Core words in semantic representation

Andrew Wang (andrew.wang@unimelb.edu.au)
Simon De Deyne (simon.dedyne@unimelb.edu.au)
Meredith McKague (mckaguem@unimelb.edu.au)
Andrew Perfors (andrew.perfors@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne

Abstract

A central question in cognitive science is how semantic information is mentally represented. Two dominant theories of semantic representation are language-based distributional semantic models (which suggest that word meaning is based on which words co-occur in language) and semantic networks based on word associations (which suggest that words are represented as a network in which words with closer meanings are more closely linked). We investigate the issue of semantic representation through the lens of core vocabulary – the set of words that are most central in the mental lexicon – which these two theories make different predictions about. We report on the results of an experiment that tests which measure of core vocabulary most closely aligns with human behaviour in a word-guessing game where the aim was to identify a target word given a set of semantically related words as hints. Target and hint words, which varied across trials, were generated from different core vocabulary lists corresponding to these different theories. Results revealed that the type of hint words did not affect performance, but that better performance was attained for target words derived from word associations than from natural language distributional statistics. Follow-up analyses ruled out several alternate explanations. Our results suggest that the semantic information reflected in word associations may be more involved in the efficient identification of lexical meaning.

Keywords: semantic representation; lexicon; distributional semantics; concepts; core vocabulary; age of acquisition; frequency; word associations

Introduction

The nature of semantic representation is a central question in cognitive science: how are the meanings of words represented in our mental lexicons? This and related questions (what is the relationship between language and thought? how much do the pressures of communication and learning shape our lexicons) have been the focus of decades of research.

One prominent approach suggests that the meaning of a word is derived in large part from the words it co-occurs with in the linguistic environment. This *language-based distributional* approach is the basis of most state-of-the-art models of language in machine learning and AI (e.g., Brown et al., 2020) and has also been shown to predict human behaviour fairly well (Mandera, Keuleers, & Brysbaert, 2017). A contrasting *semantic network* approach based on word associations suggests that we represent words in an interconnected way, such that words with stronger links are more semantically related; as such, the meaning of a word is derived from its position relative to the other words in the network.

Although the distributional approach is closely tied to natural language and the semantic network approach to word associations, this is not always the case: word associations can be encoded as a distributional model and semantic networks can be built from language co-occurrence data (Bel-Enguix, Gómez-Adorno, Reyes-Magaña, & Sierra, 2019; Rotaru, Vigliocco, & Frank, 2018; Steyvers, Shiffrin, & Nelson, 2005). For this reason, our aim is to investigate differences between these approaches not in terms of the *format*, but in the actual *content* they encode. One of the main ways this content differs is in the extent to which each incorporates pragmatic information about what tends to be *said* instead of what is merely sensed or thought about. In language-based approaches, the meaning of a word is based entirely on how words are used during communication. As a result, their semantic representations are shaped more by pragmatic factors of conversation and less by information that is sensory in nature (Vankrunkelsven, Verheyen, Storms, & De Deyne, 2018; De Deyne, Navarro, Perfors, & Storms, 2016; De Deyne, Navarro, Collell, & Perfors, 2021). For example, during language use a word like *yellow* might not be associated with *banana* as often as *green* is, because *yellow* is sensed and taken for granted as part of the common ground (and thus unstated), whereas *green* is unusual and thus important to mention.

Despite their differences, both approaches incorporate the insight that words depend on each other for their meaning (Kang, 2018; Schulte Im Walde & Melinger, 2008; Gruenenfelder, Recchia, Rubin, & Jones, 2016). This means that in both, some words are depended on more than others. We call these *core words*: the words that are representationally central to the mental lexicon. In models built from word associations, the core words are those that are linked to the most other words, which we identify here based on a measure of centrality called InStrength (defined below).

Another way to identify core words is to focus on the *process* rather than the *outcome* of building the network. For instance, the preferential attachment hypothesis suggests that semantic networks are built up by attaching new words to existing ones (Brysbaert, Van Wijnendaele, & De Deyne, 2000; Hills, Maouene, Maouene, Sheya, & Smith, 2009; Steyvers & Tenenbaum, 2005). It thus implies that the core words are the ones that are acquired earliest and have a lower age of acquisition (AoA). For distributional language models, we define core words as those that occur most frequently in natural language, measured using corpus word frequency (WF). Word frequency is directly analogous to InStrength: the central words in a graph derived from distributional statistics are the most frequent ones. It is also highly correlated with other possible measures of coreness like contextual diversity (Hollis, 2020).

As Table 1 reveals, the core words picked out by these different approaches capture their essential characteristics. High

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Table 1: Top 10 core words in each core word list.

	AoA	WF	InStrength
1	mom	go	money
2	potty	know	food
3	water	come	water
4	wet	like	car
5	spoon	right	music
6	nap	think	bird
7	dad	good	sex
8	grandma	want	love
9	hug	see	dog
10	shoe	say	old

frequency words tend to be more semantically depleted and polysemous, reflecting their versatile use in many communicative contexts (Jorgensen, 1990; Tragel, 2001). Words that are central in semantic networks tend to reflect psychologically important categories (Steyvers & Tenenbaum, 2005; De Deyne, Navarro, et al., 2016), and early-acquired words tend be salient to children (Bates et al., 1994).

The study of core words has a long tradition in linguistics, ranging from creating vocabulary lists for pedagogical purposes (Carter, 1987; Ogden, 1930; West, 1953) to studies of centrality in dictionary definitions (Vincent-Lamarre et al., 2016), to the search for universal semantic primitives (Wierzbicka, 1996). These treatments of core words do not always focus on psychologically-motivated theories of meaning discussed earlier. As such, they have not often been empirically compared to assess how well they account for human behaviour, nor have core words been used to experimentally adjudicate between the different psychological theories of meaning they reflect. This is the gap our work fills. We present people with a simple word-guessing game involving hints and targets from different core word lists. Our question is which core word list – and thus which psychological theory of meaning – best explains human behaviour in this game.

The notion of using word games to investigate the mental lexicon is not new (Moskvichev & Steyvers, 2019; Kim, Ruzmaykin, Truong, & Summerville, 2019; Shen, Hofer, Felbo, & Levy, 2018; Xu & Kemp, 2010; Heath, Norton, Ringger, & Ventura, 2013). Not only are such games cognitively natural and even fun, they provide data that can be used to quantitatively compare the predictions made by different theories of semantic meaning. In our task, which varies which core word lists (InStrength, WF, or AoA) provide the hints and target words, we focus on two main questions. First, which type of core words are the most effective hints? And secondly, which type of core words are the easiest-to-guess targets? Based on the logic that more peripheral words should be harder to guess and/or harder to guess with, if any of the core word lists results in higher performance, that may be an indication that those words are core in people's *actual* lexicons, and thus that the theory of meaning they correspond to offers a better account of human semantic representation.

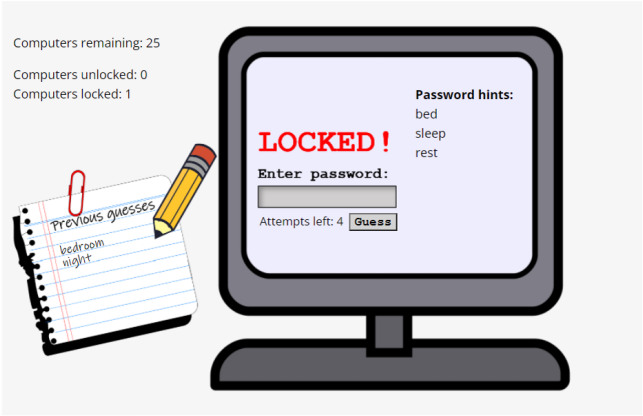

Figure 1: **Sample trial.** People were given up to six hints to guess the target word (a simple English word they were told was a password for unlocking a computer). Targets and hints were drawn from different core word lists, and participants tried to guess each password with as few hints as possible.

Method

Participants

500 participants were recruited from Amazon Mechanical Turk and paid $4.16 for the 20-25 minute task; of these, 487 had non-corrupted data files, and 479 passed the pre-registered[1] check trials (described below). Ages ranged from 20 to 79 years old (M = 40.47) and 46% were female. 90.5% reported being native English speakers, and all passed a qualification assessing English proficiency.

Procedure

Participants completed the task online after giving consent, providing optional demographic information, reading the instructions, and answering two questions about them.

The task was set up as a game in which participants were asked to unlock computers by guessing their passwords based on a series of related hint words. Each trial corresponded to a computer with a different password and the goal was to unlock as many computers in as few attempts as possible. Participants were informed that each password was a simple, common English word. One hint word was revealed at a time, with people making a guess at the password after each one. Hints were presented in order of their similarity to the target word, with the most similar hint shown first. A trial ended (and the password revealed) either if the password was successfully guessed or six hints had been provided with no success. As Figure 1 shows, participants were able to see the hints they had seen so far, their previous guesses, and the number of attempts remaining on that computer. They were also shown a running tally of how many computers were successfully unlocked and how many computers were remaining.

After a practice trial, each participant completed 24 experimental trials and two non-experimental catch trials that were

[1] https://aspredicted.org/blind.php?x=LZR_XCT.

designed to be substantially easier to guess than the experimental ones (their target words were *fish* and *hour*). As preregistered, we excluded any participant who failed to guess either of these words (eight in total). Except for the catch trials, which were the same and always on the 10th and 20th trials, the targets, hints, and order was randomised for each person. We manipulated three factors within subject in a $4 \times 3 \times 2$ design: target condition, hint condition, and similarity type (all described below). These factors were fully counter-balanced such that each of the 24 combinations of factors was seen exactly once by each participant.

Materials

Core word lists Word association data was sourced from the Small World of Words project, which contains associations for over 12,000 English words (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019). From this, each word was assigned an INSTRENGTH coreness score, calculated as the sum of the weights of all edges directed toward that node, where edge weights represent associative strengths; words that are commonly given as associates of other words have higher INSTRENGTH and thus are more core. The AOA coreness score was assigned based on norms sourced from Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012), such that earlier-learned words are more core. Finally, the WF coreness score, corresponding to the language-based approach, was based on the SUBTLEX database (Brysbaert & New, 2009), such that more frequent words were more core. INSTRENGTH and WF measures were log-10 transformed.

Once these lists were identified, we grouped together inflectional forms of the same lemma (e.g., *run*, *runs*, *running*) in order to focus on lexical concepts. For the same reason, function words (e.g., determiners, auxiliary verbs, and prepositions) were excluded. Finally, a few additional words had to be removed because they were not in the word association lexicon or text corpus and it was therefore impossible to match hint words to targets (see below for details); this mainly affected early-acquired words like *oink* or *popsicle*. Each of the three core word lists consisted of the top 300 words on that measure (see Table 1 for some examples).

In order to enable comparison of coreness scores across lists, each of the three measures was normalised by computing the difference between each word and the first word, divided by the difference between the first and 300th word. As a result, 0 corresponds to the word that is the most core on a given list and 1 corresponds to the 300th word on it.

Target conditions There are four target conditions, each corresponding to a list with 24 words (see Table 2). Target words in the AOA, WF, and INSTRENGTH conditions were selected because they were more core on that measure and less (but equally) core on the other two, while those in the EQUAL condition were selected to be equally core on all three measures.[2] The first three conditions allow us to explore whether it is easier to guess target words that are core under different theories of meaning (WF for language-based, AOA for preferential attachment, INSTRENGTH for word associations). The EQUAL condition allows us to ask whether different hint words are more or less useful (see below for a more complete description of hint word selection).

Table 2: Target words in each target condition.

AOA	EQUAL	WF	INSTRENGTH
rice	park	ready	anger
doll	block	hope	music
bite	middle	send	pain
plate	cross	use	paper
tail	stop	know	religion
grandma	roof	thing	round
pillow	face	stuff	sea
arm	chain	trouble	sick
crayon	stick	go	beach
brush	push	take	snake
bathroom	head	find	strong
boot	sound	spend	boring
snack	story	marry	tool
butt	age	keep	warm
hungry	tie	follow	white
hug	tear	way	wood
door	mess	pick	book
breakfast	storm	call	car
neck	parent	room	clean
hill	repeat	look	dirty
kitchen	cute	die	drink
bottle	choose	make	fat
towel	low	remember	horse
cookie	big	wait	light

Hint conditions Each target word was associated with three possible sets of hints corresponding to the three hint conditions (AOA, INSTRENGTH, WF). Thus, one of the hint lists was congruent with the target word (with hints and targets selected from the same list) and the other two were incongruent. Each participant who saw a given target word was shown hints from one of the three hint lists, counterbalanced so that no participant saw any target word or target / hint condition combination more than once. This ensured that over all trials, any differences in performance between target condition or hint condition could not be the result of differences in congruency between targets and hints.

[2]The mean coreness of words on their own lists is: AOA 0.73, WF 0.59, INSTRENGTH 0.59. The mean coreness of those words on the other lists is: AOA 1.44, WF 1.29, INSTRENGTH 1.2. This means, for instance, that WF target words had an average coreness of 0.59 on the WF list, but 1.29 on the other two (AOA 1.31, INSTRENGTH 1.26). That is, they are words that the language-based approach predicts are more core (because they are higher in frequency) but the other approaches predict are less (because they are less central to the semantic network and not learned as early). The words in the EQUAL target condition had an average coreness of 1.12 on all three lists (AOA 1.11, WF 1.11, INSTRENGTH 1.13).

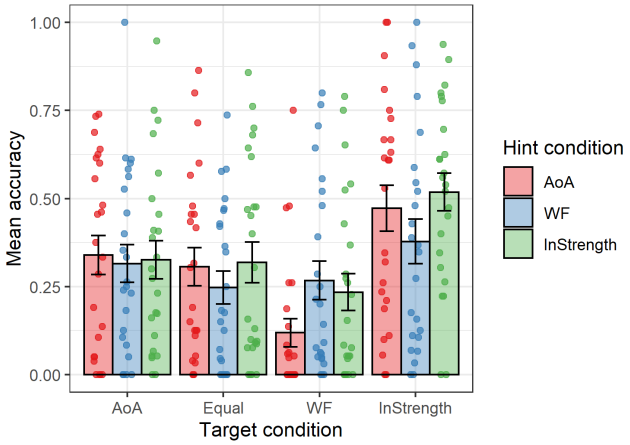

Figure 2: **Accuracy by hint and target condition.** Mean accuracy in guessing the target word as a function of what core word list the targets and hints were generated from. The y axis shows the proportion of time each target word was successfully guessed. Dots indicate target words; error bars are standard error. There was a significant effect of target condition, with INSTRENGTH target words guessed most accurately and WF targets guessed least. There was no significant effect of hint condition.

Similarity type For each target word, the hints corresponding to each core word list were selected to be the top six most semantically similar words to the target. After this process, 66 hint words (1.91%) were removed because they were variants of other hints or targets (e.g. *grandmother* and *grandma*). In these cases, the hint word with the lower semantic similarity was replaced with the next-most similar word.

Since similarity is a function of the semantic space being assumed, we used two different methods to calculate similarity, each corresponding to a different theoretical approach: based either on random walk distributions extracted from a word association network (RW similarity), or based on embeddings in the language-based model `word2vec`. RW similarity is closely related to the Katz Index and is calculated as the cosine similarity between the distribution of all weighted indirect paths between two words in a directed weighted graph. As in De Deyne et al. (2019), associative strength (i.e. the proportion of participants producing a word as an associate to another) was transformed to positive pointwise-mutual information and the decay parameter α that determines the contribution of longer paths was set at the default 0.65. Word embeddings were taken from publicly available fastText vectors (Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018), which were trained on the CommonCrawl corpus with 630B word tokens. For each pair of words, the similarity between the two was obtained by calculating the cosine of their 300-dimensional embeddings.

The difference between these two methods of calculating similarity is not the topic of current investigation, but we implemented both measures since performance can vary depending on the measure (Heath et al., 2013; Shen et al., 2018). This ensures that any differences in performance between hint and target conditions are not the result of the particular similarity metric assumed. All analyses were computed separately for each similarity type, but for space reasons we report only the RW hints here; performance was better across the board for them, and none of the main qualitative results vary for different similarity measures. The supplemental materials contain all analyses as well as the full set of hint lists.[3]

Results

Main analyses

We preregistered two outcome measures: accuracy (the proportion of time a target word was correctly guessed) and number of guesses (for those targets that were correctly guessed, how many attempts it took on average). We only report the results for accuracy here, in part because of space limitations, and also because number of guesses was difficult to interpret since it was conditioned on successfully guessing the word. The complete set of analyses involving both measures can be found in the Supplemental materials.

Overall, the task was difficult, with 27.8% of all trials resulting in a successful guess. Participants varied widely, from an overall accuracy of 4.17% to 58.3%. The high level of difficulty likely reflects the fact that the hints were drawn from a restricted set of core words rather than the full vocabulary, as is typical in similar word games. Still, the high accuracy on the easier catch trials (88% on average) indicates that people understood the task and were completing it as intended.

The mean accuracy for each target word in each hint condition was computed by averaging over individual trials. A 4×3 two-way ANOVA was conducted at the target-word level comparing mean accuracy across hint condition and target condition. As Figure 2 illustrates, there was a significant main effect of target condition, $F(3,276) = 10.82, p < .001$, but no significant effect for hint condition and no interaction. Tukey post-hoc pairwise comparisons showed that INSTRENGTH, WF, and AOA target conditions all significantly differed from each other (all ps $< .05$), with the highest accuracy achieved for INSTRENGTH and the lowest accuracy for WF. Accuracy was also significantly higher ($p = .001$) for INSTRENGTH than EQUAL target words. Figure 3 shows the accuracy for each of the target words in each target condition.

Exploratory analyses

Why were the INSTRENGTH target words consistently easier to guess, and the WF words more difficult? Our experiment was motivated by a desire to compare the psychological theories of semantic representation that each core word list reflects, but in order to draw conclusions about those theories it is necessary to delve further into these results. After all, the conditions differ in several ways, reflecting many of the natural differences between the core word lists and the approaches they reflect. Which of them drove this effect?

[3]Supplemental Materials are here: `https://github.com/andreww3/CogSciCoreWords`

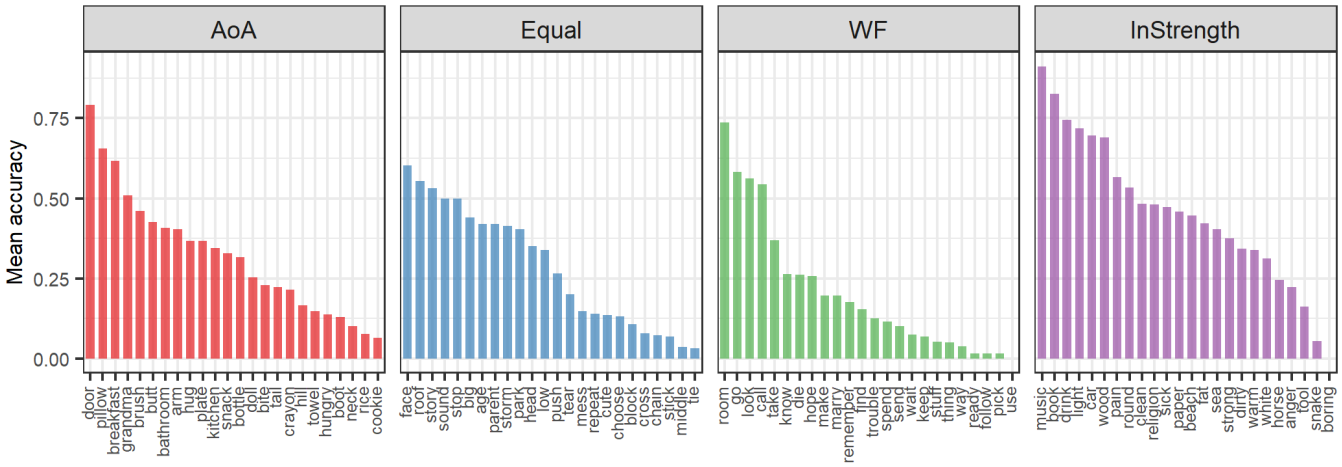

Figure 3: **Individual word accuracy by target condition.** Histograms show the mean accuracy for each target word, split by target condition (panels). Words in all conditions varied considerably in how easy they were to guess, but the INSTRENGTH target words consistently had higher accuracy and the WF target words had lower.

Looking at the target words, one immediate possibility is that the target conditions vary markedly in their part of speech (see Table 3). If some parts of speech (like nouns) are easier to guess than others (like verbs), might this explain the difference between target conditions? Indeed, a one-way ANOVA showed that the target word part of speech significantly affected accuracy, $F(2,93) = 4.24, p = .02$, with post-hoc tests indicating that accuracy was significantly higher for nouns compared to verbs, $p = .01$. Hence, it is possible that the better performance for INSTRENGTH target words and worse performance for WF arises from the differing numbers of nouns and verbs in those conditions.

While this is almost certainly part of the explanation, several considerations suggest that it is not the complete picture. As can be seen in Figure 4, differences between the target conditions remain even after taking part of speech into account. For example, considering only nouns, accuracy is still higher in the INSTRENGTH condition: a Kruskal-Wallis ANOVA for the noun target words only shows a significant difference in accuracy by target condition $H(3) = 8.99, p = .03$, with a significant difference between INSTRENGTH and WF, $p = .04$. While we caution against strong inferences due to the unbalanced design and low numbers of observations in some cells, this tentatively suggests that although differences between target conditions may be partly driven by part of speech, this is probably not the full picture.

Table 3: Part of speech distribution across target conditions.

	N	V	Adj
INSTRENGTH	15	0	9
AOA	21	2	1
WF	5	18	1
EQUAL	12	9	3
Total	53	29	14

Another possibility is that target conditions might differ in their relationship to the hints. For instance, it is probably easier to guess targets when the hints are very similar to them. If INSTRENGTH target words for some reason tend to have hints that are more similar to them than other target words, that might explain the improved performance. However, we found that the INSTRENGTH target words did not have the highest similarity to their respective hints (Figure 5, left panel). In fact, the highest similarity between hint and target words occurred in the AOA condition, and similarity was comparable between WF and INSTRENGTH conditions. Higher semantic similarity of target and hint words therefore cannot explain the improved performance for INSTRENGTH target words.

Still another hypothesis is that it may be easier to guess a word if hints are more unrelated to each other and thus span the semantic space better: *apple* is easier to guess from *red* and *fruit* than it is from *red* and *green*. However, as shown in Figure 5 (right panel), the pairwise similarity among the hints in the INSTRENGTH condition was not significantly lower than the pairwise similarity among the hints in the WF condition ($p = .55$). Thus, we cannot explain the INSTRENGTH target word advantage as arising due to differences in relatedness of hints to each other in that condition.

Discussion

We used a word-guessing game to compare different psychological theories of semantic representation, evaluating whether people performed better using words from different core word lists corresponding to different theories. We found that although the type of hints did not make a difference to performance, there was a difference with regards to the type of target words: accuracy was highest for the INSTRENGTH targets, followed by the AOA targets, with the lowest obtained for the WF targets. Additionally, this result

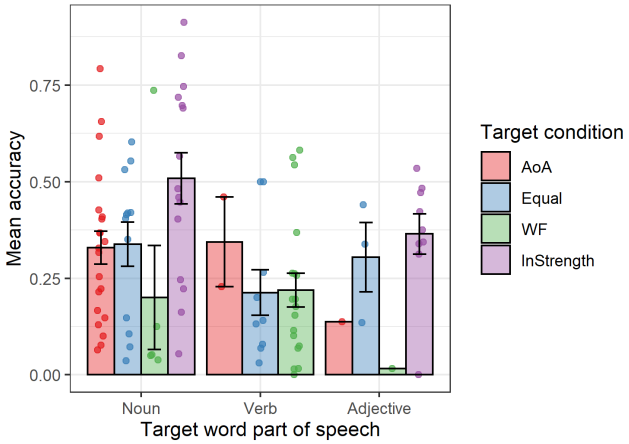

Figure 4: **Differences in accuracy by part of speech of target word.** The *y* axis shows the proportion of time each target word was successfully guessed. There was a significant effect of part of speech, with nouns guessed most accurately. However, even within nouns, INSTRENGTH target words were guessed more accurately, suggesting that part of speech alone cannot explain the difference between target conditions.

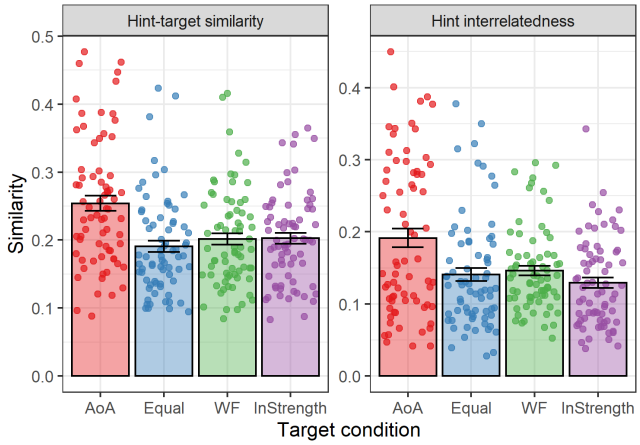

Figure 5: **Differences in hint-target similarity and hint-hint interrelatedness by target condition.** The *y* axis shows the average similarity between all pairings of targets and hint lists (left) and the average similarity between each of the pairs of hints in each hint list (right), split by target condition. AOA targets had the highest similarity to their hints, with INSTRENGTH not substantially different than the others. Hints were most related to each other in the AOA condition, with no significant differences among the others. Together, these results suggest that the improved performance on INSTRENGTH target words cannot be explained by differences in similarity or hint relatedness.

could not be explained based only on part of speech or similarity between targets and hints.

The differing effects of target and hint condition illustrate a clear asymmetry between the matter of which words are *hard to guess* and which are *hard to guess with*, where only the former mattered for the task, an unexpected result. This suggests that as participants search through semantic memory, the points from which the search process begins (i.e., the hints) are much less important than how readily available the target words are and how easily they emerge along the search process, suggesting that they occupy central or prominent positions in the semantic representation. It is in this sense that INSTRENGTH words are more core: they are more cognitively prominent and come to mind more easily compared to the AOA and WF words.

Overall, these results suggest that the content from which the INSTRENGTH words are derived – namely, word associations rather than external language use – may provide a better account of lexical representation, at least as accessed in this game. The second-best performing target condition, AOA, is like INSTRENGTH in that it also reflects a conceptual-semantic account of mental representation. By contrast, the lower performance obtained for the WF target words suggests that language-based models may not provide as good an account for human semantic representations. This finding is in line with previous research showing that word association models, compared to language-based models, better predict key semantic properties (Vankrunkelsven et al., 2018) and human similarity judgements (De Deyne, Perfors, & Navarro, 2016), and incorporate more non-linguistic information (De Deyne et al., 2021).

One concern might arise from low accuracy overall, which may suggest that the hints in general were not very effective. Why might this be? One possibility is that there is a large component of word meaning that is not accounted for by our set of 300 core words. Given that most adult vocabularies contain tens of thousands of words, this seems plausible: although core words are at the heart of the lexicon (Vincent-Lamarre et al., 2016; Wierzbicka, 1996), more peripheral words still contribute in an important way. A second non-mutually exclusive possibility is that the relatively crude presentation format of the hint words – as discrete lists of words with an unspecified relation to the target – did not allow people to meaningfully combine the hint words to construct word meanings and fully leverage the purported "core" properties of these hints. Further research may involve hint lists that are constructed in different ways. Is the INSTRENGTH advantage for target words retained if hints can be drawn from all vocabulary items, or from a restricted set of non-core words?

We conclude by adding two caveats to our main conclusion. Firstly, as discussed earlier, our claim is less about the *format* of the semantic representation (e.g., distributional vs. network), and more about the *content* that those representations encode: associative content provides a better account than linguistic content. Secondly, we have tested a limited number of measures of coreness and there are many alternative measures that could be considered for both representations. While the current comparisons have been useful as a first approximation of the research question, further studies should broaden the scope of the measures under investigation.

References

Bates, E., Marchman, A., Thal, D., Fenson, L., Dale, P., Reznick, J., ... Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(1), 85–123.

Bel-Enguix, G., Gómez-Adorno, H., Reyes-Magaña, J., & Sierra, G. (2019). Wan2vec: Embeddings learned on word association norms. *Semantic Web*, *10*(6), 991–1006.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *arXiv:2005.14165*.

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

Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, *104*(2), 215–226.

Carter, R. (1987). Is there a core vocabulary? Some implications for language teaching. *App Ling*, *8*(2), 178-193.

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, *45*(1).

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006.

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Jn of Experimental Psychology: General*, *145*(9), 1228–1254.

De Deyne, S., Perfors, A., & Navarro, D. J. (2016). Predicting human similarity judgments with distributional models: The value of word associations. In *26th International Conference on Computational Linguistics* (pp. 1861–1870).

Gruenenfelder, T. M., Recchia, G., Rubin, T., & Jones, M. N. (2016). Graph-theoretic properties of networks based on word association norms: Implications for models of lexical semantic memory. *Cognitive Science*, *40*(6), 1460–1495.

Heath, D., Norton, D., Ringger, E., & Ventura, D. (2013). Semantic models as a combination of free association norms and corpus-based correlations. *2013 IEEE Seventh International Conference on Semantic Computing*, 48–55.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science*, *20*(6), 729–739.

Hollis, G. (2020). Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability. *Journal of Memory and Language*, *114*, 104146.

Jorgensen, J. C. (1990). Definitions as theories of word meaning. *Journal of Psycholinguistic Research*, *19*(5), 293–316.

Kang, B. (2018). Collocation and word association. *International Journal of Corpus Linguistics*, *23*(1), 85–113.

Kim, A., Ruzmaykin, M., Truong, A., & Summerville, A. (2019). Cooperation and Codenames: Understanding natural language processing via Codenames. In *Proceedings of the 15th AAAI* (pp. 160–166).

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation.*

Moskvichev, A., & Steyvers, M. (2019). Word Games as milestones for NLP research. In *Proceedings of Workshop on Games and Natural Language Processing.*

Ogden, C. K. (1930). *Basic English: A general introduction with rules and grammar*.

Rotaru, A. S., Vigliocco, G., & Frank, S. L. (2018). Modeling the structure and dynamics of semantic processing. *Cognitive Science*, *42*(8), 2890-2917.

Schulte Im Walde, S., & Melinger, A. (2008). An in-depth look into the co-occurrence distribution of semantic associates. *Italian Journal of Linguistics*, *20*(1), 89–128.

Shen, J. H., Hofer, M., Felbo, B., & Levy, R. (2018). Comparing models of associative meaning: An empirical investigation of reference in simple language games. *22nd Conf on Computational Natural Language Learning*, 292–301.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healey (Ed.), *Experimental cognitive psychology and its applications.*

Steyvers, M., & Tenenbaum, J. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78.

Tragel, I. (2001). On Estonian core verbs. In I. Tragel (Ed.), *Papers in Estonian cognitive linguistics.*

Vankrunkelsven, H., Verheyen, S., Storms, G., & De Deyne, S. (2018). Predicting lexical norms: A comparison between a word association model and text-based word co-occurrence models. *Journal of Cognition*, *1*(1), 1–14.

Vincent-Lamarre, P., Blondin-Massé, A., Lopes, M., Lord, M., Marcotte, O., & Harnad, S. (2016). The latent structure of dictionaries. *Topics in Cognitive Science*, *8*(3), 625-659.

West, M. (1953). *A general service list of English words*. Longman, Green and Co.

Wierzbicka, A. (1996). *Semantics: Primes and universals*. Oxford University Press.

Xu, Y., & Kemp, C. (2010). Inference and communication in the game of Password. *NIPS 23*, 1–9.