Are the most frequent words the most useful? Investigating core vocabulary in reading

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 text coverage and comprehension may not be straightforward – some words may provide more information than others. In this study, we explore alternative methods of defining *core vocabulary* in addition to word frequency (e.g., words that are central hubs in semantic association networks). We report on the results of an empirical test of communicative utility using a text-based guessing game. We show that core words that reflect corpus-based distributional statistics (like frequency or co-occurrence centrality) were less useful for communication than others. This was evident both in terms of the size of the vocabulary that must be known *and* the proportion of the text that must be covered for successful communication.

Keywords: core vocabulary; communication; word frequency; co-occurrence; word associations; semantic representation; information theory;

Introduction

What are the most useful words for communication? If only a limited number of words could be used, which would achieve the greatest communicative success? Answers to these questions, which centre around the idea of a "core vocabulary", have much theoretical and practical significance. For example, when learning a language, it is important to focus on the words that will be most useful to people – not just because of the practical every-day utility, but because learning begets learning, and early success can speed later acquisition. An appropriate core vocabulary is also crucial for applications like simplifying complicated texts for different audiences (Siddharthan, 2014).

Within applied linguistics, much work is centred around the assumption that word frequency is a good measure of the most useful words for communication, as, by definition, they provide the greatest text coverage (e.g., Nation & Waring, 1997). One of the key questions in this research is what proportion of the words in a text needs to be known for adequate comprehension to occur. Many agree that between 95-98% coverage allows most learners to sufficiently understand the text (Hu & Nation, 2000; Schmitt, Jiang, & Grabe, 2011); exactly how many of the most frequent words are needed to achieve this varies by domain, register, and genre (Nation, 2006; Schmitt et al., 2017).

However, coverage may not be the main thing that matters for communication – some words in a text may play a more important role than others in terms of conveying the intended message; in other words, they may provide more *information*. If so, such words may not necessarily be those that are important in a distributional sense (such as high-frequency words, e.g., *go* or *thing*), which occur often, but tend to be semantically empty. Instead, they may be those that are important in terms of semantic content (e.g., *person* or *happy*). In this paper, we present an empirical test of this idea. We first outline several distinct ways of quantifying "core vocabulary". We then present an information-theoretic framework for evaluating the communicative utility of the core vocabulary sets, and test these using a text-based gist guessing game.

Conceptions of core vocabulary

From a broader perspective, word frequency (WF) captures just one aspect of *distributional information* in the linguistic environment – that is, how words are used, and what words they tend to be used with, in natural language. Thus, in addition to defining core vocabulary using word frequency (i.e., based on individual words), we can also consider how words co-occur with each other. Communication, after all, does not involve using words in isolation, but rather combining words *together* to express an idea. This suggests that another plausible way of identifying the most useful words is to look at the words that occur most often with other words, which we call co-occurrence centrality (CC). Both word frequency (WF) and co-occurrence (CC) measure the distributional statistics of language use on some level, providing high text coverage and, potentially, high communicative utility.

However, as argued above, communicative utility may also be reflected in semantic rather than distributional importance. An alternative way of defining core vocabulary, then, is based on the words that are central to people's mental representations of word meaning. To do this, we rely on semantic networks derived from word associations (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019). The underlying idea behind these networks is that words that are linked to a greater number of other words or hold more central positions in the network are more semantically prominent; thus, semantic networks provide a window into how word meaning is organised and structured in the mental lexicon. A straightforward way to identify core words in semantic networks is by calculating different network centrality measures. One such measure is in-strength (INS), which identifies hubs that connect lots of

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.*

different words; words that have higher in-strength are connected to more words, and thus might be more central.

An alternative way of identifying representationally central words is by considering how semantic networks develop over time. The preferential attachment hypothesis proposes that networks grow by attaching new words to existing ones. As such, early-acquired words serve as an anchor for new knowledge (Brysbaert, Van Wijnendaele, & De Deyne, 2000; Hills, Maouene, Maouene, Sheya, & Smith, 2009; Steyvers & Tenenbaum, 2005). This suggests that age-of-acquisition (AoA) might be a good measure to identify central words in mental representation. In previous work, we have shown that Ins and AoA words are more representationally central in terms of relationships between words (Wang, De Deyne, McKague, & Perfors, 2022) and word meaning derived from context (Wang, De Deyne, McKague, & Perfors, 2024).

Beyond coverage alone

To properly evaluate what kind of words are most useful for communication, it is necessary to be clear about what is meant by communication. From an information-theoretic perspective, communication can be defined as the transfer of *information* from one point to another (Shannon & Weaver, 1949). Communication is fundamentally limited: in using language, speakers' ultimate goal is to convey a certain idea to another person, but they cannot do so directly, mind-to-mind – they must rely on language and words. There will thus always be a certain amount of error between the speaker's intended message and the message that is interpreted by a recipient. Uncertainty in the original message can only be reduced but never eliminated; this reduction of uncertainty is information. Successful communication thus occurs to the extent that the meaning recovered by the recipient is the same as the intended meaning by the source.

We can therefore evaluate the communicative utility of words based on the extent to which they support the accurate reconstruction of the intended meaning. Of course, some words carry more information than others: in the sentence, "In Rome, I wanted to go to the Colosseum", the words *wanted* and *go* do not provide as much information as the specific place, *Colosseum*. Informativeness also depends on the surrounding words: *Colosseum* is much less surprising if one knows that *Rome* is also in the sentence. Thus, informativeness is a complex function of intended meaning and context. The question of what words are most useful for communication, then, can be reframed as what kind of words, or set of words working together, generally provide the most *information*, in terms of accurately conveying the intended meaning?

From this perspective, high-frequency words may be less useful: part of why they are frequent is that they tend to be polysemous, with many different and fuzzy senses (Tragel, 2001). They also tend to have depleted usages, drawing their meaning mostly from the surrounding context in which they are used, thereby contributing little independent information to the sentence (Jorgensen, 1990). Hence, even though high-frequency or co-occurrence centrality words provide the greatest text coverage, they may not necessarily be the words that provide the most information in a text. High word association in-strength or early-acquired words, on the other hand, may provide more information for less coverage.

Leaving aside the issue of text coverage altogether, we can also consider how the size of the vocabulary affects informativeness. The amount of information that a set of words can provide necessarily depends on the size of the set: the more words there are in the set, the more informational ground they can cover as a whole. In work looking at the evolution of semantic category systems across cultures (Kemp & Regier, 2012; Kemp, Xu, & Regier, 2018; Regier, Kemp, & Kay, 2015), this problem is formulated as a trade-off between the competing factors of *informativeness* and *simplicity*. Informativeness, as discussed previously, is the amount of information that a given system can provide. Simplicity, in this case, refers to the number of words in the system. Communicative systems support *efficient communication* to the extent that they optimally trade-off informativeness and simplicity.

A similar trade-off exists for core vocabulary. The smaller the vocabulary size, the easier it will be for speakers to learn and use, but the informativeness of the set as a whole will suffer. The larger the size of the core vocabulary, the more information can be accurately conveyed, but the harder it is to learn. An optimally useful core vocabulary, therefore, should provide the *most* information from the *smallest* number of words. Taking these considerations into account, our question becomes: what set of core words enables the most efficient communication, in terms of being maximally informative while being minimal in size?

The aim of the study is to compare different types of core vocabulary in terms of how well they facilitate communication; that is, which types of core vocabulary provide the most information? We have two main questions. Firstly, which types of core vocabulary are most communicatively efficient, in the sense that they provide the most information for the smallest size? And secondly, given that different types of core words provide more or less text coverage, does the amount of information afforded by a given amount of coverage vary between different types of core vocabulary?

Method

Participants

214 people (21-77 years, $M = 43.7$; 43% female) were recruited via Prolific. 92% were native English speakers. 29 people were excluded for not passing the pre-registered[1] attention check, leaving 185 people in the analyses.

Conditions

There were four within-participant conditions, each corresponding to a different core vocabulary list. Each list reflects the top 1000 words ranked according to the four different measures being compared in this study. Because our interest is in lexical concepts, function words (including deter-

[1] https://aspredicted.org/B9G_VLP

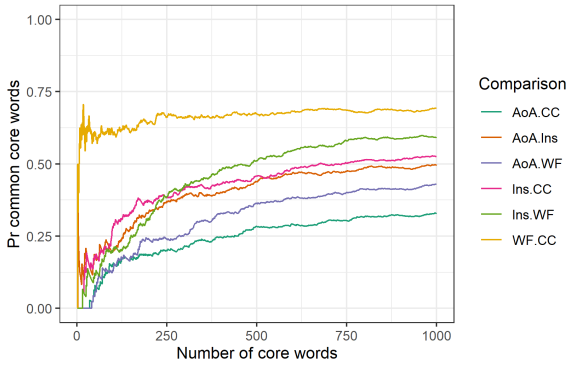

Figure 1: **Overlap between core vocabulary lists.** The *y*-axis shows the proportion of words shared between all possible pairs of core vocabulary lists as a function of the number of core words (*x*-axis). The two distributional measures, WF and CC (orange line) show the most overlap, but still contain over 25% of unique words. AoA overlaps with the distributional measures least.

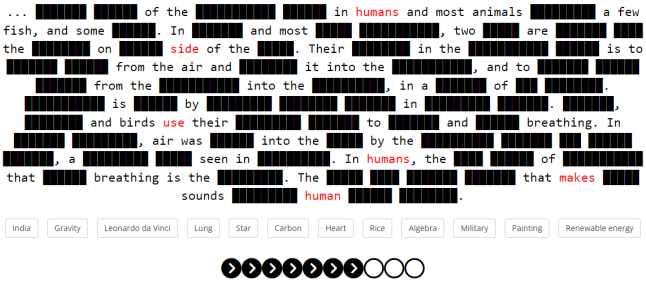

Figure 2: **Screenshot from example trial.** On each trial, people saw a single text showing only function words and core words. Their goal was to guess the topic of the article from the 12 possible answers at the bottom. After each guess, another increment of that core word list was shown (with new words in red). This is the 7th of 10 increments; for this increment, the new words are *human*, *side*, *use*, and *make*. In this trial, the correct answer was *lung*.

miners, auxiliary verbs, prepositions, conjunctions, and pronouns) were excluded from the core vocabulary lists. Moreover, words for all lists were lemmatised by grouping word forms under the same lemma (e.g., *want*, *wants*, *wanted*).

The in-strength (INS) measure, which captures the words that are most central in people's lexical representations, was computed over word associations to over 12,000 English words (De Deyne et al., 2019). The in-strength of a word is the sum of the weights of all incoming edges directed to that word, where edge weights represent the strength of association between words; words with higher in-strength (e.g., *love*, *food*) are connected to more words and are more core.

The AOA measure, which reflects words that are learned first, was sourced from the Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) norms. Words like *mom* or *wet* that were acquired earlier in life are more core according to AOA.

The word frequency measure (WF) reflects one kind of distributional information, that is, data about which words are used most often. WF data was based on the SUBTLEX database (Brysbaert & New, 2009), with more frequent words like *know* or *good* being more core.

Our final measure is co-occurrence centrality (CC), calculated based on the Corpus of Contemporary American English (Davies, 2008-). Like WF, it captures distributional information; unlike WF, the information is less about raw frequency and more about which words occur most often with other words. CC was computed as strength centrality for co-occurrences. Co-occurrence strengths were computed by normalizing the raw co-occurrence counts as a proportion of word frequency, and then CC was calculated for each word as the sum over its co-occurrence strengths. This yields a measure directly analogous to INS; the calculation is the same, but over distributional co-occurrence data rather than word associations. Words that are more core on the CC measure include *time* and *people*.

As Figure 1 shows, there is some overlap in the words on each core vocabulary list. The WF and CC lists share the most, but even there, 25% or more of the words are unique regardless of core vocabulary size. The overlap between other lists is much less. This variation suggests that core words from different lists may differ in the information they convey as well as how they are distributed within naturalistic text.

Procedure

Participants played a game where they read extracts from the beginnings of Wikipedia articles. We chose Wikipedia because it is a general-purpose and easily accessible repository of general knowledge informational content. Articles were presented with only core words shown and other words masked, and participants had to guess the topic of the article (see Figure 2). The topics corresponded to the article titles (e.g., *Gravity* or *The Renaissance*), and the beginning of the article typically provided a general description of that topic.

Each text was demasked incrementally, with the increments controlling the set size of the core vocabulary shown. At the first increment (0 core words), only function words were shown; in the next increment, the top 50 core words from a given list that were contained in the text were shown, and so forth, through the top 100, 200, 300, 500, 750, and 1000 core words (eight increments total). This allowed us to investigate how the core vocabulary size was associated with accurately identifying the article's topic (or gist).

At each increment, the newly revealed words were shown in red, and people made one guess at the topic by selecting from a set of 12 possible topics (described below). New words were added at each increment until the topic was correctly guessed or the maximum number of guesses had been reached. To make the task more challenging, the selection of the correct answer was not counted as correct until it was selected three consecutive times. This also ensured that an accidental selection of the correct answer would not artificially inflate a person's accuracy. The final increment (1000 core words) was repeated an additional two times more times to allow participants to be successful on the last two increments (for a maximum of 10 guesses per trial).

After viewing instructions and completing a practice trial, each person completed 24 trials. Each trial corresponded to

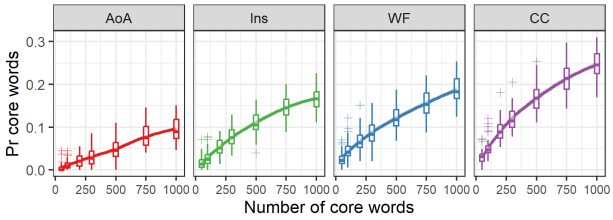

Figure 3: **Text coverage given by each core word list.** The proportion of each text covered by the core words (y axis) is shown as a function of increment (number of core words; x axis). Boxplots show the coverage distribution at each increment and LOESS regression lines show the trend of coverage across increments. The CC words achieve the highest coverage, and AoA words the least.

one text in which core words from one of the four conditions were revealed over successive increments. The texts were randomly selected from the set of text stimuli (see below), and were randomly divided into the four core word conditions so that there were six trials per condition.

Stimuli

All texts were selected from the Wikipedia vital articles[2], which is a classification of what are considered to be the most important articles across a broad range of categories: we selected articles from *People*, *Geography*, *Arts*, *Philosophy and Religion*, *Everyday life*, *Society*, *Health*, *Science*, *Technology*, and *Mathematics*. From these categories, we selected articles that captured a wide variety of general-interest topics. For each article, the first paragraphs before the first subheading were extracted (up to 300 words), and the beginning part of the entry (containing the title) was replaced with an ellipsis.

Four versions of each of these texts occurred in the experiment (one for each condition). The increment-0 text, which was the same for each condition, was created by replacing everything aside from function words and punctuation with blanks. Subsequent increments were created by revealing the core words of the rank given by that increment, for each condition (e.g., increment 50 for WF would reveal the top 50 items on the WF core word list in the text).

As one would expect, different texts contained different specific core words and a different degree of *coverage*: some texts contained many words from some lists, and some contained few. As Figure 3 shows, the coverage given by each of the full 1000 core word lists did vary substantially between the conditions. CC core words gave the most coverage, followed by WF, then INS, and AoA.

We opted *not* to control for coverage, because the natural variation in the amount of coverage afforded by the different types of core vocabulary is one of the important features. To select texts that reflected this natural variation, we chose texts whose coverage (the proportion of the text covered by the 1000 core words) was within 1.5 standard deviations of the mean for all four core word lists. Articles that had titles

[2]Articles consisted of cleaned-text versions of the Wikipedia Vital articles (https://en.wikipedia.org/wiki/Wikipedia:Vital_articles) sourced from Hugging Face (https://huggingface.co/datasets/wikipedia).

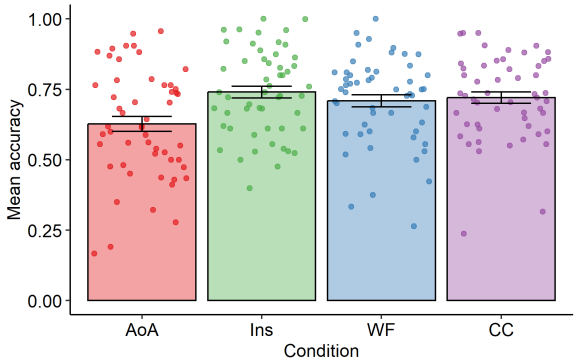

Figure 4: **Mean accuracy for texts by condition.** Each dot represents one text in a given condition whose mean accuracy (y axis) is calculated by averaging over all trials and participants. INS, WF, and CC core words made texts significantly easier to guess compared to AoA.

containing any of the 1000 core words, did not clearly describe the topic, or were not likely to be familiar to a general audience were also not selected. In the end, 53 articles were selected (between 3-7 articles from each category).[3]

The 12 answer choices for each text were the same in all four core word conditions. Choices were selected from the titles of the 1000 Wikipedia vital articles. Of the 12 possible answers people could select, one corresponded to the correct answer (the title of the target article), one was the title of an article in the same subcategory as the target article, three were titles from the same category but a different subcategory, and seven were titles drawn from different categories. The order of the answer choices was randomised for each participant.

Results

Overall accuracy

Accurate identification of the topic of the text on a given trial was counted as choosing the correct answer choice (out of 12 possibilities) three times in a row. The mean accuracy of each text for each core word list was computed by averaging across all trials from all participants. As Figure 4 reveals, the topic of the texts was correctly identified an average of around 75% of the time in all of the conditions except AoA, where accuracy was slightly lower (63%). However, there was substantial variation across texts, with some almost always guessed correctly, and a few identified less than 25% of the time.

A one-way repeated measures ANOVA revealed significant differences between conditions, $F(2.56, 133.04) = 8.27, p < .001$. Post-hoc pairwise comparisons with Holm corrections indicated that texts in the AoA condition had significantly lower mean accuracy than the INS ($p < .001$), WF ($p = .029$), and CC ($p = .003$) conditions. Even though the WF and CC core words had significantly greater text coverage than the INS core words, their accuracy was not significantly better.

[3]A one-way repeated measures ANOVA confirmed that the mean text coverage on the texts we chose significantly differed between all four conditions, $F(2.36, 122.73) = 396.62, p < .001$ (all Holm-corrected pairwise ps $< .001$).

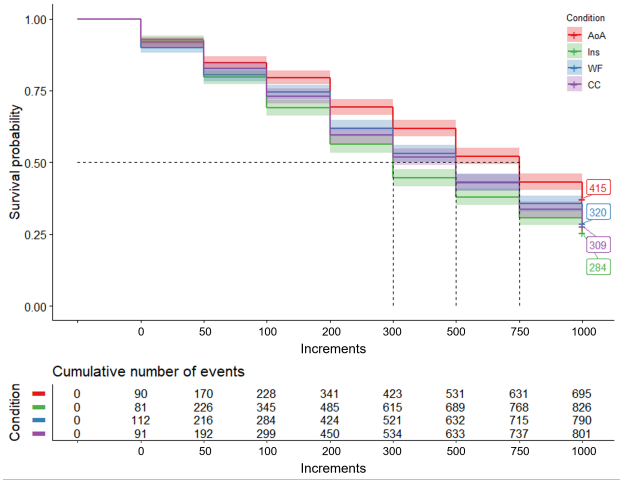

Figure 5: **Kaplan-Meier survival analysis.** A text "survives" at a given increment if its topic is not guessed. Higher survival indicates lower accuracy. Lines indicate the median survival for each condition across increments, with 95% confidence intervals. The boxed label shows the number of unidentified trials after the final guess, and the table at the bottom shows the cumulative number of trials accurately identified at each increment. AoA texts were guessed the slowest, while Ins ones were guessed most quickly.

Accuracy over increments

As well as analysing overall accuracy, we also explored whether conditions differed in how quickly people were able to identify the topic. We counted the increment which led to successful identification as the increment corresponding to the first of the three consecutive times on which the correct answer was identified. Was there a difference across conditions in terms of the size of the core vocabulary needed for accurate identification of the text? We addressed this question using a Kaplan-Meier survival analysis on the occurrences of correct responses over each successive increment, with unguessed trials right-censored at the final increment.

As Figure 5 shows, texts appeared to be guessed more slowly in the AoA condition, with a median survival of increment 750; this means that half of the texts needed 750 or more AoA core words to be correctly identified. By contrast, texts in the Ins condition were guessed quickest, with a median survival indicating that only 300 Ins core words were needed to accurately identify at least half of the articles. In the middle were the WF and CC conditions, with a median increment of size 500. These survival times were significantly different according to a log-rank test comparing the survival curves, $\chi^2 = 52.9, p < .001$. Pairwise comparisons with Holm corrections showed that survival was significantly lower (i.e., texts were guessed faster) in the Ins, WF, and CC conditions compared to AoA (all ps $< .001$). Survival was also significantly lower for Ins compared to WF, $p = .044$.

Overall, then, performance was better in the Ins, WF, and CC conditions compared to the AoA condition. Additionally, Ins core words allowed for the most accurate identification for the smallest size, while AoA core words needed a comparatively larger size for accurate identification.

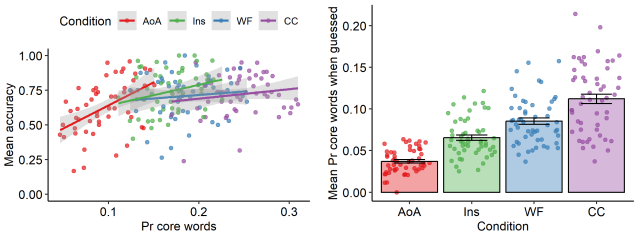

Figure 6: **Relationship between core word coverage and performance.** (a) *Left panel*. Linear regression lines for each condition predicting mean text accuracy (y axis) from core word coverage (x axis), measured as the proportion of each text covered by 1000 core words. Ins and AoA core words had higher accuracy compared to the WF and CC core words, beyond what core word coverage alone would suggest. (b) *Right panel*. For each text, the average core word coverage required for successful identification of topic (y axis). Relatively lower coverage from AoA and Ins core words was required to correctly guess a text compared to the amount of coverage required from WF and CC core words.

The role of coverage

The results so far show that the different types of core vocabulary provide different amounts of information for different sizes. But as Figure 3 shows, the amount of text coverage provided by a given vocabulary size varies widely between the different types of core words. Therefore, we also ask whether the different types of core vocabulary provide different amounts of information for the same level of coverage – do they provide different amounts of information *per word*?

We addressed this question with a linear mixed model whose outcome variable was mean text `accuracy` and whose predictor variables were `coverage` and `condition`, with `text` included as a random effect. As Figure 6a shows, core word coverage significantly predicted accuracy, $b = 1.20$, 95% CI = [0.48, 1.92]: texts with a higher proportion of core words were guessed more easily. However, there were condition differences over and above the effect of coverage: accuracy was significantly higher for Ins core words (the reference category) than for both WF, $b = -0.06$, 95% CI = [-0.11, -0.006], and CC core words, $b = -0.11$, 95% CI = [-0.19, -0.04], but not compared to AoA core words. That means, for the same amount of text coverage, the Ins and AoA core words were more successful at communicating the meaning of the text, compared to the WF and CC core words.

The above analysis measures coverage for each text based on the full 1000-item core word list, but in many cases a text was guessed correctly before all core words were seen. To account for this, we used a linear mixed model analysis in which the outcome variable was, for each trial, the amount of `coverage` at the increment at which the correct answer was identified – that is, the amount of coverage that was *needed* for accurate identification.[4] The predictor was `condition`, with `text` and `participant` included as crossed random factors. The linear model showed significant differences between conditions, with the Ins condition (the reference category) needing less core word coverage for a successful guess compared to both the WF, $b = 0.02$, 95% CI

[4]Trials that did not result in a successful guess were excluded.

= [0.015, 0.026], and CC conditions, $b = 0.048$, 95% CI = [0.043, 0.054]. However, it needed more coverage for accurate identification than the AoA condition, $b = -0.024$, 95% CI = [-0.03, -0.019].

We conducted a similar analysis on the level of the texts, by averaging the core word coverage needed across all trials for each text in each condition (Figure 6b). A one-way repeated measures ANOVA showed significant differences between conditions, $F(2.17, 113.09) = 95.06, p < .001$, with all post-hoc pairwise comparisons significant (all ps $< .001$, Holm corrections). This indicates that relatively less coverage was required to correctly guess the topic for AoA core words, followed by Ins. Accurate identification required relatively higher coverage of WF and especially CC core words.

Discussion

This study aimed to compare different conceptualisations of core vocabulary in terms of their usefulness for communicating the gist of short texts. Accuracy and survival analyses showed that Ins core words provided the most information for the smallest size, whilst AoA core words required the largest size vocabulary to convey the same amount of information. Investigating the coverage given by each type of core vocabulary revealed that AoA core words, and Ins core words to a lesser extent, provided the most information for the same amount of textual coverage.

Measures of distributional information such as word frequency (WF) or co-occurrence centrality (CC) have been argued to be a good way of identifying the most useful words for communication, as they provide the greatest text coverage (Nation & Waring, 1997). However, the current results suggest that coverage is not the only – or even the main – thing that matters for communication: *semantic* factors matter, too. Accordingly, words that are central to mental representations of word meaning, such as those that are central in word association networks (Ins) or acquired early in life (AoA), provided *more information* than WF and CC core words. In particular, we found that Ins core words especially, compared to WF and CC core words, provided more information both in terms of the size of vocabulary and the amount of textual coverage required to accurately convey an idea.

AoA core words, interestingly, while providing more information for less coverage, were counter-intuitively *less* communicatively efficient in the sense that they provided less information for the same vocabulary size. One possible explanation for this is that the AoA core vocabulary contains a lot of words that are highly relevant for young children (e.g., *potty*, *doll*), which are much less communicatively useful for adults. But it also contains a smaller number of words that have very high informational value, such as *people*, *animal*, *place*, number words, and useful descriptors like *big* and *small* (these tend to be common in other core word lists, especially Ins). This may reflect the way in which children learn fundamental, basic concepts early in life, and these serve as a basis on which new word meanings can be built (Brysbaert et al., 2000; Steyvers & Tenenbaum, 2005). It is interesting to note that this effect of AoA core words persisted even for encyclopedic texts, in which the subject matter and linguistic complexity are quite distinct from child-directed speech.

One limitation of this study is that we only looked at one genre and source of texts: expository content using Wikipedia articles. We did aim to ensure that the chosen articles covered a broad variety of domains, which were likely to draw on a wide range of knowledge, and found no consistent differences across domain. However, our general framework can be applied to any kind of communication, including different genres (e.g., narratives, news articles), modalities (written and spoken), and styles (e.g., dialogues). We aim to investigate these in future work.

There are several important differences between our study and work in applied linguistics that investigates coverage and comprehension – differences in both the aim of the research and the definition of "communicatively useful". While we evaluated the ability of the core words to communicate a very simple idea (the overall topic or gist), much of this other work investigates what is required for *adequate comprehension*. This involves a much deeper understanding of the meaning of a text. Hu and Nation (2000), for instance, assessed learners' comprehension using multiple-choice questions and cued written recall of the details of the story, with a certain bar required to be met for "sufficient understanding" which goes far beyond the ability to simply guess the topic.

Secondly, the number of core words and the level of text coverage being investigated are very different between our work and much of the applied linguistics literature. In it, 95% and 98% coverage are extremely important and oft-cited figures, with between 3000-9000 of the most frequent word families required to achieve that level of coverage (Nation, 2006; Schmitt et al., 2011). By contrast, we only investigated up to 1000 core words for each type of core vocabulary, and the amount of text coverage given by the core words in our experiment was much less than 95-98% (though this was still sufficient for identifying the gist of most of the articles).

Ultimately, all language learners have to start somewhere, and different vocabulary sizes and levels of coverage, as well as different levels of comprehension, are just quantitatively different points along the spectrum of the same qualitative goals. With those considerations, our results suggest that learners may benefit by starting out by learning words that are like Ins and AoA core words, as they can provide higher communicative value. In future work, we will also aim to extend the scope of the vocabulary size we investigate and the method of assessing comprehension.

A core vocabulary for communication has high practical utility for learners, educators, and publishers, as well as great theoretical significance in how it relates to issues such as semantic representation. We have shown that there is more to communicative usefulness than frequency or distributional information, and have identified other potential and promising ways that can be used to identify a core vocabulary.

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.

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

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

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

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049-1054.

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

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

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.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. *The Handbook of Language Emergence*, 237-263.

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.

Shannon, C. E., & Weaver, W. (1949). A mathematical model of communication. *Urbana: University of Illinois Press*.

Siddharthan, A. (2014). A survey of research on text simplification. *International Journal of Applied Linguistics*, *165*(2), 259–298.

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

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

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).

Wang, A., De Deyne, S., McKague, M., & Perfors, A. (2024). Word prediction is more than just predictability: An investigation of core vocabulary. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).