Word Prediction in Context: An Empirical Investigation of Core Vocabulary

Anonymous CogSci submission

Abstract

Core vocabulary is a topic of huge interest in linguistics and has been studied from a wide variety of perspectives, such as language learning, dictionary studies, and cross-linguistically. In many of these conceptions, word frequency is widely considered the conventional measure of a word's coreness; however, this approach overlooks important aspects of mental representation like centrality in an associative semantic network. In this experiment, we compare different approaches to defining core words in a task that involves predicting missing words in sentences. Results showed that core words (regardless of definition) were easier to guess than non-core words, but that frequency-defined ones did not perform as well as expected given their higher predictability and the nature of the task. Analysis of incorrect responses also showed that people preferred to guess core words, simple synonyms, and words that are taxonomically related to the target. The findings suggest that how core vocabulary is defined depends in part on the nature of the task and that aspects of both mental representation and the linguistic environment play an important role.

Keywords: core vocabulary; word frequency; word associations; semantic representation; language models; distributional semantics; age of acquisition

Introduction

Language has many and varied uses, ranging from classic literature and news reporting to blogs and social media, covering the myriad concepts that we encounter in our everyday lives. However, among the extensive collection of words that make up the vocabulary of a language, the idea that some words are in some way more central or important feels very natural. What constitutes a *core vocabulary* is a long-standing question in linguistics. It can illuminate the basic concepts of the mental landscape (e.g. Hsu & Hsieh, 2013), and has practical applications in both language learning and teaching.

Various ways of defining core vocabulary have been explored in the literature. Some of the more straightforward approaches involve handpicked lists of vocabulary items for pedagogical purposes (Carter, 1987; Ogden, 1930; West, 1953). Others have investigated the dependencies of words in dictionaries to identify a set of words that are sufficient to define all other words (Vincent-Lamarre et al., 2016).

In many existing conceptions of core vocabulary, word frequency is assumed to be a good measure of the coreness of words. However, words might play an important role in ways that go beyond how frequent they are in a particular language. For instance, taking a cross-linguistic perspective to define core vocabulary – that is, examining the words that are contained across the world's languages – leads to differentiation from frequency-based methods (Wu, Nicolai, & Yarowsky, 2020). And in historical linguistics, resistance to borrowing is taken to measure a word's coreness, which is useful for considering questions of language typology (Borin, 2012; Zenner, Speelman, & Geeraerts, 2014). As Stubbs (1986) argues, frequency is best thought of as a consequence of coreness, rather than a way to define it.

In this study, we investigate the question of core vocabulary from a psychological perspective – that is, taking into account the ways that words are represented in the mental lexicon and accessed during language use. From this perspective, core vocabulary concerns the words that are central to people's mental representations, according to different theories of word meaning. This has the advantage of grounding the study of core vocabulary in well-established theory and incorporates mental representation into the way it is defined.

From this perspective, word frequency is implicated in a class of models, called *Distributional Semantic Models* (DSMs), which situate word meaning in the linguistic environment. According to these models, context plays a key role in determining meaning: how we use words in communication, and the words we tend to use them with, make up a big part of what they mean. Words that are more frequent occur in more contexts, co-occur with a greater number of words, and thus contribute to the meanings of more words.

Another class of models is based on data from subjective methods that measure mental representations more directly. These methods include feature generation, word associations, and other verbal fluency tasks. The assumption is that words that are produced more often are the ones that are more mentally salient. This content can then be represented in a format such as a semantic network. Here, we measure core words in mental representations of word meaning using word association data, by deriving a graph-theoretic measure of centrality called INSTRENGTH (a weighted version of in-degree) from networks based on word associations.

We can also conceptualise core words by considering the developmental *process* by which the network is built up. For instance, the preferential attachment hypothesis suggests that semantic networks are built up by attaching new words to existing ones (Brysbaert, Van Wijnendaele, & De Deyne, 2000; Steyvers & Tenenbaum, 2005), suggesting that the core words are the ones that have a lower age of acquisition (AOA).

These contrasting approaches to semantic representation differ fundamentally in terms of the content on which they are based. Linguistic representations of word meaning, because they are based on information from the linguistic environment, are governed by factors that shape language *as used for communication*. This means that such representa-

tions are fundamentally contextual and governed by principles of pragmatics, such as considerations of conciseness and informativeness along with assumptions about what speakers already know. On the other hand, mental representations of word meaning, such as data based on word associations, are more context-independent, and are not clearly governed by communicative factors. They are also more likely to tap different sources besides linguistic information, especially experiential information such as sensory information and affect. Both factors are, of course, heavily intertwined.

In short, although core vocabulary is typically defined based on linguistic information – in particular, frequency – this may come at the cost of overlooking other important aspects of mental representation. Are definitions of core vocabulary that more directly tap our representation useful when compared to standard frequency-based definitions?

Previous work by Wang, De Deyne, McKague, and Perfors (2022) used a word-guessing game consisting of hint words and target words to investigate this question. They explored which type of core words provided the most effective hints and which were the most easily guessed targets. They found no differences with regard to the hint words, but INSTRENGTH target words were the easiest to guess. This suggests that these core words occupy a more central position in the mental lexicon and are more representationally accessible. However, their task bears similarity to the word-association method through which INSTRENGTH is measured, so it may not be completely clear whether the superior performance of INSTRENGTH core words is due to their status in the mental lexicon or because of task overlap between generating word associations and guessing them.

The current study examines whether the superior performance of INSTRENGTH target words can also be found when the task is more naturalistic and better captures how language is used for communication. We do this using a cloze-style word prediction task in which people guess missing words from sets of sentences. This task closely matches the learning objective of modern DSMs including neural network models like word2vec (Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018) and transformer-based models like BERT (Devlin, Chang, Lee, & Toutanova, 2018). Because the use of distributional information to predict words in context is a central feature of our task, it should therefore provide a fair test to evaluate the word frequency view of coreness.

Method

Participants

200 participants (20-72 years, $M = 37.6$; 34% female) were recruited from Amazon Mechanical Turk and paid $4 for the 20 minute task; of these, 199 completed the experiment. The pre-registered[1] catch trials (described below) were passed by 195 participants who were included in the analyses. 85% reported being native English speakers, and all had previously passed a qualification assessing English proficiency.

[1] https://aspredicted.org/FS2_DTR

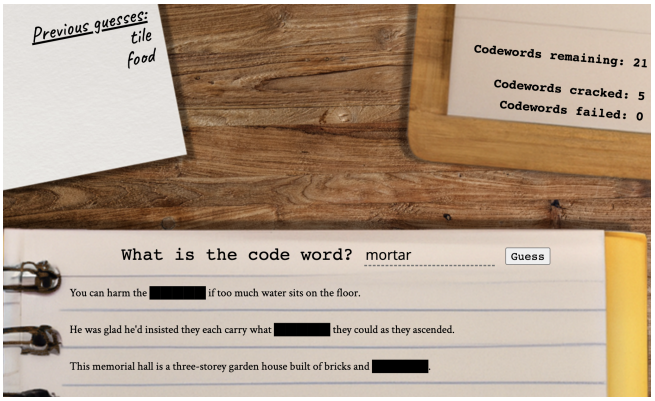

Figure 1: **Screenshot from example trial.** On each trial, participants were shown a series of sentences one after the other. Each sentence contained a black box in the location of a target word they had to guess. In this trial, the target word was *wood* and the person is guessing *mortar* on their third attempt.

Procedure

Participants completed the task online after giving consent, providing optional demographic information, reading the instructions, and answering check questions about them.[2]

The task was set up as a game in which people were told that they were cracking coded messages sent between spies. The messages were ordinary sentences taken from sources such as websites, books, and newspapers containing a blank in place of a missing "code word" (see Figure 1). On each trial, people saw up to six sentences with the same target word missing. Each time they were shown a new sentence, they had another chance to guess the missing word; this continued until they were either successful or all six attempts had been exhausted, at which point they were told the correct answer.

In order to make the task sufficiently challenging, sentences were presented from lowest to highest predictability (participants were not told this). This was accomplished by dividing the 36 sentences for each target word into six bins by predictability (calculated as described below) and randomly selecting one sentence from each bin. As a result, each participant saw a different random combination of sentences for any given target word but the difficulty of the task was approximately similar for everyone.

After a practice trial, each participant completed 24 experimental trials plus two unanalysed catch trials that were designed to be substantially easier than the experimental ones: their target words were *age* (trial 10) and *head* (trial 20). As pre-registered, we excluded the participants who failed to guess either of the words on the catch trials ($N = 4$). Target condition was manipulated within-subject, with each person seeing a random selection of six target words from each of the four conditions described below (AOA, WF, INSTRENGTH, and NONCORE). Except for the catch trials, the targets, sentences, and order were randomised for each person.

[2] Extra info about stimuli and analyses are in Supplemental Materials: https://figshare.com/s/ba5ab024d990f74388c3

Materials

Target Words. The words to be guessed came from four target conditions, each of which contained 24 words. In the INSTRENGTH, WF, and AOA conditions, the target words were selected from the corresponding core word list, as explained more fully in Wang et al. (2022) and detailed below. The NONCORE condition was designed as a comparison to these and contained words that were not on any of the three core word lists, as we describe later.

Core words were defined as the 300 most core words as defined by each of the condition-specific measures. The WF measure, corresponding to the DSM approach, is based on the SUBTLEX database (Brysbaert & New, 2009), with more frequent words being more core. The INSTRENGTH measure, reflecting mental representations of meaning, is based on word associations to over 12,000 English words (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019). It reflects the sum of the weights of all incoming edges directed towards the target word, where edge weights represent associative strengths. Common associates have higher INSTRENGTH and are more core. The AOA measure is sourced from the Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) norms, with earlier-acquired words being more core.

In line with previous work, the INSTRENGTH and WF measures were log-10 transformed and then all words were normalised by grouping together inflectional forms of the same lemma (e.g., *run*, *runs*, *running*). Function words like determiners, auxiliary verbs, and prepositions were removed. To compare the coreness of words across the three lists, we normalized each of the coreness measures by computing the difference between each word and the most core word, and scaling that proportional to the difference between the first and last (i.e., 300th) word on the list. This results in an inverted "coreness" metric where the most core word on each list has a value of 0, the last word on each core word list has a value of 1, and less core words have values beyond 1.

The AOA, WF, and INSTRENGTH target words were selected to be words that were core on their respective lists as well as less (but equally) core on the other two measures.[3] These target conditions allow us to investigate whether the words that are core under different theories of meaning (WF for DSMs, AOA for preferential attachment, INSTRENGTH for word associations) are easier to predict in sentence contexts. One word in the AOA condition was different from Wang et al. (2022) (*reindeer* instead of *crayon*) because *crayon* does not exist in the BERT vocabulary (see below).

The NONCORE target words were not on any of the three core word lists. To ensure that these would still be familiar words, we selected only items that prevalence data suggests are known by nearly everyone, i.e., of maximum prevalence (Brysbaert, Mandera, & Keuleers, 2018). The target words we chose had similar coreness on all three measures.[4] This condition provides a baseline for comparison and allows us to ask whether all core words (regardless of definition) are easier to guess or more accessible than non-core words.

Table 1: Target words in each condition.

AOA	WF	INSTRENGTH	NONCORE
rice	ready	anger	atlas
doll	hope	music	arise
bite	send	pain	noticeable
plate	use	paper	vocabulary
tail	know	religion	bloom
grandma	thing	round	quiz
pillow	stuff	sea	frequent
arm	trouble	sick	hive
reindeer	go	beach	maze
brush	take	snake	monopoly
bathroom	find	strong	gigantic
boot	spend	boring	fictional
snack	marry	tool	cube
butt	keep	warm	refusal
hungry	follow	white	substitute
hug	way	wood	tablet
door	pick	book	unwilling
breakfast	call	car	wrestler
neck	room	clean	evergreen
hill	look	dirty	backbone
kitchen	die	drink	tighten
bottle	make	fat	floppy
towel	remember	horse	shallow
cookie	wait	light	athletics

Measure of Predictability. Words vary in how predictable they are given the sentence contexts they appear in: the word *music* is highly predictable in the sentence "the album received mixed reviews from music critics" and much less so in "this music is not that great". The probability of a word given a sentence can be calculated using BERT (Devlin et al., 2018). BERT is a transformer-based neural language model that is trained with a masked language modelling objective: words are masked in the sentence input, and the model predicts the masked words based on both the left and right context. The model also calculates confidence scores for the words it predicts. The confidence score for a particular word corresponds to its predictability in that sentence context. For example, the predictability for the word *music* above is calculated by BERT to be .99 for the first sentence and .0009 for the second. We used the base, uncased version of BERT (110m parameters). The model was trained on BookCorpus (Zhu et al., 2015), a dataset consisting of 11,038 unpublished books, and English Wikipedia. The model was accessed through the *Transformers* Python package (Wolf et al., 2020).

[3] The mean coreness of the selected words on their own lists is: AOA 0.73, WF 0.59, INSTRENGTH 0.59. The mean coreness of words on the other lists is: AOA 1.43, WF 1.29, INSTRENGTH 1.2. This means, for instance, that WF target words were more core on the WF list, but less core on the INSTRENGTH and AOA measures.

[4] The mean coreness of the NONCORE words on each measure is: AOA 2.33, WF 2.21, INSTRENGTH 2.28.

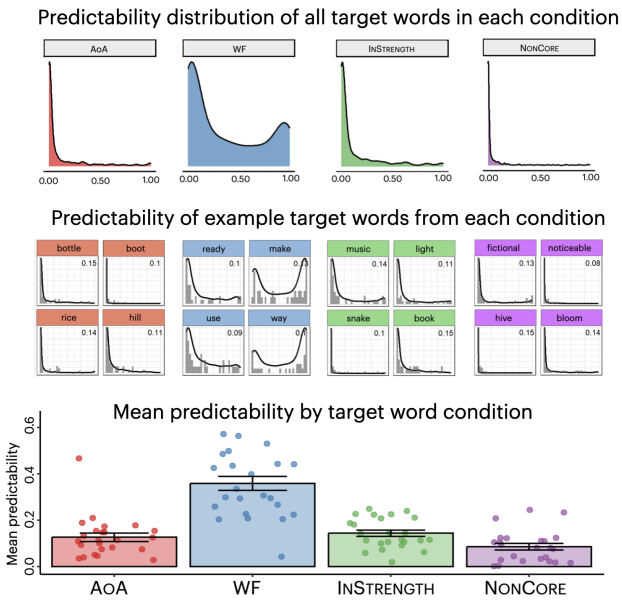

Figure 2: **Predictability distributions of target words.** *Top panel*. Distribution of predictability (x axis) of all words in each of the four conditions. The majority of InStrength, AoA, and NonCore words had very low predictability, but for WF words a considerable proportion was highly predictable. *Middle panel*. These differences in predictability are visible when considering representative example words from each condition; black lines show the distribution of that word as estimated based on 10,000 samples, while the bars indicate the predictability of the 36 sentences in the experiment. The number in the upper left is the K-S statistic capturing the difference between the distributions, which was constrained to be 0.15 or less. *Bottom panel*. Reflecting their real-world distributions, conditions varied significantly in predictability (each dot is the mean predictability of one word based on the 36 sampled sentences).

Sentences. It is important that the sentences used in this task are representative of the sentences that each word actually occurs in. To achieve this, we sourced sentences with the target words from the enTenTen web corpus (Jakubíček, Kilgarriff, Kovář, Rychlỳ, & Suchomel, 2013), which contains 36B words taken from the Internet between 2019 and 2021 with content from Wikipedia, news sites, blogs, books, and forums. It is thus a reasonable approximation to the sort of language many adult English speakers are exposed to.

We constructed a predictability distribution for each target word by sampling 10,000 sentences containing that word using Sketch Engine (Kilgarriff et al., 2014). Figure 2 shows the predictability distribution over all of the words in each condition (top panel), as well as some specific examples (middle panel). Words in the WF condition are more predictable than words from the other conditions (bottom panel).

The 36 sentences corresponding to each target word were selected from the original sample of 10,000 so as to match its predictability distribution as closely as possible. We did this by repeatedly sampling sentences until the Kolmogorov–Smirnov statistic comparing the original and sample distributions was less than .15. Sentences were also manually filtered to remove jargon or duplicate instances of the target, sensitive content, and ungrammaticality.

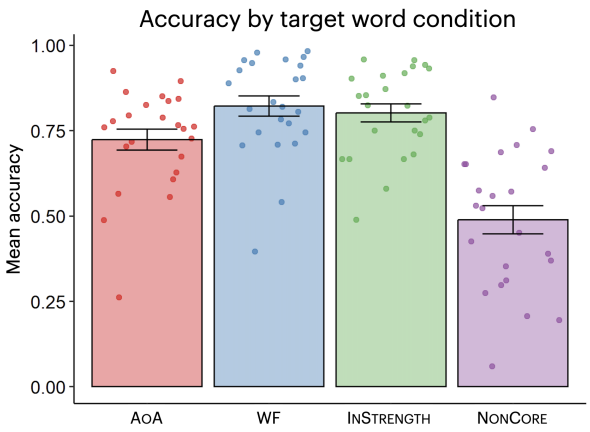

Figure 3: **Accuracy by target word condition.** Each dot represents one target word whose mean accuracy (y axis) is calculated by averaging over all trials and participants. NonCore words were significantly harder to guess. AoA words were much more accurate but still significantly worse than InStrength and WF words.

Results

How well did people do?

The main dependent measures in this task are accuracy (whether the correct word was guessed at all) and the number of guesses (when it was guessed correctly, how many sentences were required). The results were qualitatively consistent for both, but number of guesses is less straightforward to interpret so we consider only accuracy here. Target word accuracy was calculated by averaging over all of the trials it occurred in (see Figure 3). A mixed-effects logistic regression was conducted to predict accuracy on each trial from condition, with `target word` and `participant` included as crossed random factors. Multiple versions of this model were compared using BIC, which penalises unnecessary complexity (see Table 2). The full model with `condition` and both random effects was favoured, suggesting that accuracy varies significantly by condition as well as by participant and target word. Under that model, accuracy for the InStrength condition (the reference category) was significantly higher than for the AoA, $p = .036$, and NonCore, $p < .001$, conditions, but not significantly different from the WF condition. This is somewhat striking given the fact that words in the WF condition were so much more predictable than the others. Why, then, were people equally accurate on InStrength words?

We can investigate this question by conducting a linear regression model on the target word level, where the outcome variable is the mean accuracy of each target word and condition and predictability are fixed effects. As Table 2 reveals, the winning model contained both `predictability` and `condition` as factors but no interaction, meaning predictability has a similar effect in each target condition. This time the InStrength condition (the reference category) had significantly higher accuracy than the WF condition, $p = .016$, as well as the NonCore condition, $p < .001$. Taken together, this suggests that although predictability does make accurate guessing easier, InStrength (and perhaps AoA) words are easier to guess for reasons that go beyond that.

Table 2: **Model comparisons for two analyses**. Models are depicted with statistical notation where `*` indicates an interaction, 1 is a constant, and `(1|x)` indicates that `x` is a random effect. `target` means target word, `subj` means participant, `condition` indicates the four conditions, and `predictability` is the average predictability of that word. Best-fit models have the lowest BIC (bold).

The role of condition: Mixed effects logistic regression		
Model	Description	BIC
M1null	`acc ~ 1`	5642
M1C	`acc ~ condition`	5273
M1RE	`acc ~ (1|target) + (1|subj)`	4238
M1full	`acc ~ condition + (1|target) + (1|subj)`	**4211**

The role of predictability: Linear regression		
M2null	`acc ~ 1`	-24
M2C	`acc ~ condition`	-63
M2P	`acc ~ predictability`	-52
M2CP	`acc ~ condition + predictability`	**-78**
M2CPI	`acc ~ condition * predictability`	-65

What mistakes did people make?

Another way to understand what people are doing in this task is to look at their errors: what answers did they give when they were wrong about the target word? Commonalities across people and systematic differences by condition can be revealing about what words are most accessible and salient, and what drives people's performance in each condition.

Table 3 shows the most common responses as a proportion of the number of trials with the target word. For ease of interpretability, we restricted our analysis to only those responses that were given on at least 25% of target-word trials (a total of 52 target-response pairs).[5] Each response was classified as *more core* if it was more core than its target on all three coreness measures, *less core* if it was not more core on any measure, and *partially more core* otherwise.

As Figure 4 shows, one overarching pattern was that responses tended to be more core than the intended targets, especially in the NONCORE condition. This suggests that people tended to guess simpler and more central words over more complex ones. There were very few common incorrect responses to WF targets, and fewer still to INSTRENGTH ones. Incorrect responses to AOA words were more common and were usually more core on one or both of the other coreness measures.

We also explored the nature of the semantic relationship between targets and responses (see examples in Table 3). Words with the same meaning as the target were classified as *synonyms*, further specified as *basic* if they were more core than the target; words from the same category as the target and that share a hypernym were classified as *taxonomic*; words that were hypernyms or holonyms were called *general*; and everything else was called *other*. Figure 4 shows the frequency of these relation types, normalised by the proportion of the trials on which the response was given (this allows us to compare across targets that by chance occurred different numbers of times). Responses to NONCORE targets tended to be basic synonyms, whereas responses to AOA targets tended to be taxonomic alternatives.

[5]To ensure that our results are not a byproduct of this choice, we repeated the analyses with a 20% threshold as well as only the top response for each target, with qualitatively similar results.

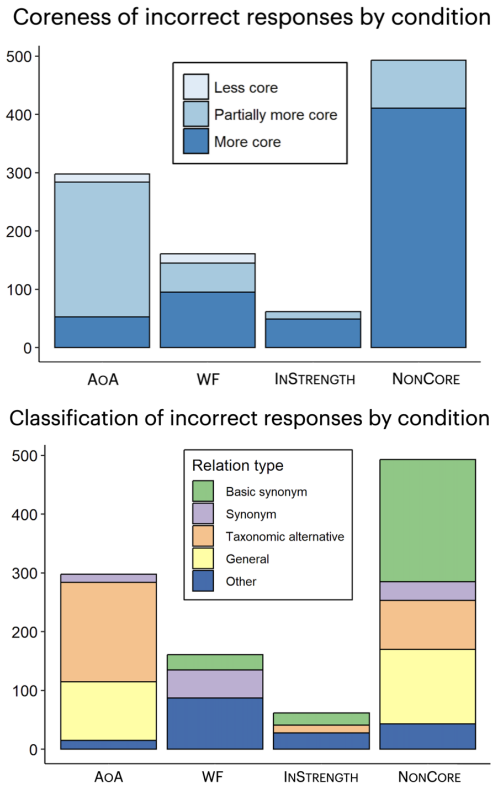

Figure 4: **Analysis of incorrect guesses.** *Top panel.* The most common incorrect responses were classified by whether they were more core than the target on each of the three core word lists. Those classified as *more core* were more core on all three, *less core* on none, and *partially more core* on one or two. It is evident that incorrect responses to NONCORE targets were more core according to most or all of our definitions. Incorrect responses to AOA targets were often more core on the other two measures but not its own. *Bottom panel.* Incorrect responses were classified by the relationship between target and response. For NONCORE targets, the guess was usually a basic synonym, demonstrating the intuition that core words are more easily accessible and central. For AOA targets, the guess was usually a taxonomic alternative, possibly reflecting the fact that many early-acquired words are used less frequently by adults (e.g., adults use *mom* more than they use *grandma*). Counts on the *y*-axis are normalised to reflect better reflect frequency.

Discussion

Given that the word prediction task centrally concerns distributional information about the linguistic environment and closely matches the training objective of DSMs, performance for the WF condition was not as good as expected. Accuracy was on par with the INSTRENGTH condition, and worse when predictability was taken into account. This suggests that performance in the task is not entirely driven by predictability, but that other factors, such as accessibility, also play an important role. This could explain the superior performance for INSTRENGTH words, since those are more central in the semantic representation. This possibility is consistent with the finding that the INSTRENGTH condition had the lowest num-

Table 3: **Most common incorrect guesses.** The response column shows the frequent wrong answers given instead of the correct answer (target), as well as the relationship between the target word and incorrect response. The most frequent incorrect responses occurred in the NONCORE condition. The *Pr* column indicates the proportion of all target word trials with that response; *big* was given on 74% of trials where *gigantic* was the target. Because there are multiple guesses per trial, $\Sigma_g Pr(g|w) > 1$ for all guesses g for target word w.

Target	Response	Relation	Condition	*Pr*
gigantic	big	basic	NONCORE	0.74
atlas	book	general	NONCORE	0.66
vocabulary	language	synonym	NONCORE	0.65
athletics	sports	general	NONCORE	0.63
atlas	map	basic	NONCORE	0.61
gigantic	large	basic	NONCORE	0.58
breakfast	dinner	taxonomic	AOA	0.57
grandma	mother	taxonomic	AOA	0.57
evergreen	green	other	NONCORE	0.51
wrestler	player	general	NONCORE	0.50
neck	head	taxonomic	AOA	0.48
wrestler	athlete	general	NONCORE	0.48
grandma	mom	taxonomic	AOA	0.46

ber of common incorrect responses: few words came to mind more easily than INSTRENGTH target words themselves. Previous work supports the idea that INSTRENGTH-defined core words are more accessible; for example, Wang et al. (2022) found that INSTRENGTH words were more easily guessed in a task that involved guessing target words from hint words.

An unexpected result occurred for AOA target words, which also had a sizeable amount of common incorrect responses, mainly of the taxonomic and general type. AOA words capture the kind of words that are common among young kids, and may therefore have been harder to guess for adults for whom such words are not as frequent or relevant. This suggests that communicative need may also play a role in this task. That is, people have a tendency to guess words reflecting concepts that are more pertinent or salient to them; while a child might be more likely to talk about a *doll* or a *grandma*, adults might instead talk about a *hobby* or *mom*. Additionally, while people can adjust to accommodate text-specific register and style, that may not be apparent given only a single sentence of context. As a result, they may not have realised that more child-like words were appropriate.

One prominent result was that *all* types of core words, regardless of definition, outperformed the non-core words. Despite the intricacies in the distinctions between different types of core words, all of them display some aspect of "coreness" that the non-core words did not. This is also evident when we look at people's incorrect responses: when people got target words wrong from any condition, they tended to guess more core words in their place. Additionally, the greatest number of common incorrect responses by far were to non-core targets, and these were mostly basic synonyms and general terms: people were able to figure out the approximate meaning of the word that was meant to go in the sentence, but guessed a simpler, more core version of it.

Taken together, this highlights the role of lexical access on this task: when people think about what the missing word is, they may formulate an idea of the rough meaning that fits the sentence and then try to access the best word with that meaning. Words that are more predictable in context are thus more easily accessed, but representationally central ones (INSTRENGTH) and core ones in general are accessed beyond what their predictability alone would suggest.

What is it about non-core words that makes them less accessible? It is difficult to tease apart frequency from representational centrality, and both of those from other factors, and to some extent, they are all intertwined and part of the reason. For instance, our non-core target words are longer and more morphologically complex compared to the core-word targets, which may have influenced performance. However, this is likely to be a natural aspect of being non-core; consider Zipf's law, which states that there is an inverse relationship between word length and frequency (Zipf, 1935).

We acknowledge that the coreness measures we have investigated here are not the only possible ones, and that alternative ways of measuring coreness exist. One notable possibility is *contextual diversity*, which could be argued to capture the idea of "occurring in more contexts" more directly than frequency. However, WF serves as a directly analogous comparison to INSTRENGTH, which reflects the number of times a word is produced in a word association task. Moreover, WF is the conventional measure used to define coreness, and it correlates highly with contextual diversity (Hollis, 2020).

Some aspects of the design may have limited our findings. Although the way we computed predictability in principle allows us to investigate the effect of the predictability of each individual sentence, because we provided multiple sentences for each target word (and thus introduced complicated dependencies between sentences), this is difficult. Additionally, although the corpus we used to derive predictability measures is large, it may not exactly approximate the language people experience given that it was sourced from the Internet.

One final limitation comes from the fact that the sentences presented in the task were decontextualised, which may make the task much harder. Although attempts were made to ensure that the stimuli could work as standalone sentences, it was often the case that they needed the broader discourse context in order to completely make sense. This suggests a future direction investigating the utility of different types of core words but scaling up to larger contexts, such as whole passages.

In sum, core vocabulary is a highly appealing notion and this work demonstrates that people's ability to predict words in context is indeed greater for core words. Although frequency-based definitions of core words are common, they may overlook other important aspects of language that are important, like centrality in the mental representation. Our results suggest that mental representation, communicative need, and the linguistic environment all play an important role in determining how core a word really is.

References

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.), (p. 53-65). Springer. doi: 10.1007/978-3-642-30773-7

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). Word prevalence norms for 62,000 english lemmas. *Behavior Research Methods*, 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., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, *104*(2), 215–226.

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

De Deyne, S., Navarro, D. J., 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.

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

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

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*, 57-84. Retrieved from https://www.aclweb.org/anthology/O13-3004

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., & Suchomel, V. (2013). The tenten corpus family. In *7th international corpus linguistics conference cl* (pp. 125–127).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, *1*, 7–36.

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

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.*

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

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 (p. 57-76). Blackwell.

Vincent-Lamarre, P., Massé, A. B., 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. (2022). Core words in semantic representation. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

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

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.emnlp-demos.6

Wu, W., Nicolai, G., & Yarowsky, D. (2020). Multilingual dictionary-based construction of core vocabulary. In (p. 4211-4217).

Zenner, E., Speelman, D., & Geeraerts, D. (2014). Core vocabulary, borrowability and entrenchment: A usage-based onomasiological approach. *Diachronica*, *31*, 74-105. Retrieved from http://www.ingentaconnect.com/content/jbp/dia/2014/00000031/00000001/art00003 doi: 10.1075/dia.31.1.03zen

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arxiv preprint arxiv:1506.06724.*

Zipf, G. K. (1935). *The psycho-biology of language*. Houghton, Mifflin.