that study, the authors reported no ambiguity effects in their fastest experiments, which concerned semantic categorizations related to narrow categories (Experiment 3, unambiguous control latencies: 566 ms; filler ‘no’ response latencies: 583 ms, average: = 575 ms;

Experiment 4, unambiguous control latencies: 559 ms; filler ‘no’ response latencies: 542 ms; average: 551 ms). In their slightly slower experiment, lexical decision, they reported an advantage for hybrids and polysemes (Experiment 1, unambiguous controls: 563 ms; filler nonword ‘no’ responses: 614 ms; average = 575 ms). In their slowest conditions, which concerned semantic categorizations related to broad categories, they only observed a hybrid disadvantage (semantic categorizations with response latencies in the 650-800 ms range). Taking only the mean response latencies from that article into account[13], these data are in line with the overall predictions of the SSD account: If responses are made very early, before semantic activity has built up to a level that is substantial enough for the effects of within-semantic cooperative dynamics to build up, no ambiguity effects are predicted. Shortly thereafter, an advantage is predicted for polysemes, and could also manifest itself for hybrid items if the cooperative dynamics are still the only major force at play. Much later, the polysemy advantage would disappear as the unambiguous word activity increases and as competition between the idiosyncratic features increases.[14]At that point, after competition has dominated for an extended period of time, only a disadvantage for homonyms and hybrid items should remain.

Of course, there are some potential issues with this interpretation which remain to be fleshed out. For example, this account assumes that the very fast semantic categorizations could be made based on a minimal amount of semantic activation, prior to the onset of strong cooperative dynamics, whereas the lexical decision results would have been influenced by slightly later semantic dynamics. This may, superficially, be an unusual prediction, but it is one that could be evaluated empirically by comparing the magnitude of other semantic effects (e.g., imageability, concreteness) across tasks and conditions. These patterns also raise questions about why our within-task manipulations were less effective at altering the overall patterns of results, a point which we elaborate on below. Furthermore, additional work will be needed to help constrain the exact patterning of hybrid items over time to validate the assumed temporal settling dynamics for this understudied set of items, although as we noted in the methods, the relatively small population of such items makes doing so with tightly-controlled sets of items difficult.

Setting these considerations aside, there are, however, also several general issues with the interpretation of the Hino et al. (2006) data that remain to be sorted out in detail as well. For instance, why is it that some semantic categorization responses, the fact that they involved narrow categories notwithstanding, could be made more quickly than lexical decisions, which at most, requires an evaluation of whether a stimulus evokes any coherent meaning whatsoever? One possibility is that participants are responding much more slowly than necessary in both tasks, as has been reported previously (for discussion, see Plaut & Booth, 2000). It would also permit a relatively early advantage for basic category information to emerge from the temporal processing in semantics (Rogers & Patterson, 2007). Such a basic category advantage could facilitate responding in semantic categorization tasks related to narrow categories. In principle, these issues could be examined by using methods that encourage participants to respond more efficiently (i.e., more quickly, without sacrificing accuracy), and this past work offers some possibilities on that front. The present results also suggest that comparisons between experiments that also involve different participants, as was the case in the Hino et al. studies, may be more noisy and underpowered. Thus, strong claims regarding between-experiment differences in mean latency, which were relatively small between the lexical decision and semantic categorization tasks involving narrow categories, should be interpreted with caution (and more caution still is needed in comparing our absolute latencies with theirs).

More broadly, the results of our experiment also challenge some basic assumptions about how lexical processing unfolds and is tapped by the response system in different contexts. For instance, staged models of written word comprehension (e.g., Borowsky & Besner, 1993, 2006) predict that stimulus degradation would extend visual/orthographic processing but not alter the amount of semantic processing that has taken place (for related work, see Yap, Lim, & Pexman, 2015). Interpreted within the context of the SSD account, our results suggest that the majority of the slow-down that has occurred in processing the degraded items occurred outside of semantics, given that the large latency increases following stimulus degradation were associated primarily with only small modulations in the semantic ambiguity effects relative to other tasks such as semantic categorization, which are associated with similar overall latencies (Hino et al., 2006). Nevertheless, our observation of modulation of ambiguity effects as a function of stimulus contrast challenge such an model and support a (possibly weakly and somewhat stage-like) cascaded theory of lexical processing. Although such a result might be argued to imply that information does not flow quite as strongly between levels of representations as implied in classic models(e.g., McClelland & Rumelhart, 1981), these results are nevertheless more in keeping with standard processing assumptions in neural/connectionist networks and run contrary to the predictions of a formal staged account (for related discussion, see Plaut & Booth, 2000; see also the discussion of temporal modularity in Carreiras et al., 2014). According to such a mechanistic account of our results, a large part (but not all) of our slow-down due to stimulus contrast would be attributable to slower early visual/orthographic processing as well, prior to the onset of semantic processing. This would also help explain why the modulation of our semantic effects, holding task constant and manipulating stimulus contrast, led to smaller changes in the overall pattern of semantic effects than a qualitative alteration of the task itself which has yielded similar overall increases in latencies (on the order of 100 ms, e.g., Hino, Lupker, & Pexman, 2002). Across those tasks, perceptual-to-semantic access should occur at the same rate, so the entire > 100 ms latency increase could be allowing for different semantic effects to emerge. In our experiment, however, a substantial portion of this latency increase is likely attributable to slower processing that is largely restricted to pre-semantic representations. Thus, our > 100 ms increases in overall latencies following the stimulus contrast manipulations should be associated with much smaller increases in the total amount of semantic processing in particular. Our stimulus contrast manipulation should therefore produce a weaker overall modulation of the semantic effects, despite similar overall increases in latency as in some previous between-task manipulations, which is what we observed.

Other manipulations, such as the use of pseudohomophone foils, were less effective than the literature might suggest. Explaining these results raises an interesting set of computational questions. In particular, the pseudohomophones were also associated with the largest increase in both word and nonword latencies but did not produce semantic effects consistent with more resolved semantic representations. Both of these effects are difficult to reconcile with an account based strictly on semantic settling dynamics but would fall out naturally from an account that involved the response selection system. Indeed, because pseudohomophones are able to engage specific semantic representations via phonology, the response system may learn to de-emphasize (although not eliminate) evidence from semantics because it is less informative in separating words from nonwords. Thus, the longer latencies in that condition may, in fact, be due to the response system needing to wait for additional non-semantic information, such as precise orthographic information, to be resolved (Hino & Lupker, 1996). A similar principle may underlie the results reported by Hino et al. (2010), in which Kanji nonwords composed of characters with meanings similar to the words weakened some ambiguity effects despite increasing overall latencies. Our other types of nonwords, which were highly wordlike in terms of bigram frequency and neighborhood size but that were less likely to engage a specific semantic representation, may therefore be better suited for eliciting strong semantic ambiguity effects.

This raises important questions for future research regarding the optimal types of nonwords that should be used to elicit powerful semantic (or, in other circumstances, non-semantic) effects using a lexical decision paradigm. It is well established that nonwords that are not very wordlike in terms of their surface features (e.g., illegal consonant strings, nonwords with very low bigram frequencies or small neighborhoods) are typically associated with small or no semantic effects and very fast overall latencies [15](and neural correlates of semantic activity) (e.g., Azuma & Van Orden, 1997; Evans, Ralph, & Woollams, 2012; Laszlo & Federmeier, 2011; Pexman & Lupker, 1999; Rodd et al., 2002)). From that lower bound, however, our results suggest that different properties of the nonwords can be adjusted to increase overall wordlikeness in ways that either emphasize or de-emphasize semantic contributions. Nonwords that are very wordlike in terms of their surface properties (legal bigrams, high bigram frequencies, large neighborhood sizes) minimize surface-level distinctions between words and nonwords in terms of, for instance, orthographic representations. These items do not, however, engage specific semantic representations, leaving the state of semantics as a strong indicator of whether a presented item was a word or a nonword. In contrast, pseudohomophones, for the reasons outlined above, will engage specific semantic representations via phonology, reducing (although not eliminating, due to inconsistencies in the orthographic-semantic mapping) the informativeness of differences in semantic activity relative to other sources of evidence (e.g., engagement of specific, previously-learned orthographic representations) in delineating between words and nonwords. Thus, even if pseudohomophones are more wordlike overall than our orthographically very wordlike nonwords (and lead to larger increases in overall latencies), leading to stronger semantic effects than nonwords that are not very wordlike, they may be less optimal for causing semantic factors to drive responding, as was desired in this particular investigation.

Nevertheless, even our other types of nonwords are not the perfect means of increasing overall latencies and semantic effects that we sought because increasing nonword difficulty had a markedly larger effect on nonword latencies than word latencies. Thus, although we set out with the aim of varying response times holding task constant and did observe effects consistent with the predictions of the SSD account, this evidence for response-specific slowing suggests the need for more integrative theories and explicit models that flesh out the specific contributions of the lexical system (including how orthographic, phonological, and semantic representations interact) and the response system to fully understand our and other researchers' results (for discussion, see Armstrong et al., 2009; Armstrong & Plaut, 2013). This is a massive undertaking, but is likely needed to fully explain all of the nuances we have observed and that have been reported elsewhere in the literature.

The tendency for the hybrid items to pattern like the polysemous items also provides interesting insight and constraint for the SSD theory, while at the same time agreeing with the results of the lexical decision study reported by Hino et al. (2006) if their homonymous items are re-interpreted as being hybrid ambiguous items, as discussed in the introduction (a similar explanation might also apply to the overall ambiguity advantage reported by Pexman et al. (2004) if their high Number of Meaning items were hybrid items and not pure homonyms). Indeed, such a re-interpretation offers a means of understanding why, in several papers by that research group, an overall ambiguity advantage is reported in lexical decision (Hino et al., 2006; Pexman et al., 2004), which may be driven by a combined polyseme/hybrid advantage; whereas in many other studies, the advantage is restricted to polysemes and homonyms are associated with a processing disadvantage or no difference from unambiguous controls (e.g., Armstrong & Plaut, 2008; Beretta et al., 2005; Klepousniotou et al., 2008; Rodd et al., 2002). It is also consistent with the re-analyses of the English Lexicon Project data reported by Armstrong, Tokowicz, and Plaut (2012). In contrast to a previous re-analysis of a smaller set of homonyms by Hargreaves, Pexman, Pittman, and Goodyear (2011), who reported a homonymy advantage (although they did not distinguish between homonym and hybrid items as we do here), Armstrong and colleagues found at best very weak evidence for a homonymy advantage specifically. In particular, they failed to replicate a significant ambiguity advantage when the data for 397 (non-hybrid) homonyms with relatively few senses were re-analyzed (nearly an order of magnitude more homonyms than are used in standard behavioral studies) and included standard psycholinguistic variables (e.g., frequency, length) as covariates. Translating these results into slightly more mechanistic terms, these data suggest that, even during (somewhat) later processing, cooperative/excitatory dynamics can still dominate competitive/inhibitory dynamics, which may relate to and provided added justification for building more biologically plausible models, particularly with respect to excitatory and inhibitory dynamics (e.g., Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012). Or, alternatively, they could indicate a potential weakness of the settling dynamics account. Before making strong claims in this regard, additional normative data related to the number and frequency of exposure to each interpretation of a hybrid item is needed to better understand how absolute interpretation frequency, relative meaning frequency, and the number and relatedness of a word's semantic features interact—particularly given that in a couple of occasions the effects of number of interpretations and relatedness of interpretations were inconsistent with those typically associated with homonymy and polysemy both in our other conditions and in other studies. More targeted experiments of this understudied type of item that involve more optimally matched sets of hybrid items may be especially informative in fleshing out these effects—especially if conducted in other languages in which the pool of such items is greater. These experiments, if coupled with more powerful manipulations of how much semantic processing takes place between conditions (such as those outlined earlier in the discussion) could also help establish how cooperative and competitive dynamics trade off for polysemous and hybrid items when very precise semantic codes are engaged (beyond those we believe to have tapped in our current investigations of within-task manipulations).

Taken together, the present results call for a refinement and a tighter integration of both theories of lexical processing and theories of response selection, over and above the descriptions that we offered in the introduction. Although the SSD account was originally developed as a direct alternative to an account based on qualitative differences in how decisions are made across tasks—and maintaining the present claim that semantic settling dynamics explain a large portion of the observed literature—it is clearly an oversimplification to ignore the contributions of the response system in determining how and why a response is made. The development of such an integrative account would not only allow for more direct contrasts of where the relevant dynamics that underlie a given empirical phenomenon are occurring, but also for a more compelling understanding of these systems as an interactive network. For instance, it remains to be formalized in explicit computational terms how it is determined when sufficient information—semantic or otherwise—has been accumulated to reliably make a response in a given task—or indeed, within different conditions of a single task, as in the study reported here.

Presenting a more comprehensive model of specific tasks that includes both a lexical system and a response system is not a trivial undertaking, however, given the contention surrounding how the response system should be integrated with the lower-level systems that provide evidence to it, how the response system learns to monitor different types of evidence as a function of experience with a given task, and how evidence that accumulates in different representation (e.g., orthography, semantics) at different rates should drive the response system (for discussion, see Armstrong & Plaut, 2013; Ratcliff, 1980; Ratcliff, Van Zandt, & McKoon, 1999; Usher &

McClelland, 2001). The empirical data, and in particular, performance in the pseudohomophone condition, also suggest that semantics, the response system, and other representations such as orthography and phonology interact in highly complex ways that involve different contributions from cooperative and competitive dynamics over time. More sophisticated modeling involving a full orthographic-phonological-semantic lexical processing model and the response system will clearly be needed to fully understand these data.

Finally, it must be acknowledged that, although the within-condition analyses were generally supportive of the predictions of the SSD account, there is some weakness in the empirical support for these conclusions: the predicted between-condition interactions often did not reach significance, and the polysemy effects appeared to be less modulated by the difficulty manipulations than the homonymy effects. Thus, although our results generally support the SSD account and undermine a strong account of the data based on a qualitative difference in the response system across tasks (given that overall task was held constant), our results, at face value, would not appear to be perfectly consistent with the SSD account, either. These not perfectly consistent findings may be due to several factors, all of which have both theoretical and methodological implications for future studies of semantic ambiguity. First, our easiest condition itself may, nevertheless, have been a relatively difficult version of lexical decision, falling near the border between the standard pattern of effects observed in lexical decision (strong polysemy advantage, weak or absent homonymy disadvantage) and the “late” semantic effects that we predicted in more difficult conditions. By moving the baseline condition closer to the more difficult conditions, the detection of strong interactions between different difficulty conditions may have been reduced. Second, the detection of these interactions was clearly more difficult because of the between-participants nature of the comparison, which was motivated by pilot studies that showed strong order effects depending on which difficulty level participants were exposed to first, as well as a general homogenization of performance across difficulty levels. Third, based on our analyses of items used in prior semantic ambiguity studies that did not explicitly control for meaning frequency (e.g., Beretta et al., 2005; Rodd et al., 2002), we expected that the number of relatively balanced homonyms would be somewhat larger in our sample, which would have increased the power and reliability of the homonymy effects. Finally, there may be an upper limit on how much lexical decision can be made to rely on higher-level semantic representations in driving the response system. Although semantic activation clearly contributes evidence that can discriminate between words and nonwords (see also Plaut, 1997), orthographic and phonological representations can obviously be used as a source of evidence to discriminate words from nonwords, as well. In visual lexical decisions, orthographic representations also benefit from receiving feedforward activation from the sensory system first and from needing to be resolved to a sufficiently precise degree to activate a relatively precise semantic representation. Collectively, this could allow for the accumulation of sufficiently discriminating non-semantic information to make lexical decisions before a reasonably strong amount of semantic activity, such as that expected to underlie a strong homonymy disadvantage in the absence of a polysemy advantage, can accumulate.

With these considerations in mind, several specific guidelines can be derived for future investigations. First, the results of the present work clearly illustrate the importance of carefully controlling for meaning frequency to maximize the magnitude of homonymy effects. The recent availability of large-scale norms of meaning frequency should make such careful control possible (Armstrong, Tokowicz, & Plaut, 2012; Armstrong et al., 2015), particularly when complemented by stimulus selection methods that control for many psycholinguistic properties on an item level (Armstrong, Watson, & Plaut, 2012). Second, it may be desirable to explore even more extreme difficulty manipulations that compare very easy versions of a given task (e.g., lexical decision with nonwords that are not very wordlike) with the current difficult version of lexical decision to boost the power of the ambiguity by difficulty interactions. Third, it is clearly desirable, in an ideal world, to hold not only items but also participants constant in comparing different ambiguity effects within a given task, particularly if there are some upper and lower limits on how easy or hard a given task can be. We originally ran a pilot version of the difficulty manipulation which showed strong order effects and homogenization of overall performance across different difficulty levels, which motivated our current between-participants design. However, in recent work, Armstrong, Barrio Abad, and Samuel (2014) showed that having participants complete multiple (~14) alternating blocks of different versions of an auditory versus visual task was able to minimize those issues and show significant interactions between task and a semantic variable—imageability—in mixed-effect regression analyses. Balancing holding items and participants constant using this type of design may therefore be a powerful general strategy for comparing different difficulty manipulations in experimental settings.

Relatedly, in post hoc analyses of the Armstrong et al. (2014) data set, which was found to contain similar numbers of homonyms and polysemes to many prior studies (e.g., Hino et al., 2010; Mirman et al., 2010), we were able to detect several significant interactions between number of meanings, number of senses, and modality of stimulus presentation, which also correlated with overall latency of responses—auditory lexical decisions were made approximately 300 ms slower than visual lexical decisions, and were associated with stronger homonymy effects and weaker polysemy effects. Although future experimental work with the a priori aim of fleshing out ambiguity effects in particular is clearly needed within this paradigm, these results parallel those observed by Rodd et al. (2002) in their auditory and visual lexical decision experiments, respectively. Specifically, using the unambiguous control, homonym, and polyseme nomenclature adopted in the present work, for visual lexical decision (Experiment 2) Rodd and colleagues reported unambiguous control latencies of 586 ms, homonym latencies of 587 ms, and polyseme latencies of 567 ms. There was therefore a 19 ms polyseme advantage and a 1 ms homonymy disadvantage. In comparison, in auditory lexical decision (Experiment 3) they reported unambiguous control latencies of 939 ms, homonym latencies of 986 ms, and polyseme latencies of 924 ms. There was therefore a 15 ms advantage for polysemes and a 47 ms disadvantage for homonyms. Numerically at least, these effects are in line with our predictions. The homonym disadvantage in particular is likely to be significantly larger in auditory lexical decision, and at the very least, these data provide no support for an increased polysemy advantage. Collectively, these findings further reinforce the notion that investigations of other tasks may be useful for building a convergent set of evidence for the SSD account (see also Mirman et al., 2010). Interestingly, one of the reasons why such clear comparisons cannot be made using the data reported by Rodd and colleagues is that non-identical sets of items were used in their auditory versus visual experiments. The use of these only partially overlapping sets was clearly necessary in that work because—among other reasons—the orthographic and phonological neighborhoods of English words can be quite different from one another and a well controlled set of items for a visual experiment may include substantial confounds if used in an auditory experiment. These problems, however, can be minimized if, like in the Armstrong et al. (2014) study, a transparent language such as Spanish is used. Additionally, given that some of the inconsistent results obtained in lexical decision have been reported in different languages (e.g., Hino et al., 2010), it is clearly valuable to assess the generality of the different ambiguity effects across languages and avoid developing general theories of ambiguous word comprehension that may, in fact, be tied to specific properties of a given language (for discussion, see Lerner, Armstrong, & Frost, 2014; Share, 2008; Frost, 2012). Insofar as these findings can be replicated in tightly controlled experimental settings, this would have critical implications and motivate new computational work studying the timecourse with which information flows not only within semantics, but within and between orthography and phonology as well.

More broadly, it may also be worth considering how other types of tasks can be used to explore the timecourse of semantic ambiguity effects. Two avenues that appear particularly worthy of investigation are contrasting go versus no-go versions of a task and tasks with explicit response deadline requirements. In both cases, such investigations would essentially explore how explicitly altering when a participant responds alters performance without modifying other core aspects of the task. With respect to go/no-go tasks, Siakaluk, Pexman, Sears, and Owen (2007) has previously shown that homophone effects are modulated by this manipulation in the context of semantic categorization. Considered in conjunction with the explicit computational modeling of yes/no lexical decision and go/no go lexical decision (Perea, Rosa, & Gómez, 2002), these results lend credibility to the notion that semantic ambiguity effects could be manipulated in this type of task and that the main difference between the conditions would be how much evidence accumulates before responding (see also Tamminen, Cleland, Quinlan, & Gaskell, 2006). Similarly, Rogers and Patterson (2007) have reported effects consistent with the earlier versus later semantic settling dynamics of a connectionist model of semantic memory in a semantic categorization task with explicit deadline requirements, although other models and data raise questions as to how general and reliable such explicit deadline methods are for tapping earlier processing dynamics versus simply speeding up the dynamics themselves [16] (Kello & Plaut, 2000, 2003). Again, studies in other languages than English, such as Japanese (e.g., Hino et al., 2006), may be particularly valuable to this end, because the relatively small number of homonyms in English makes selecting large groups of well controlled homonyms difficult (but not necessarily impossible; Hargreaves et al., 2011). In these and other possible investigations, it would also be desirable to collect convergent measures of neural correlates of processing as well to gain additional insight into the involvement of brain regions involved in semantic processing and response selection at different points in time (e.g., Beretta et al., 2005; Hargreaves et al., 2011; Pylkkänen et al., 2006; Taler et al., 2009).

Conclusion

The SSD account represents a novel alternative to accounts of semantic ambiguity effects based on the configuration of the decision system. Although this account is clearly still in its infancy, by being based in the domain-general principles of the connectionist framework—particularly with respect to how cooperative and competitive dynamics unfold over time—it holds considerable promise in providing a comprehensive account of semantic ambiguity effects, as well as connecting with a broad set of phenomena in other related domains, such as the temporal dynamics associated with orthographic and phonological processing. In the current work, we tested several of the core predictions of the account and found effects that were largely—albeit, not perfectly—consistent with its predictions. The results also help motivate several refinements to the theory—including the importance of integrating lexical processing with a response system to understand and simulate in explicit computational terms how these effects emerge in specific tasks. The findings also help to identify a number of targeted methodological improvements and related tasks that can be used to further inform theories of semantic ambiguity resolution. The SSD account should thus continue to make valuable contributions to our understanding of these issues and other related issues in future work as it matures.

Acknowledgements

A portion of the work presented in this paper formed part of BCA's doctoral thesis in the Department of Psychology at Carnegie Mellon University and the Graduate Training Program in the Center for the Neural Basis of Cognition, and a preliminary analysis of a partial data set was reported in Armstrong and Plaut (2011). We thank Marlene Behrmann, Brian MacWhinney, and Natasha Tokowicz for useful feedback throughout the development of this project, and Yevdokiya Yermolayeva for comments on a draft of this manuscript. Funding was provided by an NSERC CGS and a Marie Curie IIF awarded to BCA, and by Severo Ochoa program grant SEV-2015-049.

References

Armstrong, B. C. (2012). *The Temporal Dynamics of Word Comprehension and Response Selection: Computational and Behavioral Studies* (Unpublished doctoral dissertation). Carnegie Mellon University Psychology Department, Pittsburgh, PA.

Armstrong, B. C., Barrio Abad, E., & Samuel, A. (2014). *Cascaded vs. Stage-like Semantic Access in Spoken and Written Word Recognition: Insights from Lexical Decision.* Poster presented at the Annual Conference of the Psychonomic Society.

Armstrong, B. C., Joordens, S., & Plaut, D. C. (2009). Yoked criteria shifts in decision system adaptation: Computational and behavioral investigations. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the* 31*st* *Annual Conference of the Cognitive Science Society* (pp. 2130–2135). Austin, TX: Cognitive Science Society.

Armstrong, B. C., & Plaut, D. C. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the* 30*th* *Annual Conference of the Cognitive Science Society* (pp. 273–278). Austin, TX: Cognitive Science Society.

Armstrong, B. C., & Plaut, D. C. (2011). Inducing homonymy effects via stimulus quality and (not) nonword difficulty: Implications for models of semantic ambiguity and word recognition. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the* 33*rd* *Annual Conference of the Cognitive Science Society* (pp. 2223–2228). Austin, TX: Cognitive Science Society.

Armstrong, B. C., & Plaut, D. C. (2013). Simulating overall and trial-by-trial effects in response selection with a biologically-plausible connectionist network. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the* 35*th* *Annual Conference of the Cognitive Science Society* (pp. 139–144). Austin, TX: Cognitive Science Society.

Armstrong, B. C., & Plaut, D. C. (under review). Semantic Ambiguity Effects in Lexical Processing.

Armstrong, B. C., Tokowicz, N., & Plaut, D. C. (2012). eDom: Norming software and relative meaning frequencies for 544 English homonyms. *Behavior Research Methods*, *44*, 1015-1027. doi: 10.3758/s13428-012-0199-8

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*, 675–705. doi: 10.3758/s13428-011-0182-9

Armstrong, B. C., Zugarramurdi, C., Cabana, A., Valle Lisboa, J., & 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. doi: 10.3758/s13428-015-0639-3

Azuma, T., & Van Orden, G. C. (1997). Why SAFE is better than FAST: The relatedness of a word's meanings affects lexical decision times. *Journal of Memory and Language*, *36*(4), 484–504. doi: doi:10.1006/jmla.1997.2502

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi: doi:10.1016/j.jml.2007.12.005

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi: doi:10.1016/j.jml.2012.11.0011

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). *lme4: Linear mixed-effects models using eigen and s4.* Retrieved from `http://arxiv.org/abs/1406.5823` (ArXiv e-print; submitted to *Journal of Statistical Software*)

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. doi: doi:10.1016/j.cogbrainres.2004.12.006

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., R., P. J., Stevens, M. H., & White, J. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution*, *24*, 127-135. doi: 10.1016/j.tree.2008.10.008

Borowsky, R., & Besner, D. (1993). Visual word recognition: A multistage activation model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 813–840. doi: 10.1037/0278-7393.19.4.813

Borowsky, R., & Besner, D. (2006). Parallel Distributed Processing and Lexical-Semantic Effects in Visual Word Recognition: Are a Few Stages Necessary? *Psychological Review*, *113*(1), 181–193. doi: 10.1037/0033-295X.113.1.181

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 63–85. doi: 10.1037/0278-7393.22.1.63

Brysbaert, M., & New, B. (2009). Moving beyond Kuc̃era and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. doi: 10.3758/BRM.41.4.977

Carreiras, M., Armstrong, B. C., Perea, M., & Frost, R. (2014). The what, when, where, and how of visual word recognition. *Trends in Cognitive Sciences*, *18*, 90–98. doi: 10.1016/j.tics.2013.11.005

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. doi: 10.1016/S0022-5371(73)80014-3

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello(and many other such concrete nouns). *Journal of Experimental Psychology: General*, *132*(2), 163. doi:

10.1037/0096-3445.132.2.163

Evans, G. A., Ralph, M. A. L., & Woollams, A. M. (2012). What's in a word? a parametric study of semantic influences on visual word recognition. *Psychonomic Bulletin & Review*, *19*(2), 325–331. doi: 10.3758/s13423-011-0213-7

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. The MIT press.

Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, *29*(2), 181–200. doi: 10.1016/0749-596X(90)90071-7

Frost, R. (2012). A universal approach to modeling visual word recognition and reading: Not only possible, but also inevitable. *Behavioral and Brain Sciences*, *35*(05), 310–329. doi: 10.1017/S0140525X12000635

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*(2), 256–281. doi: 10.1037/0096-3445.113.2.256

Hargreaves, I. S., Pexman, P. M., Pittman, D. J., & Goodyear, B. G. (2011). Ambiguous words recruit the left inferior frontal gyrus in absence of a behavioral effect. *Experimental Psychology*, *58*(1), 19–30. doi: 10.1027/1618-3169/a000062

Hino, Y., Kusunose, Y., & Lupker, S. J. (2010). The relatedness-of-meaning effect for ambiguous words in lexical-decision tasks: when does relatedness matter? *Canadian Journal of Experimental Psychology*, *64*(3), 180–196. doi: 10.1037/a0020475

Hino, Y., & Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(6), 1331–1356. doi: 10.1037/0096-1523.22.6.1331

Hino, Y., Lupker, S. J., & Pexman, P. M. (2002). Ambiguity and synonymy effects in lexical decision, naming, and semantic categorization tasks: Interactions between orthography, phonology, and semantics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 686–713. doi: 10.1037/0278-7393.28.4.686

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. doi: 10.1016/j.jml.2006.04.001

Jager, B., & Cleland, A. A. (2016). Polysemy advantage with abstract but not concrete words. *Journal of Psycholinguistic Research*, *45*, 143–156. doi: 10.1007/s10936-014-9337-z

Jager, B., Green, M. J., & Cleland, A. A. (2015). Polysemy in the mental lexicon: relatedness and frequency affect representational overlap. *Language, Cognition and Neuroscience*, *31*(3), 425–429. doi: 10.1080/23273798.2015.1105986

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology*, *13*(2), 278–305. doi: 10.1016/0010-0285(81)90011-6

Joordens, S., & Besner, D. (1994). When banking on meaning is not (yet) money in the bank: Explorations in connectionist modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1051–1062. doi: 10.1037/0278-7393.20.5.1051

Joordens, S., Piercey, C. D., & Azarbehi, R. (2009). Modeling performance at the trial level within a diffusion framework: A simple yet powerful method for increasing efficiency via error detection and correction. *Canadian Journal of Experimental Psychology*, *63*(2), 81–93. doi: 10.1037/a0015385

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, *32*(4), 474–516. doi: 10.1006/jmla.1993.1026

Kello, C. T., & Plaut, D. C. (2000). Strategic control in word reading: evidence from speeded responding in the tempo-naming task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 719. doi: 10.1037/0278-7393.26.3.719

Kello, C. T., & Plaut, D. C. (2003). Strategic control over rate of processing in word reading: A computational investigation. *Journal of Memory and Language*, *48*(1), 207–232. doi: 10.1016/S0749-596X(02)00512-0

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, *45*(2), 259–282. doi: 10.1006/jmla.2001.2779

Klepousniotou, E., & Baum, S. R. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics*, *20*(1), 1–24. doi: 10.1016/j.jneuroling.2006.02.001

Klepousniotou, E., Pike, G. B., Steinhauer, K., & Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language*. doi: 10.1016/j.bandl.2012.06.007

Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1534–1543. doi: 10.1037/a0013012

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. doi: 10.1126/science.7350657

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2014). *LmerTest version 2.0-6 http://CRAN.R-project.org/package=lme4*.

Laszlo, S., & Armstrong, B. (2014). Applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended erp reading data. *Brain and Language*, *132*, 22-27. doi: 10.1016/j.bandl.2014.03.002

Laszlo, S., & Federmeier, K. D. (2011). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology*, *48*(2), 176–186. doi: 10.1111/j.1469-8986.2010.01058.x

Laszlo, S., & Plaut, D. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain and Language*, *120*(3), 271–281. doi: 10.1016/j.bandl.2011.09.001

Lerner, I., Armstrong, B. C., & Frost, R. (2014). What can we learn from learning models about sensitivity to letter-order in visual word recognition? *Journal of Memory and Language*, *77*, 40–58. doi: 10.1016/j.jml.2014.09.002

Locker, L., Simpson, G. B., & Yates, M. (2003). Semantic neighborhood effects on the recognition of ambiguous words. *Memory & Cognition*, *31*(4), 505. doi: 10.3758/BF03196092

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences*, *2*, 49–55.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of the effect of context in perception: Part 1. *Psychological Review*, *88*, 375–407. doi: 10.1037/0033-295X.88.5.375

Mirman, D., Strauss, T. J., Dixon, J. A., & Magnuson, J. S. (2010). Effect of representational distance between meanings on recognition of ambiguous spoken words. *Cognitive Science*, *34*(1), 161–173. doi: 10.1111/j.1551-6709.2009.01069.x

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation*, *1*(2), 263–269. doi: 10.1162/neco.1989.1.2.263

Perea, M., Rosa, E., & Gómez, C. (2002). Is the go/no-go lexical decision task an alternative to the yes/no lexical decision task? *Memory & Cognition*, *30*(1), 34–45. doi: 10.3758/BF03195263

Pexman, P., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, *15*(1), 161–167. doi: 10.3758/PBR.15.1.161

Pexman, P., Hino, Y., & Lupker, S. J. (2004). Semantic ambiguity and the process of generating meaning from print. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1252–1270. doi: 10.1037/0278-7393.30.6.1252

Pexman, P., & Lupker, S. J. (1999). Ambiguity and visual word recognition: Can feedback explain both homophone and polysemy effects? *Canadian Journal of Experimental Psychology*, *53*(4), 323–334. doi: 10.1037/h0087320

Pexman, P., Lupker, S. J., & Hino, Y. (2002). The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin & Review*, *9*(3), 542–549. doi: 10.3758/BF03196311

Piercey, C. D., & Joordens, S. (2000). Turning an advantage into a disadvantage: Ambiguity effects in lexical decision versus reading tasks. *Memory & Cognition*, *28*(4), 657–666. doi: 10.3758/BF03201255

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Langugage and Cognitive Processes*, *12*(5/6), 765-805. doi: 10.1080/016909697386682

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, *107*(4), 786. doi: 10.1037/0033-295X.107.4.786

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*(5), 377–500. doi: 10.1080/02643299308253469

Pylkkänen, L., Llinás, R., & Murphy, G. L. (2006). The representation of polysemy: MEG evidence. *Journal of Cognitive Neuroscience*, *18*(1), 97–109. doi: 10.1162/089892906775250003

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425. doi: 10.1016/j.jml.2008.02.002

Raaijmakers, J. G. W., Schrijnemakers, J., & Gremmen, F. (1999). How to deal with "The language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*, 416–426. doi: 10.1006/jmla.1999.2650

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1339–1362. doi: 10.1080/02724980244000099

Ratcliff, R. (1980). A note on modeling accumulation of information when the rate of accumulation changes over time. *Journal of Mathematical Psychology*, *21*(2), 178–184. doi: 10.1016/0022-2496(80)90006-1

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*(2), 261–300. doi: 10.1037/0033-295X.106.2.261

Rodd, J. M., 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. doi: 10.1006/jmla.2001.2810

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, *28*(1), 89–104. doi: 10.1016/j.cogsci.2003.08.002

Rogers, T. T., & Patterson, K. (2007). Object categorization: Reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology: General*, *126*, 451–469. doi: 10.1037/0096-3445.136.3.451

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 487–494. doi: 10.1016/S0022-5371(70)80091-3

Schneider, W., Eschman, A., & Zuccolotto, A. (2010). *E-prime, Version 2.0.8.90 [Computer Software].* Pittsburgh, PA: Psychology Software Tools.

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. M., & Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology*, *14*(4), 489–537. doi: 10.1016/0010-0285(82)90017-2

Seidenberg, M. S., Waters, G. S., Sanders, M., & Langer, P. (1984). Pre-and postlexical loci of contextual effects on word recognition. *Memory & Cognition*, *12*(4), 315–328. doi: 10.3758/BF03198291

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. doi: 10.1037/0033-2909.134.4.584

Siakaluk, P. D., Pexman, P. M., Sears, C. R., & Owen, W. J. (2007). Multiple meanings are not necessarily a disadvantage in semantic processing: Evidence from homophone effects in semantic categorisation. *Language and Cognitive Processes*, *22*(3), 453–467. doi: 10.1080/01690960600834756

Simpson, G. B. (1994). Context and the processing of ambiguous words. In Morton Ann Gersnbacher (Ed.), *Handbook of Psycholinguistics* (pp. 359–374). San Diego, CA: Academic Press.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*(4), 421–457.

Swinney, D. A. (1979). Lexical access during sentence comprehension:(Re) consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 645–659. doi: 10.1016/S0022-5371(79)90355-4

Tabossi, P. (1988). Accessing lexical ambiguity in different types of sentential contexts. *Journal of Memory and Language*, *27*(3), 324–340. doi: 10.1016/0749-596X(88)90058-7

Taler, V., Klepousniotou, E., & Phillips, N. A. (2009). Comprehension of lexical ambiguity in healthy aging, mild cognitive impairment, and mild alzheimer's disease. *Neuropsychologia*, *47*(5), 1332–1343. doi: 10.1016/j.neuropsychologia.2009.01.028

Tamminen, J., Cleland, A. A., Quinlan, P. T., & Gaskell, M. G. (2006). Processing semantic ambiguity: Different loci for meanings and senses. In *Proceedings of the Twenty-eighth Annual Conference of the Cognitive Science Society* (pp. 2222–2227).

Twilley, L. C., Dixon, P., Taylor, D., & Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition*, *22*(1), 111–126. doi: 10.3758/BF03202766

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. doi: 10.1037/0033-295X.108.3.550

Van Petten, C., & Kutas, M. (1987). Ambiguous words in context: An event-related potential analysis of time the course of meaning activation. *Journal of Memory and Language*, *26*(2), 188–208. doi: 10.1016/0749-596X(87)90123-9

Watson, C. E. (2009). *Computational and Behavioral Studies of Normal and Impaired Noun/Verb Processing* (Unpublished doctoral dissertation). Carnegie Mellon University, Pittsburgh, PA.

Williams, J. N. (1992). Processing polysemous words in context: Evidence for interrelated meanings. *Journal of Psycholinguistic Research*, *21*(3), 193–218. doi: 10.1007/BF01068072

Yap, M. J., Lim, G. Y., & Pexman, P. M. (2015). Semantic richness effects in lexical decision: The role of feedback. *Memory & Cognition*, 1–20. doi: 10.3758/s13421-015-0536-0

Yarkoni, T., Balota, D. A., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971-979. doi: 10.3758/PBR.15.5.971

Footnotes

[1] The definitions of homonymy and polysemy differ depending on where a given group of researchers draws the line between items that are (relatively) unambiguous, purely homonymous purely polysemous, or ‘hybrid’ items that are both homonymous and polysemous. Thus, other researchers might classify the ‘homonymous’ words in the Hino et al. (2006) study as being closer to ‘hybrid’ ambiguous items than pure homonyms (e.g., Armstrong & Plaut, 2008; Beretta et al., 2005; Rodd et al., 2002). A similar re-labeling might also be applicable to the high Number of Meaning items in Pexman et al. (2004), although we cannot make strong claims either way because detailed counts of meanings and senses were not reported in that paper. Comparisons between item types reported by different research groups should therefore be made with caution. Nevertheless, these issues of nomenclature in no way question the fact that very different ambiguity effects were observed in different tasks using the same items in the work by Hino and colleagues, but bear only on whether the overall ambiguity advantage that they report is across homonyms and polysemes or, per the definitions that we introduce later, across hybrid items and polysemes.

[2] An additional reason for dismissing the potential role of these dynamics was that these dynamics did not seem to agree with the homonymy advantage in lexical decision and the homonymy disadvantage in semantic categorization. As we elaborate on in the discussion based on the results of our study, there may be reason to view some of the “homonyms” in previous studies as also being polysemous. This opens up new interpretational possibilities on that front.

[3] For example, we do not rule out the possibility that semantic activity could feed back and shape orthographic processing to some degree, as is assumed in other accounts, particularly of lexical decision (e.g., Hino et al., 2006). However, given that such accounts assume that there are differences in semantic activity, either overall or in how it is distributed, and the response system can attend to semantic activity directly in other tasks that depend on semantics (e.g., semantic categorization), it appears more parsimonious to assume that semantics can influence the response system directly in all tasks.

[4] This assumption also helps dissociate our account, which is directly concerned with cooperative and competitive dynamics due to ambiguity, from the separate (and potentially confounding) issue that more semantically rich items are often shown to be associated with a processing advantage (Pexman, Lupker, & Hino, 2002; Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008; Plaut & Shallice, 1993). Unless it can be demonstrated that particular interpretations always vary in semantic richness depending on whether the item is a homonym, polyseme, or unambiguous word, this appears to be the most neutral assumption for developing theories of semantic ambiguity. Insofar as imageability ratings can viewed as a proxy for estimating featural richness, the norming that we conducted in our own study also suggests that our word types did not differ on this measure (see Table 1; participants were asked to give an "average" imageability across all interpretations of ambiguous words). Of course, this is not to say that there may

not be interactions between richness, number of interpretations, and other psycholinguistic variables—recent work suggests that there are; Jager & Cleland, 2016; Jager, Green, & Cleland, 2015—but probing this additional interaction is outside the scope of the present work. The present, computationally explicit account of semantic settling dynamics may be a useful platform for conducting these studies given that such complex interactions are difficult to capture with underspecified verbal theories.

[5] An empirical test of these ideas is also likely to prove an important major project in and of its own work simply because it is not clear how to generate appropriate subjective measures of featural overlap and relate those to the featural overlap in computational models. For example, the results of the (Cree & McRae, 2003) feature norming study suggest that having participants list the features associated with specific interpretations may bias participants to generate distinctive features as opposed to shared features. Alternatively, asking participants to rate subjective relatedness on a Likert scale would not ensure that their ratings could be mapped onto a linear measure of featural overlap.

[6] This description abstracts away from several other complicating subtleties such as nonlinear input-output functions in neural networks, the fact that the account assumes that competitive/inhibitory effects are assumed to be weak but not entirely absent, and the noisy nature of neural computation may further bias the response system to only consider a given feature to be active once it has exceeded some minimum activation level. However, these subtleties are considered to be secondary factors relative to the core claims outlined in the main text.

[7] When the items were selected, the only published large-scale database of homonym meaning frequencies was an older study by Twilley, Dixon, Taylor, and Clark (1994), and our preliminary investigations of the predictive validity of those norms called into question their utility in selecting our items. Indeed, this motivated our development of a new approach to relative meaning frequency estimation and the creation of the largest corpus of meaning frequency estimates in English (Armstrong, Tokowicz, & Plaut, 2012; see also Armstrong et al., 2015). However, because the validity of our new norming method had not been established at the time, we estimated that a sample of 100 homonyms would contain at least enough homonyms with balanced meaning frequencies to allow for meaning frequency to be controlled for in the analyses.

[8] By basing our statistical analyses on all of the data from the hybrid and homonym items and including the numerous other confounding variables as predictors, we ensure that our significant effects were not caused by only analyzing and plotting the results of a subset of the homonyms and hybrids that could have been driven by group-level confounds in the subsets.

[9] Additional information regarding the models, methodology and results are available from the first author.

[10] $\log_{10}(1+f)$ was used instead of raw frequency f. Residual familiarity, for which the effects of the meaning and sense variables at the heart of these investigations were first removed, was employed instead of raw familiarity. Raw and residual familiarity correlated strongly in the different analyses (r's $\geq$ .94). The number of WordNet definitions

was not included because it represents a sum of both unrelated and related interpretations, which are the critical psycholinguistic variables under study

[11] Note that the sign of the disadvantage in the statistical tests of the pseudohomophone-degraded condition, which includes all of the experimental items and in which all variables were controlled for via item-level matching, indicates a homonymy disadvantage, as in the other cases. However, non-perfect matching between the unambiguous controls and subset of the most balanced homonyms included in this figure on other control variables initially gives the visual impression that a homonymy advantage was detected in the figure. In turn, this plot suggests that differences in other psycholinguistic variables (e.g., slight differences in word frequency, neighborhood size, etc., across conditions), which are controlled for in the actual analyses, as well as other sources of variability (e.g., different amounts of variability in the magnitude of homonymy effects across participants; different overall amounts of participant variability) are responsible for this apparent discrepancy.

[12] We also examined the main effects of stimulus contrast and nonword difficulty for the analysis of the word data reported in the main text, as well as in simplified analyses of the nonword data, which only included the difficulty manipulation and random intercepts for participant and item. Because the results of these analyses parallel those reported in the main text, we summarize them, in brief, below. Nonword latencies always increased significantly when difficulty was increased (all p's $< .05$). Word latencies always increased significantly across contrast levels for each nonword type (all p's $> .001$). Weaker marginal trends were detected for the nonword difficulty manipulations within each contrast level (full contrast - hard vs. very hard nonwords: $p = .09$; hard vs. pseudohomophones $p = .07$. Degraded contrast - hard vs. very hard nonwords: $p = .15$; hard vs. pseudohomophones: $p = .12$). As was our aim in providing feedback after each block to reduce between condition differences in accuracy, overall accuracy did not decrease significantly except in the nonword data when comparing hard and very hard nonwords under full feedback ($p = .003$), although there were a few marginal trends (nonword data: hard vs. pseudohomophones, full contrast: $p = .10$; full vs. degraded contrast, hard nonwords, $p = .11$; full vs. degraded contrast, very had nonwords: $p = .09$, although this trend is not in the predicted direction; word data: full contrast, hard vs. very hard nonwords: $p = .04$; degraded contrast, hard vs. very had nonwords: $p = .08$; all other p's $> .15$; tests were one-tailed). Overall, these results are consistent with the notion that the stimulus contrast manipulation had a stronger influence on overall performance—and latency in particular—than the nonword difficulty manipulation.

[13] Note that if one focuses only on the unambiguous control word latencies, the lexical decision task (563 ms) appears to generate identical response latencies as the two narrow semantic categorization tasks (Experiment 3: 566 ms; Experiment 4: 559 ms). Thus, based on those data in isolation (and ignoring the weak and noisy nature of between-participant comparisons noted later in the discussion), it would be reasonable to assume that overall latencies were constant and therefore that ambiguity effects should be constant according to the SSD account. Given that different ambiguity effects were observed, by this logic the semantic settling dynamics account does not successfully capture

these data. That said, we view the averaged latencies as a reasonable measure because they should not only be more stable estimates of mean latencies but also ensure that a response bias was not keeping some responses at a relatively constant speed by slowing other responses. More work will be needed, however, to flesh out which interpretation is correct; computational models of response selection will likely be particularly valuable to this end.

[14] Indeed, a polysemy disadvantage is predicted to emerge during very late processing, but its magnitude should be very small compared to the homonymy disadvantage, particularly for polysemes with reasonably high amounts of feature overlap across interpretations. The absence of this effect in the Hino et al. (2006) study may therefore be due to the task still not being hard/slow enough and/or to statistical power issues—both of which could be the subject of additional targeted empirical study.

[15] In abstract, this logic is also similar to that used by Pexman et al. (2004) to explain why they failed to detect ambiguity effects of high-frequency words in lexical decision. Specifically, they state that "For high-frequency words, responses are sometimes made before semantic feedback can have a significant impact on orthographic activation, and, therefore, there will be situations where no ambiguity effect is observed for high-frequency words in [lexical decision]" (p. 1255). Although we propose a slightly different account where semantic activation can, in principle, drive the response system directly, these results emphasize the importance of making the task hard and slow (both in terms of the words and nonwords) for substantial semantic processing to take place so as to observe ambiguity effects.

[16] Similarly, in a pilot study, we failed to observe different ambiguity effects in a lexical decision task with explicit response deadlines, which may be due to how participants speed up to meet the deadline or due to noise in how well participants matched the deadline.

Appendix A. Neural Network Simulation

Given the complex set of interactions assumed to underlie the SSD account, we have also sought to go beyond a simple verbal account to show that the relevant principles actually give rise to an extensive number of empirically observed patterns of results. Accordingly, we carried out a small-scale simulation to examine how the cooperative and competitive dynamics within a biologically-motivated connectionist network interact with representational differences between polysemous, homonymous, and unambiguous words to produce different ambiguity effects at different points in time (see Armstrong & Plaut, 2008, for related work using a model which makes similar but non-identical assumptions regarding featural overlap, number of features, certain aspects of the neurobiology, etc., which rules out the possibility that the following results are specific to only this network).

Model Architecture. The simulation was instantiated in a more neurobiologically realistic variant of the standard connectionist formalism (see also Armstrong & Plaut, 2013; Laszlo & Plaut, 2012; Laszlo & Armstrong, 2014; Plaut & Booth, 2000; Watson, 2009) with the following properties: 1) there are distinct populations of excitatory and inhibitory units, as reflected by constraints on the sign of the unit's outgoing weights; 2) there are far fewer inhibitory units than excitatory units; 3) the connections between layers (e.g., orthography and semantics) are only excitatory, whereas connections within a layer are both excitatory and inhibitory; and 4) the density/strength of between-layer connections is weaker compared with within-layer connections. Each of these properties contribute to encouraging stronger/faster excitatory dynamics and weaker/slower inhibitory dynamics.

The model architecture is presented in Figure 3. In keeping with its minimalist design, the network was composed of four orthographic input units, each of which was used to represent an individual word, two context units, each representing one of two separate contexts, and 22 units representing semantics (the training patterns are described in detail later). Of these 22 units, 21 were excitatory units (i.e., it could have positive outgoing weights only) and were used to code the distributed semantic representations of the words presented during training. The remaining unit was inhibitory (i.e., it could have negative outgoing weights only) and did not correspond to an explicitly-specified semantic feature.

- - - - - - -

Insert Figure 3 Here

- - - - - - -

Both the orthographic and the context units were connected to the semantic units. These between-layer connections were restricted to be excitatory only (non-negative). The excitatory semantic units were connected to one another with excitatory connections, with the exception that units did not connect to themselves. All of the excitatory weights were initialized to a value of 0.05, such that the initial magnitudes of the orthographic and context inputs were equal. Thus, any differences that emerge between the orthographic and context weights should be due to how critical each type of information is to activating the representation of each word class. The inhibitory weights were initialized to a value of -0.2. This balanced, in approximate terms, the amount of excitation that a semantic unit was expected to receive early in training with an equivalent amount of inhibition. Additionally, all of the units in the network had a bias connection—equivalent to a connection from an additional input unit that is always active—that was set to an initial value of 0. There was no other variability in the weight values to simplify the interpretation of the weight structure of the network. To constrain the magnitudes of the weights that could be learned for connections between versus within layers, the between-layer connections were subjected to weight decay after each weight adjustment ($\lambda = 0.00025$). Within-layer connections, including the connections to and from the inhibitory unit, were not subject to any weight decay.

Training patterns. The training representations used in the simulation are presented in Table 3. Localist representations were used to encode each word form and context. This allows for a simplified interpretation of the weight structure, presented later. A distributed semantic representation was associated with each word/context input pair, and different groups of units were used to represent the semantic features associated with each interpretation of a word. Note that none of these specific architectural decisions is not key to obtaining the observed effects, as illustrated by Armstrong and Plaut (2008), who used distributed representations different amounts of featural overlap, and different absolute numbers of features. Future work will be needed, however, to flesh out exactly how much featural overlap is needed to obtain a processing advantage or disadvantage at a particular point in time, and how that overlap relates to empirical measures of featural overlap, as described in the introduction.

The training set consisted of two unambiguous words, one homonymous word, and one polysemous word, each presented in two contexts. Each unambiguous word consisted of the pairing of a single orthographic unit with a group of four active semantic units. These orthographic/semantic input/output pairs were presented in two different contexts, represented as the activation of one of the two context units. The pattern structure for ambiguous words differed as follows: For the homonyms, two different groups of four semantic units were associated with the orthographic input, depending on which context was activated. For the polysemes, three semantic units were associated with the orthographic input regardless of the context but the fourth semantic unit was different for each context. Note that the choice of using only four semantic units was made strictly for simplicity. Additionally, considerable past work has shown that using such small raw numbers of semantic features can lead to relatively precise simulations of a range of different phenomena linked to semantics (Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012; Plaut, 1997).

The presentation of each word was structured so that the orthographic input would be activated for 20 unit updates, after which both the orthographic and context outputs were activated for an additional 20 updates. This was intended to reflect the weak/late effect of context in as simple a way as possible and is not to be interpreted as implying a formally staged theory of contextual access (cf. Sternberg, 1969; Swinney, 1979).

- - - - - - -

Insert Table 3 Here

- - - - - - -

Training. The network was trained using recurrent back-propagation through time (Pearlmutter, 1989) and momentum descent, so that it activated the correct semantic features once both the orthographic input and context had been presented. A learning rate of 0.0001 and momentum of 0.9 were employed (momentum was set to 0 for the first update). A relatively small learning rate was selected to maintain a positive gradient linearity across the weights that resulted from each weight update, which is more difficult if excitation and inhibition are represented in distinct pools of units. Units were considered to be correctly activated when they were within 0.15 of their target activation. Error was computed for the last 5 unit updates. Between each training pattern, the net input was reset to -1.4 and the net output was reset to 0.2 for all of the units. Weights were updated after each sweep through the training vocabulary. The model was trained until all semantic units were within 0.5 of their targets for each of the input patterns.

Results

Figure 4 plots the sum of semantic activations that exceeded a threshold value of 0.5 after training. This value was chosen for simplicity because it also corresponds to the standard binary threshold used to determine if a unit is on or off, however, the results do not hinge on this specific choice of measure (for a few examples using other measures, see, e.g., Armstrong, 2012; Armstrong & Plaut, 2008). Examining the activation trajectories at successive time intervals, the model exhibits an early polysemy advantage (time A), followed later by a homonymy advantage and a polysemy disadvantage (time B), a homonymy disadvantage (time C), and finally, by a disadvantage for all ambiguous words (time D). Note that, as described in the introduction, the choice of measuring slices of time on the x-axis was intentional because although we assume that semantic activity drives the response system directly, other sources of evidence (which do not differ between ambiguous and unambiguous words, e.g., orthographic neighborhood in lexical decision) can still provide overall evidence for responding in general, without differing between word types. Here, we focus only on the unique driving force that semantics supplies.

To relate the slices of time on the x-axis to the empirical study of lexical decision that we report, we make the following assumptions. First, by instructing participants to maintain a specific accuracy level across our nonword difficulty and stimulus contrast conditions, we assume that the total amount of evidence indicating that a word versus a nonword was presented remains (approximately) constant, as well. However, different sources of evidence may be differentially informative for making fast and accurate responses in different conditions. For example, using more orthographically wordlike nonwords could reduce the degree to which orthography can serve as basis for discriminating words from nonwords, leading to less (and by proxy, slower) accumulation of the total evidence in favor of a particular response. In turn, more overall processing would take place, including semantic processing, which allows for the sampling of different slices of the semantic settling dynamics and shape the patterns of responses.

- - - - - - -

Insert Figure 4 Here

- - - - - - -

A.4 Weight structure. An analysis of the pattern of learned weights in the network provides some insights into how the network's knowledge of the words could interact with the representations of the different words and cooperative and competitive dynamics to give rise to the observed semantic activity. Of course, the weights only form a part—albeit a very important one—of understanding the full complexity of the settling dynamics within the network. Note that because there was no random variance in the initial weights, interpreting this weight structure is greatly simplified because the reported values are identical for all units within each sub-network.

Within semantics, the analysis revealed strong excitatory sub-networks for each interpretation of a word. Numerically, the weights among the core features were higher for the polyseme (weight value = 2.93) than for either the homonym or the unambiguous words, which had approximately equal intra-semantic excitatory weights (2.38 and 2.37, respectively). The distinct features of the polyseme had weaker incoming weights from the core features of the polyseme (1.08), but sent relatively stronger outgoing weights to the core features (1.23), and were disconnected from the distinct feature with which they were not consistent (0.0); the same was true for the connectivity between the semantic features used to code each meaning of the homonym. The learned weight structure associated with the polysemes is also similar to that which underlies basic-level category organization and the associated basic-level processing advantage (Rogers & Patterson, 2007).

The input from the inhibitory unit was largest for the homonym's features (-2.21), followed by the input to the core features of the polyseme (-2.00) and unambiguous words (-1.87); the least inhibition was associated with the distinguishing features of the polyseme (-1.66). These results are consistent with the predictions of the account, in terms of how much competition should exist for homonyms (high competition), polysemes (low competition), and unambiguous words (no competition). The fact that the distinct features of the polyseme receive less inhibition is commensurate with the reduced excitation that they receive from the core features and the other distinct features.

The incoming weights from orthography and context were an order of magnitude smaller than the weights within semantics, consistent with the assumption (and imposed model architecture) that such connectivity better reflects the actual connectivity structure in cortex. Feeding into semantics, the orthographic inputs with the strongest weights projected to the core features of the polyseme (0.14), followed closely by the weights to the unambiguous words' semantic features (0.13), by the weights to the semantic features of the homonym (0.10), and finally by the weights to the distinguishing features of the polyseme (0.07). With respect to the context units, the semantic features of the unambiguous words and the core features of the polysemous word were not sensitive to this information (weights = 0). The distinguishing features of the polyseme had moderate incoming weights from context (0.14), whereas the semantic units associated with the homonym received a relatively strong weight from context (0.25). These connectivity patterns are consistent with the differential need to rely on context to activate an appropriate semantic representation.

Overall, this connectivity structure is consistent with the predicted relative contributions of word form and context, and of cooperation and competition among consistent and inconsistent semantic features, that underlie the SSD account. It also helps flesh out how the presentation of a given input representation can interact with the network's memory, as encoded in the weights, to give rise to the simulated semantic settling dynamics.

Discussion

Collectively, these results substantiate the verbal description of the SSD account in explicit mechanistic terms, in that the observed changes in semantic activation over time exhibit the core properties of the account. Experiments associated with fast responses, such as visual lexical decision (e.g., Beretta et al., 2005; Klepousniotou & Baum, 2007; Rodd et al., 2002), could reflect the dynamics near time A in Figure 4, leading to a polysemy advantage; extremely easy versions of lexical decision with very fast responses would tap semantics before this point when there is no substantial activity for any word class (Azuma & Van Orden, 1997; Rodd et al., 2002). Tasks that are associated with slower responses, such as semantic categorization involving broad categories

(Hino et al., 2006), are consistent with the dynamics at time C, at which only a homonymy disadvantage is present. Tasks that require even more processing, such as those involving the activation of a contextually appropriate interpretation are typically associated with a processing disadvantage for ambiguous words, consistent with time D (e.g., Frazier & Rayner, 1990; Piercey & Joordens, 2000; Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982; Swinney, 1979; Tabossi, 1988; Van Petten & Kutas, 1987, although, as in the model, the disadvantage is smaller for polysemes, Williams, 1992).

Table 1

Properties of the Word Stimuli

	Unambig.	Poly.	Homon.	Hybrid
example	tango	blind	yard	stall
Word Frequency	20.5	21.1	20.8	21.2
Length in Letters	4.5	4.4	4.4	4.4
Number of Meanings	1	1	2.1	2.4
Number of Senses	5.6	12.9	6.2	14
WordNet Definitions	5.9	12.3	6.7	12.6
Positional Bigram Frequency	174	192	201	191
N	11.1	11.0	12.3	13.8
LD	1.4	1.3	1.3	1.3
Phonemes	3.6	3.7	3.6	3.7
Syllables	1.2	1.1	1.2	1.1
Familiarity	4.9	4.9	4.7	4.7
Imageability	4.7	4.8	4.8	4.6
Dom. Mean. Freq.	100*	100*	82	77

Note. Word frequency data were obtained from the SUBTL database (Brysbaert & New, 2009). Positional bigram frequency and orthographic neighborhood metrics were also calculated based on the words in SUBTL with frequencies greater than or equal to one. Familiarity, imageability, and meaning frequency were normed after the stimuli were selected and were not matched across quadruplets. *Dominant meaning frequency was assumed to be maximal for these items in the regression analyses. Unambig. = Unambiguous. Poly. = Polysemous. Homon. = Homonymous. WordNet Definitions = number of definitions in wordNet (Fellbaum, 1998). N = Coltheart's N (Coltheart, Davelaar, Jonasson, & Besner, 1977). LD = orthographic Levenshtein distance (Yarkoni, Balota, & Yap, 2008). Dom. Mean. Freq. = frequency of dominant meaning.

Table 2

Properties of the Nonword and the Word Stimuli

	Stimuli											
	Hard NWs			V. Hard NWs			Pseudo. NWs			Words		
Length	bi	N	LD	bi	N	LD	bi	N	LD	bi	N	LD
3	14	15	1.1	29	31	1.0	25	28	1.0	24	26	1.0
4	121	10	1.4	180	16	1.1	125	15	1.1	125	15	1.1
5	261	4	1.7	608	13	1.3	246	6	1.6	228	6	1.5
6	625	2	1.9	1789	9	1.5	377	4	1.7	603	3	1.8
7	1000	1	2.4	3190	10	1.4	429	1	2.2	766	1	2.1
8	1355	1	2.6	3777	3	1.8	678	0	2.6	806	1	2.3

Note. Four- and five-letter strings made up 85% of the items. bi = positional bigram frequency. N = Coltheart's N. LD = orthographic Levenshtein distance.

Table 3

Orthographic, Context, and Semantic Representations used in the Simulation

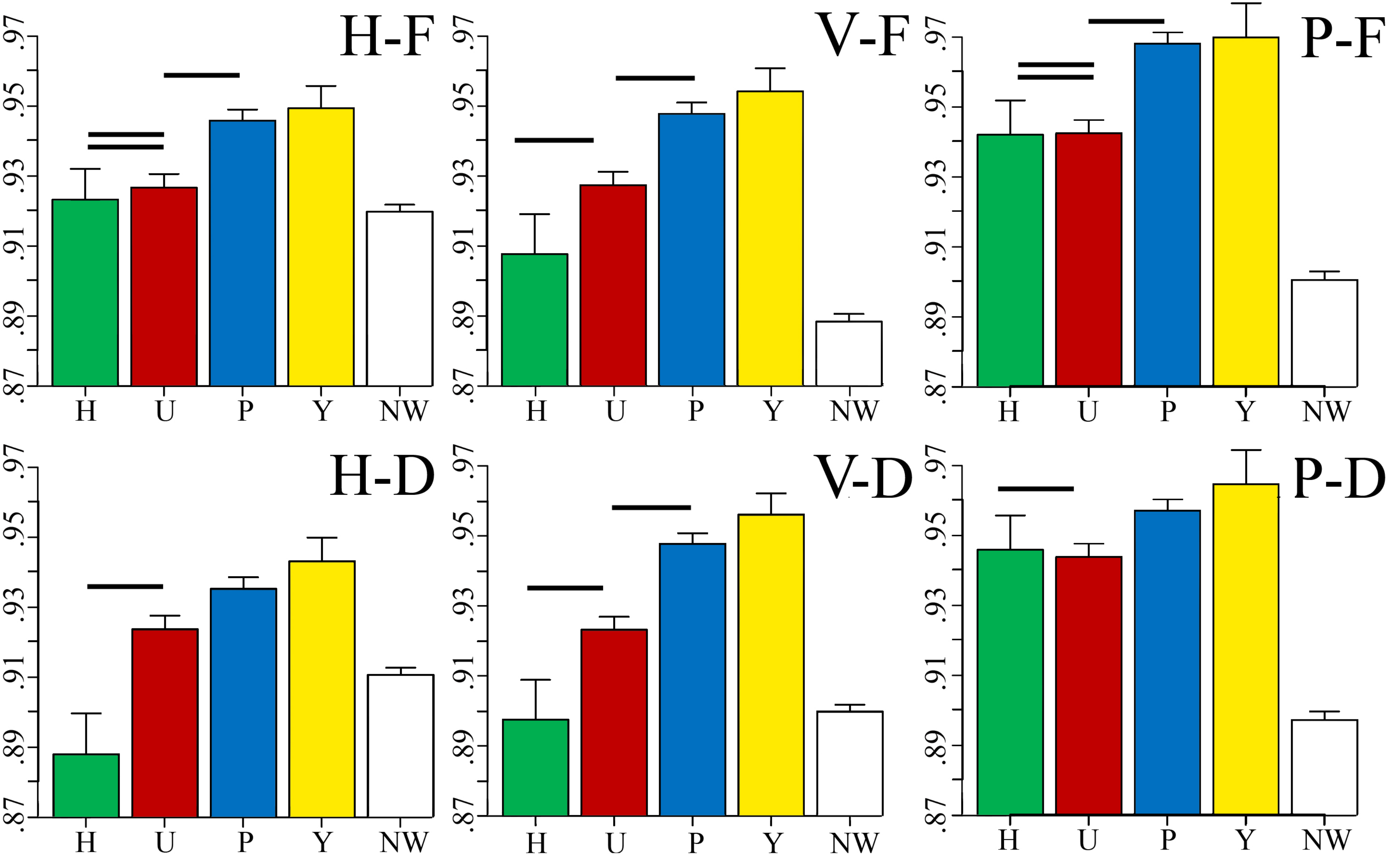
Item Type	Orthography	Context	Semantics
unambiguous	1 0 0 0	1 0	1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	1 0 0 0	0 1	1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
unambiguous	0 1 0 0	1 0	0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
	0 1 0 0	0 1	0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
polysemous	0 0 1 0	1 0	0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0
	0 0 1 0	0 1	0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0
homonymous	0 0 0 1	1 0	0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0
	0 0 0 1	0 1	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1

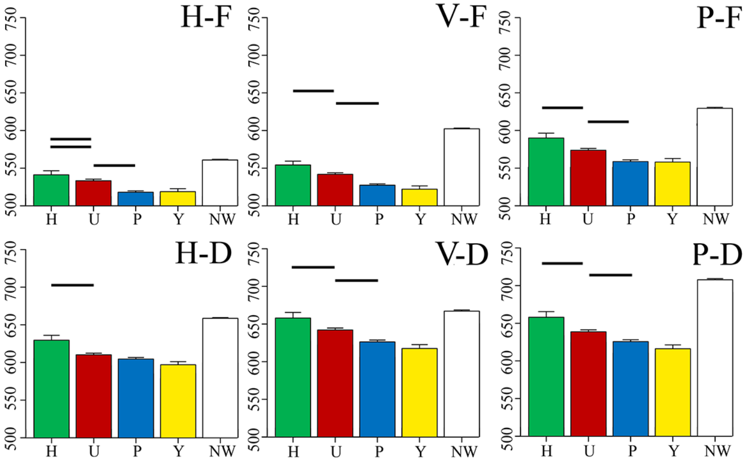
Figure 1. Accuracy data for each word type in each condition. H-F = hard-full. V-F = very hard-full. P-F = pseudohomophone-full. H-D = hard-degraded. V-D = very hard-degraded. P-D = pseudohomophone-degraded. H = homonym. U = unambiguous. P = polyseme. Y = hybrid. NW = nonword. Significant ($p < .05$) and marginal ($p < .1$) differences between homonyms and polysemes relative to unambiguous items are denoted by single and double lines, respectively. Note that the statistical tests are sensitive to meaning frequency and control for many psycholinguistic properties.

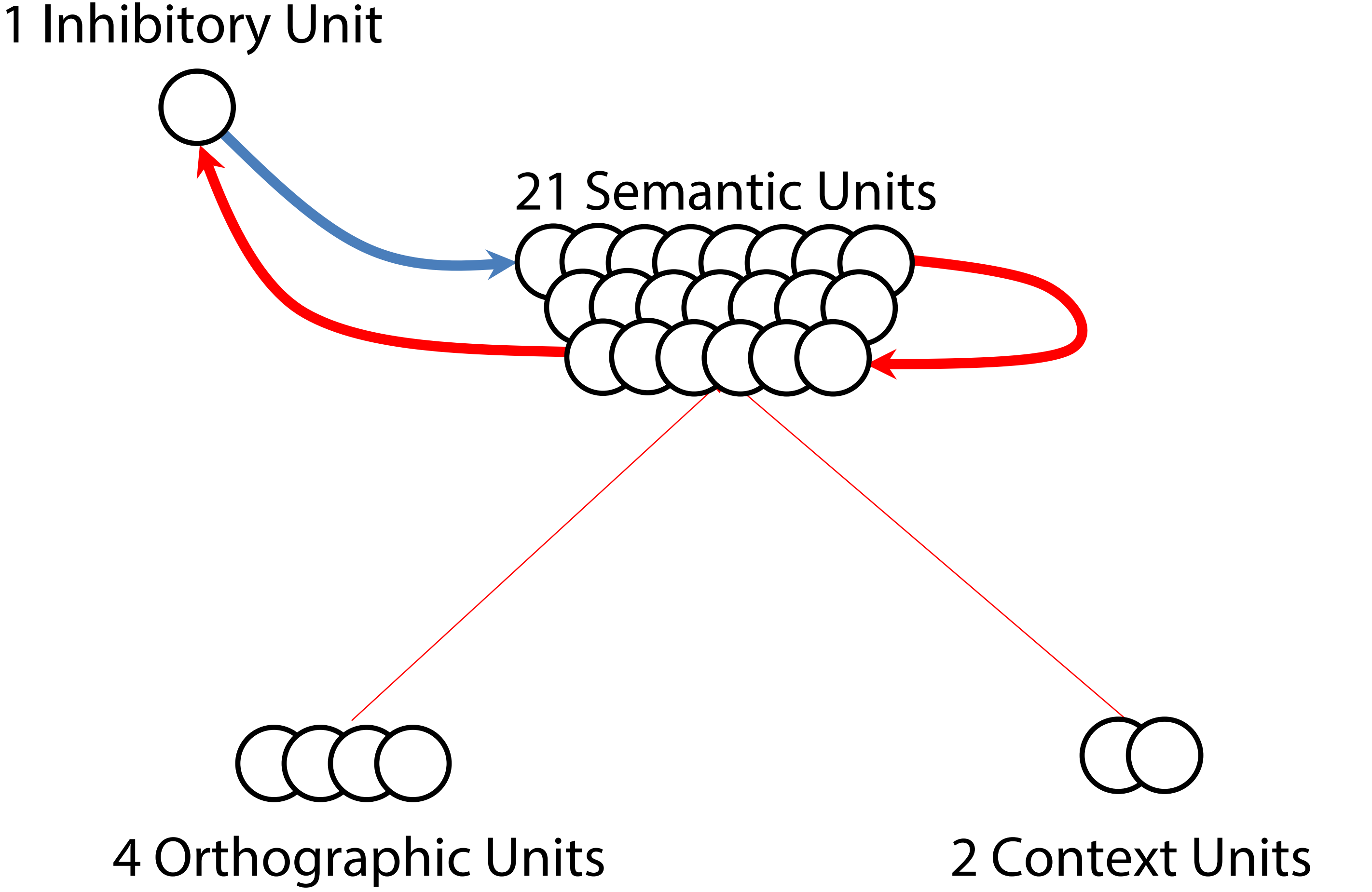
Figure 2. Correct latency for each word class in each condition. H-F = hard-full. V-F = very hard-full. P-F = pseudohomophone-full. H-D = hard-degraded. V-D = very hard-degraded. P-D = pseudohomophone-degraded. H = homonym. U = unambiguous. P = polyseme. Y = hybrid. NW = nonword. Significant and marginal differences between homonyms and polysemes relative to unambiguous items are denoted by single and double lines, respectively.

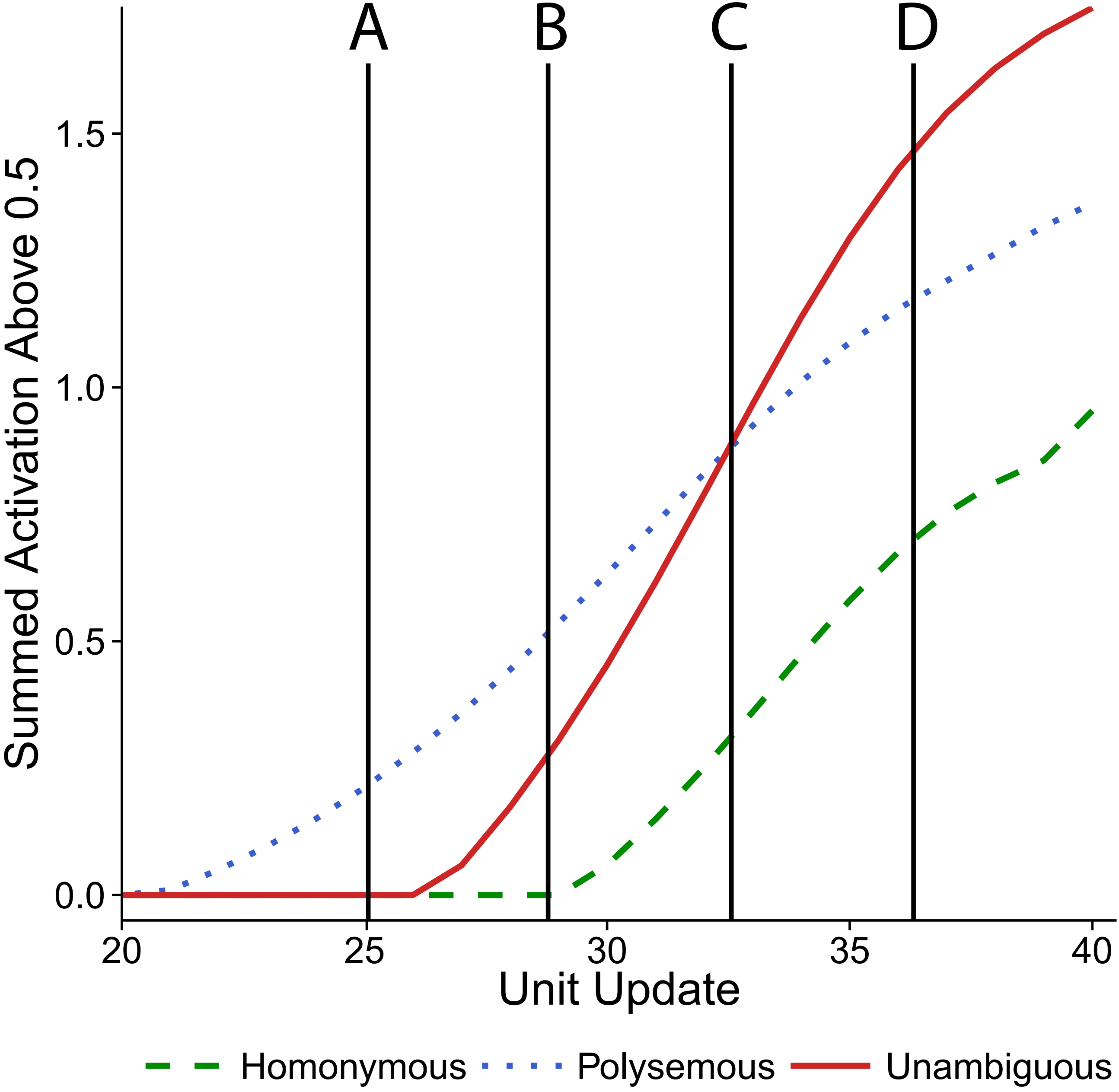
Figure 3. Architecture for the network from the simulation. Solid red (dark) arrows indicate excitatory connections and blue (light) arrows indicate inhibitory connections. Connections depicted by thinner arrows were subject to weight decay.

Figure 4. The average amount of semantic activation exceeding a threshold of 0.5 for polysemous, unambiguous, and homonymous words. Activation did not exceed the threshold during earlier processing. Vertical slices A-D correspond to points at which the model exhibits several important patterns of effects reported in the literature. The y-axis corresponds to the sum of the semantic activation that exceeds a 0.5 activation threshold across all semantic units in the simulation.


H-F
V-F
P-F
H-D
V-D
P-D
.87
.89
.91
.93
.95
.97
H
U
P
Y
NW

H-F
V-F
P-F
H-D
V-D
P-D
500
550
600
650
700
750
H
U
P
Y
NW


1 Inhibitory Unit
21 Semantic Units
4 Orthographic Units
2 Context Units


A
B
C
D
Summed Activation Above 0.5
0.0
0.5
1.0
1.5
20
25
30
35
40
Unit Update
Homonymous
Polysemous
Unambiguous