COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 49 (2025) e70151
© 2025 Cognitive Science Society LLC.
ISSN: 1551-6709 online
DOI: 10.1111/cogs.70151

# Core Vocabulary in Language Representation and Processing

Andrew Wang, Simon De Deyne, Meredith McKague, Andrew Perfors

*Complex Human Data Hub, University of Melbourne*

Received 23 December 2024; received in revised form 19 September 2025; accepted 19 November 2025

**Abstract**

The question of which words are most important or fundamental to a language has been explored in many ways. However, many of these approaches place little emphasis on how humans learn, represent, and process language from a psychological perspective. In this study, we define and compare three distinct psycholinguistic measures of core vocabulary—word frequency, age-of-acquisition, and centrality in semantic networks—and test how well these *core words* capture human performance in a word-guessing game. In two experiments, 1000 participants were given different core words as both hint and target words, with the aim of identifying the target as quickly as possible. We found that while core words in general did not make very effective hints, they were effectively guessed as targets when using hints beyond the sets core words, and furthermore, were better guessed when the core word targets were defined based on centrality in semantic networks rather than linguistic factors like frequency. This finding was consistent across a range of experimental conditions and analyses. We discuss the implications of these findings for representation and processing in semantic memory, and what factors should constitute a human core vocabulary.

*Keywords:* Core vocabulary; Word frequency; Age of acquisition; Word associations; Distributional semantic models

## 1. Introduction

By the time we reach adulthood, we know tens of thousands of words. Most people intuitively agree that some of those words are more important or fundamental to the language as a whole. But which ones are they, and on what basis? This question about the nature and existence of *core vocabulary* has been scientifically explored for many years.

Correspondence should be sent to Andrew Wang, Complex Human Data Hub, University of Melbourne, 700 Swanston Street, Carlton, VIC 3053, Australia. E-mail: andrew.wang@unimelb.edu.au

One of the key issues in the study of core vocabulary is whether it exists at all as a useful explanatory construct: do human languages have a consistent core set of lexical items that are central in terms of meaning? If so, how are they identified? One motivation behind the question is that core words would be potentially revealing about the basic concepts of the mental landscape—the fundamental primitives out of which meaning is constructed (Hsu & Hsieh, 2013).

Wierzbicka's (1996) work on semantic primitives has been highly influential in this area. This work aims to uncover a set of basic concepts that are sufficient to define the rest of the words in the language. The approach consists of constructing exhaustive paraphrases of word meanings written in terms of semantic primitives; according to the theory, these primitives are irreducible and conceptually basic elements that cannot be further defined and, further, that are universal across human languages. In a similar vein, Vincent-Lamarre et al. (2016) analyzed the graph structure of dictionary definitions to uncover the core vocabulary of dictionaries—the set of words that, if known by a reader, would enable them to understand the rest of the words in the dictionary through definition lookup alone.

Other proposals for a core vocabulary have focused less on identifying a specific set of core words, but rather on proposing tests for the coreness of vocabulary items (Carter, 2012; Lee, 2001; Stubbs, 1986). The underlying idea is that it should be possible to derive criteria for core words that are internal to the structure of the English language, in terms of syntactic and semantic relations or discourse use. The motivation for the existence of such core vocabulary items is that speakers are able to draw on processes of simplification of language when engaging in tasks such as speaking in simple terms, summarizing a text, or defining a complex word (Carter, 2012). In this approach, coreness of words is not defined by the presence or absence of any one property: rather, the more tests a word passes, the greater degree of coreness it can be considered to have. Examples of these tests include superordinateness (core words are superordinate words rather than hyponyms), syntactic substitution (core words easily substitute noncore words, e.g., "amble," "stroll," "saunter," and "trudge" can all be replaced by "walk"), productivity (core words can form compounds, collocations, expressions, and idioms with other words more easily), and discourse neutrality (core words are unmarked and do not convey expressive connotations). However, it is not entirely clear where such properties come from, or why they should necessarily be considered as indicating coreness. For example, there is ample experimental evidence from psychology that indicates that people readily think in terms of basic-level concepts rather than the superordinate level (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976).

### 1.1. *Word frequency*

The idea of a core vocabulary is of practical importance in language learning and teaching, where a key question is which words are the most useful for second-language learners to learn first. In this area, word frequency is an extremely common measure used to indicate the importance of words: the core vocabulary for a language is just the most frequent words in that language (Bell, 2012; Borin, 2012; Lee, 2001). Many word lists have been devised for pedagogical purposes that use word frequency as the key criterion for vocabulary selection. Some prominent examples of such word lists include Ogden's (1930) Basic English, the General Service List (West, 1953), and the Council of Europe's Framework of Reference word lists.

Such frequency-based word lists are routinely used in projects that aim to identify which words are most valuable for learners and speakers alike. Nation's (2006) analysis of frequency lists constructed from the British National Corpus estimates that 6000–7000 of the most frequent words are needed to adequately comprehend spoken language (8000–9000 for written language). High-frequency words also have wide-ranging applications in the creation of learning materials, such as learner dictionaries and simplified texts at different skill levels for use in coursebooks and graded readers (Bell, 2012).

Evidence from cross-linguistic data also suggests that there is something special about high-frequency words. Calude and Pagel (2014) studied the frequencies of basic words across 18 languages from the Swadesh and Leipzig-Jakarta lists. They found that word frequency was highly correlated across languages, despite the fact that the languages had very heterogeneous cultural origins. Their results imply that there is a regularity in the patterns of language use in groups of people all over the world, and that there may be a core set of basic meanings that is shared universally, at least to some extent.

These results also link frequency as a lexical characteristic to the previously discussed conceptualization of core vocabulary as the central or fundamental words in terms of meaning. The latter, though useful in many ways, places relatively little emphasis on the question of how humans represent language and word meanings psychologically, which limits their applicability. For example, the meaning paraphrases proposed by Wierzbicka (1996) are long and unwieldy, and do not offer a psychologically plausible or realistic account of how humans learn and mentally represent the meanings of words. Similarly, humans do not represent word meanings as dictionary definitions.

Even frequency-based measures, though hugely influential and widespread, leave open the question of exactly *why* the most frequent words should be considered the core ones. One possibility is that word meaning exerts a causal force on word frequency, such that word frequency reflects the need distribution for how often we need to communicate each meaning (Piantadosi, 2014). In line with this, Stubbs (1986) argues that frequency may be best thought of as a *consequence* of coreness rather than its essence. Is what makes words core the fact that they are frequent, or is it the fact that words are core that makes them frequent? As we shall see, considering psychological theories about human language learning, representation and processing can shed some light on this important question.

In this study, our aim is to approach the study of core vocabulary from a psychological perspective. This has the advantage of providing definitions of core vocabulary grounded in well-established psychological theory, which permits new insight into what makes words important. It also results in more precise measures of coreness, which allow us to make quantitative predictions in a behavioral task.

From a psycholinguistic perspective, word frequency is one of the most important lexical variables for understanding language processing and language representation. It is a robust predictor of performance in a wide variety of psycholinguistic tasks, including word naming, word recognition, lexical decision, and semantic decision, such that processing is faster and more accurate for high-frequency words. Frequency typically explains around 30–40% of variance in such tasks, and is often the most important predictor of performance (Brysbaert, Mandera, & Keuleers, 2018). For example, megastudies of visual word naming and lexical decision tasks have consistently shown that word frequency is one of the strongest predictors of response latencies (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Brysbaert, Stevens, Mandera, & Keuleers, 2016; Yap & Balota, 2009). Word frequency also has a strong effect in memory, with high-frequency words leading to better recall (Poirier & Saint-Aubin, 1996). This relationship between frequency and psycholinguistic task performance is known as the *word frequency effect*.

The importance of word frequency in psycholinguistics suggests that it plays a key role in lexical access: words that are more frequently encountered become more entrenched in memory, which then become activated more easily, leading to better processing and more efficient access. High-frequency words may be more richly encoded in our semantic representations: the idea is that repeated exposure leads people to form higher quality and more stable representations of highly frequent words compared to words which are rarely encountered. Highly frequent words may also have more and stronger associations with other words in memory (Adelman, Brown, & Quesada, 2006; Brysbaert et al., 2018; Monaghan, Chang, Welbourne, & Brysbaert, 2017).

However, another interpretation of the word frequency effect is that word frequency might be a proxy for other lexical variables that are themselves the more important factors. As Brysbaert et al. (2018) argue, word frequency is confounded with many other different characteristics, such as word length, the age at which a word is acquired, and similarity measures. This means that word frequency effects could simply be the most observable "symptom" of any of those variables in disguise (or vice versa). Indeed, the fact that many low-frequency but highly familiar words (such as "toothbrush" and "unicorn") are also processed efficiently in psycholinguistic tasks suggests that factors besides frequency may be at play (Brysbaert, Mandera, McCormick, & Keuleers, 2019).

### *1.2. Centrality in semantic networks*

What other factors besides frequency, then, might measure the coreness of words in psycholinguistic terms? One possibility is that the processing advantages attributed to frequency may be better explained by underlying characteristics of the structure and organization of meaning in the mental lexicon; this is consistent with the idea that frequency is predicted by word meaning rather than vice-versa (Piantadosi, 2014). One way to explore what these characteristics of our semantic representation might be is to look at word association data (De Deyne & Storms, 2008). Word associations measure the extent to which two words (e.g., "red" and "rose") are related in meaning. The larger-scale properties of the semantic networks derived from these associations thus reflect the global structure of the mental lexicon as a whole as well as the direct associations between pairs of words (De Deyne & Storms, 2015).

Certain measures of network structure—specifically, the centrality of words within the semantic network—are strongly related to word frequency. Centrality within a network can be measured in different ways. An example of a centrality measure related to frequency is obtained by counting the number of links attached to a node, or if network links are directed (for instance, links pointing from cue words to response words), the number of incoming or outgoing links. The number of links a node has is known as *Degree* centrality, or *InDegree* and *OutDegree* if counting incoming or outgoing links. Network links can also be weighted, reflecting the strength of association between two words, with centrality calculated as the sum of the strength of the links connected to a node. This is known as *Strength* centrality, or *InStrength* and *OutStrength* if considering the direction of the links.

De Deyne and Storms (2008) showed that central words in word association networks tend to be both highly frequent and early acquired. Similarly, network properties of word associations predict word frequency. Differences in frequency between antonym pairs (e.g., “hot” and “cold”) are predicted by the number of connections a word and its neighbors have as well as the amount of influence that words have on other words, such as by facilitating connections (Liu, De Deyne, Jiang, & Lupyan, 2023).

Not only do these network properties correlate highly with word frequency; some studies also suggest that network models can provide an underlying account for word frequency effects. According to these network accounts, it is the structural properties of the mental lexicon that determine how efficiently people can process words, rather than word frequency itself (Beckage & Colunga, 2016; De Deyne & Storms, 2015). For instance, performance in a word fluency task is better predicted by a measure of network centrality known as *PageRank* (which captures the dynamics of information flow within a network) than it is by frequency (Griffiths, Steyvers, & Firl, 2007). Words with a high number of associates in a semantic network have higher and faster lexical access as well (Duñabeitia, Avilés, & Carreiras, 2008). This finding can be interpreted in terms of increased activation for words which are well connected in the mental lexicon, gathering activation from neighboring nodes.

### *1.3. Age of acquisition*

But another question raised by these findings is why are central nodes more central? Steyvers and Tenenbaum (2005) suggest that the statistical properties of semantic networks can be accounted for by a model of semantic growth in which new words are added to existing words (nodes in the network) with a probability proportional to the number of connections the existing node already has. This preferential attachment model implies that the age at which a word was learned might matter separately from its frequency and degree of connectivity. In addition, age of acquisition (AoA) provides a mechanistic account that takes into account order-based dependencies in learning.

Age of acquisition is a strong predictor of performance in psycholinguistic tasks, where responses are typically faster and more accurate for earlier-acquired words compared to later-acquired ones, even after controlling for word frequency and other important lexical characteristics (Brysbaert, Van Wijnendaele, & De Deyne, 2000; Brysbaert & Biemiller, 2017; Brysbaert & Cortese, 2011). That said, the AoA effects reported in the literature are somewhat nuanced and can be divided into two kinds: those that are linked to the effects of frequency, and those that are independent from frequency.

The frequency-related AoA effects emerge in tasks such as naming and lexical decision (Brysbaert & Cortese, 2011; Brysbaert & Ghyselinck, 2006). It may be the case that the key

Table 1
Top 10 core words in each core word list

| | AoA | WF | InStr |
|---|---|---|---|
| 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 |

*Note*: AoA reflects the earliest-acquired words, WF reflects the highest frequency words, and InStr reflects the most central words in word association networks. The Method contains full details of how coreness was calculated.

factor that matters for word processing tasks such as these is actually *cumulative* frequency, which combines both how often a word is encountered day-to-day (frequency) and how long the word has been known over a lifetime (AoA). These effects can also be explained by neural network accounts in which certain items have a greater influence on the resulting state of the network (Ellis & Lambon Ralph, 2000); these are items which either entered the network earlier (AoA) or were presented to the network more often (frequency).

By contrast, frequency-independent AoA effects appear to be present in tasks where the activation of semantic information is more important, such as picture naming (Brysbaert & Ghyselinck, 2006). Taken together, these effects suggest that AoA is a key organizational principle of the semantic system, and that the meanings of later-acquired words are built on, and depend on, the meanings of earlier-acquired words (Brysbaert et al., 2000). This idea also corresponds closely with the model of semantic growth proposed by Steyvers and Tenenbaum (2005).

### *1.4. The current study*

In summary, word frequency (WF), age-of-acquisition (AoA), and network centrality (InStr) are closely interrelated but conceptually distinct, and offer differing psychological accounts of the nature of core vocabulary. Indeed, as Table 1 shows, the top 10 core words on each of the three measures have little overlap.

In this study, we operationalize and compare these three measures of coreness based on how well they capture human performance in a psychological task that taps into lexical representation and processing. This task is similar to other game-style tasks that have previously been used to investigate other aspects of word meaning (Heath, Norton, Ringger, & Ventura, 2013; Kim, Ruzmaykin, Truong, & Summerville, 2019; Moskvichev & Steyvers, 2019; Shen, Hofer, Felbo, & Levy, 2018; Vankrunkelsven, Vankelecom, Storms, De Deyne, & Voorspoels, 2021; Xu & Kemp, 2010). The different types of core words are used as target words to be guessed and hint words on the task.

We have two main research questions. First, which type of core words are the most effective hints? If the core words according to a given approach effectively comprise the meanings of other words in the lexicon, then they should allow target words to be successfully guessed. And second, which type of core words are the easiest to guess targets? If core words are those that are central or important in the mental lexicon, then they should be more successfully guessed.

## 2. Experiment 1

### 2.1. Method

#### 2.1.1. Participants

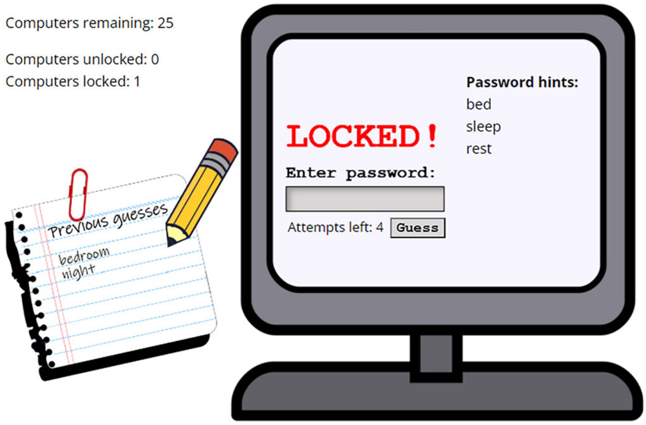
Five hundred participants were recruited from Amazon Mechanical Turk and paid US$4.16 for the 20–25 min task; of these, 487 had complete data files,[1] and 479 passed the preregistered[2] check trials (described below) and thus were included in the data analysis. Ages ranged from 20 to 79 years old ($M = 40.47$) and 46% were female. 90.5% reported being native English speakers, and all had previously passed a qualification assessing English proficiency and detecting bots. The research was approved by the Human Research Ethics Committee of The University of Melbourne (ID 20935).

#### 2.1.2. Procedure

Participants completed the task online after giving consent, providing optional demographic information, reading the instructions, and answering comprehension questions about them. The task was programmed using jsPsych (de Leeuw, Gilbert, & Luchterhandt, 2023) and was set up as a game in which the aim was to guess computer passwords from a series of hint words which were related in meaning to the password. Each trial corresponded to a computer with a different password, which participants were told was a common English word. Each password was presented with up to six hint words, revealed one at a time; participants guessed the password after each hint was presented. Hints were presented in order of their semantic similarity to the target word (see below), with the most similar hint shown first. A trial ended (and the password was revealed) either when the password was successfully guessed or after all six hints had been provided with no success. Performance on each trial was measured by computing whether the target was correctly guessed (accuracy), and if so, the number of attempts it took to get (number of guesses).

Participants were told that their goal was to unlock as many computers as possible using as few guesses as possible. The instructions provided to participants were:

> All of the computers have been locked and their passwords reset to a random word in the English language. To make matters worse, each computer has been locked with a different password. You manage to create an algorithm that makes each computer generate 6 password hints for each password. Each hint is related in meaning to the password, but is *not* the password itself. For each computer, you'll see the hints one by



Computers remaining: 25

Computers unlocked: 0
Computers locked: 1

Previous guesses
bedroom
night

LOCKED!

Enter password:

Attempts left: 4

Password hints:
bed
sleep
rest

Fig. 1. Screenshot of an experimental trial. Participants were instructed that they were to guess a computer password on each trial, which was a simple English word related in meaning to the password hints. The hints were displayed one at a time, with an attempt to guess the password after each one. The previous guesses for the current trial, the number of trials succeeded/failed, and the remaining number of trials were also shown. In this example, the target word is "pillow" and the person has been given three hints so far.

> one, and will have a chance to guess the password after each one. If you can't guess the password after 6 hints, the computer will be locked permanently. Your goal is to unlock as many computers as you can. It is crucial that you crack each password in as few guesses as possible.

Before the main experiment began, participants were shown 10 example passwords to illustrate the fact that they were simple, common English words (*family*, *table*, *jump*, *great*, *math*, *tell*, *thirsty*, *matter*, *want*, and *smell*), but were informed than none of the examples were any of the to-be-guessed words. They then completed a practice trial (the target word was *cat*) before continuing on to the 24 experimental trials. As Fig. 1 shows, on each trial, participants were able to see the hints they had been shown up to the current hint, their previous guesses, and the number of attempts remaining for the computer on that trial. They were also shown a running tally of how many computers had been successfully unlocked thus far and how many computers/trials were remaining.

Each participant completed 24 experimental trials plus two nonexperimental catch trials that were 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 people in total); this was done in order to remove people who were not trying or did not understand the task. Except for the catch trials, which were the same and always on the 10th and 20th trials, the target words and order of the trials were randomized for each person.

We manipulated three factors within subject in a 4 × 3 × 2 design: target condition, hint condition, and hint similarity type (all described below). These factors were fully counterbalanced, such that each of the 24 combinations of factors was seen exactly once by each participant. Each participant completed six target words from each target condition (which were randomly selected from the target words in each condition), which were randomly paired with the six 3 hint condition × 2 similarity type combinations.

### *2.1.3. Materials*

**Core word lists**. Here, we describe the methods for measuring coreness and identifying core words under the three different approaches considered in the introduction.

- InStr. Word association-based core words were measured by In-Strength, a graph-theoretic measure of centrality. Word association data was sourced from the Small World of Words (SWOW) project, which contains associations for over 12,000 English words (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019). The InStr measure for each word was computed as the log-10 transformed sum of the weights of all edges directed toward a given node, where nodes represent words and edge weights represent associative strengths. Words that are more core according to InStr are those that are commonly given as associates to other words (e.g., *money, car, love, dog*).
- AoA. Core words based on those that are earliest to be learned were measured using age-of-acquisition norms sourced from Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012). These norms provide the mean age-of-acquisition in years for over 30,000 English words, and are among the most widely used norms and also correlate well with more objective measures of AoA (Brysbaert & Biemiller, 2017). According to this definition of coreness, earlier-learned words are more core (e.g., *mom, potty, nap, hug*).
- WF. Coreness corresponding to the distributional language-based approach was measured using log-10 word frequency from the SUBTLEX norms (Brysbaert & New, 2009), such that more frequent words are more core (e.g., *go, like, good, say*).

For all original data sources, we standardized all words to US spelling and lemmatized words by aggregating inflectional forms of the same lemma (e.g., *run*, *runs*, *running*) using the AntBNC Lemma List, which was created based on the BNC corpus (Anthony, 2022). Words were lemmatized in the SWOW data by summing response counts and recalculating associative strengths for lemmas before computing InStr. For the AoA data, we took the lowest age-of-acquisition among all words for a given lemma, and for the word frequency data, we summed the raw frequencies for all the word forms linked to a lemma. We excluded function words, including determiners, particles, negations, conjunctions, auxiliary verbs, prepositions, pronouns, and interjections, along with proper nouns, taboo words, and non-words. Some additional words (all AoA) were also removed because they were not in the datasets used to select hint words for targets.[3]

Following this, we designated the top 300 words on each measure to be the "core words" for that measure, which we used to select target and hint words. Table 1 contains examples from each list. There is modest overlap between the core word lists: InStr shares 40% and 43% of its words with the AoA and WF lists, respectively, and the overlap between AoA and WF is 26%. And although these psycholinguistic measures are well known to correlate highly on the whole (Brysbaert et al., 2018), the correlations for the core words themselves are much lower ($|\rho| < .26$).

Table 2
Target words in each target condition

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

*Note*: Each participant saw a random subset of six words from each target condition (24 experimental trials total).

In order to compare word coreness across the three measures, the three coreness measures were normalized to a common coreness metric. For each of the coreness measures, we computed a word's coreness as the difference between each word and the most core word on that measure, divided by the difference between the most core word and the last (i.e., 300th) core word. This results in a coreness metric where the top 300 words in the list have values between 0 and 1, with 0 representing the most core word on each list. Words that are not in the top 300 are considered noncore words (see below) and have values above 1.

**Target conditions**. There were four target conditions, each corresponding to a list of 24 target words taken from the top 300 core words (see Table 2). Target words in the AoA, WF, and InStr conditions were selected to be more core on the corresponding coreness measure and less (but equally) core on the other two. As a baseline, we also created an Equal target condition composed of words that were equally core on all three measures. Each participant had to guess six random words from each of the four target conditions.

The target words in each of the three core-word conditions were chosen so that, as much as possible, they (1) maximized coreness based on the measure for that condition; (2) maximized the difference between that measure and the coreness measures for the other two conditions; and (3) were as similar as possible for the coreness measures of the other two conditions. The resulting target words were more core on the measure of the condition they were in, and less (but equally) core on the other coreness measures (see Fig. 2). The target words in the Equal condition were chosen to have similar coreness on all three measures (e.g., *big* scored around 0.5 on all measures, *chain* around 1.5 on all, etc.).

As shown in Table 3, the target words varied considerably in their part-of-speech makeup across conditions. The WF condition target words had more verbs relative to the other conditions, which is sensible given that verbs are highly frequent in language use. Conversely, the InStr and AoA conditions had more nouns, possibly reflecting the fact that they are more imageable and mentally salient. We chose not to control for part-of-speech because these differences reflect real distinctions between what each approach predicts should be core. That said, exploratory analyses suggest that part-of-speech differences are not the primary driver of our findings (see Results).

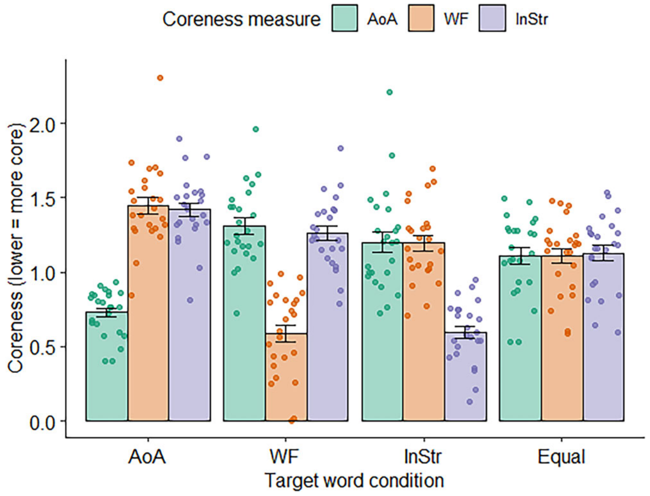


Fig. 2. Coreness of each target word in the four target conditions (AoA, WF, InStr, and Equal) on each of the three coreness measures (AoA, WF, and InStr). Coreness values closer to 0 represent words that are more core. Targets were more core on the coreness measure corresponding to their condition and less (but equally) core on the other two. For example, WF target words had an average coreness of 0.59 on the WF measure, but 1.29 on the other two (AoA 1.31, InStr 1.26). Targets in the Equal condition had as close to equal coreness on all three measures as possible.

Table 3
Number of target words in each of the four conditions (rows) corresponding to each part of speech

| | Noun | Verb | Adjective |
|---|---|---|---|
| AoA | 21 | 2 | 1 |
| InStr | 15 | 0 | 9 |
| Equal | 12 | 9 | 3 |
| WF | 5 | 18 | 1 |
| Total | 53 | 29 | 14 |

*Note*: WF target words were more likely to be verbs than the other conditions.

The target-word conditions allow us to explore whether it is easier to guess words that are more central according to different theories of meaning. The Equal condition allows us to ask whether different hint words are more or less useful. We turn to this next.

**Hint conditions**. The three hint conditions (AoA, InStr, WF) corresponded to the three core word lists from which hint words for the target words were drawn. Each target word was paired with sets of six hints for each measure. For instance, the target word *rice* (from the AoA list) was associated with six hints from the AoA list (*noodle, bread, breakfast, sugar, dinner, bean*), six hints from the InStr list (*bread, cook, chicken, food, vegetable, cake*), and six hints from the WF list (*dinner, eat, hot, honey, whole, minute*). Thus, for each target, one of the hint lists was congruent (with hints and targets selected from the same core word list) and the other two were incongruent.

Each target word that a participant saw was paired with a set of hints from one of the three hint conditions. This was counterbalanced so that each participant was presented with every target / hint condition combination (e.g., WF target word with AoA hints) the same number of times. 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.

**Similarity type**. For any given hint condition and target word, the hints were the six words in that hint list that were most semantically similar to the target. Since similarity is a function of the semantic representation 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 the Small World of Words word association network (SWOW similarity), or based on embeddings in a distributional language-based model (`word2vec`).

SWOW 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 $\alpha$ 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. The similarity between each pair of words was obtained by calculating the cosine of their 300-dimensional embeddings.

The difference between these two methods of calculating similarity (word2vec and SWOW) is not the topic of current investigation itself, but we implemented both metrics since performance on these kinds of task can vary depending on the measure (Heath et al., 2013; Shen et al., 2018; Vankrunkelsven et al., 2021). We counterbalanced the target, hint, and similarity metric conditions such that each target / hint condition combination was presented once using hints selected via the SWOW method, and once using word2vec. This ensures that performance between hint and target conditions can be disentangled from the particular similarity metric assumed.

After selecting hints based on the highest semantic similarity, 66 hint words (1.91%) were replaced because they were close variants of the target word or other hint words (e.g., *grandmother* and *grandma*). In these cases, the hint word with the higher semantic similarity was retained, and the other hint was replaced with the next-most similar word.

## 2.2. Results

### 2.2.1. Accuracy

As Fig. 3 shows, the task was difficult: only 27.8% of all trials resulted in a successful guess. However, performance varied widely. The lowest-scoring participant guessed only one word (4.2% accurate), while the best one guessed 14 words (58.3%). Nobody guessed the hardest word (*boring*), while 86% guessed the easiest (*music*). The high level of difficulty likely reflects the fact that the hints were drawn from a restricted set of core words, rather

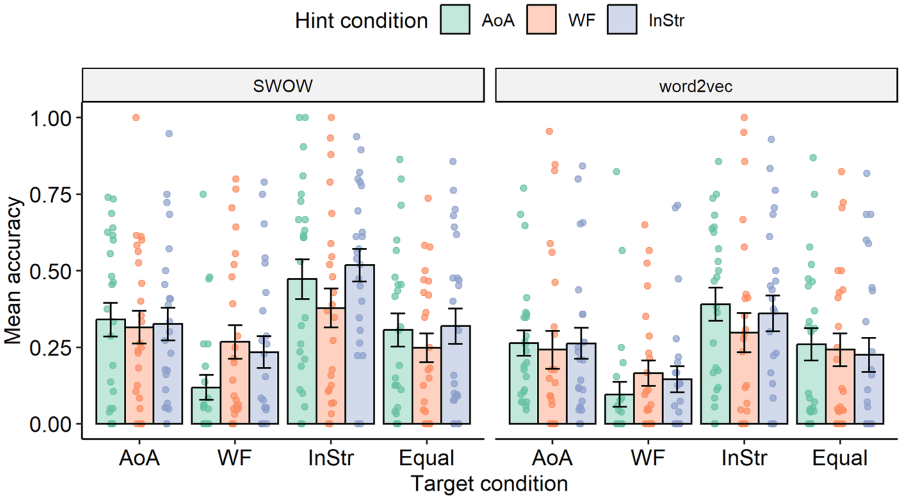


Fig. 3. Experiment 1: Mean accuracy for target words by hint and target condition. The results are displayed separately for each similarity type used to select hints (SWOW and word2vec). Each dot represents the mean performance for a target word in that hint (color) and target (*x* axis) condition. Overall, there is little systematic accuracy difference between the hint conditions, but the InStr target words tend to be guessed with the most accuracy and the WF target words appear to be correctly guessed least often.

than being taken from the full vocabulary, as is typical in similar word games. That said, the fact that only eight people were excluded for failure to guess the check trials indicates that people understood the task and were completing it as intended.

As preregistered, we first conducted preliminary analyses to assess whether the similarity metric that was used to generate the hints had an effect on accuracy. Two 2 × 3 two-way repeated measures ANOVAs were conducted with mean accuracy as the outcome measure and hint condition and similarity metric as the two predictors. There was a significant difference between the similarity metrics on accuracy, $F(1, 95) = 28.05$, $p < .001$, with SWOW hints yielding higher accuracy compared to word2vec hints. The interaction between hint condition and similarity metric was not significant. In light of this effect, analyses were conducted separately for the SWOW and word2vec hints.

For each similarity metric, we conducted a 4 × 3 two-way mixed ANOVA with mean accuracy as the outcome and target condition and hint condition as the predictors (see Fig. 3). For the trials with SWOW hints, there was no significant effect of hint condition on mean accuracy, $F(1.81, 166.59) = 1.39$, $p = .253$; however, there were significant differences between the target conditions, $F(3, 92) = 6.14$, $p < .001$. Post-hoc tests with Holm corrections showed that accuracy for the InStr condition was significantly higher than all others (all $p$s $< .05$), and accuracy for AoA target words was higher than WF target words ($p = .02$). The interaction between hint condition and target condition was nonsignificant, $F(5.43, 166.59) = 1.84$, $p = .101$, indicating that there was no systematic advantage for congruent hint and target conditions.

Table 4
Experiment 1 model comparison: Accuracy

| | | BIC | |
|---|---|---|---|
| Model | Description | SWOW | word2vec |
| mNull | `correct ~ 1` | 7159 | 6372 |
| mHxT | `correct ~ hint * target` | 6972 | 6245 |
| m1p | `correct ~ 1 + (1 | participant)` | 7158 | 6381 |
| mH1p | `correct ~ hint + (1 | participant)` | 7156 | 6394 |
| mT1p | `correct ~ target + (1 | participant)` | 6973 | **6213** |
| mHT1p | `correct ~ hint + target + (1 | participant)` | 6971 | 6226 |
| mHxT1p | `correct ~ hint * target + (1 | participant)` | **6964** | 6252 |

*Note*: We compare a set of preregistered mixed-effects logistic regression models using BIC (lowest is best, shown in bold). We perform separate analyses for each similarity metric (SWOW vs. word2vec). All models have accuracy as an outcome variable (`correct`, which is 1 if the word was guessed and 0 if not). Models vary in the presence of a random effect for participant, target word condition (`target`), hint condition (`hint`), and interaction(s). For both similarity metrics, the best-fit model contained target word condition as a predictor. For SWOW trials, the best-fit model also contained hint condition and an interaction between hint and target.

For the word2vec trials, once again there was a significant effect of target condition, $F(3, 92) = 4.41$, $p = .006$, but not hint condition, $F(1.88, 172.97) = 0.18$, $p = .819$, or the interaction, $F(5.64, 172.97) = 0.85$, $p = .528$. Post-hoc tests with Holm corrections showed that accuracy for the INSTR and AOA target words was significantly higher than the WF target words ($p < .001$ and $p = .02$, respectively). Additionally, the INSTR accuracy was significantly higher than the accuracy for EQUAL words ($p = .04$), and the WF target word accuracy was significantly lower than EQUAL ($p = .04$).

These ANOVA analyses show the average performance for each target word, but collapse over individual trials and provide no way to capture participant-level effects. We, therefore, also conducted a set of preregistered mixed-effects logistic regressions with hint condition and target condition as fixed effects and participant included as a random effect. As Table 4 shows, multiple versions of each model were compared using BIC, which penalizes unnecessary complexity.

For the SWOW trials, the best-fitting model (mHxT1p) contained both target and hint condition, a random effect for participant, and an interaction between hint and target condition. Table 5 reports the main parameters of the best-fit model,[4] revealing that accuracy was significantly higher for the INSTR target words and significantly lower for the WF target words, compared to the EQUAL condition (the reference category). This is consistent with the ANOVA results. The model also showed a slight effect of hint condition, with lower accuracy for WF hints compared to AOA (the reference category).

For the word2vec trials, the best-fitting model (mT1p) contained only target word condition as a fixed effect. Its parameters (Table 5) reveal that accuracy was significantly higher for INSTR target words and significantly lower for WF targets compared to the reference category (EQUAL).

Table 5
Experiment 1: Parameters of the best-fitting mixed-effects logistic regressions in Table 4

| | SWOW | | | | word2vec | | | |
|---|---|---|---|---|---|---|---|---|
| Term | *b* | 95% CI | *z* | *p* | *b* | 95% CI | *z* | *p* |
| Intercept | −0.87 | [−1.07, −0.67] | −8.49 | <.001 | −1.19 | [−1.31, −1.06] | −18.62 | <.001 |
| WF hint | −0.40 | [−0.69, −0.11] | −2.67 | .008 | | | | |
| InStr hint | 0.15 | [−0.13, 0.42] | 1.06 | .291 | | | | |
| AoA target | 0.16 | [−0.12, 0.43] | 1.13 | .260 | 0.15 | [−0.02, 0.32] | 1.69 | .091 |
| WF target | −1.22 | [−1.57, −0.88] | −6.94 | <.001 | −0.75 | [−0.95, −0.55] | −7.42 | <.001 |
| InStr target | 0.69 | [0.42, 0.96] | 5.04 | <.001 | 0.52 | [0.36, 0.69] | 6.25 | <.001 |
| ICC | 0.036 | | | | 0.013 | | | |

*Note*: The best-fitting model for SWOW similarity (`mHxT1p`) contains both hint condition and target condition and an interaction. The best-fitting model for word2vec similarity (`mT1p`) contains target condition but not hint condition. The reference categories are AoA (hint condition) and Equal (target condition). *ICC* = Intraclass Correlation Coefficient.

Taken together, the results for both SWOW and word2vec trials suggested a consistent pattern of performance, in which InStr targets were guessed more often than AoA words, and accuracy for both was better than WF target words. There were no consistent differences found between the hint conditions.

This is interesting, but one limitation is that our accuracy measure does not differentiate between guessing correctly on the first hint or the sixth. One way to address this limitation would be to perform similar analyses on only the correctly guessed words, and ask what factors predict fewer guesses. Supplementary Appendix A has the results of this analysis, which found few systematic effects. In hindsight, this may not be a surprise: it is not ideal to condition only on correct guesses, both because of limited power and because of the difficulty interpreting any results. A more sensible approach is to conduct survival analyses, which take into account both accuracy and number of guesses. We turn to this next.

### *2.2.2. Survival analysis*

We performed Kaplan–Meier survival analyses, which allow us to estimate the “survival” of target words over the course of a trial. They thus allow us to take into account time (in this case, number of guesses) in addition to overall accuracy. Here, successful guesses are the “death” events in the analysis, meaning that lower survival corresponds to more successful guessing. As Fig. 4 shows, the qualitative pattern of results is similar to what we saw before: InStr target words tended to be guessed most quickly, while WF target words were guessed most slowly.

A log-rank test confirmed that for both SWOW and word2vec similarity, the survival curves by hint condition and target condition differed significantly from each other (both *p*s $< .001$). A Cox proportional-hazards model analyzed the effect of each factor on overall survival (see Table 6). For the SWOW trials, InStr hints were associated with significantly better guessing than AoA hints (the reference category), but the effect was small. Bigger differences were observed between the target conditions: relative to the Equal condition (the reference

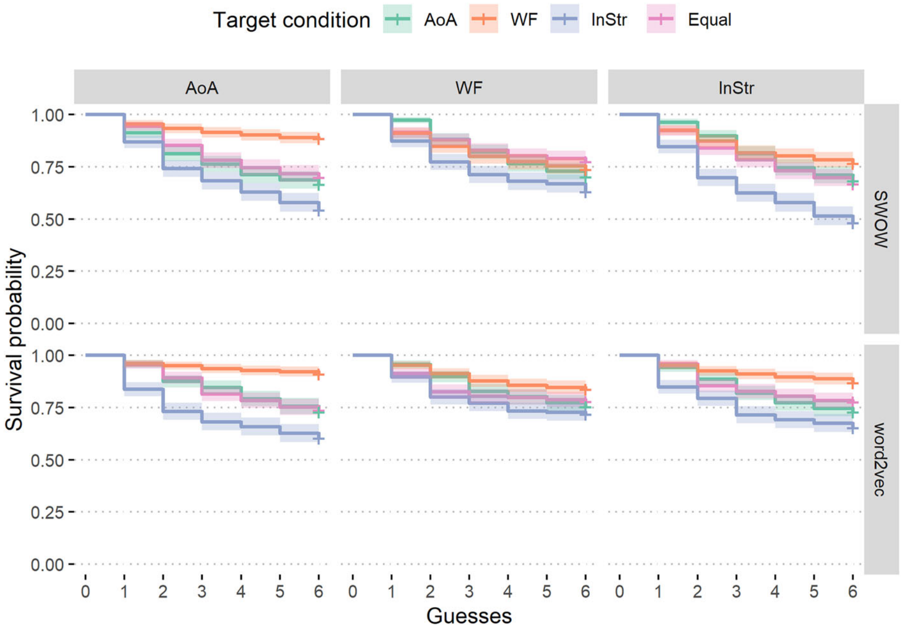


Fig. 4. Experiment 1: Kaplan–Meier survival curves by hint and target condition. Each line shows the survival probability of the trials in each target condition (colors) over successive guesses. Events were defined as successful guesses, so lower survival is better performance. Column panels represent each hint condition and rows show the similarity type. Across all conditions, the best performance occurs with the INSTR target words (lowest curves), and the worst performance for the WF target words (highest curves).

Table 6
Experiment 1: Cox proportional-hazards model predicting target word survival from hint condition and target condition

| | SWOW | | | word2vec | | |
|---|---|---|---|---|---|---|
| Term | HR | 95% CI | *p* | HR | 95% CI | *p* |
| WF hint | 0.94 | [0.83, 1.05] | .283 | 0.88 | [0.77, 1.00] | .046 |
| INSTR hint | 1.19 | [1.06, 1.33] | .003 | 0.94 | [0.83, 1.07] | .381 |
| AOA target | 1.10 | [0.97, 1.26] | .146 | 1.10 | [0.95, 1.28] | .183 |
| WF target | 0.68 | [0.59, 0.80] | <.001 | 0.51 | [0.43, 0.61] | <.001 |
| INSTR target | 1.78 | [1.57, 2.02] | <.001 | 1.59 | [1.39, 1.83] | <.001 |

*Note*: Results are displayed separately for each similarity type condition. A hazard ratio (HR) below 1 indicates higher survival, which in this experiment indicates worse guessing. The reference categories are AOA (hint condition) and EQUAL (target condition). There are consistent large differences between the target conditions, with WF having worse performance and INSTR having better performance. The differences between the hint conditions are smaller and less consistent.

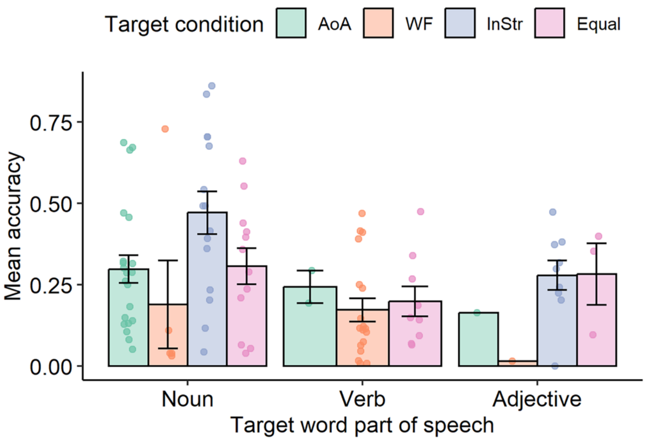


Fig. 5. Experiment 1: Part-of-speech effects on mean accuracy by target condition. Each dot represents the mean accuracy for a target word in a given target condition, separated by its part-of-speech. There are differences in the overall accuracy of part-of-speech categories, but differences between the target conditions within each part-of-speech category also exist.

category), InStr target words were guessed significantly better and WF target words were guessed significantly worse. The same pattern of differences for the target words was observed for the word2vec trials, but for the hints, WF hints were associated with significantly worse guessing than AoA.

Overall, regardless of the method of analysis, we see the same consistent pattern: target word performance varies substantially depending on what core word list the targets are from. People guessed InStr targets most accurately, followed by AoA and Equal. The highly frequent target words from the WF core word list were hardest to guess. However, there were no consistent differences between hint conditions.

What might explain the differences in target word performance? To what extent are these differences due to factors that naturally differ between core word types, like part of speech or similarity structure? We explore these questions in the next section.

### 2.2.3. *Exploratory analyses*

**Part of speech**. As mentioned before, one of the obvious differences between the target conditions was in their part-of-speech distributions, with WF words containing many more verbs. Indeed, target word part-of-speech did influence performance: a Welch test showed significant differences in mean accuracy, $F(2, 40.20) = 6.70$, $p = .003$. Games–Howell post-hoc tests showed significantly higher accuracy for nouns ($M = .34$) compared to verbs ($M = .19$), $p < .001$, with adjectives in between ($M = .25$).

That said, while part-of-speech differences may partially explain the performance differentials between target word conditions, Fig. 5 demonstrates that it is not the complete story. For example, when considering only target words that are nouns, accuracy is still higher for the InStr condition: a Kruskal–Wallis ANOVA shows a significant difference in accuracy

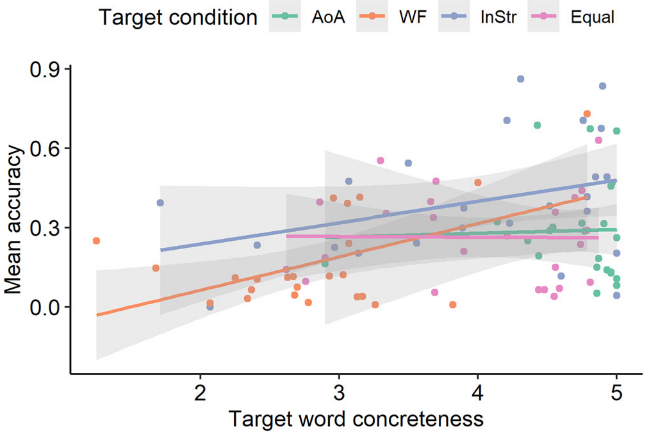


Fig. 6. Experiment 1: Concreteness effects on mean accuracy by target condition. Each dot represents the concreteness ($x$ axis) and mean accuracy ($y$ axis) for a target word in a given target condition. Concreteness predicts accuracy overall, but differences between the target conditions over and above the effect of concreteness also exist.

by target condition, $H(3) = 8.58$, $p = .035$, with Dunn's post-hoc tests showing a significant difference between InStr and WF, $p = .039$. Although the unbalanced design and low numbers of observations in some cells make it difficult to conduct a complete analysis, this tentatively suggests that differences between the target conditions remain even after taking part-of-speech into account.

**Concreteness**. One factor that is closely related to part-of-speech is concreteness, which could influence nameability and salience, making abstract words harder to guess. To explore this, we sourced concreteness for each target word based on the Brysbaert, Warriner, and Kuperman (2014) norms. Indeed, WF target words had the lowest concreteness ($M = 2.84$, $SD = 0.73$), which could have contributed to their lower performance. By contrast, the AoA target words had the highest average concreteness ($M = 4.69$, $SD = 0.45$), which reflects the fact that they contain lots of early acquired words which tend to be more physical in nature, while the InStr ($M = 4.00$, $SD = 0.99$) and Equal ($M = 4.02$, $SD = 0.74$) target words were in between. A simple linear regression showed that higher concreteness did significantly predict higher target word accuracy, $b = 0.07$, 95% CI $[0.03, 0.11]$, $p < .001$.

However, as before, the target word condition differences remained over and above the effect of concreteness when it was added to the analysis (see Fig. 6). A linear regression model predicting target word accuracy based on both concreteness and target condition showed that the InStr target words had significantly higher accuracy than Equal target words even with concreteness taken into account, $b = 0.14$, 95% CI $[0.03, 0.25]$, $p = .014$. By comparison, accuracy in the AoA and WF target conditions was not significantly different from the Equal target condition.[5]

**Hint similarity and interrelatedness**. Another factor that is likely to influence performance on this task is semantic similarity, as demonstrated in many previous studies (Heath et al., 2013; Shen et al., 2018; Vankrunkelsven et al., 2021). We explored two possibilities

Table 7
Experiment 1: Model comparisons for semantic similarity analyses

| | | BIC | |
|---|---|---|---|
| Model | Description | SWOW | word2vec |
| mNull | `accuracy ~ 1 + (1 \| targetWord)` | 62 | 1 |
| mS | `accuracy ~ hintSim + (1 \| targetWord)` | 24 | −11 |
| mR | `accuracy ~ hintRel + (1 \| targetWord)` | 66 | 8 |
| mRS | `accuracy ~ hintSim + hintRel + (1 \| targetWord)` | −3 | **−31** |
| mRxS | `accuracy ~ hintSim * hintRel + (1 \| targetWord)` | **−12** | −28 |

*Note*: We compare a set of mixed-effects linear regression models using BIC (lowest is best, shown in bold). We perform separate analyses for each similarity metric (SWOW vs. word2vec). All models have a random effect for target word (`targetWord`) and use mean `accuracy` of each target word for a given hint type as the outcome variable. Models vary in the presence of predictors for average hint-target similarity (`hintSim`) and hint interrelatedness (`hintRel`) for that word. Both best-fitting models contain both predictors, but for SWOW, there is also an interaction between hint similarity and hint relatedness.

in this regard. First, we evaluated whether accuracy was improved on trials where the hints had a higher average similarity to the target word (*hint-target similarity*). Second, we tested whether accuracy was higher for sets of hints that were more diverse and less related to each other (*hint interrelatedness*), with the reasoning that more semantically distinct hints might provide more independent information about the target. There is an inherent link between these two factors: sets of hints that are semantically closer to their respective target will necessarily be constrained to be more similar to each other as well.

In order to analyze the effects of hint-target similarity and hint interrelatedness, we performed model comparison between a set of mixed-effects models which all had target word as a random effect and an outcome variable of mean accuracy for each target word in a given hint condition and similarity type. Both predictors were mean-centered. The results, shown in Table 7, reveal that for both similarity types, accuracy was best predicted by both hint-target similarity and hint interrelatedness. Hint-target similarity had a significant positive relationship with mean accuracy on SWOW trials, $b = 2.90$ $[2.33, 3.47]$, as well as word2vec ones, $b = 1.92$ $[1.39, 2.45]$. This indicates that, as one would expect, targets with more semantically related hints were guessed more easily. Conversely, hint interrelatedness had a significant negative effect on mean accuracy on both SWOW trials, $b = -1.39$ $[-1.93, -0.86]$ and word2vec ones $b = -1.10$ $[-1.50, -0.69]$, indicating that more semantically distinct hints were better. Also, for the SWOW trials, the significant interaction, $b = -5.59$ $[-8.78, -2.42]$, indicated that there was an added benefit of having hints that were more similar to the target when they were more distinct from each other.

These results are reassuring given their concordance with our intuitions of how hint similarity and interrelatedness should affect guessing. However, the more pertinent question here is: would the effect of target word condition remain even after taking hint similarity and interrelatedness into account? Put another way, does the accuracy advantage for InStr target words go beyond the similarity and/or interrelatedness of its hints? In order to explore this, we evaluated a set of linear models with target condition as a fixed effect instead of target

Table 8
Experiment 1: Model comparisons for target word and hint similarity

| | | **BIC** | |
|---|---|---|---|
| Model | Description | SWOW | word2vec |
| mNull | `accuracy ~ 1` | 97 | 58 |
| mRS | `accuracy ~ hintSim + hintRel` | 34 | 33 |
| mRxS | `accuracy ~ hintSim * hintRel` | 26 | 39 |
| mRST | `accuracy ~ hintSim + hintRel + target` | 18 | **22** |
| mRxST | `accuracy ~ hintSim * hintRel + target` | **8** | 26 |

*Note*: We compare a set of linear regression models using BIC (lowest is best, shown in bold), with separate analyses for each similarity metric (SWOW vs. word2vec). All models use mean `accuracy` for each target word for a given hint type as the outcome variable. They vary in the presence of predictors for average hint-target similarity (`hintSim`), hint interrelatedness (`hintRel`), and target condition (`target`) for that word. Both best-fitting models contain target word condition as a predictor in addition to hint-target similarity and hint relatedness.

word as a random effect (we cannot include both because no target words are shared between target conditions). The outcome variable of accuracy and the other two predictors are the same as before, but the question is whether models that contain target word condition provide additional predictive power over models which do not.

As Table 8 indicates, the best-fit models for both SWOW and word2vec contained target word condition as well as both similarity predictors. The parameters for hint-target similarity and hint interrelatedness are similar in magnitude and direction as before, but our focus here is on the parameters for target word condition of the best-fit models. Relative to the reference category of Equal, the InStr target words had significantly higher accuracy on both SWOW trials, $b = 0.11$, $t = 2.84$, $p = .005$ and word2vec ones, $b = 0.08$, $t = 2.08$, $p = .038$. In addition, the WF target words had significantly lower accuracy on both SWOW trials, $b = -0.12$, $t = -3.13$, $p = .002$ and word2vec ones, $b = -0.14$, $t = -3.34$, $p < .001$. Overall, this suggests that the differential accuracy for different core target words cannot be fully explained by factors like semantic similarity.

### *2.3. Discussion*

Experiment 1 compared different types of core words in a word-guessing game in terms of (1) their effectiveness as hints, and (2) their ability to be successfully guessed as target words. We found that performance was affected by the nature of the target words but not the hint words. The InStr target words were guessed most easily, followed by AoA, with WF guessed correctly least often. This pattern of results remained even after accounting for factors like part-of-speech, concreteness, and semantic similarity.

These results reveal an asymmetry between our two questions: in this task, what mattered were the target words being guessed, rather than the hint words from which the guesses were made. The higher accuracy for InStr target words (and to a lesser extent, AoA), regardless of hint type, suggests that words which are highly connected in word association networks may be more accessible or mentally salient in some way, perhaps occupying more central or prominent positions in the mental lexicon. This is in spite of well-documented word frequency effects in language processing, which would benefit the WF target words (Brysbaert et al., 2018; Ellis, 2002).

In contrast, core words in general were not very effective as hints; this is evident in the overall low accuracy on the task (with only around a third of words being successfully guessed) as well as the lack of difference between the hint conditions. This may indicate that the restricted set of words which we have defined as core words may not be sufficient to fully account for word meaning in this task, and that words from beyond our core word lists could make important contributions to accessing or providing indicators to word meaning. It also raises the question of how robust the InStr target word advantage is: would InStr core words still be guessed more easily if the hints were less restricted? Experiment 2 addresses these issues. We have two aims: first, to investigate the effectiveness of noncore word hints, and, second, to test whether the differences between types of target words remain for those hints as well.

## 3. Experiment 2

### *3.1. Background*

The aim of this experiment was to investigate how noncore word hints would affect performance on our password game. Given that we found no differences between hint conditions in Experiment 1, we wanted to find out how the core-word hints compared to a comparably sized list of noncore word hints (the Random vocabulary hint condition) as well as to the best possible hints regardless of coreness (the Full vocabulary hint condition). We were also interested in whether the differential performance by target word in Experiment 1 would remain even for these different hints. We, therefore, kept the same target words and basic experimental paradigm from Experiment 1, but changed the hint words, as described below.

### *3.2. Method*

#### *3.2.1. Participants and procedure*

Five hundred participants were recruited from Amazon Mechanical Turk and paid \$3.33 for the 20 min task. Four hundred and ninety-seven participants passed the preregistered[6] check trials. Ages ranged from 20 to 74 years old ($M = 38.27$) and 40% were female. Eighty-eight percent reported being native English speakers, and all had previously passed a qualification assessing English proficiency.

The task was the same as in Experiment 1. Participants completed 24 experimental trials and two nonexperimental catch trials (for this experiment, the target words were *apple* and *cake*). As before, each participant completed six randomly selected target words from each target condition. The catch trials were kept in the same position (the 10th and 20th trials), and the target words and the order of the trials were randomized for each participant. Target condition and similarity type were fully manipulated within-subject such that each target condition had an equal number of trials with hints selected using each similarity metric. Hint condition was manipulated between-subject such that each participant only got hints from either the RANDOM or FULL vocabulary hint condition.

### 3.2.2. Materials

**Hint conditions**. The target words were the same as in Experiment 1, but two new hint conditions replaced the core-word ones. In the RANDOM vocabulary condition, the hints were drawn from a pool of 300 randomly selected noncore words; this was done for comparability to the restricted range of the 300-item core word lists in Experiment 1. In the FULL vocabulary condition, the hints were drawn from the entire available vocabulary. The complete vocabulary consisted of the set of words that were in both the SWOW and word2vec data used to select hints for target words (i.e., all the words for which semantic similarity data was available to select hints). Similarly to Experiment 1, words were lemmatized to group inflectional forms of the same lemma together, and function words, proper nouns, taboo words, and nonwords were removed.

For each participant in the RANDOM vocabulary hint condition, the AOA, WF, and INSTR core words (i.e., all of the candidate hint words from Experiment 1) were removed from the complete vocabulary before a unique set of 300 noncore words was randomly sampled. The most semantically similar words to the target in that set were then chosen as the hints in the same way as in Experiment 1 (with half of the trials using the SWOW similarity metric and half using word2vec). This meant that two participants guessing the same target word would tend to see different hint words; on average, 73% of the hints that were presented for a given target word across participants were unique.

For the FULL vocabulary condition, the pool of candidate hints consisted of the complete vocabulary, regardless of whether the words were core words or not. The six hints were the most semantically similar out of this entire set (for each target word, half of the trials used the SWOW measure and half used the word2vec). Of the final selected hints in this condition, 10.9% were AOA core words, 12.6% were WF core words, and 8.9% were from the INSTR core word list.

As in Experiment 1, 257 hints (22%) were replaced for being close variants of the target word or other hint words in the FULL vocabulary condition. This was not done in the RANDOM condition because hint selection took place during the experiment for each participant based on their unique set of words.

## 3.3. Results

### 3.3.1. Main analyses

The two conditions varied widely in terms of difficulty: only 25.3% of trials were guessed successfully in the RANDOM condition, but the FULL condition was much easier with 70.3% of trials correct overall. Performance also varied widely, with the lowest-scoring participant in the RANDOM condition getting no trials correct, and the lowest-scoring participant in the FULL condition guessing only three correct words (12.5%), but the highest score in both conditions being 100%. The hardest word in both conditions was *crayon*, while the best-guessed words were *book* and *clean* in the FULL condition (100%) and *car* in the RANDOM condition (79.7%).

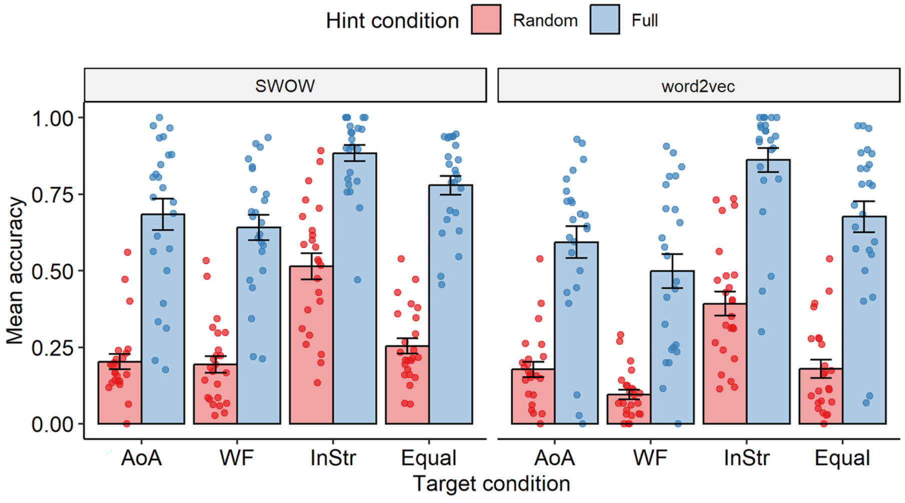


Fig. 7. Experiment 2: Mean accuracy for target words by hint condition and target condition. The results are displayed separately for each similarity type condition used to select hints, SWOW and word2vec. Each dot represents the mean accuracy for a target word in a given condition, averaged across all trials for that word. Performance is consistently better for the FULL vocabulary hint condition compared to the RANDOM vocabulary hint condition. Additionally, the best performance was attained for INSTR target words compared to the other target word conditions.

Similarly to Experiment 1, preliminary analyses assessing the effects of similarity metric showed that there was a significant effect on mean accuracy, $F(1, 95) = 44.21$, $p < .001$, in which better performance was attained for SWOW hints compared to word2vec hints. Once again, there were no significant interactions between hint condition and similarity metric.

For each similarity metric, we conducted a 4 × 2 two-way mixed ANOVA with mean accuracy as the outcome and target condition and hint condition as the predictors (see Fig. 7). There was a significant effect of target condition, $F(3, 92) = 19.81$, $p < .001$, with Holm-corrected post-hoc tests showing that the INSTR condition had significantly higher accuracy than all other target conditions (all $ps < .01$), with no significant differences between the other three. There was also a significant effect of hint condition, $F(1, 92) = 506.06$, $p < .001$; accuracy was much higher for the FULL vocabulary hints compared to the RANDOM vocabulary hints. The lack of a significant interaction between hint and target condition, $F(3, 92) = 2.62$, $p = .055$, indicates that the pattern of results for the target conditions held across both hint conditions, and vice versa.

The same qualitative patterns of results were observed for the word2vec trials: the ANOVA showed significant effects of hint condition, $F(1, 92) = 436.44$, $p < .001$, and target condition, $F(3, 92) = 16.10$, $p < .001$, but no interaction $F(3, 92) = 1.06$, $p = .369$. Accuracy was significantly higher for INSTR targets compared to all other target types (all $ps < .01$), and for FULL hints compared to RANDOM ones.

Mixed-effects logistic regression models supported the ANOVA results: as Table 9 shows, the best-fitting models for both similarity metrics contained both hint condition and target

Table 9
Experiment 2 model comparison: Accuracy

| Model | Description | BIC | |
|---|---|---|---|
| | | SWOW | word2vec |
| mNull | `correct ~ 1` | 8267 | 8185 |
| mHxT | `correct ~ hint * target` | 6685 | 6512 |
| m1p | `correct ~ 1 + (1 \| participant)` | 7452 | 7390 |
| mH1p | `correct ~ hint + (1 \| participant)` | 6963 | 6886 |
| mT1p | `correct ~ target + (1 \| participant)` | 7086 | 6939 |
| mHT1p | `correct ~ hint + target + (1 \| participant)` | **6596** | **6433** |
| mHxT1p | `correct ~ hint * target + (1 \| participant)` | 6617 | 6452 |

*Note*: We compare a set of preregistered mixed-effects logistic regression models using BIC (lowest is best, shown in bold). We perform separate analyses for each similarity metric (SWOW vs. word2vec). All models have accuracy as an outcome variable (`correct`, which is 1 if the word was guessed and 0 if not). Models vary in the presence of a random effect for participant, target word condition (`target`), hint condition (`hint`), and interaction(s). For both similarity metrics, the best-fit model contained both target and hint condition as predictors, with no interaction.

Table 10
Experiment 2: Parameters of the best-fitting mixed-effects logistic regressions in Table 9

| | SWOW | | | | word2vec | | | |
|---|---|---|---|---|---|---|---|---|
| Term | *b* | 95% CI | *z* | *p* | *b* | 95% CI | *z* | *p* |
| Intercept | −1.01 | [−1.16, −0.86] | −13.02 | <.001 | −1.54 | [−1.70, −1.38] | −18.78 | <.001 |
| FULL hint | 2.24 | [2.08, 2.41] | 26.74 | <.001 | 2.31 | [2.14, 2.48] | 27.13 | <.001 |
| AOA target | −0.43 | [−0.60, −0.26] | −4.99 | <.001 | −0.24 | [−0.42, −0.07] | −2.80 | .005 |
| WF target | −0.57 | [−0.74, −0.40] | −6.57 | <.001 | −0.79 | [−0.96, −0.61] | −8.77 | <.001 |
| INSTR target | 1.02 | [0.84, 1.19] | 11.38 | <.001 | 1.11 | [0.94, 1.29] | 12.43 | <.001 |
| ICC | 0.089 | | | | 0.086 | | | |

*Note*: The best-fitting model for both SWOW and word2vec similarity contains hint and target condition as predictors with no interaction. The reference categories are RANDOM (hint condition) and EQUAL (target condition). *ICC* = Intraclass Correlation Coefficient.

condition as predictors and no interaction. Table 10 reports the parameters of the best-fit models, showing that accuracy was significantly higher for the FULL compared to the RANDOM vocabulary hint condition (the reference category), and that accuracy was significantly higher for the INSTR target words compared to the EQUAL ones (the reference category). Accuracy was significantly lower for AOA and WF target words.

So far, these results indicate that people were able to guess significantly more target words on the basis of FULL vocabulary hints than RANDOM vocabulary ones. In addition, we see the same pattern we saw in Experiment 1, where INSTR target words were easiest to guess accurately and WF target words the most difficult.

In order to ensure that this qualitative finding is robust, we again follow up with Kaplan–Meier survival analyses. As Fig. 8 shows, the results remain consistent. Log-rank tests showed

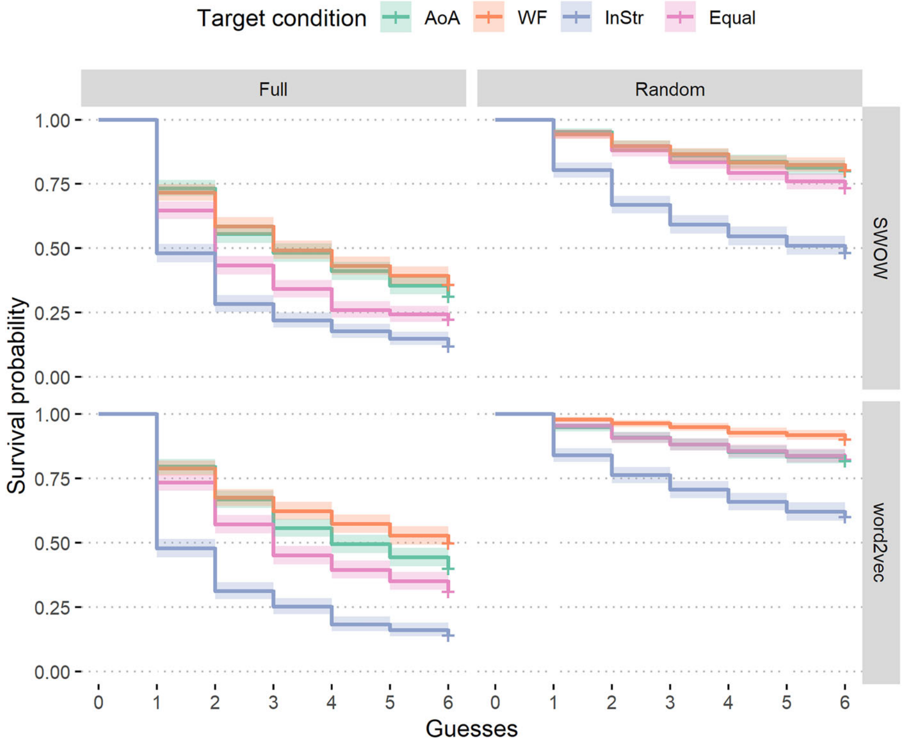


Fig. 8. Experiment 2: Kaplan–Meier survival curves by hint condition and target condition. Each line shows the survival probability of the trials in each target condition over successive guesses. Events were defined as successful guesses, so lower survival indicates better performance. Column panels represent hint condition and row panels show similarity type. People performed best for FULL vocabulary hints and INSTR target words.

that the survival curves by hint condition and target condition differed significantly from each other, for both the SWOW and word2vec trials (both $ps < .001$). Cox proportional-hazards models (see Table 11) showed that for both similarity types, the FULL vocabulary hints were guessed significantly better than the RANDOM ones. Additionally, the INSTR target words were guessed significantly more often than the EQUAL target words, and the AOA and WF significantly less often. The hazard ratios indicate that all of these effects were strong.

Taken together, similar results were attained as for Experiment 1, with the best performance for INSTR target words and the worst for the WF target words. Consistent differences in performance were also observed between the hint conditions, with FULL vocabulary hints robustly outperforming RANDOM ones. In the next section, we conduct a series of exploratory analyses designed to better understand this pattern of results.

### *3.3.2. Exploratory analyses*

In this section, we investigate whether performance in Experiment 2 was affected by the same factors we identified in Experiment 1 (i.e., part of speech, concreteness, hint

Table 11
Experiment 2: Cox proportional-hazards model predicting target word survival from hint condition and target condition

| | SWOW | | | word2vec | | |
|---|---|---|---|---|---|---|
| Term | HR | 95% CI | *p* | HR | 95% CI | *p* |
| Random hint | 0.24 | [0.22, 0.26] | <.001 | 0.20 | [0.19, 0.22] | <.001 |
| AoA target | 0.74 | [0.67, 0.82] | <.001 | 0.82 | [0.74, 0.92] | <.001 |
| WF target | 0.69 | [0.62, 0.77] | <.001 | 0.60 | [0.53, 0.68] | <.001 |
| InStr target | 1.72 | [1.56, 1.88] | <.001 | 2.04 | [1.84, 2.25] | <.001 |

*Note*: Results are displayed separately for each similarity type. A hazard ratio (HR) below 1 indicates higher survival, which here means worse guessing. The reference categories are Full (hint condition) and Equal (target condition). The Full vocabulary hint condition consistently outperforms the Random one. Performance is also consistently lowest for the AoA and WF target words, and highest for the InStr targets.

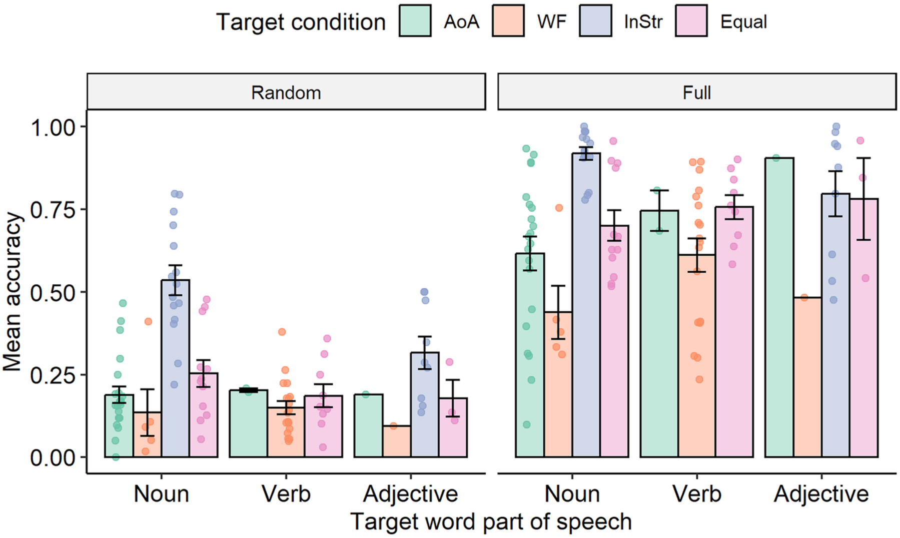


Fig. 9. Experiment 2: Part-of-speech effects on mean accuracy by target condition, displayed for both the Random and Full vocabulary hint conditions. Each dot represents the mean accuracy for a target word in a given target condition, separated by its part-of-speech. There are differences in the overall accuracy of part-of-speech categories, but differences between the target conditions within each part-of-speech category also exist.

interrelatedness, and hint-target semantic similarity). However, because of the large differences we observed between hint conditions on this experiment, we conduct separate analyses for the Random vocabulary and Full vocabulary hint conditions.

**Part of speech**. Fig. 9 shows accuracy broken down by part of speech for different target word conditions and hint conditions. There was no significant difference between target word parts-of-speech for Full vocabulary hints (Kruskal–Wallis, $H(2) = 3.36$, $p = .19$), possibly because of ceiling effects from high accuracy in this condition, but a significant one for Random ones (Welch test, $F(2, 35.37) = 8.99$, $p < .001$). Games–Howell post-hoc tests

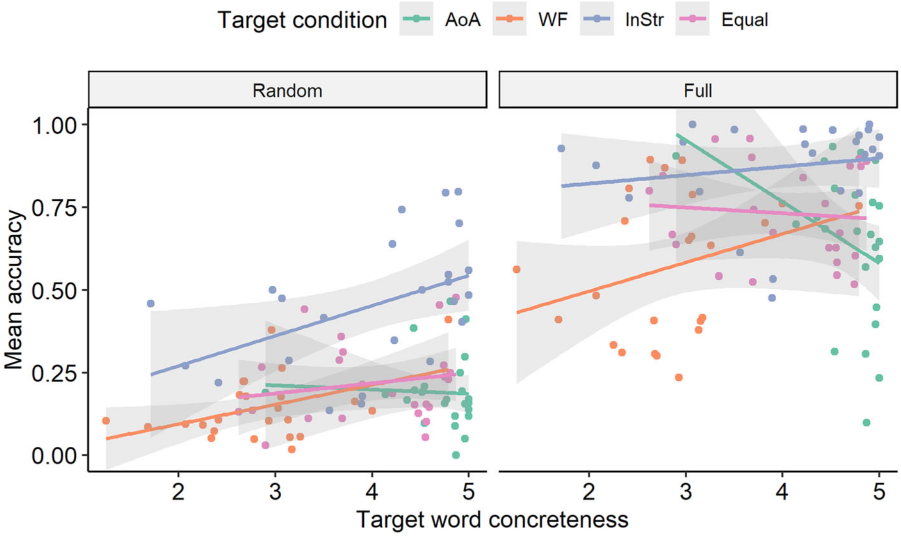


Fig. 10. Experiment 2: Concreteness effects on mean accuracy by target condition, displayed for both the RANDOM and FULL vocabulary hint conditions. Each dot represents the concreteness ($x$ axis) and mean accuracy ($y$ axis) for a target word in a given target condition. Concreteness predicts accuracy in the RANDOM condition, but not the FULL condition. In both cases, differences between the target conditions over and above the effect of concreteness also exist.

revealed that for RANDOM hints, accuracy was significantly higher for nouns than verbs, with adjectives in between.

As we did in Experiment 1, we can also investigate whether the differences between target conditions remain after taking part-of-speech into account by focusing only on nouns (the only part of speech with enough data from each target word condition to compare). Consistent with what we see in Fig. 9, separate Kruskal–Wallis ANOVAs showed a significant effect of target condition for both RANDOM hints, $H(3) = 27.50$, $p < .001$, and FULL vocabulary ones, $H(3) = 26.15$, $p < .001$. Dunn's post-hoc tests revealed that accuracy for INSTR target words was significantly higher than all other target word types (all $p$s $< .01$ for both RANDOM vocabulary hints and FULL vocabulary hints).

**Concreteness**. Linear regression models predicting target word accuracy from concreteness showed that concreteness significantly predicted accuracy for the RANDOM vocabulary hint condition, $b = 0.05$, 95% CI [0.02, 0.09], $p = .003$. However, it was not significant for the FULL hint condition, $b = 0.03$, 95% CI [−0.01, 0.08], $p = .131$, again possibly because of high accuracy for target words overall. As before, we are interested in what the relationship between target condition and accuracy looks like after accounting for concreteness effects. Fig. 10 shows the relationship between concreteness and accuracy, separated by target condition. For both the RANDOM and FULL vocabulary hint conditions, linear regression models predicting target word accuracy from concreteness and target condition showed in both cases that the INSTR target words still had higher accuracy even after taking concreteness into account.

Table 12
Experiment 2: Model comparisons for target word and hint similarity

| | | BIC | |
|---|---|---|---|
| Model | Description | SWOW | word2vec |
| mNull | accuracy ~ 1 | 100 | 119 |
| mRS | accuracy ~ hintSim + hintRel | −28 | −11 |
| mRxS | accuracy ~ hintSim * hintRel | −40 | −8 |
| mRST | accuracy ~ hintSim + hintRel + target | −52 | −31 |
| mRxST | accuracy ~ hintSim * hintRel + target | **−71** | **−36** |

*Note*: We compare a set of linear regression models using BIC (lowest is best, shown in bold), with separate analyses for each similarity metric (SWOW vs. word2vec). All models use mean accuracy for each target word for a given hint type as the outcome variable. They vary in the presence of predictors for average hint-target similarity (hintSim), hint interrelatedness (hintRel), and target condition (target). Both best-fitting models contain target word condition as a predictor in addition to hint-target similarity and hint relatedness.

For the Random vocabulary hints, relative to the Equal target condition, only the InStr target condition was significantly different, $b = 0.24$, 95% CI $[0.16, 0.31]$, $p < .001$, having higher accuracy. The AoA, $b = -0.07$, 95% CI $[-0.15, 0.01]$, $p = .085$, and WF, $b = -0.004$, 95% CI $[-0.09, 0.08]$, $p = .930$, target conditions were not significantly different to the Equal target condition. Concreteness remained a significant predictor of accuracy, $b = 0.06$, 95% CI $[0.02, 0.10]$, $p = .001$.

For the Full vocabulary hints, relative to the Equal target condition, InStr target words had significantly higher accuracy, $b = 0.14$, 95% CI $[0.03, 0.25]$, $p = .011$, while WF target words had significantly lower accuracy, $b = -0.15$, 95% CI $[-0.27, -0.02]$, $p = .019$. The AoA target condition was not significantly different, $b = -0.10$, 95% CI $[-0.21, 0.01]$, $p = .086$, and concreteness remained a nonsignificant predictor of accuracy, $b = 0.01$, 95% CI $[-0.04, 0.06]$, $p = .695$.

**Hint similarity and hint interrelatedness**. In Experiment 1, we performed two analyses to explore the relationship between target word accuracy, hint similarity, and hint interrelatedness. The first focused on the question of the relative contribution of hint similarity and hint interrelatedness and involved mixed-effects models with target word as a random effect. We performed the same analyses here with the same qualitative results, but defer detailed explanation of these findings to Supplementary Appendix B for space reasons.

Instead, our primary focus here is on the second question: whether the differences in target word accuracy remain even after accounting for hint-target similarity and hint relatedness. As before, we explored this by evaluating a set of linear models with target condition as a fixed effect. Do models that contain target word condition provide additional predictive power over models which do not?

As Table 12 indicates, the best-fit models for both similarity metrics contained target word condition as well as both hint predictors. The parameters for hint-target similarity and hint interrelatedness are similar in magnitude and direction as those reported in Supplementary Appendix B, but we are most interested in the target word condition parameters. For SWOW

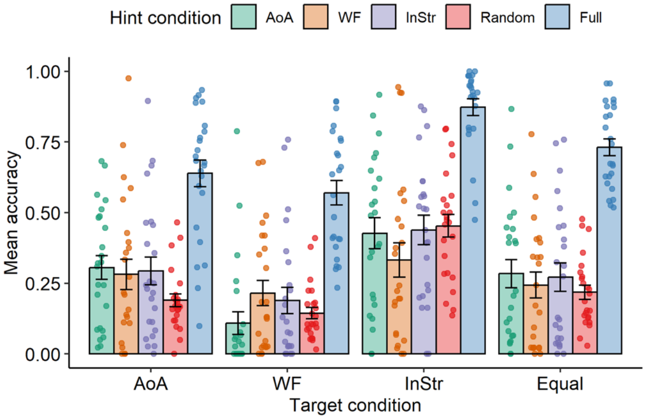


Fig. 11. Experiments 1 and 2: Accuracy for target words by hint condition and target condition. Each dot represents the mean performance for a target word in a given condition, averaged across all trials for that word. The FULL vocabulary hint condition has the best performance, while all other hint conditions are comparable. The INSTR target words yield consistently better accuracy compared to the other target word conditions.

trials, relative to the reference category of EQUAL, the AOA target words had significantly lower accuracy, $b = -0.17$, $t = -4.54$, $p < .001$, as did the WF target words, $b = -0.12$, $t = -3.28$, $p = .001$. The same was true for word2vec trials: the AOA target words had significantly lower accuracy, $b = -0.10$, $t = -2.52$, $p = .013$, and so did the WF target words, $b = -0.18$, $t = -4.29$, $p < .001$. The INSTR target words did not differ significantly from the reference category for SWOW trials, but were slightly more accurate for the word2vec ones, $b = 0.10$, $t = 2.27$, $p = .025$.

Overall, this result is consistent with what we saw in Experiment 1: even after accounting for hint-target similarity and hint interrelatedness, differences remained in terms of how easy different types of target words were to guess. In the next section, we integrate the results across Experiments 1 and 2 so we can compare the effects of core-word and noncore words hints more directly.

### 3.3.3. *Combined analyses across Experiments 1 and 2*

Fig. 11 shows target word mean accuracy for all five hint conditions across both experiments (collapsed across similarity metrics). As in both experiments individually, there was a significant main effect of target condition, $F(3, 92) = 11.22$, $p < .001$, with post-hoc tests showing significant differences between all conditions except AOA and EQUAL (all *p*s $< .05$). There was also a significant main effect of hint condition, $F(3.25, 298.66) = 131.41$, $p < .001$, with post-hoc tests showing that the FULL vocabulary hints had the highest accuracy (all *p*s $<.001$), while all other hints were comparable.

In addition, our combined analysis also showed a significant interaction between hint condition and target condition, $F(9.74, 298.66) = 2.05$, $p = .029$. This suggests that the *magnitude* of the target condition differences varied depending on the hint condition. Indeed, the effect sizes were larger for RANDOM and FULL vocabulary hints ($\eta^2 = .44$ and $\eta^2 = .28$,

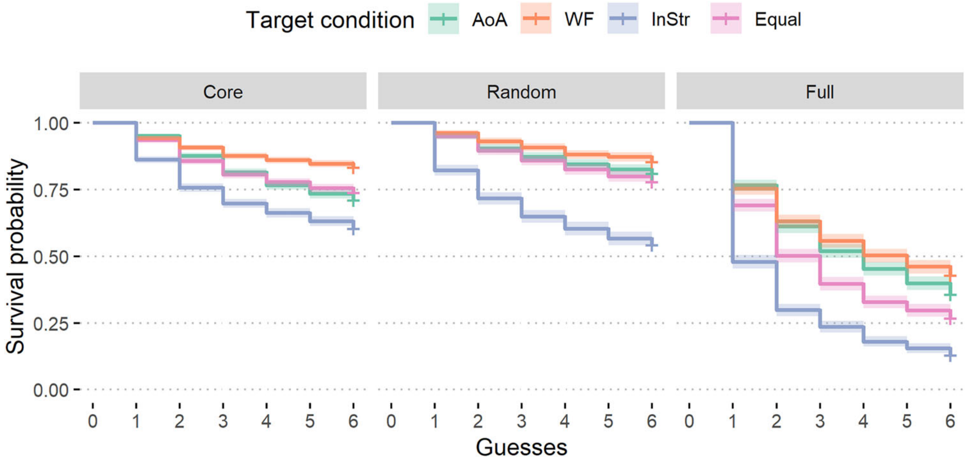


Fig. 12. Experiments 1 and 2: Kaplan–Meier survival curves by hint condition and target condition. Each line shows the survival probability over successive guesses. Events were defined as successful guesses, so lower survival indicates better performance. Panels represent the hint conditions (where CORE collapses across AOA, WF, and INSTR). Performance is better for the FULL vocabulary hints and the INSTR target words.

respectively) compared to the other hint conditions (AOA $\eta^2 = .20$, INSTR $\eta^2 = .12$, and WF $\eta^2 = .03$).

We can explore whether there is a difference between the core word and non-core-word hint conditions more fully using Kaplan–Meier survival analysis. Since performance for the three core word hint conditions (AOA, WF, and INSTR) was so similar, we combined them together into one CORE vocabulary condition, which we then compared to the RANDOM and FULL noncore vocabulary hints. The results, shown in Fig. 12, demonstrate that the target words in the FULL vocabulary hint condition were guessed much more quickly than any of the other hints. These results also suggest, consistent with our previous analysis, that differences between INSTR target words and the other target words are larger with the FULL and RANDOM vocabulary hints than with the CORE vocabulary hints. Indeed, a log-rank test showed that the survival curves differed significantly from each other ($p < .001$).

Table 13 reports the output of a Cox proportional-hazards model demonstrating that FULL vocabulary hints led to much better guessing than the CORE vocabulary hints (the reference category). Moreover, the RANDOM vocabulary hints led to slightly worse guessing, with a small effect on survival. Consistent with our findings elsewhere, INSTR target words were guessed significantly better than EQUAL words (the reference category), and AOA and WF targets were guessed significantly worse.

Overall, across both experiments, the best performance was obtained for INSTR target words, followed by AOA, and lastly WF. The differences between target words were larger for the noncore word hint conditions in Experiment 2 compared to the core-word hint conditions in Experiment 1. Finally, FULL vocabulary hints were by far the best, with core-word hints being comparable to or slightly better than RANDOM ones.

Table 13
Cox proportional-hazards model predicting target word survival from hint condition and target condition

| Term | HR | 95% CI | *p* |
|---|---|---|---|
| RANDOM hint | 0.91 | [0.85, 0.97] | .002 |
| FULL hint | 4.06 | [3.87, 4.25] | <.001 |
| AOA target | 0.88 | [0.83, 0.94] | <.001 |
| WF target | 0.64 | [0.60, 0.68] | <.001 |
| INSTR target | 1.79 | [1.70, 1.90] | <.001 |

*Note*: A hazard ratio (HR) below 1 indicates higher survival, which here means worse guessing. The reference category for word hint condition is CORE (collapsed across the AOA, WF, and INSTR hint conditions) and for target word condition is EQUAL. Performance is best for the FULL vocabulary hint condition and the INSTR target condition.

### 3.4. Discussion

Experiment 2 introduced hint words chosen either at random or from the full vocabulary set, rather than being restricted to just our core word types. Performance with the FULL vocabulary hints was much higher than for either the RANDOM, non-core-word hints *or* any of the core-word hints from Experiment 1. This indicates that our password task is doable as long as the hints and targets are well-aligned.

The pattern of differences between the target conditions we observed in Experiment 2 was fairly consistent with what we saw in Experiment 1: in both, people found it easiest to guess the INSTR target words. This time, however, there was a smaller difference in performance between the AOA and WF target words, and accuracy was lower on both compared to the baseline EQUAL target word condition. Overall, the advantage for INSTR target words appeared to be greater in Experiment 2 than in Experiment 1.

## 4. General discussion

Across two experiments, we compared three different approaches to core vocabulary in terms of how well they accounted for human performance in a word guessing game. The results were highly consistent across a range of analyses, revealing that performance varied based on the type of core words that served as the target words but not the type of core words that made up the hints. Hints from the full vocabulary (drawing from all possible words, core or noncore) were the most effective by far, whereas the different types of core-word hints and randomly sampled sets of noncore word hints were all comparable in performance. In contrast to hints, there were substantial differences between core-word targets, with INSTR target words being the easiest to guess across the board, followed by AOA targets, and lastly WF targets.

While there were no differences between the core words in terms of effectiveness as hints, there were clear differences between the types of target words. We suggest that this might reveal something about the search process through semantic memory: the points from which the search process begins may be less important than how readily available the target words are and how easily they emerge along the search process. This is consistent with Xie, Bainbridge, Inati, Baker, and Zaghloul (2020), who suggest that memory search is guided by highly memorable words, which serve as key locations from which the search begins, occupying central and prominent positions in the semantic representation. This leads them to be better retrieved even when paired with arbitrarily related cues. Our results suggest that it is the words that are core in InStr (and to a lesser extent, early acquired in life), which are highly connected in word association networks, that occupy these central and prominent positions, and are, therefore, good candidates for core words in psycholinguistic terms, at least as measured by our word-guessing task.

One implication of these results is that although frequency is a widely used and theoretically valuable measure in the area of core vocabulary (Bell, 2012), it may be useful to consider other factors as well: word associations (and, to a lesser extent, age of acquisition) may be more important to understanding lexical coreness than frequency alone. This is consistent with earlier arguments that frequency is perhaps better thought of as a consequence of coreness rather than a way to define it (Stubbs, 1986).

By contrast, our results for hint words underscore that important aspects of lexical meaning or access are probably *not* sufficiently embodied by a restricted set of core words. The lack of difference between the core word types as hints, combined with their ineffectiveness compared to full-vocabulary hints, suggests that *none* of the types of core words fully incorporates the factors that make something a good hint. Perhaps not surprisingly, it turns out that 300 binary semantic features are nowhere near sufficient to span the richness of semantic space for an entire lexicon. Consistent with this, words proved to be most useful as hints when they were closely related to the meaning of the target word, which is why hints drawn from the full vocabulary were often more useful. In fact, for the same reasons that central target words are easier to guess because of their prominent and easily accessible positions in the lexicon, central hints may be unhelpful hints because they do not provide much specific or disambiguating information about other words *because* of their centrality. If one is searching semantic space through something like a random walk, it may in fact be detrimental to begin in a central node because there are so many directions out of it.

Of course, even though we compared a discrete set of the 300 words that were "most" core to those that were not, it is important to remember that coreness itself is a spectrum, not a binary. It is possible that a larger set of core words would show a different outcome. It is also possible that the artificial nature of the task may not be fully suited to assessing the contribution of core words to word meanings. It involved the relatively simplistic presentation of discrete lists of hint words that were intended to cue target words, with otherwise no specified relation to either the target word or to each other. This is quite different from the comparably richer ways in which the meanings of core words are leveraged and combined in explanations that rely on semantic primitives (Wierzbicka, 1996), where they are used to construct long and comprehensive definitions. More complex tasks like the semantic navigation task (Beckage, Steyvers, & Butts, 2012) or the mini-dictionary game proposed by Vincent-Lamarre et al. (2016) would be a useful way of studying other aspects of core word semantics.

Given that the task involved the simple presentation of hints selected by semantic similarity, one may wonder whether the advantage for InStr arose because of this, especially since the

SWOW similarity metric and the InStr measure were computed on the same data. However, there are several strong reasons that the selection of hints using similarity did not privilege InStr, the first of which being that the InStr targets do not have the highest hint-target similarity to begin with—in fact, the AoA target words have the highest similarity for the SWOW hints, and the WF target words have the highest similarity for word2vec hints (see Supplementary Appendix C). This means that the selection of hints from the core words using SWOW and word2vec similarity did not lead to hints that were more similar to InStr targets. Despite this, our results showed that InStr targets were guessed more accurately overall, and even remained more accurate after taking semantic similarity into account, meaning that regardless of whether InStr targets had an advantage due to semantic similarity, they were still guessed more accurately for reasons that go beyond this. Furthermore, association, on which InStr is based, and similarity are psychologically different constructs—thus, they are only weakly related, even though both are computed on the SWOW data (see Supplementary Appendix C). This was also the reason for including a text-based semantic similarity metric, namely, word2vec, which is distinct from data collected from humans; the advantage for InStr was consistently found for both similarity metrics.

More broadly, the InStr core words may have had an advantage because of the nature of the task we used; a word-guessing game is similar in many ways to the word-association task from which they were derived. Moreover, the task implicitly assumes that word meaning is a function of relationships between individual words. It is, therefore, useful for investigating how people conceptualize words in isolation, but the results might be different in tasks that incorporate more of the communicative and contextual elements of language. We are testing these possibilities in additional experiments. Our preliminary findings suggest that when the task is to guess target words in a context-rich cloze prediction task, the advantage to InStr words remains (Wang, De Deyne, McKague, & Perfors, 2024).

Some potential implications of our work are relevant to an ongoing debate about whether human notions of meaning are derived primarily from the distribution of language in the environment (e.g., Firth, 1957; Günther, Rinaldi, & Marelli, 2019; Kumar, Steyvers, & Balota, 2022; Landauer & Dumais, 1997). Since word frequency is one of the most important distributional language measures, our results finding that people performed better with InStr targets might suggest that distributional language models fall short of capturing important aspects of human semantic representation. This is in line with previous research showing that word-association-based models provide a better account than distributional language-based models in predicting semantic properties (Vankrunkelsven, Verheyen, Storms, & De Deyne, 2018), human similarity judgments (De Deyne, Perfors, & Navarro, 2016), and incorporating visual and affective conceptual information (De Deyne, Navarro, Collell, & Perfors, 2021). This finding is also supported by our results for the similarity type that was used to select hints for targets: although not one of the main questions of interest, we found that hints that were selected using SWOW similarity (instantiating the word-association-based model) consistently led to better performance than hints selected using word2vec similarity (distributional language-based model), which replicates the finding from Vankrunkelsven et al. (2021).

That said, such a conclusion would be premature. For one thing, word frequency is only one possible measure that one might extract from distributional information, and it is perhaps the simplest possible measure, only considering unigram counts. It is more common to think of distributional language models as reflecting co-occurrence information between words, or counts transformed into salience weights using algorithms like positive pointwise-mutual information (PPMI) (Bullinaria & Levy, 2007). Many distributional semantic models (DSMs) also make use of neural networks trained to predict words from context to learn vector representations for words; these include models like word2vec (Mikolov, Chen, Corrado, & Dean, 2013) and GloVe (Pennington, Socher, & Manning, 2014), as well as Transformer-based models like GPT (Brown et al., 2020) and BERT (Devlin, Chang, Lee, & Toutanova, 2018). Might our results be different if we used a richer distributionally based measure of coreness than word frequency?

While evaluating the effectiveness of distributional language models *in general* is far beyond the scope of this paper, we have performed exploratory analyses of this question (see Supplementary Appendix D). In it, we recalculated the coreness of our target words using a variety of common distributional measures *other* than word frequency; these include co-occurrence counts, PPMI centrality, distance in vector space, and cosine similarity, among others. Our qualitative findings remain the same: higher accuracy is significantly predicted by higher InStr coreness, but not by any of the DSM-based coreness measures we tested. It is an open question whether the same result would occur with a wider selection of distributional language measures or in a task that was specifically designed to compare them.

## Notes

[1] Thirteen data files were lost due to a server issue.

[2] https://aspredicted.org/9pwc-xpqs.pdf

[3] The words that were excluded were: *oink, granddad, suppertime, firetruck, belly-button, popsicle, highchair, dinnertime, schoolteacher, sandbox, Christmas tree*, and *shoelace*.

[4] For interested readers, the parameters for the interaction are reported in Supplementary Appendix A.

[5] AoA: $b = -0.02$, 95% CI $[-0.13, 0.10]$, $p = .743$; WF: $b = -0.02$, 95% CI $[-0.14, 0.11]$, $p = .790$. Concreteness as well still significantly predicted accuracy, $b = 0.07$, 95% CI $[0.01, 0.12]$, $p = .015$.

[6] Random condition: https://aspredicted.org/qhsp-pccx.pdf
Full condition: https://aspredicted.org/c4nr-9cpy.pdf

## References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823.

Anthony, L. (2022). AntBNC lemma list. Available online at. https://www.laurenceanthony.net/software/antconc/.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*(2), 283.

Beckage, N. M., & Colunga, E. (2016). Language networks as models of cognition: Understanding cognition through language. In A. Mehler, A. Lücking, S. Banisch, P. Blanchard, & B. Job (Eds.), *Towards a theoretical framework for analyzing complex linguistic networks*, Understanding complex systems (pp. 3–28). Springer.

Beckage, N. M., Steyvers, M., & Butts, C. T. (2012). Route choice in individuals–semantic network navigation. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 108–113).

Bell, H. (2012). Core vocabulary. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (1st ed.) (pp. 1132–1136). Wiley.

Borin, L. (2012). Core vocabulary: A useful but mystical concept in some kinds of linguistics. In D. Santos, K. Lindén, & W. Ng'ang'a (Eds.), *Shall we play the Festschrift game?: Essays on the occasion of Lauri Carlson's 60th birthday* (pp. 53–65). Springer.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *arXiv:2005.14165*.

Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods*, *49*, 1520–1523.

Brysbaert, M., & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *Quarterly Journal of Experimental Psychology*, *64*(3), 545–559.

Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition*, *13*(7–8), 992–1011.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50.

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467–479.

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., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 441.

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

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*, 510–526.

Calude, A. S., & Pagel, M. (2014). Frequency of use and basic vocabulary. In L. Filipović & M. Putz (Eds.), *Multilingual cognition and language use: Processing and typological perspectives* (pp. 45–72). John Benjamins Publishing Company.

Carter, R. (2012). *Vocabulary: Applied linguistic perspectives*. Routledge.

Davies, M. (2008). The Corpus of Contemporary American English (COCA). Available online at. https://www.english-corpora.org/coca/.

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), e12922.

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., 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).

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, *40*(1), 213–231.

De Deyne, S., & Storms, G. (2015). Word associations. In J. R. Taylor (Ed.), *The Oxford handbook of the word* (pp. 465–480). Oxford University Press.

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*(85), 5351.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*.

Duñabeitia, J. A., Avilés, A., & Carreiras, M. (2008). NoA's ark: Influence of the number of associates in visual word recognition. *Psychonomic Bulletin & Review*, *15*(6), 1072–1077.

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1103.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Firth, J. (1957). A synopsis of linguistic theory 1930–1955. *Studies in Linguistic Analysis, Special Volume/Blackwell*.

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). SimVerb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychological Science*, *18*(12), 1069–1076.

Heath, D., Norton, D., Ringger, E., & Ventura, D. (2013). Semantic models as a combination of free association norms and corpus-based correlations. In *2013 IEEE 7th International Conference on Semantic Computing* (pp. 48–55).

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Hsu, C.-C., & Hsieh, S.-K. (2013). Back to the basic: Exploring base concepts from the Wordnet glosses. *Computational Linguistics and Chinese Language Processing*, *18*(2), 57–84.

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).

Kumar, A. A., Steyvers, M., & Balota, D. A. (2022). A critical review of network-based and distributional approaches to semantic memory structure and processes. *Topics in Cognitive Science*, *14*(1), 54–77.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Lee, D. Y. W. (2001). Defining core vocabulary and tracking its distribution across spoken and written genres: Evidence of a gradience of variation from the British National Corpus. *Journal of English Linguistics*, *29*(3), 250–278.

Liu, Q., De Deyne, S., Jiang, X., & Lupyan, G. (2023). Understanding the frequency of a word by its associates: A network perspective. In *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*, volume 45.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

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*.

Monaghan, P., Chang, Y. N., Welbourne, S., & Brysbaert, M. (2017). Exploring the relations between word frequency, language exposure, and bilingualism in a computational model of reading. *Journal of Memory and Language*, *93*, 1–21.

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

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82.

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

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130.

Poirier, M., & Saint-Aubin, J. (1996). Immediate serial recall, word frequency, item identity and item position. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *50*(4), 408.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

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

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.

Stubbs, M. (1986). Language development, lexical competence and nuclear vocabulary. In M. Stubbs (Ed.), *Educational linguistics* (pp. 57–76). Blackwell.

Vankrunkelsven, H., Vankelecom, L., Storms, G., De Deyne, S., & Voorspoels, W. (2021). Guessing words. In G. Kristiansen, K. Franco, S. De Pascale, L. Rosseel, & W. Zhang (Eds.), *Cognitive sociolinguistics revisited* (pp. 572–583). De Gruyter.

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.

Wang, A., De Deyne, S., McKague, M., & Perfors, A. (2024). Word prediction is more than just predictability: An investigation of core vocabulary. In *Proceedings of the 46th Annual Conference of the Cognitive Science Society* (pp. 4106–4112).

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

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

Xie, W., Bainbridge, W. A., Inati, S. K., Baker, C. I., & Zaghloul, K. A. (2020). Memorability of words in arbitrary verbal associations modulates memory retrieval in the anterior temporal lobe. *Nature Human Behaviour*, *4*(9), 937–948.

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

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, *60*(4), 502–529.

## Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information