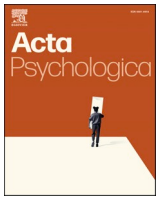
Acta Psychologica 238 (2023) 103985

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



Pupillometric and behavioural evidence shows no differences between polyseme and homonym processing

Juan Haro [a,*], Natalia López-Cortés [b], Pilar Ferré [a]

[a] Universitat Rovira i Virgili, Department of Psychology, Research Center for Behaviour Assessment (CRAMC), Tarragona, Spain
[b] Universidad de Zaragoza, Zaragoza, Spain

ARTICLE INFO

Keywords:
Lexical ambiguity
Semantic ambiguity
Ambiguous words processing
Pupillometry
Pupillary response
Polysemes
Homonyms
Experimental task difficulty

ABSTRACT

Ambiguous words can have related meanings (polysemes, e.g., *newspaper*) or unrelated meanings (homonyms, e.g., *bat*). Here we examined the processing of both types of ambiguous words (as well as unambiguous words) in tasks of increasing level of semantic engagement. Four experiments were conducted in which the degree of semantic engagement of the task was manipulated: lexical decision task (Experiments 1 and 2), semantic categorization task (Experiment 3) and number-of-meanings task (Experiment 4). RTs and pupillary response were recorded. To our knowledge, pupillary response had never been used before to study ambiguous words processing in isolation. Results showed faster RTs for ambiguous words with respect to unambiguous words in LDT, and larger pupil dilation was observed for ambiguous words in comparison to unambiguous ones in number-of-meanings task. However, differences between polysemes and homonyms were not observed in any task. These results provide no evidence that polysemes and homonyms are processed differently.

1. Introduction

All languages have many words with multiple meanings, i.e., words that are semantically ambiguous (e.g., *bat*). Ambiguous words do not comprise a uniform set of words, but rather differ in many ways. One of the aspects of these words that has most attracted the attention of researchers is the relatedness between their meanings. Broadly speaking, ambiguous words can be categorized as polysemes, if their meanings are related (e.g., *newspaper*), or as homonyms if their meanings are unrelated (e.g., *bank*). Experimental research in this field has sought to better understand how polysemes and homonyms are processed and represented in the mind (see Eddington & Tokowicz, 2015, for a review). The results of some studies using different experimental tasks have shown a disadvantage for homonyms in comparison with polysemes using different experimental tasks, where homonyms have been associated with slower responses (Armstrong & Plaut, 2008, 2011, 2016; Brown, 2008; Klepousniotou & Baum, 2007; Rodd et al., 2002). To account for this homonymy disadvantage, it has been argued that the meanings of homonyms are represented differently from those of polysemes. One of the most influential models in this field is that of Rodd et al. (2004), which maintains that each meaning of an ambiguous word is represented as an attractor basin within a semantic network. The semantic network starts in a random state of activation, and word recognition takes place when the network accesses an attractor basin. The structure and location of these basins differ for polysemes and homonyms. Related meanings of polysemes are represented by neighbouring basins, forming a single, wide, shallow basin; in contrast, unrelated meanings of homonyms are represented by competing basins located in disparate regions of the semantic network. Hence, according to Rodd et al. (2004), the homonymy disadvantage arises from the competition between unrelated meanings, inhibiting the recognition of these words during word processing.

However, there is evidence incompatible with Rodd et al. (2004), mostly from studies that have used the lexical decision task (hereafter, LDT), a task where participants are presented with strings of letters corresponding to words or nonwords, and they have to decide if the string is a word or not. In fact, some studies have reported no differences between polysemes and homonyms in the LDT (Haro, Demestre, et al., 2017; Hino et al., 2006; Hino et al., 2010; Pexman et al., 2004), showing an advantage for both types of ambiguous words compared to unambiguous words. This result is, on the other hand, consistent with the ambiguity advantage effect, that is, the finding of faster response times (RTs) for words with multiple meanings (regardless of the semantic relatedness between them) over words with a single meaning in the LDT

* Corresponding author at: Department of Psychology and CRAMC, University Rovira i Virgili, Crta. de Valls s/n., 43007 Tarragona, Spain.
E-mail address: juan.haro@urv.cat (J. Haro).

https://doi.org/10.1016/j.actpsy.2023.103985
Received 7 March 2023; Received in revised form 22 June 2023; Accepted 10 July 2023
Available online 13 July 2023

(e.g., Borowsky & Masson, 1996; Hino & Lupker, 1996; Rubenstein et al., 1970). In addition to behavioural data, there is also neurophysiological evidence in this line (but see Beretta et al., 2005, for a different pattern of findings). In an LDT study conducted in our laboratory (Haro, Demestre, et al., 2017), we observed similar event-related potential (ERP) amplitudes for polysemes and homonyms in the N400 component, an ERP component associated with semantic processing (e.g., Laszlo & Federmeier, 2011). Both types of ambiguous words showed higher N400 amplitudes than unambiguous words, as well as faster RTs, which we interpreted as evidence of greater semantic-to-orthographic activation for ambiguous words compared to unambiguous words.

The failure of some studies to find a homonymy disadvantage in LDT could be due to the fact that the task is not sufficiently sensitive to semantic competition processes. Indeed, it has been argued that the disadvantage for homonyms would be larger as semantic processing goes on (e.g., Armstrong & Plaut, 2008, 2011). Experimental tasks that require more semantic engagement or lead to longer response times would allow the unrelated meanings of a homonym to accumulate more activation and, thus, increase the competition between them. This increased competition would ultimately result in greater inhibition for homonyms with respect to polysemes. This may explain why some studies have observed longer response times for homonyms compared to polysemes in tasks with high level of semantic engagement (e.g., semantic categorization task), but not in the LDT (Hino et al., 2006; Pexman et al., 2004), as well as an increase in the homonymy disadvantage by manipulating the difficulty of nonwords and stimulus contrast in the LDT (Armstrong & Plaut, 2016). It should be noted, however, that some studies have observed both a homonymy disadvantage as well as an absence of differences between polysemes and homonyms in tasks involving a high level of semantic engagement (Hino et al., 2006; Pexman et al., 2004). These inconsistencies have been attributed to the decision-making component involved in such tasks, rather than to representational differences between these words. For example, Hino et al. (2006) found that homonyms showed a disadvantage in a semantic categorization task (hereafter, SCT) only when broad categories were used (e.g., "living beings"), but not with narrower categories (e.g., "vegetables"). This led them to suggest that the disadvantage for homonyms occurs within the decision-making process, during which the system checks for category-congruent information in the semantic representation of the word. If the category has clear boundaries, i.e., defined properties (e.g., "plant" or "animal"), the system only checks for that specific set of properties in the semantic representation of the word (e.g., those that represent the characteristics of an animal or plant). In this case, according to Hino et al. (2006), the comparison should be fast, and no differences are expected between polysemes and homonyms. On the other hand, if the category is broad or its boundaries are diffuse, the semantic analysis must be more exhaustive. The disadvantage for homonyms in a SCT therefore stems from the fact that their meanings, unlike those of polysemes, do not include overlapped information and this delays their analysis.

In view of the above, it seems that a more extensive and in-depth study of the processing and representation of homonyms and polysemes is needed, which would also shed some light on the effect of task type. The aim of this study was to address these questions. To do so, we explored the role of semantic ambiguity as well as of the relationship between the meanings of ambiguous words in tasks involving an increasing level of semantic engagement. Furthermore, we recorded pupillary response during the tasks, a neurophysiological measure that seems very promising for the study of word processing.

For many years, pupillary response has been widely used by psychologists as a measure of mental workload during cognitive tasks (see Laeng et al., 2012, for an overview). Pupillary response has certain advantages over behavioural measures. For example, it provides time-related information on how cognitive processes develop, and it is sensitive to processes that are only partially activated, but which never pass the threshold for eliciting overt behaviour or reaching consciousness (Laeng et al., 2012). In addition, since it does not require overt responses, it avoids the influence of response execution during processing.

Furthermore, recent studies have shown that pupillary response is sensitive to several psycholinguistic variables, for instance, word frequency (Haro, Guasch, et al., 2017; Kuchinke et al., 2007; Papesh & Goldinger, 2012), cognate status (Guasch et al., 2017), and emotional valence (Schmidtke, 2014). These studies have reported that pupil dilation increases as a function of word-processing difficulty; for example, low frequency words are associated with greater pupil dilation than high frequency words (e.g., Haro, Guasch, et al., 2017). More relevant to the present study is the evidence that pupillary response is sensitive to inhibitory processes during word recognition (Geller et al., 2016), and to semantic ambiguity during auditory sentence comprehension (Kadem et al., 2020). In a masked priming LDT study, Geller et al. (2016) observed that responding to words preceded by primes of similar orthography (e.g., cold-CORD) elicited greater pupillary responses than words preceded by primes with different orthography (e.g., rest-CORD). The authors interpreted this result as evidence of the inhibition exerted by the lexical competitor (prime) on the target during recognition. Moreover, they found that pupillary response appeared to be more sensitive than RTs to inhibition processes. In addition, Kadem et al. (2020) recently found that participants presented greater pupil dilation when listening to sentences containing a word with more than one meaning (e.g., "The shell was fired towards the tank") than when listening to sentences that did not include an ambiguous word (e.g., "Her secrets were written in her diary"). Although Kadem et al. (2020) did not examine polysemes and homonyms separately, this result seems to suggest that the pupillary response reflects the activation (and possible competition) of different meanings of ambiguous words during sentence processing. In summary, pupillary response shows great promise as a measure for studying the processing and representation of semantic ambiguity, especially for the study of the distinction between polysemy and homonymy, since this measure appears to reflect inhibition and semantic competition during word processing. In addition, it complements behavioural data and could provide novel evidence in this field of research.

In this study, we recorded the pupillary response to polysemes and homonyms (as well as unambiguous words) in several tasks involving different levels of semantic engagement. We presented the same set of polysemes, homonyms and unambiguous words to different groups of participants in four experiments. In Experiment 1 we presented the words in an LDT, in which we collected only behavioural data. The purpose of this first experiment was to provide a behavioural validation of the selected stimulus set. The subsequent experiments focused on the pupillary response. In Experiment 2, we recorded pupillary responses during an LDT. In Experiment 3, the participants completed an SCT (i.e., "Does the word belong to the category jobs, professions, and ranks?"), during which we collected pupillary data. Finally, in Experiment 4, we recorded pupillary responses in a task where participants had to indicate whether the words had one or more meanings (number-of-meanings task).

This sequence of experiments allowed us to examine whether differences between polysemes and homonyms appear, or become more pronounced, in tasks that involve increased semantic processing compared to less semantically engaging tasks, as well as if such differences are reflected in pupillary responses. According to theoretical proposals suggesting that the meanings of homonyms are represented differently from those of polysemes (Armstrong & Plaut, 2008, 2011; Rodd et al., 2004), we predicted a disadvantage for homonyms with respect to polysemes in all the experimental tasks. This disadvantage would be represented by slower RTs (Experiment 1) and increased pupil dilation (Experiments 2, 3 and 4) for homonyms compared to polysemes. However, we also considered the possibility that this disadvantage would not be apparent in LDT (Experiments 1 and 2), but rather in tasks that entail increased semantic processing (Experiments 3 and 4), according to the hypothesis that unrelated meanings accumulate more

activation and, therefore, lead to greater levels of competition in the later stages of semantic processing (Armstrong & Plaut, 2008, 2011).

2. Experiment 1: lexical decision task (response times)

2.1. Materials and methods

2.1.1. Participants

Forty Spanish speakers took part in this experiment. The participants were undergraduate students at Rovira i Virgili University (Tarragona, Spain), all of whom had either normal or corrected-to-normal vision. They received academic credits for their participation. Prior to the experiment, the participants signed an informed consent document. A local ethics committee at the Universitat Rovira i Virgili approved this research (CEIPSA-2021-PR-0044).

2.1.2. Materials

We selected 112 Spanish words: 56 ambiguous words and 56 unambiguous words (the full set of stimuli is included in the Appendix). We used the ratings of two different ambiguity measures from Haro, Ferré, et al. (2017) to categorize the words as ambiguous or unambiguous. The first was number-of-meanings (NOM). To obtain NOM ratings, participants are asked to think about all the meanings of a word and indicate if the word has no meaning (0), one meaning (1), or more than one meaning (2) (e.g., Hino et al., 2006; Hino et al., 2010; Pexman et al., 2004). The unambiguous words included in this experiment were rated as having one meaning (mean NOM = 1.09) and ambiguous words were rated as having more than one meaning (mean NOM = 1.77). The second ambiguity measure requires the participants to generate word associates for a set of words. Then four judges classify the words as unambiguous or ambiguous according to the associates generated (see Haro, Ferré, et al., 2017, for more details). All the unambiguous words in this experiment had an "unambiguous" rating in this measure, and all the ambiguous words had an "ambiguous" rating.

The set of 56 ambiguous words included 28 homonyms and 28 polysemes. The words were classified as homonyms or polysemes according to the relatedness-of-meanings (ROM) ratings from Haro, Ferré, et al. (2017). To collect this measure, Haro et al. presented participants with two pairs of words for each ambiguous word. Each pair contained the ambiguous word and a strong associate to one of its meanings (as a way to indicate one of the meanings of the ambiguous word). The participants were asked to judge the relatedness of the meanings indicated by the pairs of words (from 1 = "unrelated meanings" to 9 = "the same meaning"). Low ROM ratings indicate that the meanings of the word are unrelated, while high ROM ratings indicate that the meanings are related. Words with ROM ratings below 2.5 were categorized as homonyms, and those with ROM ratings above 2.5 were categorized as polysemes (see Haro, Demestre, et al., 2017, Hino et al., 2010, and Hino et al., 2006, for similar procedures). Homonyms in this experiment had an average ROM of 1.81, and polysemes had an average ROM of 3.73, $t(54) = 10.20$, $p < 0.001$.

All the experimental conditions (ambiguous vs. unambiguous words, as well as homonyms vs. polysemes) were carefully matched in a large set of lexical and semantic variables (all $p_s > 0.05$; the descriptive statistics are presented in Table 1). Orthographic variables were taken from EsPal (Duchon et al., 2013). Concreteness, familiarity, and subjective age of acquisition ratings were retrieved from Haro, Ferré, et al. (2017). As age-of-acquisition ratings for 17 words were not available in Haro et al.'s database, we asked a sample of 20 participants to provide them.

Finally, we generated a set of 112 pronounceable nonwords from the 112 experimental words. To do this, we employed the Wuggy pseudoword generator (Keuleers & Brysbaert, 2010). Nonwords were matched with words in terms of length, number of syllables, subsyllabic structure, and transition frequencies.

2.1.3. Procedure

The LDT included 224 experimental trials. DMDX software (Forster & Forster, 2003) was used to present the stimuli and record the responses. Each trial started with a fixation point (i.e., "+") that appeared in the middle of the screen for 500 ms. After that, a string of letters (either a word or a nonword) replaced the fixation point, and the participants had to decide if the string was a Spanish word or not. They had to press the "YES" button on a keypad with their preferred hand if it was a word, or the "NO" button of the keypad with their non-preferred hand if it was not a word. The string of letters disappeared after either the participant's response or 2000 ms. The participants received a feedback message ("ERROR" or "CORRECT") after responding. The interval between trials was 750 ms. The order of the experimental trials was randomized for each participant. Before starting the experiment, the participants completed a practice block.

2.1.4. Data analyses

Here we detail the data cleaning and analyses performed in this experiment. We obtained a total of 8960 RTs from the 40 participants who responded to the 224 stimuli (112 words and 112 nonwords). We removed any RTs under 300 ms, as well as those that were 2 standard deviations above or below the mean for each participant (480 RTs). The overall accuracy was high (mean = 94.77 %), so we decided to not analyze the error data, and we removed the RTs from the error responses (397 RTs). Altogether we removed 877 RTs (9.79 % of the total), leaving a total of 8093 RTs for analysis.

We analysed the data using linear mixed-effect models (e.g., Baayen, 2008; Baayen et al., 2008) and a Bayes Factor analysis. We used the lme4 package from R for the linear mixed-effect models (Bates et al., 2019). Different linear models were generated to independently examine the effect of the variables of interest: ambiguity (ambiguous words vs. unambiguous words) and semantic relatedness (homonyms vs. polysemes) on inverse RTs (−1000/RT). First, we created a model in which we introduced the variable of interest (ambiguity or semantic relatedness) and several covariates for which there is evidence of their effect on LDT: log frequency, log contextual diversity, familiarity, age of acquisition, number of letters, number of syllables, concreteness, old20, number of

Table 1
Descriptive statistics of the stimuli used in the experiments.

		NOM	ROM	FRE	CTD	FAM	AoA	LNG	SYL	CON	OLD	NEI	NHF	BFQ
Unambiguous words	Mean	1.09	–	1.22	0.84	5.56	6.15	5.82	2.46	4.94	1.62	6.79	0.86	5458.11
	SD	0.07		0.66	0.52	1.05	2.30	1.77	0.76	1.15	0.45	7.24	1.83	3228.13
Ambiguous words	Mean	1.77	2.77	1.16	0.79	5.55	6.34	5.57	2.30	4.73	1.53	9.25	1.29	5584.63
	SD	0.15	1.19	0.37	0.30	0.66	1.73	1.06	0.57	0.59	0.41	10.01	2.11	3072.41
Polysemes	Mean	1.74	3.73	1.21	0.84	5.50	6.21	5.64	2.29	4.64	1.51	9.04	0.93	5282.75
	SD	0.15	0.94	0.39	0.30	0.82	1.71	0.99	0.53	0.59	0.38	9.65	1.90	2349.92
Homonyms	Mean	1.79	1.81	1.12	0.74	5.60	6.48	5.50	2.32	4.82	1.54	9.46	1.64	5886.51
	SD	0.14	0.34	0.36	0.29	0.48	1.77	1.14	0.61	0.59	0.45	10.54	2.28	3676.66

Note. NOM = subjective number-of-meanings ratings; ROM = subjective relatedness-of-meanings ratings; FRE = log frequency; CTD = log contextual diversity; FAM = familiarity; AoA = subjective age-of-acquisition; LNG = number of letters; SYL = number of syllables; CON = concreteness; OLD = old20; NEI = number of substitution neighbors; NHF = number of higher frequency substitution neighbors; BFQ = mean bigram frequency.

neighbors, number of higher frequency neighbors, and mean bigram frequency (see Table 1). We calculated the multicollinearity among the fixed effects introduced in the models (R VIF function) and removed those with a VIF > 3 (Zuur et al., 2010); in particular, we removed contextual diversity, number of syllables, and old20. Each model included random intercepts for participants and words. To analyze the effect of the variables of interest, we compared a model that included as fixed effects the variable of interest plus the covariates to one that only included the covariates as fixed effects. Comparisons were made using log-likelihood ratio tests (R ANOVA function). All covariates were centred and converted to Z scores. Ambiguity and semantic relatedness were encoded using sum contrast coding (ambiguity: −0.5 [unambiguous words], +0.5 [ambiguous words]; semantic relatedness: −0.5 [homonyms], +0.5 [polysemes]). We also report the results of the *t*-test analyses of the coefficient estimates for each fixed effect. To this end we used Satterthwaite's approximations to the degrees of freedom of the denominator (*p*-values were estimated by the lmerTest package, Kuznetsova et al., 2019).

In addition, the data were examined using Bayesian analyses. In order to do so, we used the Bayes Factor (BF_{10}). BF_{10} allowed us to quantify the amount of evidence for (H_1) and against (H_0) the effect of the variables of interest (i.e., ambiguity and semantic relatedness). The magnitude of this evidence is presented as an odds ratio (H_1 evidence/H_0 evidence), which can range from 0 to infinite. If the value increases, it provides evidence in favour of H_1; if it approaches 0, it provides evidence in favour of H_0. Values close to or equal to 1 indicate that both H_1 and H_0 are equally probable. By convention, values above 3 can be interpreted as moderate evidence supporting H_1, and values below 1/3 provide moderate support for H_0 (Dienes, 2014; Jeffreys, 1961). We used the BayesFactor package in R (Morey & Rouder, 2015) to perform these analyses. We compared a model that included the factor of interest (H_1) with one that did not (H_0) using the lmBF function. BF_{10} indicates the amount of evidence for H_1 relative to H_0, and its inverse, BF_{01}, is the evidence in favour of H_0 relative to H_1 (in which the interpretation is inverted, i.e., a BF_{01} of 3 suggests moderate evidence for H_0, whereas 1/3 is moderate evidence for H_1, and so on.). A Jeffreys-Zellner-Siow (JZS) prior with a scaling factor of r = 0.707 was used in all the analyses (Rouder et al., 2009).

2.2. Results and discussion

The mean RTs for each condition are shown in Table 2. The analyses showed a significant ambiguity effect: ambiguous words were recognized faster than unambiguous words (estimate = −0.08, $SE = 0.02$, $t = -4.64$, $p < 0.001$, $\chi^2\ (1) = 21.27$, $p < 0.001$; see Table 3). The BF_{10} was 372.53 (±0.60 %), suggesting that the data are 373 times more likely under H_1 (existence of an ambiguity effect), than under H_0 (absence of an ambiguity effect). According to Jeffreys' classification (1961), this constitutes "decisive evidence" ($BF_{10} > 100$) in favour of an ambiguity effect. On the other hand, the effect of semantic relatedness was not significant (estimate = 0.01, $SE = 0.02$, $t = 0.60$, $p = 0.551$, $\chi^2\ (1) = 0.43$, $p = 0.514$; see Table 4). The BF_{01} was 8.31 (±0.67 %), suggesting moderate evidence (Jeffreys, 1961) in favour of H_0; in other words, RTs showed no differences between polysemes and homonyms.

On the one hand, the results of this experiment evidenced a robust ambiguity effect, i.e., we found a processing advantage for words with multiple meanings over words with only one meaning, which is consistent with some studies that have observed a similar ambiguity advantage using LDT (e.g., Haro et al., 2019; Hino et al., 2010; Jastrzembski, 1981; Rubenstein et al., 1970). On the other hand, we found a null effect for semantic relatedness, which was strongly supported by the Bayesian analysis. This result was in line with previous studies that have reported no differences between polysemes and homonyms in LDT, showing a facilitation for both types of words relative to unambiguous words (Haro, Demestre, et al., 2017; Hino et al., 2006, 2010; Pexman et al., 2004). Conversely, such evidence contrasts with other reports of a homonym disadvantage in LDT (e.g., Klepousniotou & Baum, 2007; Rodd et al., 2002), and it seems incompatible with theoretical accounts suggesting that homonym meanings compete during lexical processing (Armstrong & Plaut, 2008, 2011; Rodd et al., 2004). However, it is possible that behavioural measures such as RTs may not be sufficiently sensitive to reflect this competition between meanings. For this reason, in the following experiment we recorded a neurophysiological measure – pupillary response – during an LDT in which we presented the same set of words as in Experiment 1.

Table 2
Mean RT and standard error (in parenthesis) of each condition.

Ambiguity	Semantic relatedness	RT
Unambiguous words		629 (3.46)
Ambiguous words		602 (3.03)
	Polysemes	601 (4.29)
	Homonyms	604 (4.29)

Table 3
Summary of effects in the ambiguity model.

	Estimate	*SE*	*t*	*p*
Intercept	−1.70	0.03	−52.67	<0.001
Ambiguity	−0.08	0.02	−4.64	<0.001
Log frequency	−0.04	0.01	−3.64	<0.001
Familiarity	−0.02	0.01	−1.86	0.062
Age of acquisition	0.03	0.01	2.25	0.025
Number of letters	0.04	0.01	3.52	<0.001
Concreteness	−0.01	0.01	−1.39	0.163
Number of neighbors	0.02	0.01	1.36	0.173
Number of higher frequency neighbors	−0.00	0.01	−0.01	0.995
Mean bigram frequency	−0.01	0.01	−0.76	0.447

Table 4
Summary of effects in the semantic relatedness model.

	Estimate	*SE*	*t*	*p*
Intercept	−1.74	0.03	−52.48	<0.001
Semantic relatedness	0.01	0.02	0.60	0.551
Log frequency	−0.03	0.02	−1.44	0.149
Familiarity	−0.02	0.02	−0.95	0.341
Age of acquisition	0.03	0.02	1.83	0.068
Number of letters	0.02	0.03	0.65	0.514
Concreteness	−0.00	0.02	−0.16	0.875
Number of neighbors	0.01	0.02	0.62	0.536
Number of higher frequency neighbors	−0.01	0.02	−0.80	0.422
Mean bigram frequency	−0.02	0.01	−1.16	0.245

3. Experiment 2: lexical decision task (pupillary response)

3.1. Materials and methods

3.1.1. Participants

Twenty-five Spanish speakers took part in this experiment. The participants were undergraduate students at Rovira i Virgili University (Tarragona, Spain), all of whom had either normal or corrected-to-normal vision. They received academic credits for their participation. Prior to the experiment, the participants signed an informed consent document. A local ethics committee at the Universitat Rovira i Virgili approved this research (CEIPSA-2021-PR-0044).

3.1.2. Materials

The same 112 words and 112 nonwords used in Experiment 1 were employed in this experiment.

3.1.3. Procedure

The participants were examined individually in a medium-

illuminated room. They sat with their head on a chinrest with forehead support. The chinrest was adjusted for each participant to stabilize their head and maintain a constant distance of 60 cm between their eyes and the monitor, a 19″ computer screen with a resolution of 1024 × 768 pixels. The diameter and position of the right pupil was recorded continuously at a sampling rate of 1000 Hz, using an EyeLink 1000 eye tracker.

Experiment Builder software was used to present the stimuli for the experiment. These were displayed in lowercase black characters (Arial font, 24 pixels) in the centre of a screen with a grey background (RGB 150). The participants performed a delayed response LDT. Each trial began with the presentation of a fixation point ("+") in the centre of the screen for 1000 ms. The fixation point was then replaced by a string of letters representing either a Spanish word or a nonword. The string of letters was displayed for 500 ms, followed by a new fixation point for 1500 ms. After that, a question mark ("?") appeared on the screen asking participants to indicate whether the string of letters was a Spanish word or not. They responded by pressing the mouse button labelled "YES" (left button) or "NO" (right button) with their right hand. The question remained on the screen for 2000 ms or until an answer was given. The 224 experimental stimuli were presented in a different random order for each participant. The trials were divided into two blocks. Between the blocks, the participants could take a short break. Before the experimental trials, the participants completed a practice session. A calibration routine was performed at the beginning of the experiment as well as after the break.

3.1.4. Data cleaning and analysis

The pupil data were processed and analysed using software designed for this purpose (CHAP, Hershman et al., 2019). This software cleans data prior to analysis, removing both outliers within each trial and any trials with a specified percentage of missing samples. Following this procedure, the samples from each trial with *Z*-scores of >2.5 from the mean pupil dilation of each participant were removed. Blinks were also detected and removed, and these data points were replaced through linear interpolation (applying the Hershman & Henik method [Hershman et al., 2018], included in the CHAP software). Trials with >20 % missing samples and those in which the participant did not respond or responded incorrectly were also discarded. In total, the cleaning process removed <5 % of the data. One participant was excluded from the analyses due to a low number of valid trials once the data had been cleaned. Baseline pupil dilation was defined by averaging the pupil dilation over the 200 ms prior to the onset of the target (while the fixation point was displayed). The pupil dilation from each trial was converted to a relative dilation expressed as a proportional difference (in percentage change) from the baseline pupil dilation.

To examine the differences between conditions, we compared the pupil dilation in each experimental condition over the course of the trial; in particular, from the onset of the target to the presentation of the question (from 0 to 2000 ms). Following the analysis procedure included in CHAP (Hershman et al., 2019), we performed Bayesian paired sample *t*-tests between the conditions over the time range (in 5 ms bins). As in Experiment 1, we used a JZS prior with a scaling factor of r = 0.707 (Rouder et al., 2009).

3.2. Results and discussion

First, to verify the reliability of the pupillary data collected in the experiment, we examined the lexicality effect – the difference in pupillary response between words and nonwords. The Bayesian paired sample t-tests showed moderate ($BF_{10} > 3$; Jeffreys, 1961) to very strong ($BF_{10} > 30$) evidence of a lexicality effect from 1032 to 1137 ms; extreme evidence ($BF_{10} > 100$) from 1137 to 1447 ms, and very strong to moderate evidence from 1447 to 1673 ms (see Fig. 1). In this time range, nonwords elicited greater pupil dilation than words. On the other hand, the analysis of the ambiguity effect revealed moderate evidence ($BF_{01} > 3$) for a null ambiguity effect from 571 to 1142 ms, and from 1508 to 1818 ms (see Fig. 2). Finally, we found moderate evidence ($BF_{01} > 3$) for an absence of a semantic-relatedness effect over the entire time range (see Fig. 3).

The results of this experiment showed null effects for both ambiguity and semantic relatedness. One explanation for these null effects may be that the pupillary response was not sensitive to the properties of the stimuli presented in the LDT. Although this is a reasonable possibility,

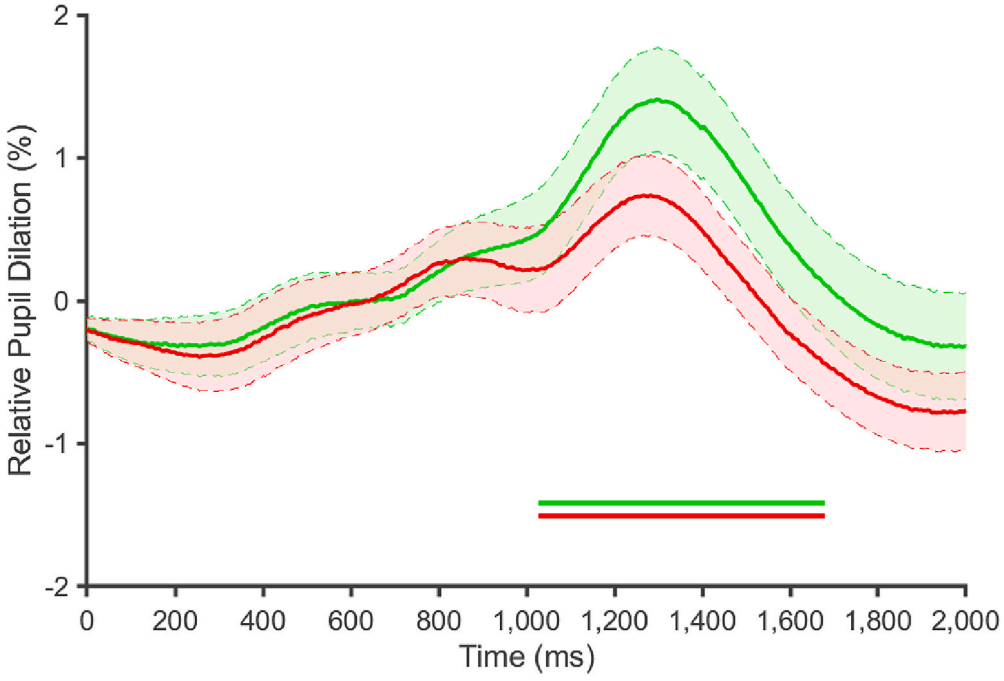


Fig. 1. Relative pupil dilation (expressed in percentage) for nonwords (green line) and words (red line) from the onset of the stimulus until response. The area around each line represents the standard error. The horizontal lines represent the time range where BF10 was above 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

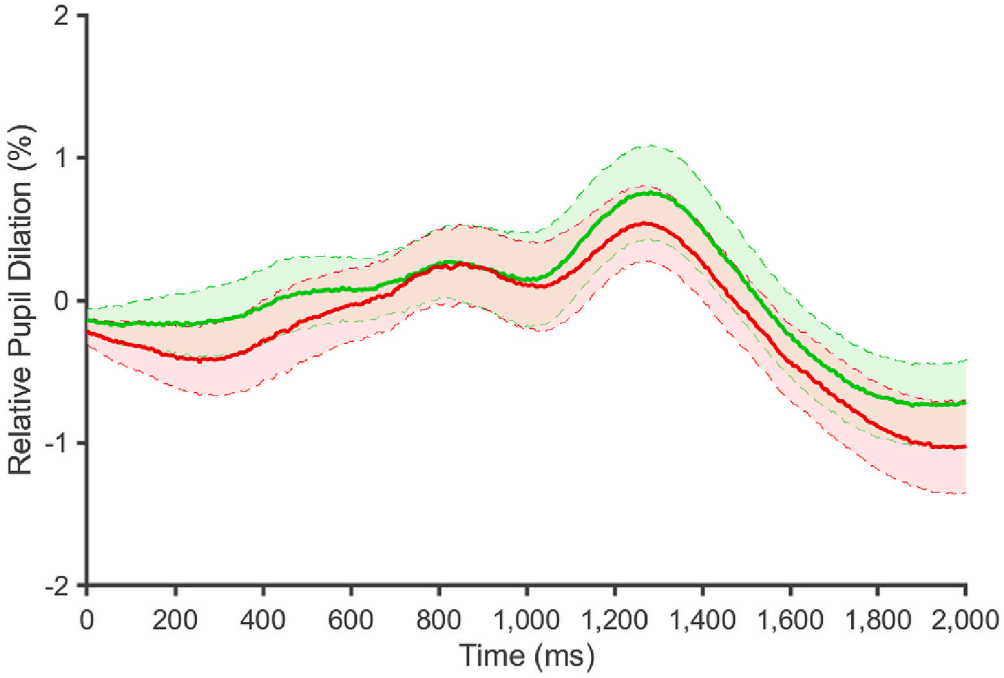


Fig. 2. Relative pupil dilation (expressed in percentage) for unambiguous words (green line) and ambiguous words (red line) from the onset of the stimulus until response. The area around each line represents the standard error. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

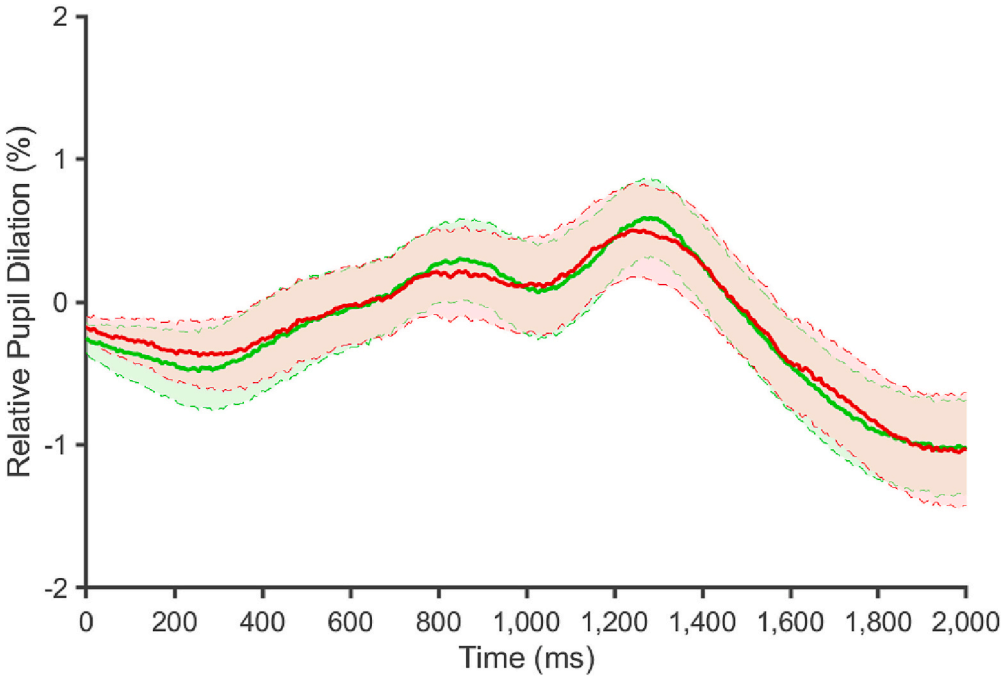


Fig. 3. Relative pupil dilation (expressed in percentage) for polysemes (green line) and homonyms (red line) from the onset of the stimulus until response. The area around each line represents the standard error. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the robust lexicality effect in the pupillary response does not support this hypothesis. The lexicality effect suggests that nonwords elicited greater processing effort than words, indicating that the pupillary response was modulated by the properties of the stimuli presented. This effect on the pupillary response might be due to the decision-making process. Checking that a nonword is not a word requires an exhaustive process of ensuring that no representation matches the stimulus, which is more cognitively demanding than checking that a word is mentally represented. However, if the relatedness between the meanings of ambiguous words influences the activation and competition between semantic representations, one would expect this to be also reflected in the word-nonword decision-making process, and thus in the pupillary response. Specifically, there should be an inhibition for homonyms compared to polysemes, since the competition between the meanings of the

homonyms would interfere with the process of checking whether their semantic representation correspond to the presented stimulus.

Nevertheless, the null effect of semantic relatedness may also be explained by the fact that the LDT is not semantically demanding enough to elicit competition between the meanings of ambiguous words (Armstrong & Plaut, 2008, 2011), and that in such a task the distinct meanings of homonyms do not accumulate sufficient activation. If this is so, a more semantically demanding task is needed to produce observable differences between homonyms and polysemes. To examine this possibility, in the subsequent experiment we presented the same words in an SCT, in which participants had to decide if each word belonged to a given category, specifically "jobs, professions and ranks". We chose this category because a previous study (Hino et al., 2006) only observed differences between homonyms and polysemes in SCT when using broad categories.

4. Experiment 3: semantic categorization task (pupillary response)

4.1. Materials and methods

4.1.1. Participants

Twenty-eight Spanish speakers took part in this experiment. The participants were undergraduate students at Rovira i Virgili University (Tarragona, Spain), all of whom had either normal or corrected-to-normal vision. They received academic credits for their participation. Prior to the experiment, the participants signed an informed consent document. A local ethics committee at the Universitat Rovira i Virgili approved this research (CEIPSA-2021-PR-0044).

4.1.2. Materials

We used the same set of stimuli as in Experiments 1 and 2, although we had to remove some words as these had a relationship with the category used in the SCT ("jobs, professions and ranks"); specifically, we removed three homonyms and, in order to match the number of stimuli between conditions, we removed three polysemes of similar lexical and semantic characteristics as the homonyms. Likewise, we removed six unambiguous words with similar properties to the aforementioned six ambiguous words. We therefore presented a total of 50 ambiguous words (25 homonyms and 25 polysemes) and 50 unambiguous words. In addition, we selected a further 100 words related to the category "jobs, professions and ranks". These words were included as fillers required to perform the task. The lexical values of these words were comparable to those of the ambiguous and unambiguous words.

4.1.3. Procedure

The procedure was identical to that of Experiment 2, with the difference being that in this case the participants had to indicate if the word belonged to the category "jobs, professions and ranks" or not. They responded by using their right hand to press the mouse button labelled "YES" (left button) or "NO" (right button). The expected response to the critical stimuli (ambiguous and unambiguous words) was "NO", and to the fillers "YES". As in Experiment 2, the response was delayed until two seconds after stimulus onset.

4.1.4. Data cleaning and analysis

The same data cleaning and analysis procedure was applied in this experiment as in Experiment 2. The cleaning removed <5 % of the data. Three participants were excluded from the analyses due to a high number of errors (>20 %). One participant was excluded from the analyses because of a low number of valid trials after data cleaning. The data from 24 participants were therefore included in the analyses.

4.2. Results and discussion

As in Experiment 2, to check the reliability of the pupillary data obtained, we examined the category factor effect using Bayesian paired sample *t*-tests to compare pupillary dilation over the time course between words belonging to the category "jobs, professions and ranks" and words not from that category (i.e., the set of ambiguous and unambiguous experimental words). The results showed moderate ($BF_{10} > 3$; Jeffreys, 1961) to very strong evidence ($BF_{10} > 30$) for a category effect from 736 to 851 ms, and extreme evidence ($BF_{10} > 100$) from 851 to 2000 ms (see Fig. 4). Ambiguous and unambiguous words elicited smaller pupil dilations than words referring to the category "jobs, professions and ranks". On the other hand, Bayesian analyses showed moderate evidence ($BF_{01} > 3$) of a null effect for ambiguity from the onset of the word up to 2000 ms, indicating that there were no differences in pupil dilation between ambiguous and unambiguous words over the time period (see Fig. 5). Finally, the analyses showed moderate ($BF_{01} > 3$) evidence of a null effect for semantic relatedness in the following time ranges: 0 to 200 ms, 1052 to 1157 ms, and 1468 to 1868 ms (see Fig. 6).

The results of this experiment are in line with those observed in Experiment 2. Firstly, as in Experiment 2, we obtained evidence suggesting that pupillary response is sensitive to word properties, as category-congruent words showed greater pupillary dilation than category-incongruent words. As with the word-nonword decision, this effect on pupillary response might be related to the decision-making process. The results seem to suggest that it was more cognitively demanding to check that a word referring to "jobs, professions and ranks" matched this category than it was to check that a word not referring to this category did not belong to it. This greater load for congruent trials in SCT seems counterintuitive at first glance, but it is worth noting that congruent trials in SCT tasks do not necessarily show faster times than incongruent trials (e.g., see Experiments 3 and 5 in Hino et al., 2006). Furthermore, the response of the SCT used in the present experiment was delayed, so the process of matching the stimulus to the category would not be directly comparable to that in immediate response SCT.

However, we again found a null effect for semantic relatedness. It should be noted that Hino et al. (2006) observed differences between polysemes and homonyms in a SCT with a broad category, like the one used in the current experiment. But these differences did not appear when they presented the same words in a SCT with a narrower category. This led them to conclude that the semantic relatedness effect in SCT was a result of the operations performed during the decision-making process. It is possible that the delayed response in the present experiment would reduce the effects of the decision component by providing more time to make a response, and this may also explain why we did not observe an ambiguity effect in SCT (this issue will be discussed in more detail later). In any case, by providing more time for their meanings to be activated and compete, this should have left more room for differences between polysemes and homonyms to emerge. Thus, these results seem inconsistent with that proposals suggesting that the meanings of homonyms compete during word processing (Armstrong & Plaut, 2008, 2011; Rodd et al., 2004).

Finally, it is also possible that, as in the case of the LDT, the SCT is not sufficiently demanding to elicit differences between homonyms and polysemes. In other words, the task may not allow the semantic representations of the homonyms to accumulate sufficient activation and hence compete with one another. To rule out this possibility, in the subsequent experiment we employed an even more semantically demanding task, in which the responses required awareness of the multiplicity of meanings of ambiguous words. Specifically, we asked the participants to indicate the number of meanings the words had. Our hypothesis was that, in order perform this task, the participants would activate all the semantic representations of the ambiguous words, which should lead to competition between the meanings of homonyms, but not of polysemes.

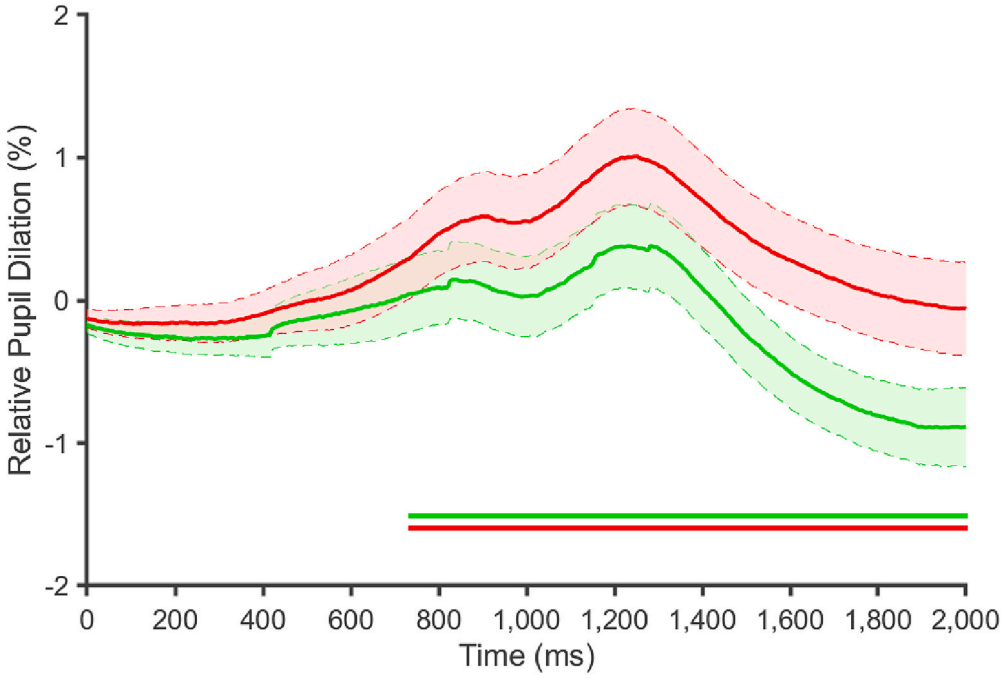


Fig. 4. Relative pupil dilation (expressed in percentage) for ambiguous and unambiguous words (green line) and words related to the category "jobs, professions and ranks" (red line) from the onset of the stimulus until response. The area around each line represents the standard error. The horizontal lines represent the time range where BF_{10} was above 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

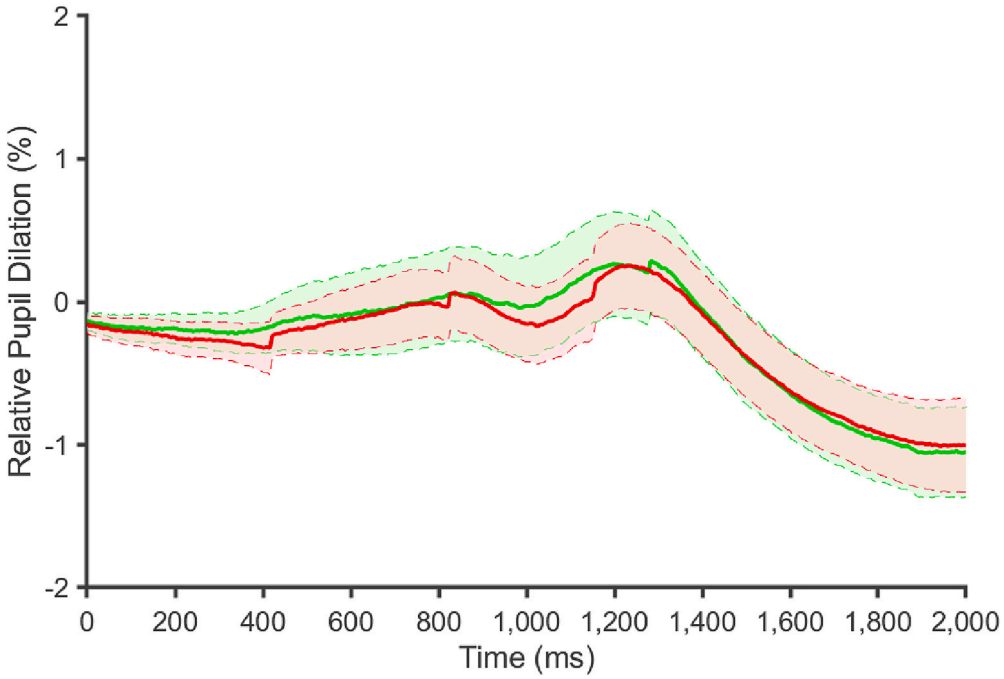


Fig. 5. Relative pupil dilation (expressed in percentage) for unambiguous words (green line) and ambiguous words (red line) from the onset of the stimulus until response. The area around each line represents the standard error. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Experiment 4: number-of-meanings task (pupillary response)

5.1. Materials and methods

5.1.1. Participants

Twenty-four Spanish speakers took part in this experiment. The participants were undergraduate students at Rovira i Virgili University (Tarragona, Spain), all of whom had either normal or corrected-to-normal vision. They received academic credits for their participation. Prior to the experiment, the participants signed an informed consent document. A local ethics committee at the Universitat Rovira i Virgili approved this research (CEIPSA-2021-PR-0044).

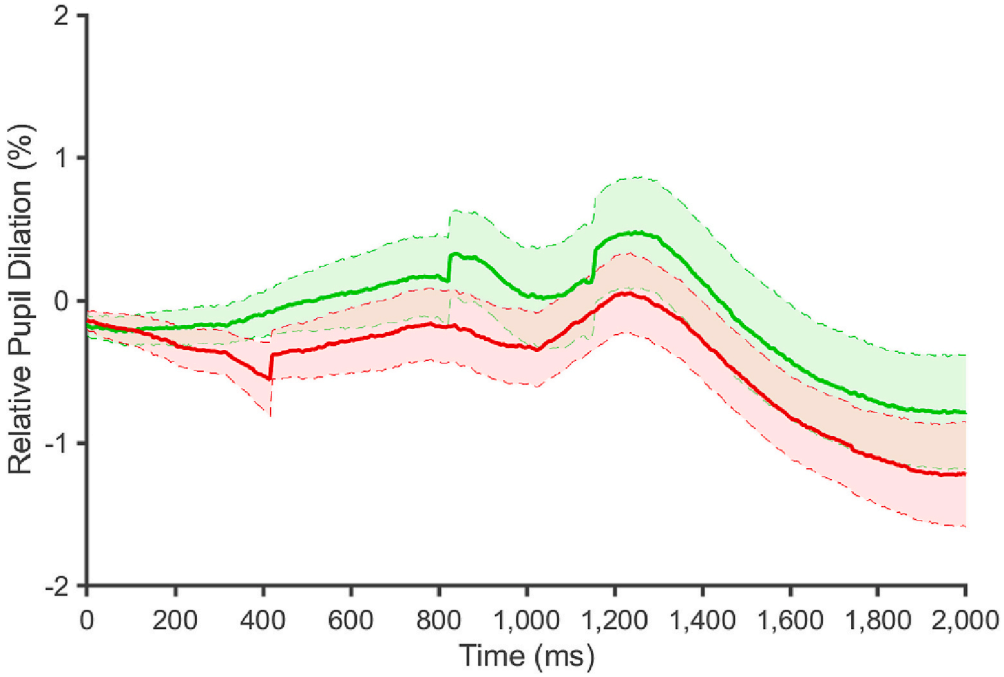


Fig. 6. Relative pupil dilation (expressed in percentage) for polysemes (green line) and homonyms (red line) from the onset of the stimulus until response. The area around each line represents the standard error. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.1.2. Materials

We used the same set of ambiguous and unambiguous words as in Experiment 3, but excluded the 100 filler words related to the category "jobs, professions and ranks".

5.1.3. Procedure

The procedure was identical to that of Experiments 2 and 3, although in this case the participants had to indicate whether the word presented to them had one meaning or more than one meaning. They responded by using their right hand to press either the left mouse button if the word had one meaning, or the right button if it had more than one meaning.

5.1.4. Data cleaning and analysis

We applied the same data cleaning and analysis procedure used in Experiments 2 and 3. The cleaning process removed <5 % of the data.

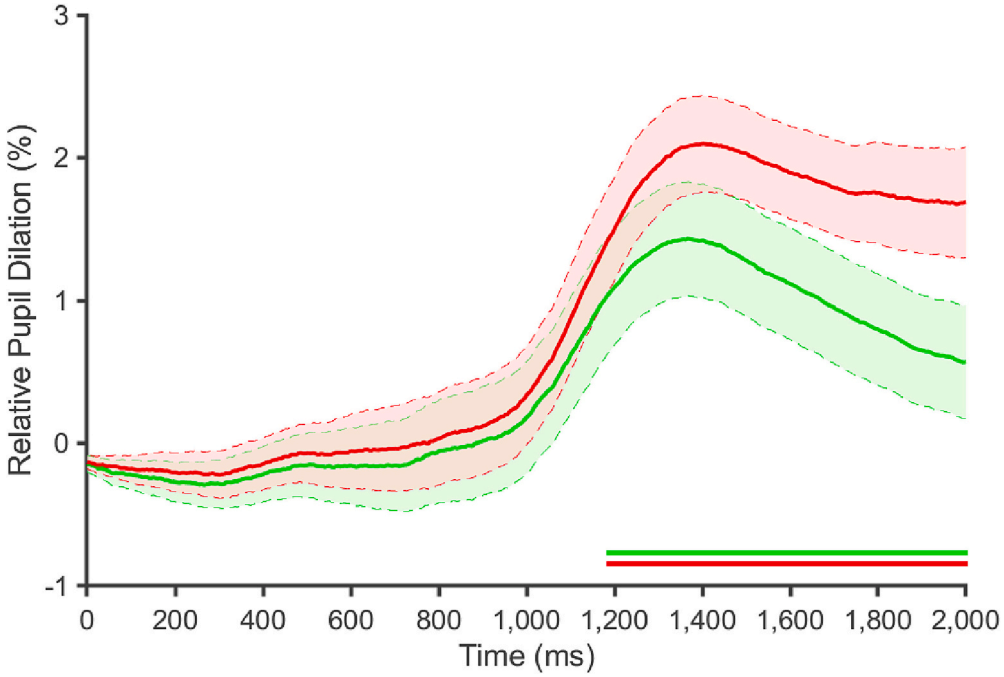


Fig. 7. Relative pupil dilation (expressed in percentage) for unambiguous words (green line) and ambiguous words (red line) from the onset of the stimulus until response. The area around each line represents the standard error. The horizontal lines represent the time range where BF_{10} was above 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.2. Results and discussion

Bayesian paired sample *t*-tests showed moderate evidence ($BF_{10} > 3$; Jeffreys, 1961) of an ambiguity effect from 1187 to 1262 ms; strong evidence ($BF_{10} > 10$) from 1262 to 1768 ms; very strong ($BF_{10} > 30$) from 1768 to 1903 ms, and extreme evidence ($BF_{10} > 100$) from 1903 to 2000 ms. Ambiguous words showed greater pupillary response than unambiguous words (see Fig. 7). On the other hand, the results of the analyses showed moderate evidence ($BF_{01} > 3$) of a null effect for semantic relatedness over almost the entire time period (except for the time range between 1303 and 1658 ms, where the evidence was anecdotal; $BF_{01} > 1$) (see Fig. 8).

In this experiment we observed a clear ambiguity effect, which provides strong evidence that the participants activated the different meanings of the ambiguous words before responding. Since ambiguous words have more semantic information than unambiguous words, the process of verifying whether they have one or more meanings must be more costly for ambiguous words compared to unambiguous words. And this higher cognitive cost would be reflected in the increased pupillary response to ambiguous words.

However, for the third time, no semantic relatedness effect was observed in the pupillary response. This is more striking than in the two previous experiments, since here the ambiguity effect suggests that semantic representations of ambiguous words were more extensively activated than in LDT and SCT. Thus, contrary to what Rodd et al. (2004) and Armstrong and Plaut (2008, 2011) hypothesised, although the meanings of ambiguous words would have been highly activated during this task, there was no evidence of competition between semantic representations of homonyms.

6. General discussion

The aim of this study was to examine how polysemes and homonyms are processed in different experimental tasks. We manipulated the semantic engagement of the tasks to test whether differences between polysemes and homonyms emerge or increase as a function of the demands of the experimental task. Specifically, the same set of polysemes, homonyms and unambiguous words was presented in two LDTs, an SCT, and a number-of-meanings task. In addition to behavioural measures (Experiment 1), we also recorded the pupillary response (Experiments 2, 3, 4) of the participants. No differences between polysemes and homonyms were observed in the LDT, in either behavioural measures or pupillary responses. Similarly, the pupillary response to polysemes and homonyms was virtually identical in the SCT and the number-of-meanings task. On the other hand, ambiguous words (i.e., polysemous and homonymous words considered together) were associated with faster responses in the LDT and greater pupillary responses in the number-of-meanings task than unambiguous words. Nevertheless, no pupillary response differences between ambiguous and unambiguous words were observed in either the LDT or the SCT.

The behavioural results are consistent with studies that found no differences between polysemes and homonyms in LDT (e.g., Hino et al., 2006, 2010; Pexman et al., 2004). However, this evidence is incompatible with reports showing a homonym-recognition disadvantage compared to polysemes in the same task (e.g., Armstrong & Plaut, 2008; Rodd et al., 2002). More importantly, the pupillary response failed to provide evidence of inhibition during the processing of homonymous words in any of the three tasks examined here. This null effect on pupillary response contrasts with that obtained through another neurophysiological measure (MEG), as used by Beretta et al. (2005), who reported longer latencies in the M350 component in homonyms compared to polysemes in an LDT. However, in a later ERP study conducted in our laboratory (Haro, Demestre, et al., 2017), we observed comparable amplitudes in the N400 component between polysemes and homonyms. As argued in Haro, Demestre, et al. (2017) and Haro and Ferré (2018), these discrepancies between the study by Beretta et al. and Haro et al. could be attributed to the categorization and word selection methods employed in the two studies. While Beretta et al. used dictionary measures to categorize and select homonyms and polysemes for their study, Haro et al. used subjective measures (provided by the participants). It is important to note that, in the present study, subjective measures were also used to determine the relatedness of ambiguous word meanings, just as they were in those studies reporting a lack of homonymy disadvantage in word recognition (Hino et al., 2006, 2010; Pexman et al., 2004). In contrast, studies that found a homonym disadvantage in LDT used dictionary measures to categorize and select polysemes and homonyms (Armstrong & Plaut, 2008, 2011, 2016; Rodd et al., 2002). It is thus plausible that the finding of a disadvantage for

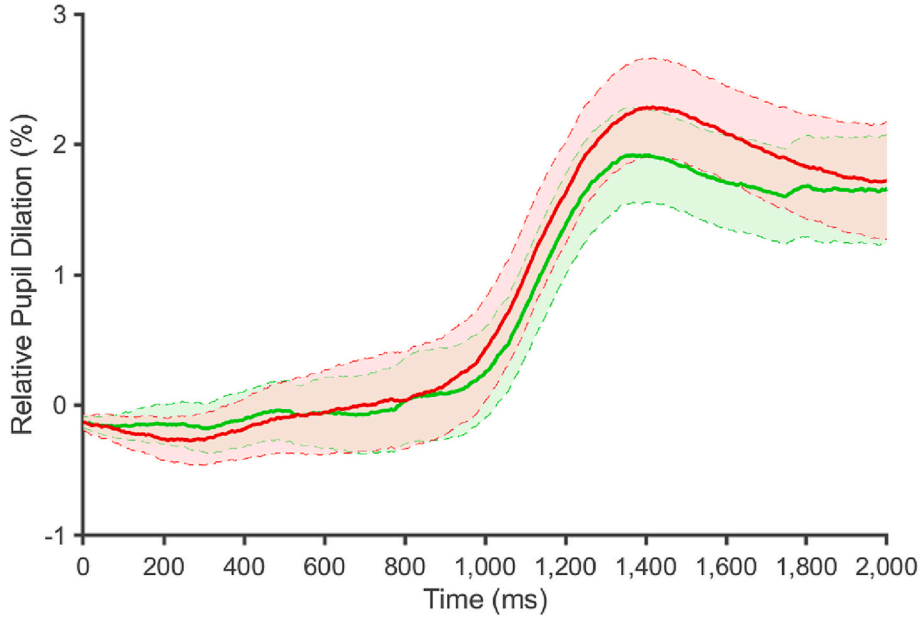


Fig. 8. Relative pupil dilation (expressed in percentage) for polysemes (green line) and homonyms (red line) from the onset of the stimulus until response. The area around each line represents the standard error. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

homonyms in some studies and a null effect in others may be due to the type of meaning-relatedness measure used in each case. This raises the question of the psychological validity of the measures, and of which is more appropriate for studying the processing of ambiguous words (see Haro & Ferré, 2018, for more on this issue).

In this study we found no evidence supporting either the model of Rodd et al. (2004) or that of Armstrong and Plaut (2008, 2011). According to these models, the meanings of homonyms, as opposed to those of polysemes, are represented independently and compete during word processing. Moreover, this competition is expected to increase according to the demand level of the experimental task (Armstrong & Plaut, 2008, 2011). Tasks that require deeper word processing, where access to the semantic properties of words is required, or where more time is needed to provide a response, should allow the meanings of homonyms to accumulate greater activation, thus increasing the competition between them. However, we observed no differences in pupillary response to polysemes and homonyms in any task, even in those that required deeper semantic processing of the stimuli (SCT and number-of-meanings task).

A possible explanation for the failure of our study to detect a semantic relatedness effect is that pupillary response is not a sensitive measure of lexical processing. Nevertheless, there is substantial evidence demonstrating that this measure is influenced by certain psycholinguistic variables (Geller et al., 2016; Guasch et al., 2017; Haro, Guasch, et al., 2017; Kuchinke et al., 2007; Papesh & Goldinger, 2012; Schmidtke, 2014). Indeed, in our study we observed that the pupillary response was sensitive to the lexical status of the stimuli in the LDT and to the semantic category of the word in the SCT. More importantly, in the number-of-meanings task we observed increased pupillary response for ambiguous words compared to unambiguous words. Despite all the above, it should be noted that the number of stimuli used to examine semantic relatedness in the present study was half that used to examine ambiguity (56 vs 112). This may have limited the statistical power to detect a semantic relatedness effect if we assume that such effect is small and that pupillometry is not as sensitive as a measure for word processing tasks as it is for more cognitively demanding tasks (e.g., Laeng et al., 2012). Finally, there is also the possibility that certain variables that have recently been shown to affect the processing of ambiguous words may also influenced the results; for example, the emotional content of each of the meanings of ambiguous words (Ferré et al., 2021). In any case, it seems that further research is needed to address these issues.

A final question we consider worth addressing is why we found no ambiguity effect in the pupillary response in either the LDT or the SCT. Several factors may be involved. First, as far as we know, this is the first time the ambiguity effect of words presented in isolation has been examined in a delayed-response task. We used this methodological approach to obtain a clearer picture of word processing, in an attempt to reduce the effects caused by decision-making and response execution. In this way, the participants are not subjected to the temporal pressure present in immediate-response tasks and, therefore, the processing and retrieval of semantic word information is likely to be spread out over time, rather than concentrated into the period immediately preceding and during the response. It is therefore likely that our use of delayed tasks attenuated the differences between unambiguous and ambiguous words in terms of the activation of their semantic representations during the LDT and SCT. Finally, it is also feasible that the LDT and SCT are not semantically demanding enough to trigger an ambiguity effect in the pupillary response. A task requiring deeper processing of the stimuli would therefore be required to elicit such an effect. Indeed, this possibility seems to be compatible with the ambiguity effect in the pupillary response observed in the number-of-meanings task (Experiment 4), as well as that found in a previous study using a sentence comprehension task (Kadem et al., 2020). Both tasks are assumed to require deeper processing than the LDT or the SCT.

In summary, in this study we found no evidence that polysemes and homonyms are processed differently. We consider that further research is needed to shed more light on how the two types of words are represented and processed. It would be interesting for future studies to examine the processing of these words in different tasks with a high level of semantic engagement.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Ministerio de Ciencia e Innovación (grant number PID2019-107206GB-I00), by the Ministerio de Ciencia, Innovación y Universidades (grant number RED2018-102615-T), and by the Universitat Rovira i Virgili (grant number 2018PFR-URV-B2-32). The second author also holds a grant from the Agencia Estatal de Investigación (FFI2017-82460-P) and from the Gobierno de Aragón (UZ-DGA-H11-17R).

Appendix A. Experimental stimuli

word	Ambiguity	Semantic relatedness
ácido	Ambiguous	Homonym
acuario	Ambiguous	Homonym
baja	Ambiguous	Homonym
burbuja	Ambiguous	Homonym
campaña	Ambiguous	Homonym
caña	Ambiguous	Homonym
colonia	Ambiguous	Homonym
cómoda	Ambiguous	Homonym
copa	Ambiguous	Homonym
ficha	Ambiguous	Homonym
fuente	Ambiguous	Homonym
guión	Ambiguous	Homonym
heroína	Ambiguous	Homonym
jota	Ambiguous	Homonym
lima	Ambiguous	Homonym
matriz	Ambiguous	Homonym

(*continued on next page*)

(*continued*)

mona	Ambiguous	Homonym
monitor	Ambiguous	Homonym
palma	Ambiguous	Homonym
pasta	Ambiguous	Homonym
patrón	Ambiguous	Homonym
pensión	Ambiguous	Homonym
perfil	Ambiguous	Homonym
pipa	Ambiguous	Homonym
recto	Ambiguous	Homonym
sirena	Ambiguous	Homonym
tabla	Ambiguous	Homonym
tanque	Ambiguous	Homonym
acento	Ambiguous	Polyseme
activo	Ambiguous	Polyseme
barra	Ambiguous	Polyseme
bestia	Ambiguous	Polyseme
billete	Ambiguous	Polyseme
bombón	Ambiguous	Polyseme
capa	Ambiguous	Polyseme
cartas	Ambiguous	Polyseme
cólera	Ambiguous	Polyseme
damas	Ambiguous	Polyseme
fracción	Ambiguous	Polyseme
genio	Ambiguous	Polyseme
globo	Ambiguous	Polyseme
grano	Ambiguous	Polyseme
letra	Ambiguous	Polyseme
manual	Ambiguous	Polyseme
marca	Ambiguous	Polyseme
pasajero	Ambiguous	Polyseme
pluma	Ambiguous	Polyseme
rosa	Ambiguous	Polyseme
señal	Ambiguous	Polyseme
solar	Ambiguous	Polyseme
sólido	Ambiguous	Polyseme
talla	Ambiguous	Polyseme
titular	Ambiguous	Polyseme
tronco	Ambiguous	Polyseme
virgen	Ambiguous	Polyseme
vocal	Ambiguous	Polyseme
aceite	Unambiguous	
acero	Unambiguous	
agua	Unambiguous	
alma	Unambiguous	
almirante	Unambiguous	
amar	Unambiguous	
barranco	Unambiguous	
baúl	Unambiguous	
biólogo	Unambiguous	
calor	Unambiguous	
casta	Unambiguous	
cerilla	Unambiguous	
cerveza	Unambiguous	
coágulo	Unambiguous	
cofre	Unambiguous	
comercio	Unambiguous	
cóndor	Unambiguous	
contusión	Unambiguous	
cuestionario	Unambiguous	
década	Unambiguous	
domingo	Unambiguous	
ecuación	Unambiguous	
error	Unambiguous	
fe	Unambiguous	
flores	Unambiguous	
geología	Unambiguous	
guitarra	Unambiguous	
hallar	Unambiguous	
hélice	Unambiguous	
hijo	Unambiguous	
himno	Unambiguous	
humo	Unambiguous	
ira	Unambiguous	
jabón	Unambiguous	
jeringa	Unambiguous	
junio	Unambiguous	
legado	Unambiguous	

(*continued on next page*)

(*continued*)

llegar	Unambiguous
lograr	Unambiguous
mente	Unambiguous
miel	Unambiguous
neutrón	Unambiguous
optar	Unambiguous
pan	Unambiguous
paraguas	Unambiguous
pensar	Unambiguous
rato	Unambiguous
recado	Unambiguous
riñón	Unambiguous
sobrina	Unambiguous
sombra	Unambiguous
tarea	Unambiguous
teclado	Unambiguous
usar	Unambiguous
vejez	Unambiguous
zona	Unambiguous

References

Armstrong, B. C., & Plaut, D. C. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 273–278).

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 *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 2223–2228).

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

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

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*, 390–412.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … Fox, J. (2019). lme4: Linear mixed-effects models using 'Eigen' and S4. R package version 1.1-21 [computer software]. http://CRAN.R-project.org/package=lme4.

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.

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

Brown, S. W. (2008). Polysemy in the mental lexicon. *Colorado Research in Linguistics, 21*, 1–12.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781.

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

Eddington, C. M., & Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing: The current state of the literature. *Psychonomic Bulletin & Review, 22*(1), 13–37.

Ferré, P., Haro, J., Huete-Pérez, D., & Fraga, I. (2021). Emotionality effects in ambiguous word recognition: The crucial role of the affective congruence between distinct meanings of ambiguous words. *Quarterly Journal of Experimental Psychology, 74*(7), 1234–1243.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35*(1), 116–124.

Geller, J., Still, M. L., & Morris, A. L. (2016). Eyes wide open: Pupil size as a proxy for inhibition in the masked-priming paradigm. *Memory & Cognition, 44*(4), 554–564.

Guasch, M., Ferre, P., & Haro, J. (2017). Pupil dilation is sensitive to the cognate status of words: Further evidence for non-selectivity in bilingual lexical access. *Bilingualism, 20*(1), 49–54.

Haro, J., Comesaña, M., & Ferré, P. (2019). Is there an orthographic boost for ambiguous words during their processing? *Journal of Psycholinguistic Research, 48*(2), 519–534.

Haro, J., Demestre, J., Boada, R., & Ferré, P. (2017). ERP and behavioral effects of semantic ambiguity in a lexical decision task. *Journal of Neurolinguistics, 44*, 190–202.

Haro, J., & Ferré, P. (2018). Semantic ambiguity: Do multiple meanings inhibit or facilitate word recognition? *Journal of Psycholinguistic Research, 47*(3), 679–698.

Haro, J., Ferré, P., Boada, R., & Demestre, J. (2017). Semantic ambiguity norms for 530 Spanish words. *Applied PsychoLinguistics, 38*(2), 457–475.

Haro, J., Guasch, M., Vallès, B., & Ferré, P. (2017). Is pupillary response a reliable index of word recognition? Evidence from a delayed lexical decision task. *Behavior Research Methods, 49*(5), 1930–1938.

Hershman, R., Henik, A., & Cohen, N. (2018). A novel blink detection method based on pupillometry noise. *Behavior Research Methods, 50*(1), 107–114.

Hershman, R., Henik, A., & Cohen, N. (2019). CHAP: Open-source software for processing and analyzing pupillometry data. *Behavior Research Methods, 51*(3), 1059–1074.

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.

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.

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.

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology, 13*(2), 278–305.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). New York: Oxford University Press.

Kadem, M., Herrmann, B., Rodd, J. M., & Johnsrude, I. S. (2020). Pupil dilation is sensitive to semantic ambiguity and acoustic degradation. *Trends in Hearing, 24*, 1–16.

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

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.

Kuchinke, L., Võ, M. L. H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology, 65*(2), 132–140.

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B., & Jensen, S. P. (2019). lmerTest: Tests for random and fixed effects for linear mixed effect models. R package version 3.1-1 [computer software]. http://CRAN.R-project.org/package=lmerTest.

Laeng, B., Sirois, S., & Gredeback, G. (2012). Pupillometry: A Window to the preconscious? *Perspectives on Psychological Science, 7*(1), 18–27.

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.

Morey, R. D., & Rouder, J. N. (2015). *{BAYESFACTOR}: Computation of Bayes factors for common designs*. R package version 0.9.12-2. Available online at: https://cran.r-project.org/web/packages/BayesFactor/.

Papesh, M. H., & Goldinger, S. D. (2012). Pupil-BLAH-metry: Cognitive effort in speech planning reflected by pupil dilation. *Attention, Perception, & Psychophysics, 74*(4), 754–765.

Pexman, P. M., 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.

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

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.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237.

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.

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: Evidence from pupillometry. *Frontiers in Psychology, 5*, 137.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*(1), 3–14.