Communicative efficiency of distributional and semantically-based core vocabularies in narrative text comprehension

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

High-frequency words are often assumed to be the most useful words for communication, as they provide the greatest coverage of texts. However, the relationship between coverage and comprehension may not be straightforward; how words relate semantically to people's mental representations is also important. In this study, we evaluate how useful different sets of "core vocabularies" are in text comprehension. The core vocabularies, which reflect different aspects of distributional and semantic information, provide different amounts of information for different vocabulary size and amount of text coverage. In our experiment, we showed people narrative texts with all but the core words removed, and measured comprehension in a variety of ways. Our results show that both distributional (e.g., frequency-based) and semantic (e.g., word association-based) core vocabularies are communicatively useful, but that the semantically-based core vocabularies provide more information when textual coverage is held constant.

Keywords: core vocabulary; comprehension; situation models; information theory; word frequency; word associations

Introduction

The question of which words are the most useful ones for communication is important both theoretically, in terms of informing theories about language processing and human communication, and practically, in terms of applications in language learning and teaching. High-frequency words are often assumed to be the most useful for communication, since by their very definition they provide the greatest coverage of texts (e.g., Nation & Waring, 1997). Much work in applied linguistics thus aims to understand what proportion of the words in a text language learners need to know to achieve "adequate" comprehension (e.g. Hu & Nation, 2000; Schmitt, Jiang, & Grabe, 2011), as well as how many of the most frequent words are needed to achieve that level of coverage (e.g. Nation, 2006; Schmitt et al., 2017).

The implicit assumption behind this work is that coverage is a primary determinant of text comprehension. However, models of reading comprehension in the psychological literature suggest that the relationship between text coverage and comprehension may not be as straightforward as one would think. For instance, according to Kintsch's (1988) Construction-Integration model of comprehension, texts are processed during reading into a network of propositions that represents the meaning of the text. From this perspective, what matters most is not the coverage of each word, but the role that each word plays in the network of propositions.

More generally, it is widely accepted that comprehension is a process of constructing mental representations of the situation *underlying* the text, rather than something reflective of the verbatim content of the text itself (Zwaan & Radvansky, 1998). This again suggests that what might matter more than individual occurrences of specific words is how the words lead a reader to construct the situation model. The question of which words are the most useful for communication, then, may be understood more precisely as which words are the most influential, or contribute the most accurately, to the *representation* that a reader constructs during comprehension.

Given this, how can we measure the contribution of a specific word? One possibility is to use information theory, which provides a model of the entire communicative process: in it, a speaker (or writer) aims to transmit a certain idea from their mind into the mind of a listener (or reader), via a message that is composed of words. The degree to which the receiver's reconstructed meaning matches the speaker's intended meaning constitutes *information*. The words that are thus the most useful for communication are those that provide the most information about what the speaker means.

The aim of this study is to compare different types of *core vocabulary* based on how successfully they facilitate communication. Our different candidate core vocabularies (described below) reflect different theories about the lexicon. Evaluating entire vocabularies rather than single words is not only useful because those vocabularies map onto psychological theory, but also because it enables us to investigate this question in terms of informational tradeoffs with both vocabulary *size* and vocabulary *coverage*.

Vocabulary **size** matters because the larger a vocabulary, the more information it can provide. However, an ideal vocabulary should also be simple – it should yield the most amount of information from the smallest number of words. Vocabulary **coverage** matters for similar reasons: if the words in a given vocabulary cover a greater proportion of what one would want to communicate (as in a text), then that vocabulary is probably more informative. However, if the same amount of information can be provided using fewer words (i.e., with less coverage of the text), then that vocabulary is simpler without sacrificing informativeness. A communication system, if it is to be efficient, should optimally trade-off between these competing factors of informativeness and simplicity (Kemp, Xu, & Regier, 2018).

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

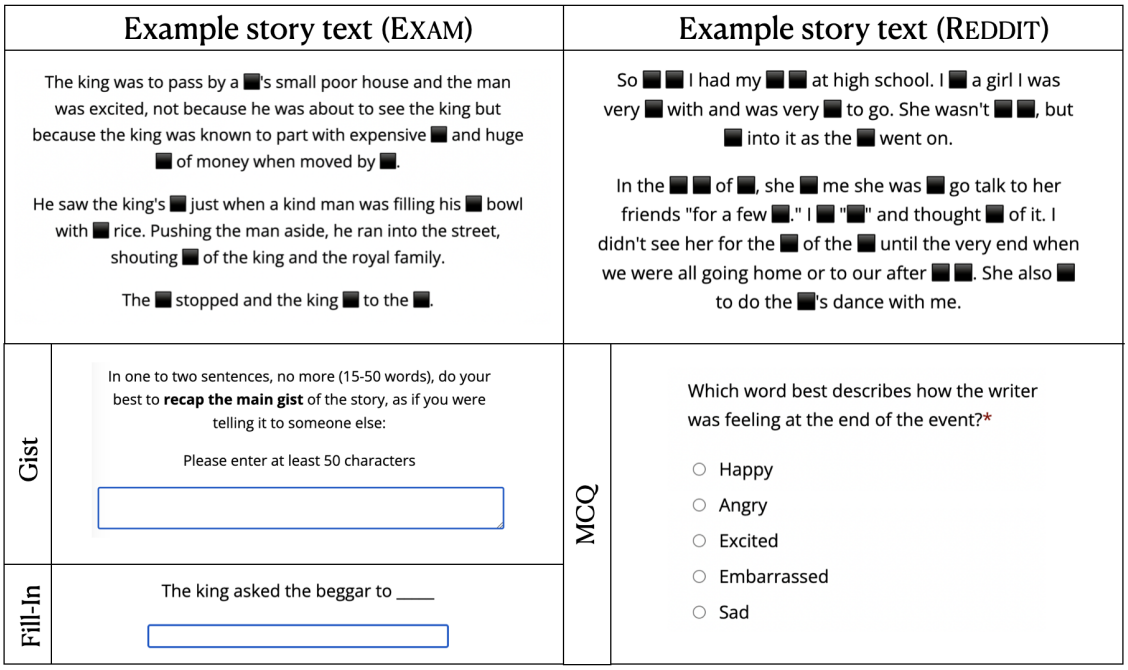

Figure 1: **Example experimental trials**. On each of the 16 trials, people read a 200-400 word story in which only some words were revealed, depending on the vocabulary Type and Size condition of that trial. Half of the texts were sourced from an English as a Second Language (ESL) EXAM (sample excerpt in left panel, corresponding to a large Size with more words included) and half from REDDIT (right panel, smaller Size). Each story was followed by five questions; the first was always Gist. The other questions, which varied for each text, included formats like fill-in-the-blank and multiple-choice/select-all. Questions for a given text were identical for all permutations of vocabulary Type and Size, and always presented in the same order so that earlier ones did not give away later answers.

Core vocabularies Based on previous work and motivated by different theories of the lexicon, we consider four ways of defining a core vocabulary. The first two approaches capture the idea that the lexicon reflects distributional information in the linguistic environment. This commonly maps onto a core vocabulary based on word frequency (WF): words that are most frequent are the most core. However, another way to capture distributional information is to look at how words occur in text *together*, using co-occurrences. We therefore also derive a core vocabulary based on co-occurrence centrality (CC), which defines core words as those that occur most often with other words.

Another approach is based on the idea that comprehension centrally involves reconstructing the mental representation of the speaker. We therefore consider core vocabularies that reflect the *semantic importance* of words. One way to measure this importance is through word association data (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019). Semantic networks constructed from word associations can be used to represent how words relate to each other in the mental lexicon. We use a measure of network centrality called in-strength (INS) which reflects the importance of words in the network.

One might also consider the core words to be those that were acquired early, either because they are conceptually most primitive or because they serve as an anchor as the semantic network grows (Brysbaert, Van Wijnendaele, & De Deyne, 2000). We thus include a final core vocabulary based on age-of-acquisition norms (AOA).

Study aims Our goal is to investigate which type of core vocabulary is most useful for communication. This means not just which vocabulary provides the most information, but crucially, which does so most *efficiently*, after taking vocabulary size and coverage into account. We therefore have two specific questions. First, which type of core vocabulary provides the most information for a given vocabulary size? And second, which type of core vocabulary provides the most information for a given amount of textual coverage?

In previous work (Wang et al., 2024) we investigated these questions by having people identify the topic of Wikipedia articles based on seeing only the core words in those articles. We found that people performed best (for a given size and coverage) when reading articles containing the core vocabularies based on semantic centrality (INS and AOA) rather than distributional information (WF and CC).

However, this work has two main limitations. First, the expository genre of Wikipedia articles is relatively far removed from the everyday uses of language that most people experience. It remains an open question as to whether those results would be observed in more typical naturalistic language use, as might be found in narratives. Moreover, identification of the topic of the articles is a fairly coarse measure that does not fully tap into comprehension of the nuances and details of a text. It is suitable for expository texts (where more detailed comprehension questions would be conflated with general knowledge), but investigating genres such as narratives allows us to use more in-depth measures of comprehension.

Method

Participants

213 people (19-80 years, $M = 42.3$; 45% female) were recruited via Prolific. 96% were native English speakers. One person was excluded for not passing the pre-registered[1] attention checks, leaving 212 people in the analyses.

Procedure

The main task consisted of reading a series of short stories in which only some of the words were revealed in the text (see Figure 1). Which words were shown varied within-subject in a 4×4 design in which we manipulated the core vocabulary Type (AoA, Ins, WF, CC) and the Size of the vocabulary from which the revealed words were selected (200, 500, 1000, and 2000 words). There were thus 16 trials, each corresponding to one story and one of the 16 possible Type×Size combinations, presented in a random order for each participant.

There were 30 texts in total, half originating from Reddit posts and half from an dataset of English Exam questions intended for students in China (see below). Each text occurred in all 16 conditions (four Type × four Size). Every participant saw a random selection of 16 of the 30 texts, constrained so that nobody saw the same text more than once and each person was shown 8 Reddit texts and 8 Exam ones.

On each of the 16 trials, people were first shown a story in which only the words corresponding to the core vocabulary Type and Size were shown and all others were replaced with a black box. After summarising the text in a few sentences, people were asked five comprehension questions of varying difficulty. Before beginning the 16 main trials, they had to pass a quiz verifying that they had understood the instructions as well as complete a practice story with questions.

Materials

Core vocabulary type The four core vocabularies were defined as the set of top n words on a given coreness measure (n corresponds to Size: 200, 500, 1000, or 2000). Because our interest is in lexical concepts, function words (including determiners, auxiliary verbs, prepositions, conjunctions, and pronouns) were excluded from the core vocabularies, and all words were lemmatised.

The in-strength (Ins) measure captures the words that are most central in people's lexical representations. We used the Small World of Words (De Deyne et al., 2019) word association dataset, and computed the In-Strength for each word as the sum of the strength of incoming links to that word, where the strength represents the association strength from a cue to a response. Words with higher in-strength (e.g., *love*, *food*) are connected to more words and are more core.

The AoA core vocabulary consisted of words with the lowest average age-of-acquisition, sourced from the Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) norms. Words like *mom*, learned earlier in life, are more core in AoA.

[1] https://aspredicted.org/38nr-kk98.pdf

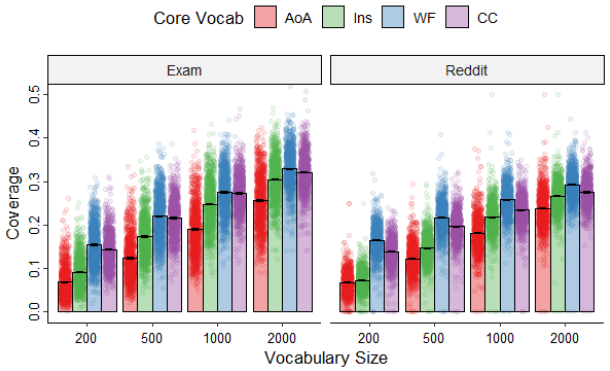

Figure 2: **Text coverage given by each core vocabulary.** The proportion of each Exam and Reddit text covered by core words (y axis) is shown as a function of vocabulary size (x axis). WF and CC consistently achieve the highest coverage at any given vocabulary size, and AoA provides the least.

The word frequency measure (WF), which reflects data about which words are used most often, used norms sourced from the SUBTLEX database (Brysbaert & New, 2009). More frequent words like *know* or *good* are more core.

Lastly, the co-occurrence centrality (CC) core vocabulary was based on co-occurrence data calculated from the Corpus of Contemporary American English (Davies, 2008-). CC represents strength centrality for co-occurrences, making it directly analogous to Ins, but based on distributional information rather than word associations. Co-occurrence strengths were computed by normalizing raw co-occurrence counts as a proportion of word frequency; strength centrality was then computed by summing over the co-occurrence strengths for a given word. The words with the highest co-occurrence centrality, like *time* and *people*, make up the CC core vocabulary.

Texts The 30 story texts used in the experiment, which ranged between 200-400 words long, included two kinds of narratives. Half followed an informal, naturalistic style, and were taken from anecdotes posted on Reddit. We sourced datasets containing posts from two subreddits built around story sharing: `r/AITA` [2] and `r/TIFU` (Kim et al., 2019). The other 15 texts were more formal, structured narratives corresponding to passages from English exams for secondary students in China, sourced from the RACE dataset (Lai et al., 2017). These passages were written by English instructors to evaluate reading comprehension.

Naturally, the amount of coverage afforded by core words varies as a function of core vocabulary Size and Type. As shown in Figure 2, in general, the distributional core vocabularies (WF and CC) have the greatest coverage: the words in those core vocabularies constitute a larger proportion of the words in a given text, compared to core vocabulary from Ins or AoA. This is expected given that WF and CC directly reflect frequency and co-occurrence frequency. We selected texts that preserved the natural coverage for each core vocabulary, choosing only stories whose coverage was within 1 SD of the mean for all four core vocabulary at size 1000.

[2] https://github.com/iterative/aita_dataset

Each text was edited slightly to remove meta-commentary (e.g., "So Reddit, AITA?") and for minor proofreading (capitalisation, misspelled words, etc). We created 16 versions of each text (corresponding to each core vocabulary Type and specific list Size) by removing all words from the text except for the core words on that list. All other words were replaced with a black block (Unicode character U+2B1B) which did not contain any information about their length (see Figure 1). We also retained punctuation, a limited set of function words (determiners, pronouns, prepositions, etc.), which were the same for all vocabulary types and sizes, and the names of the characters in the story.

Comprehension questions Each text was followed by two types of questions. For every text we first asked participants to summarise the *Gist* of the story. The purpose was to have one question that was the same for all texts, and which was revealing about what essential information people were able to extract from the story given only the core vocabulary they saw. The full text of the question is shown in Figure 1 and participants were unable to copy-paste into the text box or move on without answering it in at least 50 characters.

After the Gist question, each text was associated with five comprehension questions targeting more specific details of the story. The goal of these questions was to assess understanding of key events, causes, and motivations, but not to focus on specific wordings or verbatim content; they were intended to measure a relatively deep level of comprehension by targeting the situation model rather than the surface form of the words (Kamalski, 2004). Some questions, usually earlier ones, asked about broad details (e.g., "Where does the story take place?") while others asked about more specific details (e.g., "What happened after the cashier gave back the writer's change?"). The format of the questions included free text response (with either words or short phrases) as well as multiple choice (some with only one possible answer, some with many). Examples are shown in Figure 1.

As much as possible, the questions and options were written without incorporating wording that matched the text (e.g., using "café" if the text used "coffee shop"). We also ordered the questions so that earlier ones gave minimal information away for later ones. The format of questions varied for different texts, depending on which were best given the content and nature of that text. Since all questions and texts were identical across any given Type×Size manipulation, we can be certain that any differences observed in core vocabulary Type or Size cannot be due to variation in text or question difficulty.

Results

Gist accuracy

Responses to the gist question were converted into vector embeddings using BERT sentence transformers (all-MiniLM-L6-v2), and compared to the vector embeddings of the corresponding full texts using cosine similarity. Thus, higher similarity indicates that the summary did a better job of capturing the gist of the text. The mean gist accuracy for each text and Type×Size condition was computed by averaging across the cosine similarities of the individual gist responses of all participants in that condition.

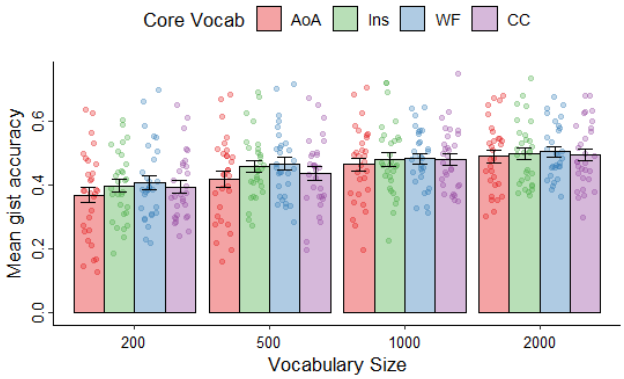

Figure 3: **Mean gist accuracy for texts by condition.** Gist accuracy for a single participant on a single trial is the cosine similarity between the full text and their attempt to summarise the gist in a few sentences. Each dot is a single text whose gist accuracy (y axis) is calculated by averaging over all trials and participants for that text in that condition. Although the differences between core vocabularies were not large, they were consistent, with INS and WF leading to more accurate gist responses, AOA the least and CC in the middle.

Role of core vocabulary Our first question is whether the accuracy of the gist summaries varied based on the nature or size of the core vocabulary corresponding to the words people were given in each text. We explored this with a two-way repeated measures ANOVA with gist accuracy as the outcome variable and Type and Size as predictor variables. As Figure 3 shows, each increase in vocabulary size was always beneficial: there was a significant main effect of Size, $F(1.92, 55.70) = 58.47$, $p < .001$, and Holm-corrected post-hoc tests showed that all vocabulary sizes were significantly different from each other (all ps $< .001$).

There was also a significant main effect of core vocabulary Type, $F(2.24, 64.91) = 4.40$, $p = .013$. Post-hoc tests with Holm corrections showed that texts with INS and WF core words yielded significantly higher gist accuracy than AOA (both ps $< .01$), with CC in between. The interaction between vocabulary Type and Size was not significant, $F(9, 261) = 0.66$, $p = .74$. Overall, these results suggest the INS and WF core vocabularies provided relatively more information about the gist of the narrative for a given vocabulary size, while the AOA core vocabulary provided relatively less.

Role of coverage As we saw earlier, the WF core vocabulary consistently had higher coverage of any given text at any given vocabulary size. We might therefore ask how much of the relatively superior performance WF is due to its superior coverage; put another way, would a WF core vocabulary still provide more information for texts of the same coverage?

We explored this question by doing model selection using BIC over a set of linear mixed models predicting mean gist accuracy from different combinations of Coverage and core vocabulary Type, with text included as a random effect. The

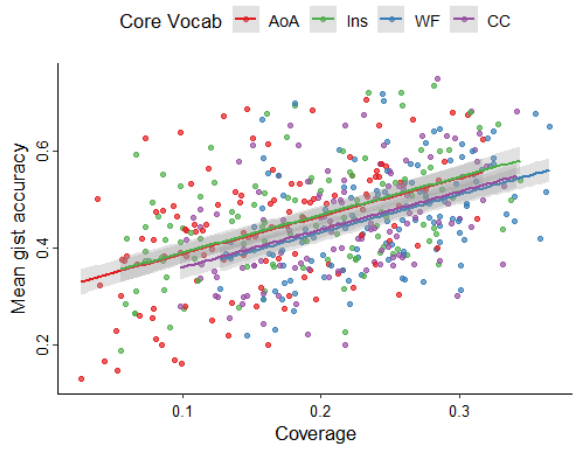

Figure 4: **Mean gist accuracy as a function of coverage.** Each dot represents one text in a given condition whose mean gist accuracy (y axis) is calculated by averaging over all trials and participants, and whose coverage is calculated as the proportion of its words in a given core vocabulary (x axis). The regression lines for each core vocabulary type from the model with Coverage and Type as predictors are shown. Coverage is the most important factor affecting gist accuracy, but the INS and AOA core words resulted in better performance for the same level of coverage, compared to WF and CC.

two best-fitting models both contained Coverage (Table 1a).

To evaluate the role of core vocabulary Type when Coverage is taken into account, we examined the model containing both factors (`M1typeCov`). Coverage significantly predicted gist accuracy, $b = 0.66$, 95% CI $[0.58, 0.74]$: as expected, greater coverage of core words led to more accurate summaries. However, core vocabulary mattered over and above the effect of coverage: compared to INS (the reference category), gist accuracy was significantly lower for WF and CC core vocabularies (WF: $b = -0.03, [-0.04, -0.02]$; CC: $b = -0.03, [-0.04, -0.01]$). These effects are not large, but they suggest that for the same amount of coverage, INS words provided more information about the overall gist of the stories than did WF and CC words (see Figure 4).

Comprehension question accuracy

Accuracy for each comprehension question was scored differently depending on the question format. MCQs with only one allowable answer scored 1 if the correct answer was selected and 0 otherwise. For those with multiple correct response options, we computed the F1 score; this took into account both recall and precision of the person's choices. The text response questions were scored automatically by querying OpenAI's GPT-4o model. Each individual text question response was provided to the model, along with the question text, and the full corresponding story text. The prompt asked the model to rate the accuracy of the response from 1 to 10 (which was then rescaled between 0 and 1). A random subset of 320 responses (20 from each experimental condition) were manually scored and compared to the LLM ratings. Spearman's r_ρ was .79 and over 80% of responses were within 2 rating points.

Table 1: **Model comparisons**. Models are depicted with statistical notation where * is an interaction and 1 is a constant. The outcome variables are gist accuracy (`gAcc`) and comprehension question accuracy (`qAcc`). `type` indicates the four core vocabulary conditions, `size` indicates the four vocabulary size conditions, `cov` is core word coverage, and `text` and `ques` indicate random effects for text and question, respectively. The best models have the lowest BIC (bold).

(a) Gist Accuracy and Coverage		
Model	Description	BIC
M1null	`gAcc ~ 1 + (1|text)`	-1041
M1type	`gAcc ~ type + (1|text)`	-1011
M1cov	`gAcc ~ cov + (1|text)`	**-1242**
M1typeCov	`gAcc ~ type + cov + (1|text)`	**-1220**
M1typeCovInt	`gAcc ~ type * cov + (1 |text)`	-1201
(b) Question Accuracy and Vocabulary Size		
M2nullT	`qAcc ~ 1 + (1|text)`	-43
M2nullQ	`qAcc ~ 1 + (1|text/ques)`	-564
M2type	`qAcc ~ type + (1|text/ques)`	-575
M2size	`qAcc ~ size + (1|text/ques)`	**-1197**
M2typeSize	`qAcc ~ type + size + (1|text/ques)`	**-1226**
M2typeSizeInt	`qAcc ~ type * size + (1|text/ques)`	-1119
(c) Question Accuracy and Coverage		
M3nullT	`qAcc ~ 1 + (1|text)`	-43
M3nullQ	`qAcc ~ 1 + (1|text/ques)`	-564
M3type	`qAcc ~ type + (1|text/ques)`	-575
M3cov	`qAcc ~ cov + (1|text/ques)`	**-1335**
M3typeCov	`qAcc ~ type + cov + (1|text/ques)`	**-1330**
M3typeCovInt	`qAcc ~ type * cov + (1|text/ques)`	-1314

Role of core vocabulary Figure 5 shows the mean accuracy for each question in each condition. Table 1b shows the result of model selection among a series of linear mixed models in which the outcome variable was question accuracy and possible factors included core vocabulary Type and Size. The two best-fitting models both contained Size, with increasing size significantly associated with higher accuracy.

We examine the model containing Type as well as Size (`M2typeSize`) in order to investigate the role of core vocabulary. Compared to INS (the reference category), WF had significantly higher accuracy, $b = 0.04, [0.02, 0.06]$ and AOA had significantly lower accuracy, $b = -0.05, [-0.07, -0.03]$. The intraclass correlations (ICCs) also showed that there was more variation due to individual questions within texts (ICC = 0.374) than between the texts as a whole (ICC = 0.096); this indicates that each text had questions that ranged substantially in difficulty. Overall, these results suggest that WF core words provide the most information for any given vocabulary size and AOA words provide the least.

Role of coverage As we did with gist accuracy, we evaluated the role of coverage by comparing a set of linear mixed models, this time with question accuracy as the outcome variable. Table 1c shows that, as before, the two best-fitting models both contained Coverage. In the model with Type as well as Coverage (`M3typeCov`), greater Coverage led to better comprehension question accuracy, $b = 1.57$, $[1.47, 1.67]$. Just as for gist accuracy, there were similar effects of vocabulary Type beyond the effect of Coverage (see Figure 6). Compared to INS, both WF and CC had significantly lower question ac-

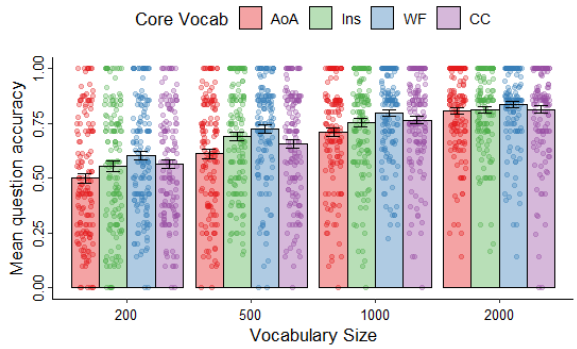

Figure 5: **Mean accuracy for questions by condition.** Each dot represents one question for a text in a given Type×Size condition. The mean accuracy on that question (*y* axis) is calculated by averaging over all trials and participants. The WF core words led to higher accuracy, and AoA words led to the lowest.

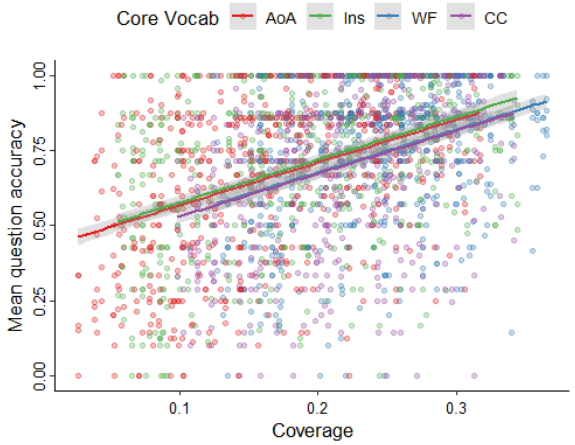

Figure 6: **Mean accuracy for questions by coverage.** Each dot represents one question for a text in each condition, whose mean accuracy is shown on the *y* axis and whose coverage is calcuated as the proportion of its words in a given core vocabulary (*x* axis). The regression lines for each core vocabulary type from the model with Coverage and Type as predictors are shown. Coverage is the most important factor affecting gist accuracy, but the INS and AoA core words resulted in better performance for the same level of coverage, compared to WF and CC.

curacy when coverage was controlled for (WF: $b = -0.05$, $[-0.07, -0.03]$; CC: $b = -0.05$, $[-0.07, -0.03]$). As before, the results suggested that INS and AoA core words yielded more information for the same coverage.

Discussion

In this study we compared different types of core vocabularies based on how well they facilitate comprehension of narrative texts. The INS and WF core vocabularies best enabled participants to summarise the gist, relative to the size of the core vocabulary. However, it was the INS and AoA core vocabularies that yielded the most information about gist for the same amount of coverage. When it came to more specific details about each story, assessed using comprehension questions, the WF core vocabulary was superior for a given vocabulary size. However, when controlling for coverage, the results were the same as for gist accuracy, with the INS and AoA core vocabularies faring best.

These results are consistent with our previous work looking at topic identification for Wikipedia articles (Wang et al., 2024). In both this and the previous study, we found that the INS and AoA core vocabularies provided more information once coverage was controlled for. In contrast to our previous results, however, this time the WF core vocabulary provided more information for a given vocabulary size; it was on par with the INS core vocabulary for gist accuracy, and outperformed all others on the comprehension questions. In combination with the fact that it did *not* perform well relative to the amount of text coverage it provided, this suggests that the utility of WF is likely because it contains commonly-occurring words. As a result of this, it provides a lot of information per text; however, each *instance* of a WF core word is less useful than each instance of an INS or AoA core word.

The pattern of results for AoA was almost the opposite of WF: each AoA core word provides relatively more information for a given coverage, but there are relatively few AoA words in a text for each given vocabulary size. This may be because early-acquired words include many child-related words (e.g., *mommy*, *potty*) which are not usually attested in everyday adult language use: as we saw in Figure 2, AoA often had the lowest coverage. However, early-learned words also include many useful and semantically basic items, like *people* or *animal*, which generally provide a lot of information. Consequently, as a *set* the AoA words may not be as useful, but individual *instances* of AoA words can have very high informational value when they do occur.

Finally, the INS core vocabulary seems to be a good "all-rounder" – it provided a lot of information for a given coverage regardless of whether gist or comprehension accuracy was assessed. It was also on par with WF words in terms of gist accuracy for a given vocabulary size. This may be because INS core words, being the most central words in word association networks, represent words that are semantically highly important and general and which can therefore provide a lot of information about the underlying ideas within texts. This semantically basic quality of INS core words may also explain why they fared better in terms of gist accuracy, where a more vague or general representation of the situation may be sufficient, rather than more specific details about the text, where WF core words did relatively better.

Overall, these results suggest that successful communication requires not just sufficient coverage of the intended meaning but also using words that individually provide the most information about the meaning. In other words, the semantics of the words, and how they relate to the situation models that people construct during comprehension, is an important factor, especially when it comes to understanding the overall gist. Our work highlights the importance of taking into account the process by which people construct mental representations from language.

References

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.

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., 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.

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403-430.

Kamalski, J. (2004). How to measure the situation model. *Yearbook Utrecht Institute of Linguistics*, 121–134.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109-128.

Kim, B., Kim, H., & Kim, G. (2019). Abstractive summarization of Reddit posts with multi-level memory networks. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 2519–2531).

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, *95*(2), 163.

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

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017, September). RACE: Large-scale ReAding comprehension dataset from examinations. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 785–794).

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

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

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, *50*(2), 212–226.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, *95*, 26-43.

Wang, A., De Deyne, S., McKague, M., & Perfors, A. (2024). Are the most frequent words the most useful? Investigating core vocabulary in reading. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, *123*(2), 162.