Word prediction is more than just predictability: An investigation of core vocabulary

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

Abstract

What words are central in our semantic representations? In this experiment, we compared the core vocabulary derived from different association-based and language-based distributional models of semantic representation. Our question was: what kinds of words are easiest to guess given the surrounding sentential context? This task strongly resembles the prediction tasks on which distributional language models are trained, so core words from distributional models might be expected to be easier to guess. Results from 667 participants revealed that people's guesses were affected by word predictability, but that aspects of their performance could not be explained by distributional language models and were better captured by association-based semantic representations.

Keywords: core vocabulary; word prediction; semantic representation; distributional semantics; word frequency; word associations; semantic networks; language models

Introduction

Which words in a language are most basic, central, useful, or important? This question about the notion of *core vocabulary* has been explored in various ways. In many areas, word frequency is assumed to be a good indicator of what makes a word core. Some of the earliest work in defining core vocabulary involved hand-crafted lists of vocabulary items – mostly based on word frequency – created by linguists for pedagogical purposes, such as the General Service List (West, 1953) and Ogden's (1930) Basic English where high-frequency words are assumed to be the most useful words for communication, as they provide greater coverage of texts (Nation & Waring, 1997). Another way of conceptualising coreness is in terms of the most basic or fundamental concepts in the mind (Hsu & Hsieh, 2013); this approach is evident in research on semantic primitives Wierzbicka (1996) and analysis of the defining vocabulary of dictionaries (Vincent-Lamarre et al., 2016) but is often vaguely defined and has not been explored in much depth.[1]

Our main goal is to extend this approach by exploring whether there is more to coreness than word frequency by considering core vocabulary as the words central to people's *mental representations*. This strengthens the study of core vocabulary by grounding it in well-established psychological theory about words and concept representations and, in doing so, formulating precise, quantitative definitions of coreness that are psychologically motivated and can be empirically tested to evaluate how well they account for behaviour on tasks that tap into language representation and use.

One prominent account of semantic representation is that it is reflected in data from subjective methods that tap mental representations, such as feature generation or word association tasks (De Deyne, Verheyen, & Storms, 2016). Representing this content in a semantic network provides a straightforward way to analyse the overall structure and organisation of the mental lexicon, and the centrality of words interconnected with other words through semantic relations. Network centrality measures such as INSTRENGTH then permit us to identify hubs of association that connect many different words; hubs are higher in INSTRENGTH and have more connectivity, thus might be considered more core.

An extension of this approach, which yields an alternate measure of coreness, is based on the preferential attachment hypothesis, which suggests that semantic networks grow over time by attaching new words to existing ones and that early-acquired words are an anchor for new knowledge (Brysbaert, Van Wijnendaele, & De Deyne, 2000; Steyvers & Tenenbaum, 2005). Under this view, core words are those that are acquired earlier (i.e., have a low age-of-acquisition, AOA).

Another prominent account of word meaning situates it in the linguistic environment, suggesting that meaning is derived from natural language usage. This theory is reflected in a class of models called *Distributional Semantic Models* (DSMs), in which context plays a key role: a word's meaning is based on the way it is used for communication and the words it tends to occur with (Lenci, 2018).

From the perspective of semantic representation, word frequency (WF) is implicated in DSMs, as it captures coreness in natural language. Although WF does not fully capture coreness in terms of the *format* of the representation in DSMs, as raw counts are often transformed during learning (Mikolov et al., 2013; Bullinaria & Levy, 2007), we argue that it captures coreness in the natural language *content* on which DSMs are based.[2] One limitation of using natural lan-

[1] Although most approaches treat coreness as a continuous measure, it is often useful to designate a subset of words as the "core words". For instance, although word frequency is a continuous measure, words are often divided into discrete lists of the top n most frequent words for language teaching or characterising texts.

[2] Other measures besides word frequency that we have used to measure core words in DSMs have either been highly similar to WF (including contextual diversity, and a measure very similar to INSTRENGTH on co-occurrence data) or have thus far not produced sensible candidates for core vocabulary (e.g., cluster analyses on

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

guage to capture mental representations is that the primary purpose of natural language is communication and is therefore subject to pragmatic constraints; for example, people are unlikely to express the fact that bananas are yellow or apples are round as often as can be expected given how central these properties are. Therefore, there are limits as to what can be learned about the mental lexicon from text corpora. By contrast, word associations are considered to be free from the intent to communicate (Szalay & Deese, 1978) and are therefore thought to directly tap the content of semantic representations (De Deyne et al., 2016).

These different theoretical approaches yield core words with different characteristics. High-frequency words like *go* or *good* tend to be more semantically depleted and polysemous, reflecting their versatile use in many communicative contexts (Jorgensen, 1990; Tragel, 2001). High INSTRENGTH words like *food* or *love* tend to reflect psychologically salient concepts (Steyvers & Tenenbaum, 2005; De Deyne et al., 2016) and AOA words like *mom* and *wet* tend to be learned early in life.

In previous work, we compared these measures using a task designed to primarily tap lexical representation: a word-guessing game in which a target word was guessed from multiple hint words (Wang, De Deyne, McKague, & Perfors, 2022). Our question was which type of core words provided the most effective hints and which were the most easily guessed targets. We found no differences concerning the hint words, but INSTRENGTH-defined target words were the easiest to guess, regardless of hint type. This suggests that INSTRENGTH core words occupy a more prominent position in the mental lexicon and are more semantically central.

However, the lexical task we used may be biased towards INSTRENGTH, by being more aligned with a network-based conception of word meaning, emphasising the interrelated nature of words in semantic network models. The task was useful for investigating how people conceptualise words in isolation but may not accurately reflect the lexical representations we access when words are used with other words in a more communicative and contextual way.

This paper aims to address that limitation and reports on an experiment focused on word meaning *in context* using a cloze-style word prediction task where people guess missing words in sentences. This task is much more closely aligned with the distributional conception of word meaning and is more of a reflection of which words are communicatively useful or appropriate given the context of the surrounding sentence. The task also matches the learning objective of word-based DSMs like `word2vec` (Mikolov et al., 2018) and context-based transformer models like GPT (Brown et al., 2020) and BERT (Devlin et al., 2018). Given a context, these models predict target words with an associated probability – $p(word|context)$, which we refer to as *predictability* – reflecting how likely a word is based on the co-occurrence statistics in the linguistic data. As word prediction is a task that is centrally about distributional information and predictability reflects people's expectations during reading (Smith & Levy, 2013; Wilcox et al., 2023), WF, being a distribution definition of coreness, would be expected to perform well.

On the other hand, humans may predict words in context in a different way to language models, by extracting different linguistic information or by using extra-linguistic information. Association-based networks make no claims of how words are acquired, but incorporate aspects of word meaning that are not present in language-based DSMs, like perceptual or affective information; as a result, they better capture key semantic properties (Vankrunkelsven et al., 2018) and human similarity judgements (De Deyne et al., 2021) compared to DSMs. It is possible that these and other factors also matter during word prediction tasks. For instance, people might prefer simpler or more basic words, which would bias words that are more representationally accessible or semantically prominent. For example, in the sentence "The creature grew to ___ proportions", the top completion from BERT is *gigantic*, whereas most people might respond *large* or *huge*. This is an important difference that highlights a discrepancy between model predictability – that is, what can be learned from co-occurrence statistics – and the way that people process language. If performance on the word prediction task depends on factors besides predictability alone, then these factors may be captured in the INSTRENGTH and AOA coreness measures.

This work has two main aims. First, we compare how well people predict the three different kinds of core words (INSTRENGTH, WF, AOA) in sentence contexts, and compare these to words that are not core (NONCORE). Second, we ask what (if anything) drives performance on this task over and above predictability, and to what extent this can be captured in the coreness measures.

Method

Participants

667 participants (19-82 years, $M = 42.6$; 53% female) were recruited via Prolific. 93% of participants reported being native English speakers. Three participants were excluded for not passing the pre-registered[3] catch trials (described below), leaving 664 participants in the full analyses.

Procedure

The task was set up as a game in which people were told they were cracking coded messages sent between spies, which were ordinary sentences hidden in sources such as websites, books, and newspapers. As Figure 1 depicts, each sentence contained a blank in place of a missing "code word" (the target word). The task on each trial was to guess the missing word. Participants were given one attempt, after which the correct answer was revealed.

word embeddings, networks derived from word embeddings). For this reason, and because of its established importance in the core vocabulary literature, we focus here on WF, but are investigating alternative measures of core words in DSMs in future work.

[3] `https://aspredicted.org/SRW_R9Z`

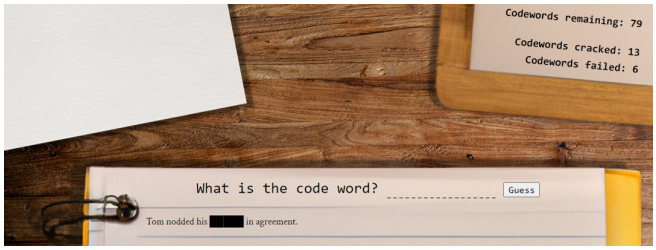

Figure 1: **Screenshot from example trial.** On each trial, participants were shown a sentence with a black box in the location of the target word to be guessed. In this trial, the target word was *head*.

After a practice trial, every participant completed one sentence for all 96 target words (24 for each of the four target word conditions). For each target, each participant saw a sentence that was randomly selected from a pool of stimulus sentences for that target (see below). In addition to the target trials, participants did four catch trials designed to be substantially easier than the experimental ones to check random guessing. These catch trials were presented in the same position for everyone; the targets were *age* (trial 10), *head* (trial 40), *phone* (trial 65), and *city* (trial 90). As pre-registered, we excluded participants who failed all catch trials. The order of the target word trials was randomised for each person.

Materials

Target Words. The target words came from four target conditions, each containing 24 words. The INSTRENGTH, WF, and AOA conditions contained words that were core according to the corresponding coreness measure. 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 in Wang et al. (2022), the core word list for each coreness measure was defined as the top 300 core words on that measure. The WF measure was based on the SUBTLEX database (Brysbaert & New, 2009), with more frequent words being more core. The INSTRENGTH measure was based on the semantic network derived from word associations to over 12,000 English words (De Deyne et al., 2019). INSTRENGTH was calculated for each word as the sum of the weights of all incoming edges directed towards that word, where edge weights represent the strength of association between words. Common associates have higher INSTRENGTH and are more core. The AOA measure was sourced from the Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) norms, with earlier-acquired words being more core.

The INSTRENGTH and WF measures were log-10 transformed in line with previous work. All words were also normalised by grouping inflectional forms of the same lemma (e.g., *run*, *runs*, *running*). Function words, including determiners, auxiliary verbs, and prepositions, were removed. To compare the coreness of words across the measures, we normalised 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) core word. This results in an inverted "coreness" metric where the most core word has a value of 0, and higher values indicate diminishing coreness.

Table 1: Target words in each condition.

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

The AOA, WF, and INSTRENGTH target words were selected to be words that were core on their respective lists while also being less (but equally) core on the other two measures.[4] This allows 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. The words were the same as in Wang et al. (2022) except for one word in the AOA condition (*reindeer* instead of *crayon*) because *crayon* did not exist in the BERT vocabulary.

The NONCORE target words were selected to be words of low coreness and not on any of the three core word lists. To ensure that these would still be familiar words, we only used words known by most native speakers as measured through word prevalence norms (Brysbaert, Mandera, & Keuleers, 2018). The selected NONCORE words had similar coreness across all three coreness measures.[5] 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.

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

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

Sentences. To ensure that the sentences used in the task were naturalistic instances of how words are used in language, we sourced sentence stimuli from the enTenTen web corpus via Sketch Engine (Jakubíček et al., 2013). The corpus used in this study 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 of the language many adult English speakers are exposed to. All stimulus sentences were filtered to remove ungrammatical or incomplete sentences, sentences with jargon or duplicate instances of the target, and sensitive content.

We used BERT (Devlin et al., 2018) to calculate the predictability of the target word in each sentence.[6] BERT is a transformer-based LLM trained by masking words in the sentence input and having the model predict the masked words based on both the left and right context. We used the base, uncased version of BERT (110m parameters), which was trained on BookCorpus (Zhu et al., 2015) and English Wikipedia. The model was accessed through the *Transformers* Python package (Wolf et al., 2020).

The target words varied in predictability, reflecting natural differences in their distributional profile in the linguistic environment. For each word, a predictability distribution was constructed by computing its predictability in 10,000 sampled sentences containing it (Figure 2). WF words had the highest average predictability, followed by the INSTRENGTH and AOA; all were more predictable than the NONCORE words.

We created two sets of stimuli based on the predictability statistics. The **corpus-matched** sentences were designed to match the corpus predictability distributions as closely as possible. This allowed us to investigate task performance while preserving the natural variation in predictability, and then to ascertain what factors affected performance after controlling for predictability. Each target word had 36 sentences, which were selected by repeatedly sampling them from the full corpus until 36 sentences matched the corpus distribution (with a Kolmogorov–Smirnov (K-S) statistic of less than .15).

In the **matched-predictability** sentences, predictability was controlled across all target words. This allowed us to investigate performance when predictability was controlled in the stimuli at the outset, and thus provided a better estimate of whether any target condition differences exist above and beyond predictability differences. For each target word, 10 sentences were selected in which the predictability was as close as possible to specified values ranging from .05 to .95 in increments of 0.1.[7] Additionally, in all sentences, the target word was the top completion; this ensured that any responses other than the target word were not predicted by BERT.

[6] All analyses here use raw predictability. Analysis of the data using log-transformed predictability did not change the qualitative findings. See the supplemental materials: https://osf.io/4bqgt/?view_only=c067008134284c40a4d41c7912162661

[7] For each of the ten values, the mean difference in predictability between any two conditions was no more than 0.01. One target word, *gigantic* (NONCORE) had only 5 stimulus sentences, as no sentences could be found with a predictability of .55 or higher.

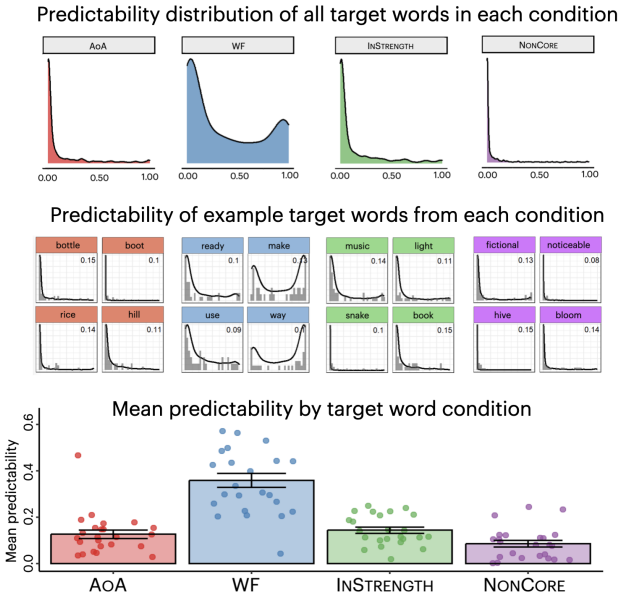

Figure 2: **Predictability distributions of target words.** *Top panel.* Distribution of predictability (x axis) of all words in each of the four conditions. INSTRENGTH, AOA, and NONCORE words had very low predictability relative to WF words, which had higher predictability. *Middle panel.* These differences in predictability are visible when considering representative example words from each condition. Black lines show the distribution for that word as estimated based on the 10,000 sentences. Bars indicate the predictability of the 36 corpus-matched stimuli, along with the K-S statistic capturing the difference between the distributions, which was constrained to be .15 or less. *Bottom panel.* Reflecting their real-world distributions, conditions varied significantly in predictability (each dot is the mean predictability of the 36 stimulus sentences for one target).

Results

Corpus-matched stimuli: Accuracy

The mean accuracy for each target word was computed by averaging over all participants and sentences, as shown in Figure 3. There were significant differences in target word accuracy between conditions, $F(3,92) = 22.61, p < .001$ (one-way ANOVA). Post-hoc pairwise comparisons with Holm corrections indicated that accuracy for the WF condition was significantly higher than the AOA ($p < .001$) and INSTRENGTH ($p = .002$) conditions and that all core-word conditions had significantly higher accuracy than the NONCORE condition (all ps $< .001$).

The higher accuracy for WF target words is expected based on the predictability differences shown in Figure 2. A more interesting question is whether coreness (by any measure) affects accuracy over and above predictability. To analyse this, we compared several different linear regression models (see Table 2). In all of the models, the outcome variable was target word accuracy. Models varied by whether and how they included the factors of target condition and target word predictability (computed as the mean predictability of each target in the stimulus sentences).

The best-fitting model (M1CP in Table 2) contained both predictors with no interaction, suggesting that predictability had a similar effect in each condition. In that model,

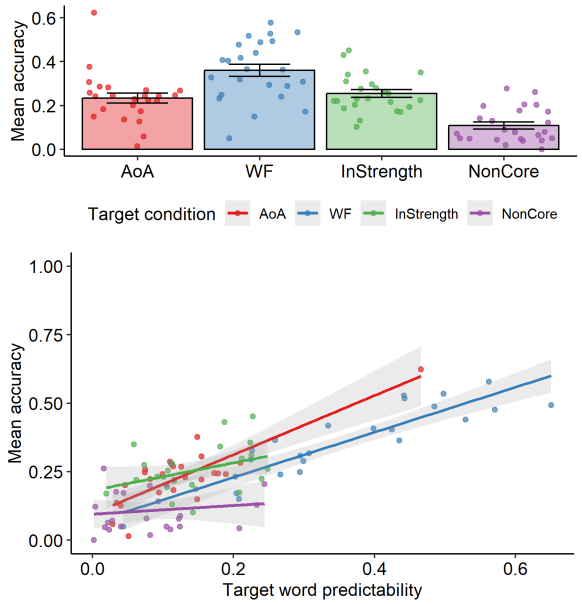

Figure 3: **Accuracy for corpus-matched stimuli.** *Top panel*. Each dot represents one target word whose mean accuracy (y axis) is calculated by averaging over all trials and participants. WF words were the easiest to guess, followed by AoA and InStrength. NonCore words were the hardest to guess. *Bottom panel*. Linear regression lines for each condition predicting mean accuracy (y axis) from target word predictability (x axis). InStrength and AoA words had an advantage and NonCore words a disadvantage relative to the others, beyond what predictability alone would suggest.

target word predictability significantly predicted accuracy, $b = 0.76, p < .001$, meaning that words that were more predictable on average were easier to guess. However, there were target condition differences over and above the effect of predictability: accuracy was significantly higher in the InStrength condition (the reference category) than in the WF ($p = .046$) and NonCore conditions ($p < .001$); it was not significantly different from AoA accuracy.

Overall, our findings indicate that when predictability is not controlled for, people found it easiest to guess the WF target words (which were far more predictable than the others by definition). But when predictability was controlled for, accuracy was actually *lower* for WF targets than for InStrength and AoA words; the latter were easier to guess for reasons that go beyond their predictability.

Matched-predictability stimuli: Accuracy

Our previous analysis suggested that some target words are easier to guess than their predictability alone would indicate. Still, there is always some uncertainty when statistically controlling for a variable. We therefore ask a similar question for the matched-predictability stimuli, which experimentally controlled for predictability across target words. A one-way ANOVA showed significant differences in sentence accuracy between conditions, $F(3,951) = 42.55, p < .001$, but this time with a different ordering: post-hoc comparisons with Holm corrections revealed that the AoA condition had higher accuracy than the InStrength condition ($p = .044$), both had higher accuracy than the WF condition ($p < .001$ and $p = .025$, respectively), and all core-word conditions had higher accuracy than the NonCore condition (all ps $< .001$).

Table 2: **Model comparisons for two analyses**. Models are depicted with statistical notation where `*` indicates an interaction and 1 is a constant. `condition` indicates the four conditions and `pred` is target word predictability. Best models have the lowest BIC (bold).

Corpus-matched stimuli		
Model	Description	BIC
M1null	`accuracy ~ 1`	-100
M1C	`accuracy ~ condition`	-139
M1P	`accuracy ~ pred`	-187
M1CP	`accuracy ~ pred + condition`	**-199**
M1CPI	`accuracy ~ pred * condition`	-198

Matched-predictability stimuli		
M2null	`accuracy ~ 1`	472
M2C	`accuracy ~ condition`	372
M2P	`accuracy ~ pred`	337
M2CP	`accuracy ~ pred + condition`	**219**
M2CPI	`accuracy ~ pred * condition`	232

As before, we explored the role of predictability by comparing a set of linear regression models (see Table 2). The best-fitting model, shown in Figure 4, contained both predictors with no interaction, suggesting that target condition differences did not change substantially for sentences with higher or lower predictability. Again, predictability significantly predicted accuracy, $b = 0.40, p < .001$. The same pattern of condition differences remained, with accuracy for the AoA condition being higher than the InStrength condition (the reference category; $p = .028$), and accuracy for the InStrength condition being higher than the WF ($p = .007$) and NonCore conditions ($p < .001$).

Overall, these results are reasonably consistent with the results for the corpus-matched stimuli. There were differences in accuracy between different core word targets even when predictability was controlled for in the stimuli. As before, the WF target words were harder to guess than target words from other core word conditions, and the NonCore target words were most difficult to guess of all.

Corpus-matched stimuli: Incorrect responses

How robust is the finding that there are factors over and above predictability that drive performance on this task? And, if so, to what extent do any of our coreness measures capture such factors? We address these questions by focusing on the *incorrect* responses: to what extent can the specific guesses people made be predicted based on their BERT predictability and/or coreness? The response probability of a word was computed as the number of times it was given for a sentence out of the total number of presentations of that sentence. This was compared to the BERT predictability of that response in that sentence. Items not in the BERT vocabulary (9%) were excluded. Linear regression models were fitted that predicted

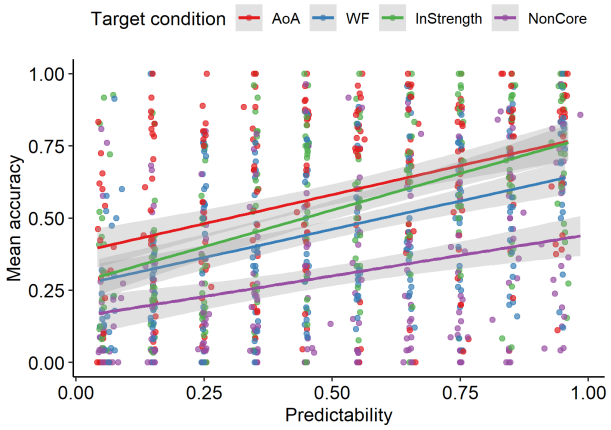

Figure 4: **Accuracy for matched-predictability stimuli.** Each dot is one sentence whose mean accuracy (y axis) was calculated by averaging over all trials and participants. The x axis shows the predictability of the target word in that sentence. Linear regression lines for each condition are shown. AoA words had higher accuracy than InStrength, followed by WF. NonCore words were least accurate, beyond what predictability alone would suggest.

response probability based on BERT predictability as well as AoA, WF, and InStrength coreness.

As expected, predictability significantly predicted response probability, $b = 0.39, p < .001$, but coreness predicted response probability over and above that. Responses that were more core in AoA, $b = -0.006, p < .001$, and InStrength, $b = -0.014, p < .001$, were more likely to be given (0 being the most core word on the normalised measures, and higher values denoting lower coreness), and InStrength coreness was a stronger predictor than AoA. Interestingly, after taking into account all the other variables, WF coreness actually predicted *lower* response probability, $b = 0.008, p = .001$. This indicates that higher frequency words were in fact *less* likely to be given as responses after controlling for everything else, and is qualitatively consistent with our accuracy analyses.

Matched-predictability stimuli: Incorrect responses

We ran the same analysis on the matched-predictability stimuli and found similar results. Predictability was significantly related to response probability, $b = 0.65, p < .001$, but coreness predicted responses over and above that. Greater coreness on AoA, $b = -0.004, p = .009$, and InStrength, $b = -0.012, p < .001$, predicted higher response probability. As before, WF coreness was associated with lower response probability, $b = 0.009, p = .002$.

Discussion

This study compared coreness measures derived from different theories of semantic representation on a word prediction task to evaluate how each theory captured the guesses people made given the rest of the sentence context. Reflecting their predictability statistics in the linguistic environment, high-frequency words (WF) were easier to guess than words that are central in association networks (InStrength) and learned early (AoA). However, people's guesses involved more than just predictability: InStrength and AoA core words were guessed more often than their predictability alone would suggest. Additionally, incorrect responses that were high in InStrength and AoA coreness were given more often than expected based only on their predictability, while high WF responses were not. These results were consistent when controlling for predictability in the corpus-matched stimuli and when predictability was matched across stimuli.

The results show that humans differ from language models in how they predict words in context in important ways. For instance, they prefer to use simpler or more basic words. To give some qualitative examples, while BERT completes "explodes with" with *rage*, most people responded with *anger*. Similarly, instead of *gigantic*, people responded with *large*, *huge*, or *epic*. Thus, people rely on more than just information captured by BERT predictability. This aspect of how humans process language was partly captured by coreness measures that tap into aspects of representational accessibility and semantic prominence, such as InStrength and AoA. This finding is consistent with work showing that association-based models incorporate information about word meaning that is not present in language-based DSMs (De Deyne et al., 2021; Vankrunkelsven et al., 2018).

These findings are also consistent with our previous work (Wang et al., 2022) showing that InStrength words are more core on a task where people guessed words in isolation. Our findings indicated that InStrength core words were more semantically accessible compared to AoA and WF core words. Even when we controlled for factors that would be expected to matter for that word-guessing task, such as semantic similarity or target word part-of-speech, InStrength core words were still easier to guess than other core word types. Thus, on tasks that tap how people conceptualise word meanings both *in isolation* and *in context*, InStrength-defined core words led to the best performance.

It is worth noting the result that all types of core words, regardless of definition, far outperformed the NonCore words – both in terms of overall performance and in terms of what was expected based on predictability – even though these words were perfectly well known according to prevalence norms (Brysbaert et al., 2018). This suggests that *all* of the conceptualisations of core vocabulary we have explored do tap into some measurable aspect of coreness, and points to the idea that there are multiple good ways of defining it.

Ultimately, a complete theory of what words are mentally core is likely to involve a combination of multiple factors, and explain how their role depends on the nature of the task. Here, we have attempted to investigate exactly how different factors contribute to what is mentally core, and have shown that there is more to coreness than distributional information, even on a task that relies on prediction in context.

References

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

Brysbaert, M., 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.

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

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*.

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

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

De Deyne, S., Verheyen, S., & Storms, G. (2016). Structure and organization of the mental lexicon: A network approach derived from syntactic dependency relations and word associations. *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, 47-79.

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

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.

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).

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

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

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151-171.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1-12.

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.*

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. *Vocabulary: Description, acquisition and pedagogy*, *14*(1), 6–19.

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

Smith, N. J., & Levy, R. (2013, 9). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302-319.

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.

Szalay, L. B., & Deese, J. (1978). *Subjective meaning and culture: An assessment through word associations*. Hillsdale: Lawrence Erlbaum.

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

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

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.

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

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470.

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.

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.*