Polysemy in the Mental Lexicon:

Relatedness and Frequency affect Representational Overlap

Bernadet Jager, Matthew J. Green, & Alexandra A. Cleland

University of Aberdeen

B. Jager (corresponding author)

School of Psychology, University of Aberdeen, William Guild Building,

Aberdeen, AB24 3FX, UK.

e-mail: bernadetjager@gmail.com

M. J. Green

Department of Computing Science, University of Aberdeen, Meston Building

Aberdeen, AB24 3UE, UK.

e-mail: mjgreen@abdn.ac.uk

A. A. Cleland

School of Psychology, University of Aberdeen, William Guild Building,

Aberdeen, AB24 3FX, UK.

e-mail: a.cleland@abdn.ac.uk

Abstract

Meaning relatedness affects storage of ambiguous words in the mental lexicon: unrelated meanings (homonymy) are stored separately whereas related senses (polysemy) are stored as one large representational entry. We hypothesized that word frequency could have similar effects on storage, with low-frequency words having high representational overlap and high-frequency words having low representational overlap. Participants performed lexical decision or semantic categorization to high- and low-frequency nouns with few and many senses. Results showed a three-way interaction between frequency, task type, and polysemy. Low-frequency words showed a polysemy advantage with lexical decision but a polysemy disadvantage with semantic categorization, whereas high-frequency words showed the opposite pattern. These results confirmed our hypothesis that relatedness and word frequency have similar effects on storage of ambiguous words.

Key words: polysemy, lexical ambiguity, relatedness, word frequency, representational overlap

Several researchers (e.g., Kawamoto, 1993; Klein & Murphy, 2001; Klepousniotou et al., 2008) have suggested that ambiguity might be a continuum ranging from words with several non-overlapping entries to those with many strongly-overlapping ones. One feature which is commonly investigated regarding this continuum is relatedness (e.g., Azuma & Van Orden, 1997; Rodd et al., 2002). On one end of the continuum are *homonyms*, words with few unrelated meanings. An example of a homonym would be the word 'bank' which can refer either to the side of a river or to a financial institution. On the other end are *polysemous words* with their many strongly-related senses.

Polysemy is a pervasive element of language, since almost every word can be interpreted in several slightly different ways. For example, the word 'hook' can refer to similar objects (compare coat hooks and fish hooks) but also to hook-shaped trajectories such as by a road or a boxer. In addition, the word can be used as verb, referring to actions performed with hook-like objects ('to hook a fish') or in hook-like trajectories ('The road hooks to the right.'). The word has even been metaphorically extended to talk about addictions ('hooked on a feeling'). Similar patterns can be found for almost any word: they form a cloud of interrelated senses.

Whereas homonymous meanings are hypothesized to be stored as separate entries, polysemous senses are thought to form one large entry due to overlap (e.g., Rodd et al., 2004). This storage difference would then result in opposite reaction time patterns for the two types of ambiguity: a polysemy advantage due to larger shared activation space but a homonymy disadvantage caused by smaller individual activation spaces. Although there has been experimental support for this hypothesis (e.g., Beretta et al., 2005; Rodd et al., 2002, Experiment 3; Tamminen et al., 2006), there have also been studies that found a polysemy advantage but no homonymy disadvantage (e.g., Klepousniotou & Baum, 2007; Rodd et al., 2002, Experiment 2). Therefore, it seems that overlap may be affected by factors other than

relatedness. The current study was conducted to investigate one of these factors: word frequency.

Kawamoto (1993) suggested that frequency may play a role in whether senses develop separate representations in the lexicon: the more often they are encountered, the more likely it is that they develop their own entries. This hypothesis fits with the findings of differential effects for homonymy and polysemy: it seems likely that unrelated meanings are encountered separately more often than related senses purely because they occur in very different contexts and are more easily distinguished from each other. However, the interesting question is whether frequency alone can influence ambiguity effects. In particular, we were interested to see whether frequency affects the processing of polysemous words because we suspected (due to the reasons stated above) that homonyms develop separate entries by default whereas this may not be the case for polysemous words. Therefore, we expected to find frequency effects for the processing of words with many related senses.

A second variable of interest was task type. Rodd et al. predicted that the effect of polysemy should reverse into a processing disadvantage (Rodd et al., 2002; 2004) or disappear (Rodd et al., 2002) when participants do not merely perform lexical decision but have to process words for meaning in a semantic categorization task. Similarly, a parallel distributed processing (PDP) model by Armstrong & Plaut (2008) predicted an ambiguity advantage when a task does not require precise interpretation of a word but an ambiguity disadvantage when a specific interpretation is needed. However, whereas Rodd et al. expected both patterns for high-overlap words (polysemy), Armstrong and Plaut found a processing advantage for high-overlap words (polysemy) when a specific interpretation was not required but a processing disadvantage for low-overlap words (homonymy) when a specific interpretation was required. We wanted to investigate whether our manipulation of sense overlap by varying frequency would also result in differential processing patterns for

lexical decision and semantic categorization. Therefore, we added task type as a second manipulation.

The current study showed a two (frequency: low/high) by two (task type: lexical decision/semantic categorization) by two (polysemy: few or many senses) design. With lexical decision, we wanted to test predictions for low- and high-frequency words. We predicted a polysemy advantage for low-frequency words because interpretations would not have been encountered often enough to develop separate representations, therefore presenting one large activation surface. This finding would be in line with both Rodd et al. (2002) and Armstrong and Plaut (2008). In contrast, predictions for high-frequency words could go two ways. Assuming that high frequency leads to several smaller individual activation surfaces, we either expected a polysemy disadvantage (Rodd et al.) or no effect at all (Armstrong and Plaut). With semantic categorization, a polysemy disadvantage for low-frequency words would support predictions by Rodd et al. whereas a polysemy disadvantage for high-frequency words would support the model proposed by Armstrong and Plaut.

Method

Participants

Sixty undergraduate students at the University of Aberdeen participated in return for course credit. Thirty of them (21 female) took part in the lexical decision task; the remaining half (17 female) performed the semantic categorization task. Age ranged from 17 to 31 (M_{LEX} = 20; M_{SEM} = 22.5). All participants had normal or corrected-to-normal vision, and were native speakers of English.

Design and materials

The study encompassed a mixed 2 x 2 x 2 design: frequency (low/high) by task (lexical/semantic) by polysemy (few/many senses). Data were analysed by means of linear

mixed-effect models (e.g., Dixon, 2008; Jaeger, 2008; Baayen, 2008). Frequency and polysemy were manipulated within participants while task type varied between participants. The target stimuli consisted of 120 concrete object words. Interested readers are referred to Supplementary Materials A for this stimulus set as well as a description of its properties. Filler stimuli were included for both task conditions. For lexical decision, the 120 filler words consisted of concrete nouns for living beings (e.g., 'snail'). Fillers for the semantic categorization condition consisted of 55 concrete object words and 175 concrete animal words. Thus, in both conditions half of the word stimuli referred to living beings, the other half to objects. Finally, the lexical decision task also required inclusion of 240 nonwords (legal nonwords and pseudo-homophones). These were matched in length to the words, and were created by replacing a letter in existing words (that were different from the word stimuli).

Procedure

Participants were presented with a series of letter strings. They responded by pressing one of two buttons: word/nonword (lexical decision condition) or object/animal (semantic categorization condition). On each trial, a fixation cross appeared for 500 ms, followed by presentation of the letter string (Courier New, 28 points). The trial ended when the participant had responded or 3000 ms after presentation of the word. Following the end of the trial, the screen remained blank for 1000 ms before presentation of the next fixation cross. Order of presentation was randomised for each participant. Prior to the experimental session, participants performed a few practice trials for which they received speed and accuracy feedback. The experiments were presented by means of a Dell PC (Windows XP), using E-Prime software, and responses were recorded via an Eprime SRBox. The experimental session took around 15 minutes to complete.

Analyses

Data were analysed by means of linear mixed-effect models (Dixon, 2008; Jaeger, 2008; an extensive description of the method can be found in Baayen, 2008; for a user-friendly overview tailored towards researchers without a strong computational background, see Cunnings, 2012). An overview of the analyses can be found in Supplementary Materials B. Analyses for both reaction times and error rates always included the three effects of interest (frequency, task type, and polysemy) as well as their interactions.

Results

Target trials were excluded from analyses if reaction times were 2.5 standard deviations above/below each participant's mean per condition (2.92% of trials). Of the remaining trials, participants' mean error rate ranged from 0% to 10% (M = 3.61%). For error rates, the model's fit was significantly increased by adding random slopes ($\chi 2$ (11) = 44.14, $p < .001$). This best-fitting model (N= 6990, log-likelihood = -956.54) showed no significant effects for any of the main effects, nor of the interactions, all ps $\geq .428$. A summary of the error rate results has been provided in Table 1. Error trials were excluded for the reaction time analyses. For those filtered data, participants' mean reaction times ranged from 424 to 720 ms (M = 547 ms). Reaction time data have been summarized in Table 2.

(tables 1 and 2 about here)

For reaction times, adding random slopes significantly increased the model's fit ($\chi 2$ (11) = 53.74, $p < .001$). Therefore, the best fitting model for reaction times (N = 6738, log-likelihood = 1887.96) included both random intercepts and random slopes. Of the main effects, only frequency reached significance, t = -4.86, $p < .001$. The effect of task type did not reach significance, t = 1.51, p = .131. The same was true of polysemy, t = -1.46, p = .144. All interaction effects reached significance. Frequency interacted with task type, t = 3.14, p = .002, as well as polysemy, t = 2.21, p = .027. In addition, task type interacted with polysemy,

t = 2.49, p = .013. Most importantly, there was a three-way interaction between all three factors, t = -3.39, p = < .001. To shed more light on this three-way interaction, we conducted separate analyses for low- and high-frequency words.

For low-frequency words, the model's fit was significantly improved by adding random slopes ($\chi 2$ (4) = 38.03, p < .001). The same was true for high-frequency words ($\chi 2$ (4) = 14.08, p = .007). The best-fitting model for low-frequency words (N = 3323, log-likelihood = 871.08) showed no main effects for either task type, t = 1.47, p = .142, or polysemy, t = -1.27, p = .204. However, there was an interaction between these two variables, t = 2.28, p = .023. As can be seen in Table 2, low-frequency words showed a 15 ms polysemy advantage with lexical decision, but a 15 ms polysemy disadvantage with semantic categorization. The best-fitting model for high-frequency words (N = 3415, log-likelihood = 959.55) showed significant main effects of task type, = 3.84, p < .001 and polysemy, t = 2.04, p = .041. Importantly, the interaction between task type and polysemy again reached significance, t = -2.60, p = .009. However, the pattern went into the opposite direction as had been found for low-frequency words. High-frequency words showed a 14 ms polysemy disadvantage with lexical decision, but a 10 ms polysemy advantage with semantic categorization.

Finally, we excluded several alternative explanations for the current findings by checking contribution to variance by any of the six matched word properties (bigram frequency, number of neighbours, familiarity, concreteness, word length, and number of syllables) as well as two unmatched word properties (age of acquisition and semantic diversity). Only two word properties significantly contributed to variance: familiarity (t = -3.14, p = .002) and bigram frequency (t = -2.38, p = .017). Effects of the remaining word properties did not reach significance (all ps ≥ .131). The extended model's fit was again significantly improved by including random slopes ($\chi 2$ (22) = 65.26, p < .001). The new model that included the two extra variables (N = 6738, log-likelihood = 1902.07) showed

effects that were very similar to those for the original model. Task type and polysemy still did not affect reaction times (both *p*s ≥ .084). The effect of frequency still reached significance, as did all interactions (all *p*s ≤ .028). Thus, most of the eight additional word properties did not affect reaction times at all, while inclusion of familiarity and bigram frequency did not affect the current findings.

Discussion

The current study was conducted to test the hypothesis that word frequency affects representational overlap in the mental lexicon. To this end, we had participants perform lexical decision and semantic categorization tasks for low- and high-frequency words with few or many senses. We found a three-way interaction between word frequency, task type, and polysemy. Low-frequency words showed a polysemy advantage with lexical decision, but a polysemy disadvantage with semantic categorization. In contrast, high-frequency words showed the opposite pattern: a polysemy disadvantage with lexical decision, but a polysemy advantage with semantic categorization.

Firstly, the current findings confirmed our prediction that lexical decision would result in a polysemy advantage for low-frequency words with lexical decision. This finding is in line with Rodd et al. (2002) as well as Armstrong and Plaut (2008): both teams predicted that high sense overlap should result in a processing advantage when words do not have to be processed for meaning. However, whereas these researchers focused on overlap caused by meaning relatedness, we posited that word frequency can affect representational overlap as well. The remaining effects provided support for this hypothesis.

Secondly, our findings supported the low-overlap prediction for high-frequency words: lexical decision for high frequency words resulted in a polysemy disadvantage. This novel pattern would be hard to explain if meaning overlap was only affected by relatedness. In that case, both low- and high-frequency words would show a polysemy advantage with

lexical decision. However, the current findings make perfect sense under the hypothesis that even related senses develop separate representations if they have been encountered frequently. Thus, results are in line with Rodd et al. (2002) who predicted that weak representational overlap should result in a processing disadvantage. Again, whereas Rodd et al. posited this pattern for unrelated meanings (homonymy), our findings support the notion that this pattern will also occur for related senses (polysemy) as long as they are encountered frequently.

Thirdly, current results supported representational overlap predictions for the semantic categorization task. Semantic categorization resulted in a polysemy disadvantage for low-frequency words. These findings fit with predictions by Rodd et al. (2002) who proposed that high representational overlap (polysemy) should turn into a disadvantage with semantic categorization due to the fact that a specific interpretation was now required. Again, the fact that we found this pattern for low-frequency but not high-frequency words supports our hypothesis that frequency affects representational overlap, with senses of low-frequency words overlapping more strongly than those for high-frequency words.

Fourthly, apart from the findings for the low-frequency words in the lexical decision task, the current results are not in line with those found by Armstrong and Plaut (2008). In their data, low representational overlap (homonymy) did not affect reaction times when no specific interpretation was required whereas high representational overlap (polysemy) did not affect reaction times with high precision requirements. Although their findings did show effects of both homonymy and polysemy with a task requiring moderate precision, assuming such moderate precision requirements for our current study would still not explain the different reaction time patterns for our lexical and semantic tasks. Thus, it seems that processing depth (current study) and processing precision (Armstrong and Plaut) affect word recognition in different ways.

Finally, our results showed a novel and unexpected pattern: high-frequency words showed a polysemy advantage with semantic categorization. As far as we are aware, similar processing advantages have not been found for homonymy when words need to be processed more deeply. Based on Armstrong and Plaut (2008), we actually suspected we might find an opposite pattern: a polysemy disadvantage for high-frequency words in the semantic categorization task. However, several theories (see Twilley & Dixon, 2000) propose that the commonly-found homonymy disadvantage may be the result of reinterpretations after early commitments. Interestingly, under our low-overlap hypothesis for high-frequency words, this processing pattern may have worked out to the readers' benefit in our semantic categorization task. If high-frequency words have developed low-overlap representations for senses, it may be that one of these senses is more dominant than the other. This seems particularly likely for our semantic categorization task, since our target stimuli were explicitly selected to have dominant object interpretations *and* participants were encouraged to categorize these words as either objects or animals. Therefore, it might be that participants first selected the most dominant interpretation, and ended up being correct in the majority of cases so reinterpretation was not needed. We wonder whether a similar processing advantage may also be found for homonyms when their dominant interpretation is contextually relevant.

The current study supported our hypothesis that word frequency affects representational overlap in a way similar to relatedness: low frequency (like polysemy) results in high representational overlap, whereas high frequency (like homonymy) leads to low representational overlap. Furthermore, it showed that interactions of frequency and polysemy result in differential processing with lexical decision and semantic categorization. Finally, we found indications that low representational overlap may not always hinder word processing; depending on meaning dominance and contextual relevance, having to commit early may actually be advantageous.

Acknowledgements

The reported research and the writing of this paper were supported by a grant awarded by the Graduate School in the College of Life Sciences and Medicine in Aberdeen.

References

Armstrong, B. C., & Plaut, D. C. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral investigations. In *Proceedings of the 31st annual conference of the cognitive science society*. Hillsdale, NJ: Cognitive Science Society.

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*, 484-504.

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

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

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research, 28*, 369-382.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59*, 447-456.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434-446.

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language, 32*, 474-516.

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language, 45,* 259-282.

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*, 1-24.

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*, 1534–1543.

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*, 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*, 89-104.

Tamminen, J., Cleland, A. A., Quinlan, P. T., & Gaskell, M. G. (2006). Processing semantic ambiguity: Different loci for meanings and senses. *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science* Society, 2222-2227. Mahwah, NJ: Lawrence Erlbaum Associates.Twilley, L. C., & Dixon, P. (2000). Meaning resolution processes for words: A parallel independent model. *Psychonomic Bulletin and Review, 7*, 49-82.

Table 1. Mean error rates.

frequency	**task type**	**polysemy**		
		few senses	many senses	**all**
low	lexical	4.8 (21.5)	5.7 (23.2)	5.3 (22.4)
low	semantic	4.0 (19.7)	4.7 (21.1)	4.4 (20.4)
low	**all**	4.4 (20.6)	5.2 (22.2)	4.8 (21.4)
high	lexical	2.3 (15.0)	3.1 (17.3)	2.7 (16.2)
high	semantic	2.2 (14.6)	2.1 (14.2)	2.1 (14.4)
high	**all**	2.2 (14.8)	2.6 (15.8)	2.4 (15.3)
all	lexical	3.6 (18.5)	4.4 (20.5)	4.0 (19.6)
all	semantic	3.1 (17.3)	3.4 (18.1)	3.2 (17.7)
all	**all**	3.3 (17.9)	3.9 (19.3)	3.6 (18.6)

Note. Error rates as percentages (with standard deviations in brackets).

Table 2. Mean reaction times.

frequency	**task type**	**polysemy**		
		few senses	**many senses**	**all**
low	lexical	547 (128)	532 (116)	539 (122)
low	semantic	572 (132)	587 (136)	579 (134)
low	**all**	560 (130)	559 (130)	559 (130)
high	lexical	505 (105)	519 (114)	512 (110)
high	semantic	564 (131)	554 (131)	559 (131)
high	**all**	535 (122)	537 (124)	536 (123)
all	lexical	526 (118)	525 (115)	525 (117)
all	semantic	568 (131)	570 (134)	569 (133)
all	**all**	547 (127)	548 (127)	547 (127)

Note: Reaction times in ms (with standard deviations in brackets).

Supplementary materials A:

Stimulus information

The target word set consisted of 120 concrete non-homonyms primarily referring to objects. A summary of their properties has been provided in Table 1. The stimuli themselves are presented in Table 2. Both tables can be found at the end of this manuscript. Below we briefly describe construction of the stimulus set.

Previous studies (e.g., Rodd et al., 2002) have shown that defining ambiguity by means of questionnaires increases the risk of conflating homonymy and polysemy because the two co-occur. Therefore, several variables were defined by means of the Wordsmyth Dictionary-Thesaurus (WDT; Parks, Ray, & Bland, 1998). This online dictionary provides separate entries for meanings and senses, and lists them in order of frequency of use (see Parks et al., 1998; http://www.wordsmyth.net/?mode=history). Words were selected if only one meaning was provided, and if the first (or only) sense entry was a concrete noun interpretation.

The first independent variable *frequency* was defined by means of lemma frequency counts taken from the CELEX database (Baayen, Piepenbrock, & Van Rijn, 1993). In text and tables, the variable will be reported as frequency per million. However, for the analyses we used the log-transformed scores (as recommended by e.g. Whaley, 1978). Word frequencies ranged from 2 per million ('leek') to 353 per million ('car'), with a mean of 38.5 per million. The second independent variable *polysemy* was defined by counting the number of sense entries in the WDT. Number of senses for the included words ranged from 1 (e.g., 'barn') to 21 ('crown'), with a mean of 4.87. After we had constructed our stimulus set, the WDT changed their formatting. Whereas transitive and intransitive verbs were originally consistently listed as separate sense entries, now they are sometimes combined into one entry and sometimes listed separately. However, when we re-counted the sense entries while

consistently treating transitive and intransitive verbs as separate senses, we found that the number of sense entries remained the same as before. Therefore, this formatting change did not affect our definitions or analyses.

Conditions were closely matched for 6 variables: bigram frequency and number of neighbours (Baayen et al., 1993), familiarity and concreteness (Coltheart, 1981), as well as word length and number of syllables. As can be seen in Table 1, quite similar numerical values were obtained for all conditions. Statistically, word properties were indeed closely matched between all polysemy conditions (all Fs < 0.19, all ps > .665). However, as mentioned in the main text, the same was not true for the frequency conditions (many ps < .05). To ensure that this issue did not distort findings, we checked whether any of these word properties contributed to models' fits. In addition, contributions were also checked for two additional variables for which information was not widely available when the stimulus set was being constructed: age of acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012) and semantic diversity (Hoffman, Lambon Ralph, & Rogers, 2013). More details about these additional analyses can be found in Supplementary Materials B.

References

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). The CELEX lexical database [CD-ROM] Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33*, 334-338. Retrieved from: http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm.

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). *Behavior Research Methods, 45*, 718-730.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*, 978-990.

Parks, R., Ray, J. & Bland, S. (1998). *Wordsmyth English dictionary – Thesaurus.* [ONLINE]. Retrieved from: http://www.wordsmyth.net, University of Chicago.

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*, 245-266.

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior, 17*, 143-154.

Table 1. Descriptive statistics for target stimuli Experiments 1 and 2

	Few senses			Many senses		
	low frequency	high frequency	all	low frequency	high frequency	all
Example	sword	knife		shield	crown	
N	30	30	60	30	30	60
Senses	2.17	3.20	2.68	6.13	7.97	7.05
Frequency	12.07	66.27	39.17	12.50	63.27	37.88
Familiarity	5.08	5.58	5.33	5.12	5.51	5.32
Concreteness	5.99	6.01	6.00	6.00	6.02	6.01
Letters	4.97	4.73	4.85	5.27	4.37	4.82
Syllables	1.33	1.33	1.33	1.50	1.10	1.30
Bigram frequency	8289	7278	7783	7806	7114	7460
Neighbours	4.93	6.83	5.88	2.83	8.50	5.67

Table 2. Target stimuli

Few Senses		Many Senses	
Low Frequency	Higher Frequency	Low Frequency	Higher Frequency
badge	apple	anchor	belt
barn	basket	balloon	bench
blouse	bell	bean	bomb
broom	blanket	birch	bone
cage	boat	bucket	boot
cigar	bottle	cherry	brick
coffin	bread	chestnut	button
couch	bullet	coin	cake
flask	car	cork	chain
grape	card	cradle	coat
helmet	carpet	diamond	crown
leek	cheek	flute	gun
mattress	chest	fork	hammer
medal	clock	glove	key
onion	desk	gown	leaf
oven	doll	horn	nail
peach	egg	jewel	pan
pear	engine	kite	plate
plank	fruit	lemon	rod
poster	hat	needle	ship
sofa	hut	olive	shoe
spool	jacket	pearl	sink
spoon	knife	plum	skirt
stair	lamp	ribbon	stone
statue	map	saddle	table
sword	missile	shield	tank
twig	phone	shovel	tent
vase	toe	ski	thumb
vine	tray	spear	train
yacht	weapon	trumpet	trunk

Supplementary Materials B:

Analyses

Data were analysed by means of linear mixed-effect models. Target responses were excluded if they were 2.5 standard deviations above/below each participant's mean for the eight conditions. In the reaction time analyses only correct trials were included. Reaction times were log-transformed (as recommended in Baayen, 2008). We were interested in the main effects of frequency (low/high), task type (lexical/semantic), and polysemy (few/many senses), as well as their interactions. These were all included by default. Eight additional ("covariate") variables were included to exclude alternative explanations and increase statistical power (by reducing noise): the six matched word properties (bigram frequency, number of neighbours, familiarity, concreteness, word length, and number of syllables) and two additional variables for which information was not available when the stimulus set was being constructed (age of acquisition and semantic diversity). However, these additional variables were only included if they significantly improved a model's fit (see below). The word properties (being continuous variables) were centred to reduce collinearity within the model (Jaeger, 2010).

Random intercepts were added for participants and items. In addition, it was checked whether a model's fit was significantly improved by including random slopes: over-item slopes (for task type) and over-subject slopes (for polysemy, frequency, and their interaction). Models were fitted by means of the forward "best-path" approach (as in e.g. Baayen, 2008; Cunnings, 2012), in which random slopes are added or subtracted on the basis of ANOVAs between models. As can be seen in the main text, contributions of random slopes were significant in all current models, so they were included.

We checked for any potential confounding effects of the matched and unmatched word properties by including them as additional variables in a second model. Inclusion of

these covariate variables was determined by first including all eight of them in an initial model. They were only included in the final model if they significantly contributed to variance. Since covariates were only added to the second analysis to ensure that they did not influence findings, their effects will not be extensively discussed in the main text. However, it will be reported which ones were included and whether their inclusion affected results.

Currently there is no agreement about the optimal way to estimate significance for effects obtained with the function lmer(), so as suggested by Cunnings (2012) we decided to use a formula from Baayen (2008, p248):

p = 2 * (1 - pt(abs(X), Y-Z)).

In this formula, X is the t-value, Y is the number of observations, and Z is the number of fixed effect parameters including the intercept (so Z comes down to the total number of fixed effects plus 1). Binomial data such as accuracy scores can be analysed with the function glmer(), which in contrast to the function lmer() does provide significance levels. Therefore, no additional calculations were needed for accuracy data.

References

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

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research, 28*, 369-382.

Jaeger, T. F. (2010). Common issues and solutions in regression modelling (mixed or not). Presentation at Brain and Cognitive Sciences, University of Rochester, UK, May 4 2010. Retrieved from: https://www.hlp.rochester.edu/resources/recordedHLPtalks/ PennStateRegression10/PennState-Day2.pdf (Feb 2014).

R Development Core Team. (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org.