Why are Some Words More Frequent than Others? New Insights from Network Science

Qiawen Liu*[1], Simon De Deyne[2], and Gary Lupyan[1]

*Corresponding Author: qliu295@wisc.edu
[1]Department of Psychology, University of Wisconsin-Madison, Madison, United States
[2]School of Psychological Sciences, University of Melbourne, Melbourne, Australia

Why are some words more frequent than others? Surprisingly, the obvious answers to this seemingly simple question, e.g., that frequent words reflect greater communicative needs, are either wrong or incomplete. We show that a word's frequency is strongly associated with its position in a semantic association network. More centrally located words are more frequent. But is a word's centrality in a network merely a reflection of *inherent* centrality of the word's meaning? Through cross-linguistic comparisons, we found that differences in the frequency of translation-equivalents are predicted by differences in the word's network structures in the different languages. Specifically, frequency was linked to how many connections a word had and to its capacity to bridge words that are typically not linked. This hints that a word's frequency (and with it, its meaning) may change as a function of the word's association with other words.

1. Introduction

Word frequencies are often used as a key predictor in studies of word recognition (Brysbaert et al., 2016; Ferrand et al., 2010; Keuleers et al., 2012), comprehension (Just & Carpenter, 1980; Halgren & Smith, 1987), production (Oldfield & Wingfield, 1965; Jescheniak & Levelt, 1994; Alario et al., 2004), recall (Arndt & Reder, 2002; Clark, 1992; Gregg, 1976; Meier et al., 2013; Yonelinas, 2002), and learning (Braginsky et al., 2019). While much is known about what word frequency predicts, much less is known about what predicts word frequencies (e.g., Calude & Pagel, 2014, 2011; Liu et al., 2023). Why are some words more frequent than others? Why do some words become more frequent while others become less frequent over time? How similar are frequencies of translation equivalents across languages and what does it mean if a word denoting a certain meaning is more frequent in one language than in another?

The question of why some words are more frequent than others suggests some obvious answers. One is that more frequent words denote meanings that are more important for people's goals and needs. 'Water' is more frequent than 'lamp','matrix', or 'abracadabra' because it is more important. This explanation only goes so far, however because important meanings are fragmented into more basic terms. Presumably 'mammals' are more important than 'dogs' or 'cats' yet

the frequency of 'mammal' is a small fraction of either of those more basic terms (in fact 'dog' is about as frequent as 'animal'). Another possibility is that more prototypical referents have higher frequencies (e.g., 'robin' compared to 'penguin') because they more closely correspond with what speakers have in mind (Rosch et al., 1976). However, what counts as a prototype can vary significantly based on context, making prototypicality an inconsistent predictor of word frequency. For example, prototypicality as a bird may explain why 'robin' is more frequent than 'penguin', but not why 'chicken' is more frequent than 'robin'. It is also possible that word frequencies might mirror the prevalence of certain objects in our surroundings. But discrepancies arise here too. 'Red' is the most frequent chromatic term even though red objects are not more common. And explanations invoking ecological frequencies cannot explain the frequencies of abstract words that refer to intangibles. Explanations that stress communicative needs: 'frequent words denote things we most want to talk about' also run into problems. First, they simply push the question of word meanings to communicative need. 'Girl' is more frequent than 'boy', but do we really have a greater need to communicate about girls than boys? All these explanations also struggle with explaining why words with similar meanings are more frequent in some languages than others.

In a recent study, Liu et al. (2023) predicted word frequencies from properties of the words' semantic networks. To rule out idiosyncratic explanations such as importance and ecological frequency, they examined pairs of antonyms which would seem to have equal communicative value but often differ in word frequency (as the example of girl/boy above). After factoring out effects like morphological complexity and polysemy, the analysis revealed two network properties that predicted word frequency especially well: the number of connections the word and its associated words have, and the word's ability to bridge otherwise sparsely linked words. As further revealed by a longitudinal analysis, these network properties didn't seem to just correlate with current word frequencies. Instead, they also predicted the way the word's frequency changed in the subsequent decades, suggesting a causal role of network properties in explaining changes to word frequencies over time.

One alternative explanation is that the more frequent words in each antonym pair correspond to a default or unmarked state (Clark, 1992). For instance, 'good' in the pair good-bad is unmarked such that asking 'how good was it?' does not imply goodness, while asking 'how bad was it?' implies badness. Markedness-focused explanations make a simple prediction: if one end of a semantic dimension denoted by an antonym pair is inherently more central to communication and/or thinking (which leads to greater frequency of the associated word), then translation equivalents of these antonym pairs should show consistent frequency differences across languages. In the next section, we examined if translation equivalents of antonym pairs in Chinese and English display analogous frequency patterns. We then sought to replicate our earlier findings (2023) concerning the in-

fluence of network properties on word frequency using Chinese word-association data.

2. Effects of network centrality on Chinese word frequencies

2.1. *Materials*

We used the English and Chinese semantic association networks from the Small World of Words (SWOW) project. In this project, crowdsourced word association responses were gathered in various languages (De Deyne et al., 2019). Participants were shown target words and asked to list the first three words that came to mind. These associations then acted as cues for subsequent participants, generating further associations. This iterative method yielded a weighted network with directed edges. The edge direction signifies forward or backward associations, while the weight represents the likelihood of each association based on the response or cue. To focus on the most robust associations, we used only the first response and excluded responses provided by only a single respondent. The Chinese SWOW network had 21434 words and 78057 directional associative links.

In Liu et al. (2023), 774 antonym pairs of English words were extracted from WordNet (Fellbaum, 1998). Among them, 661 pairs were considered having appropriate Chinese translations, and were translated by a professional translator, resulting in 761 Chinese translation equivalents as some words had multiple valid translations. Word frequencies for both English and Chinese pairs were sourced from the Exquisite Corpus using the 'wordfreq' Python package (Speer, 2022).

2.2. *Variables*

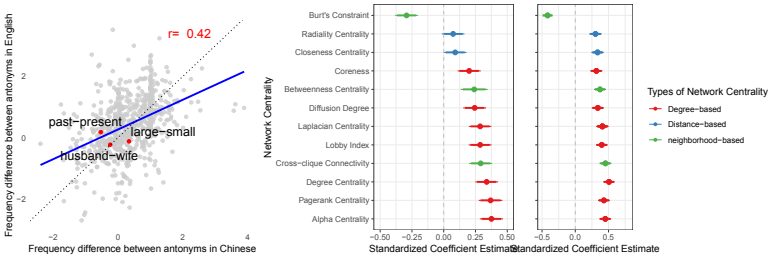
Using a linear regression model, we predicted the difference in Zipf frequency (calculated as the base-10 logarithm of occurrences per billion words) from different types of network centrality measures. These centralities are grouped into *degree-based* centralities which emphasize the number of connections a word and its neighbors have, *neighborhood-based* centralities which measure how well a word bridges between less-connected words, and *distance-based* centralities which consider words with short paths to others in the network as more central. We also included three covariates: the difference in morpheme count (more complex derived words may be less frequent), the differences in number of word senses (operationalized as Chinese Wordnet synsets) (Wang & Bond, 2013), and how often the word was mentioned as a cue, i.e., its frequency in SWOW. This allows us to discern the impact of the word's network above and beyond the frequency effect that some centrality measures may inevitably capture.

2.3. *Analysis & Result*

First, we examined the word frequency patterns between English and Chinese. As shown in Figure 2a, there's a moderate correlation ($r = .42$) between English-

Chinese differences in antonym pairs. This indicates that although there's a relationship in frequency patterns between the two languages, the prominence of a word in one language doesn't necessarily denote its inherent significance in meaning. For example, while 'small' is more frequent than 'large' in English, the Chinese counterpart xiǎo (small) is less frequent than dà (large); Do such frequency variations align with the word's network properties?

Because network centralities are highly inter-correlated, we regressed the difference in word frequency on each network centrality individually. We also controlled for differences in morpheme count, sense count, and SWOW frequency. In cases where an English antonym pair matched multiple Chinese antonym pairs, we averaged the measures for various translated pairs. Some words were absent from the Chinese WordNet or Chinese SWOW network, leaving us with 381 pairs for the analysis. We log-transformed notably skewed predictors. Figure 1b shows that, after controlling for other variables, degree-based, neighborhood-based, and distance-based network centralities all significantly predict frequencies (all significant at $\alpha = .01$, except closeness which significant at $\alpha = .05$). Radiality is a marginally significant predictor ($p = .1$). The predictions were all in the expected direction, with the only negative coefficient for Burt's constraint indicating that words with fewer redundant neighbors have higher frequencies, replicating Liu et al. (2023)'s results with the English SWOW network using the Chinese SWOW network. Words that are more associated with others, closer to others, and bridge otherwise less connected words are more frequent.



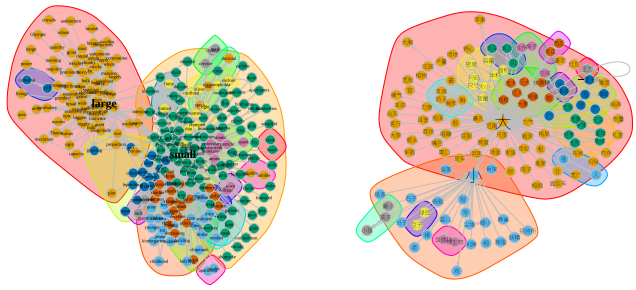
(a) (b) (c)

Figure 1.: (a) Which end of an antonym pair is more frequent is only moderately correlated between English and Chinese, e.g., 'past-present' favors 'present' in English and 'past' in Chinese. 'large-small' favors 'small' in English 'large' in Chinese. (b) Network centrality measures significantly predict Chinese word frequencies. (c) The differences between Chinese and English word frequencies for matched word pairs are predicted by differences in centrality measures.

3. Do cross-linguistic differences in network centralities predict cross-linguistic differences in word frequencies?

Are differences in word frequencies between English and Chinese associated with differences in the word's respective semantic networks? We predicted cross-linguistic differences in word frequencies between English and Chinese antonyms

from the cross-linguistic differences in network centralities between English and Chinese antonyms, controlling for the same covariates as above. As shown in Figure 1 c, all cross-linguistic differences in centralities remain significant predictors of cross-linguistic differences in word frequency ($p < .01$).



(a) (b)

Figure 2.: Subnetworks of a translation-equivalent word pair. In English (a), *small* is more frequent than *large* and is also connected to more words & diverse neighborhoods (colored clusters). (b) In Chinese *dà, large* is more frequent and central than *xiǎo, small*.

4. Discussion

We analyzed the cross-linguistic frequency differences between English and Chinese antonyms, replicating the previous finding from English, in Chinese. The results reveal that words with greater degree-based, neighborhood-based, and distance-based centrality are, on average, more frequent. Specifically, words with more neighbors, especially more influential neighbors as determined by measures like PageRank or alpha centrality (which assign weights based on the importance of neighbors), tend to be more frequent (Fig 1 b). Words bridging less connected areas (Burt's constraint, Cross-clique connectivity, Betweenness centrality) and those with shorter paths to other words (closeness centrality) also tend to be more frequent. Moreover, the frequency differences across these languages can be attributed to differences in the words' network centrality. Overall, we show that variations in word frequency can be linked to the structural properties of the semantic network rather than solely to the inherent conceptual prominence of their denoted meanings. Differences in association network dynamics may underpin language evolution and influence patterns of word usage across languages.

How do cross-linguistic differences in network centralities inform our understanding of cultural variations in word meanings? Returning to the previously used example of cross-linguistic differences between large and small, we show that the more frequent word in both cases, despite having opposite meanings, is more centrally located and is a better 'bridge' to other meanings (Fig 2). One hypothesis linking centrality and frequency is that a more centrally located word

has a higher base-level of activation during speech comprehension and production due to receiving more input from its neighbors, leading to a greater likelihood of a user producing it–a rich get richer type phenomenon.

Future research may further elucidate how fluctuations in network connectivity and cognitive accessibility influence the dynamics of competing synonyms (Karjus et al., 2020). In addition, words that connect otherwise less connected neighborhoods indicate they may have higher contextual diversity and larger semantic extensions compared to words that are surrounded by more redundant interconnected neighbors. Again, as shown in Figure 2, 'small' in English is more frequent and associates with a more diverse set of neighbors than 'large'; while this pattern is reversed in Chinese. For instance, the English 'small' is incorporated into phrases like 'small talk', helping to increase its opportunities for use compared to 'large' which lacks similar sense-extension. Conversely, in Chinese 'dà' (large) can refer to generality and lack of precision, as in 'dà gài' (probably or approximately), 'dàjú' (overall situation), 'dàyì' (careless), while 'xiǎo' (small) does not have analogous extension to higher certainty or precision.

Finally, what causes a word to occupy a more or less central location in a semantic network? Studies suggest that the structure of semantic networks is shaped by both external interactions with the environment (Hills et al., 2009b; Laurino et al., 2023) and internal linguistic structures (Steyvers & Tenenbaum, 2005). For example, the 'preferential acquisition hypothesis' (Hills et al., 2009a) suggests that children learn new words based on their prevalence and connections within the surrounding environment, independent of their existing vocabulary. On the other hand, the 'preferential attachment' theory (Steyvers & Tenenbaum, 2005) suggests that new words are more likely to be integrated into a child's vocabulary if they connect to already well-connected words, reinforcing the significance of these central nodes. Given that individual-level cognitive selection can predict global language change (Li et al., 2024), a word's centrality might stem from its relevance in real-world contexts and its connectivity within the language structure.

Many questions remain. What is the causal direction between network centrality and frequency we observe here? Although Liu et al., 2023 found that changes in network centrality predicted subsequent changes in the words' frequencies, the changes in frequencies of the words in the sample were quite small for the tested time period. More insight can be gained from analysis of words that have undergone rapid changes in frequency. These are sometimes accompanied by shifts of meaning: The frequency of 'broadcast' hugely increased when its meaning shifted from sowing seeds by scattering, to radio and TV transmissions. It is difficult to tell how its semantic network changed at that time, but it is possible to study the semantic networks of words currently undergoing rapid changes in frequency such as those studied by Grieve, Nini, and Guo (2017). It is also possible to experimentally manipulate a word's location in a participant's semantic network to see if it causes changes in likelihood of production.

References

Alario, F. X., Ferrand, L., Laganaro, M., New, B., Frauenfelder, U. H., & Segui, J. (2004). Predictors of picture naming speed. *Behavior research methods, instruments, & computers*, *36*, 140–155.

Arndt, J., & Reder, L. M. (2002). Word Frequency and Receiver Operating Characteristic Curves in Recognition Memory: Evidence for a Dual-Process Interpretation. *Journal of experimental psychology. Learning, memory, and cognition*, *28*(5), 830–842.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and Variability in Children's Word Learning Across Languages. *Open Mind: Discoveries in Cognitive Science*, *3*, 52–67.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 441–458. (Place: US Publisher: American Psychological Association)

Calude, A. S., & Pagel, M. (2011). How do we use language? Shared patterns in the frequency of word use across 17 world languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1101–1107.

Calude, A. S., & Pagel, M. (2014). *Frequency of use and basic vocabulary.* John Benjamins Publishing Company.

Clark, S. E. (1992). Word frequency effects in associative and item recognition. *Memory & Cognition*, *20*(3), 231–243.

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*, 987–1006.

Fellbaum, C. (1998). *WordNet: An electronic lexical database and some of its applications.* MIT press Cambridge.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488–496.

Gregg, V. (1976). Word frequency, recognition and recall. In *Recall and recognition* (pp. x, 275–x, 275). Oxford, England: John Wiley & Sons.

Grieve, J., Nini, A., & Guo, D. (2017). Analyzing lexical emergence in Modern American English online1. *English Language & Linguistics*, *21*(1), 99–127.

Halgren, E., & Smith, M. (1987). Cognitive evoked potentials as modulatory processes in human memory formation and retrieval. *Human neurobiology*, *6*(2), 129–139.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009a). Lon-

gitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological science*, *20*(6), 729–739.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009b). Longitudinal Analysis of Early Semantic Networks: Preferential Attachment or Preferential Acquisition? *Psychological Science*, *20*(6), 729–739. (Publisher: SAGE Publications Inc)

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of experimental psychology: learning, Memory, and cognition*, *20*(4), 824.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, *87*(4), 329.

Karjus, A., Blythe, R. A., Kirby, S., & Smith, K. (2020). Communicative need modulates competition in language change. *arXiv preprint arXiv:2006.09277*.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304.

Laurino, J., De Deyne, S., Cabana, Á., & Kaczer, L. (2023). The pandemic in words: Tracking fast semantic changes via a large-scale word association task. *Open Mind*, *7*, 221–239.

Li, Y., Breithaupt, F., Hills, T., Lin, Z., Chen, Y., Siew, C. S., & Hertwig, R. (2024). How cognitive selection affects language change. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(1).

Liu, Q., De Deyne, S., Jiang, X., & Lupyan, G. (2023). Understanding the frequency of a word by its associates: A network perspective. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Meier, B., Rey-Mermet, A., Rothen, N., & Graf, P. (2013). Recognition memory across the lifespan: the impact of word frequency and study-test interval on estimates of familiarity and recollection. *Frontiers in Psychology*, *4*.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*(4), 273–281.

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

Speer, R. (2022, September). *rspeer/wordfreq: v3.0.* Zenodo.

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

Wang, S., & Bond, F. (2013). Building the Chinese open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th workshop on Asian language resources* (pp. 10–18). Nagoya, Japan: Asian Federation of Natural Language Processing.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517. (Place: Netherlands Publisher: Elsevier Science)