Psycholinguistic Correlates of Symbol Grounding in Dictionaries

Philippe Vincent-Lamarre[1], Alexandre Blondin-Massé[1], Mélanie Lord[1],
Marcos Lopes[2] & Stevan Harnad[1,3]
Université du Québec à Montréal[1], Universidade de São Paulo[2], University of Southampton[3]

UQÀM

Introduction

We can learn the meaning of a word from a definition. A definition is a string of further words, each of whose meanings we can likewise learn via definition. This recursive process can be repeated indefinitely, but it cannot convey meaning at all unless the meanings of some words, at least, are already known by some other means than verbal definition. This is called the *Symbol Grounding Problem:* the meanings of some words must be grounded in prior direct sensorimotor experience (Harnad 1990). How many words? And which ones?

A dictionary can be represented as a directed graph with links from defining to defined words. The *minimal feedback vertex sets* (*MinSets*, Ms) of a dictionary graph are the smallest sets of words from which all the rest can be defined. We computed Ms for four English dictionaries.

Data

Psycholinguistic data

Three psycholinguistic variables were used : **frequency** (SUBTLEXus corpus, Brysbaert & New, 2009), **age of acquisition** (Kuperman *et al.,* 2012) and **concreteness** (Brysbaert *et al.,* 2013).

Structures

- The **Kernel** (K) (~10% of the dictionary) is extracted by recursively removing all words that can be reached by definition but that do not define any further words.
- The **Rest** (R) (~90% of the dictionary) is the part of the dictionary removed to get K.
- The **Core** (C) (6-9% of the dictionary) is the biggest *strongly connected component* (SCC), in which every node can be reached from every other node.
- The **Satellites** (S) (1-4% of the dictionary) are the smaller SCCs in K surrounding C.
- The **MinSets** (Ms) (~1% of the dictionary) are the smallest sets of words from which all the rest can be defined.

	Cambridge	Longman	Webster	WordNet
MinSets	373 (1%)	452 (1%)	1396 (2%)	1094 (1%)
Core	2009 (8%)	1786 (6%)	7977 (9%)	6392 (8%)
Satellites	232 (1%)	540 (2%)	2978 (3%)	3410 (4%)
Kernel	2241 (9%)	2326 (7%)	10955 (12%)	9802 (12%)
Rest	22891 (91%)	28700 (93%)	80433 (88%)	75393 (88%)
Total nodes	25132	31026	91388	85195

Method

We compared the words in the dictionary's three components (C, S, R) on our three psycholinguistic variables. The psycholinguistic databases are large enough to cover most of our words (~90%) for each variable. We only report effects that showed the same pattern for all four dictionaries.

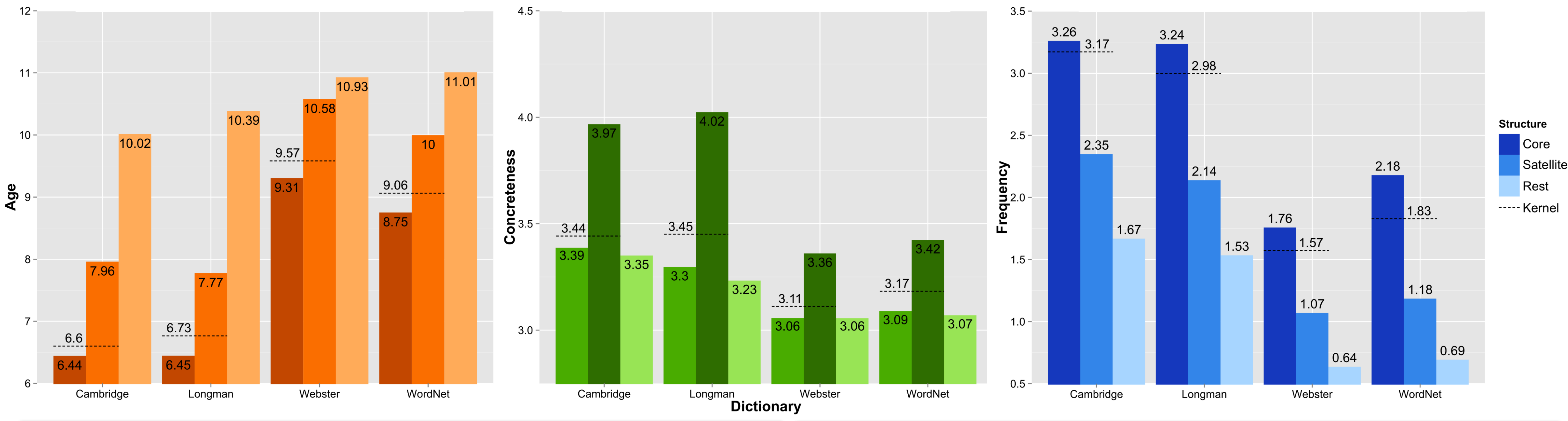
To test whether words in each M differ from words in the rest of K in frequency, age or concreteness, we generated random samples including the same S/C ratio of words for each dictionary. We had multiple Ms for the two smaller dictionaries only (Cambridge (n = 20) and Longman (n = 19)), and only one for each bigger dictionary because computing Ms is still too hard. The differences between the Ms and the random samples were compared with t-tests.

Results

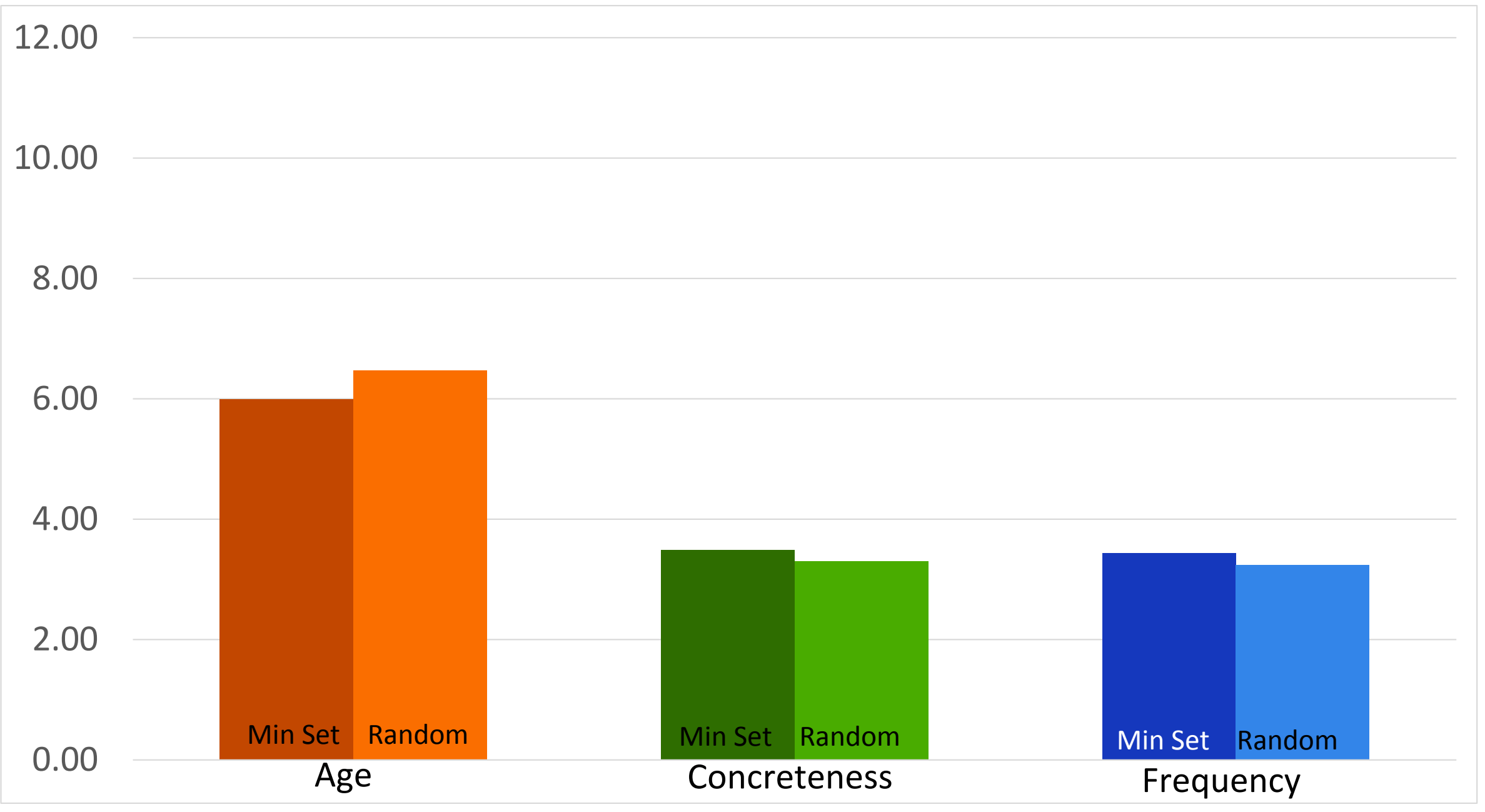
The deeper the words in the RKSC hierarchy, the younger and more frequent they are (C>S>K>R). For concreteness, the pattern is somewhat different: S>C and S>R. The S and C parts of Ms also differ. The C-parts of Ms are more concrete, frequent and younger than the random samples. The S-parts of Ms are significantly less frequent and older than the random samples. Although these effects are small, all are highly significant (p <0.0001) and consistent for all four dictionaries.

Results

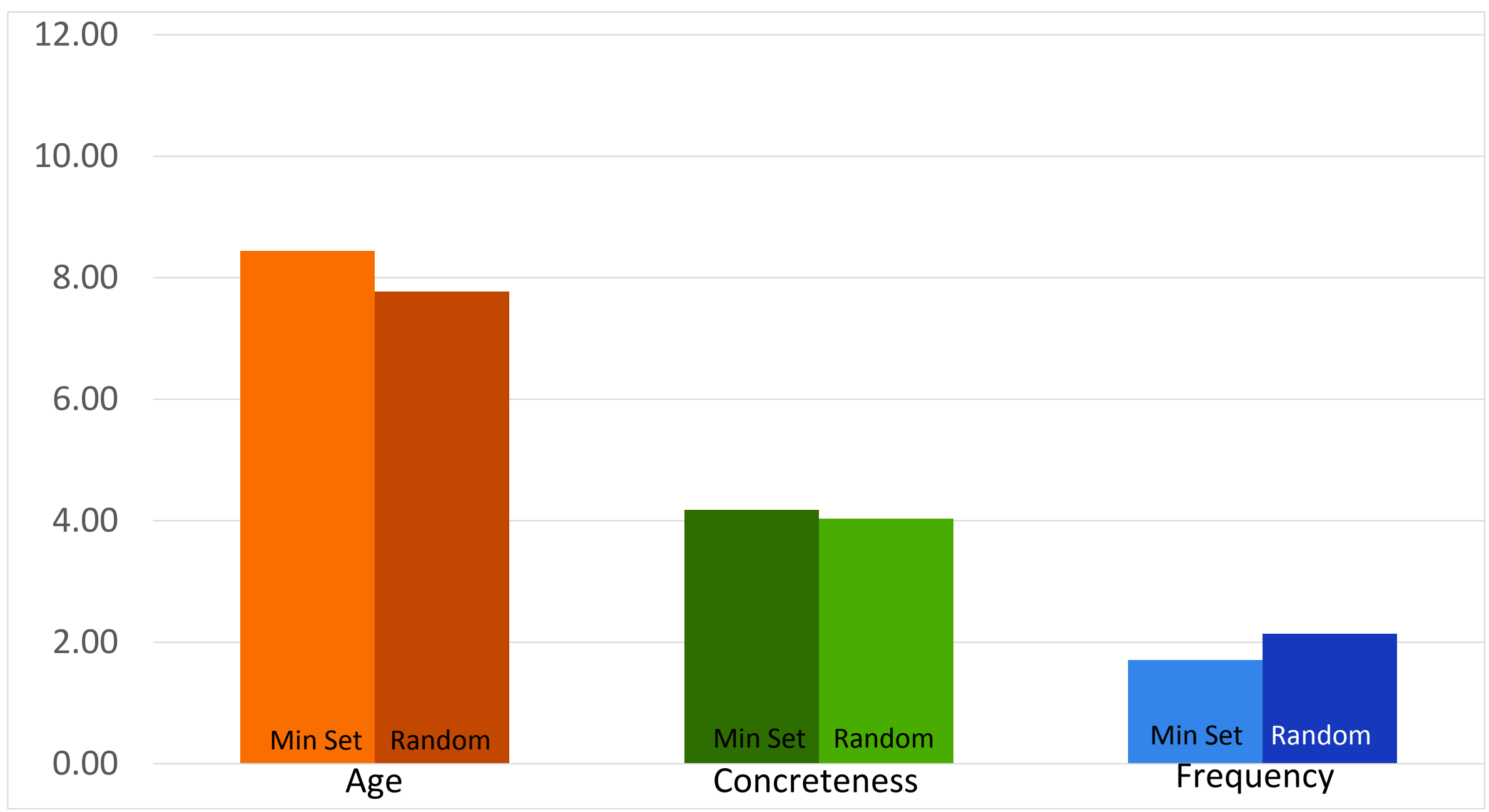
Average by components



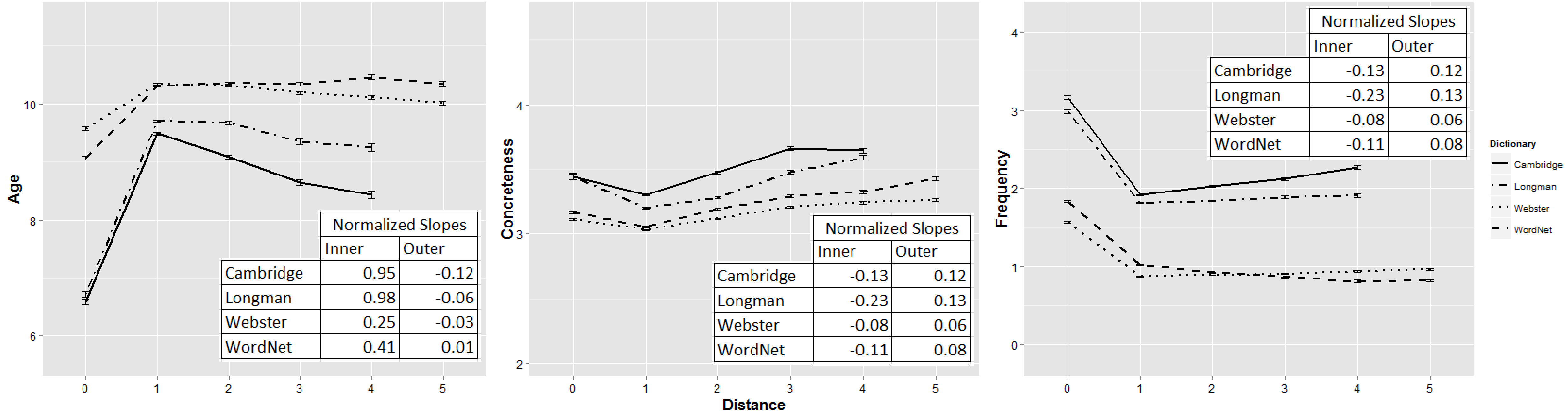
MinSets (Core)



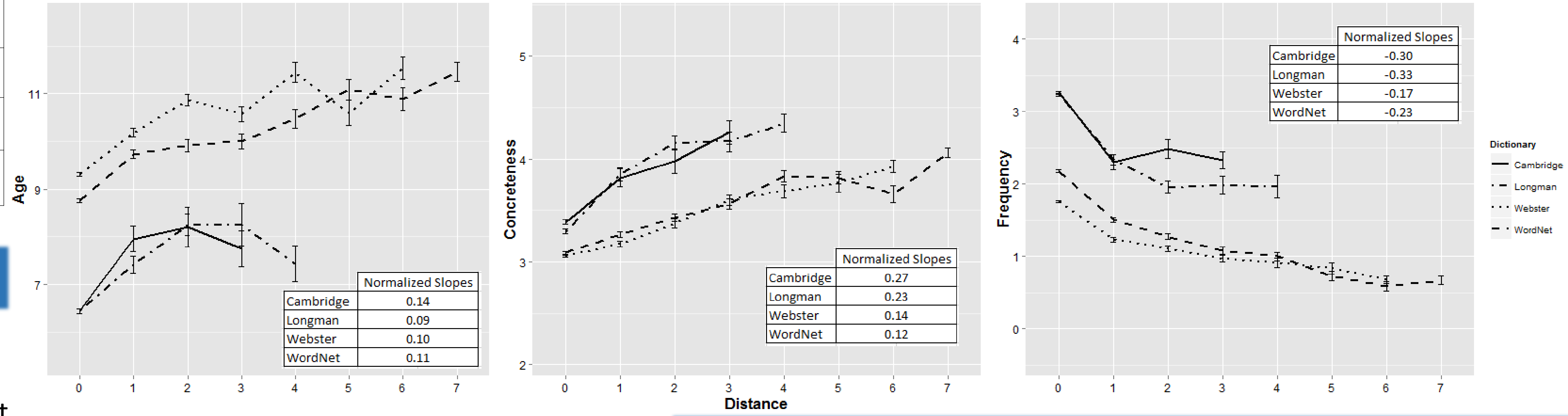
MinSets (Satellites)

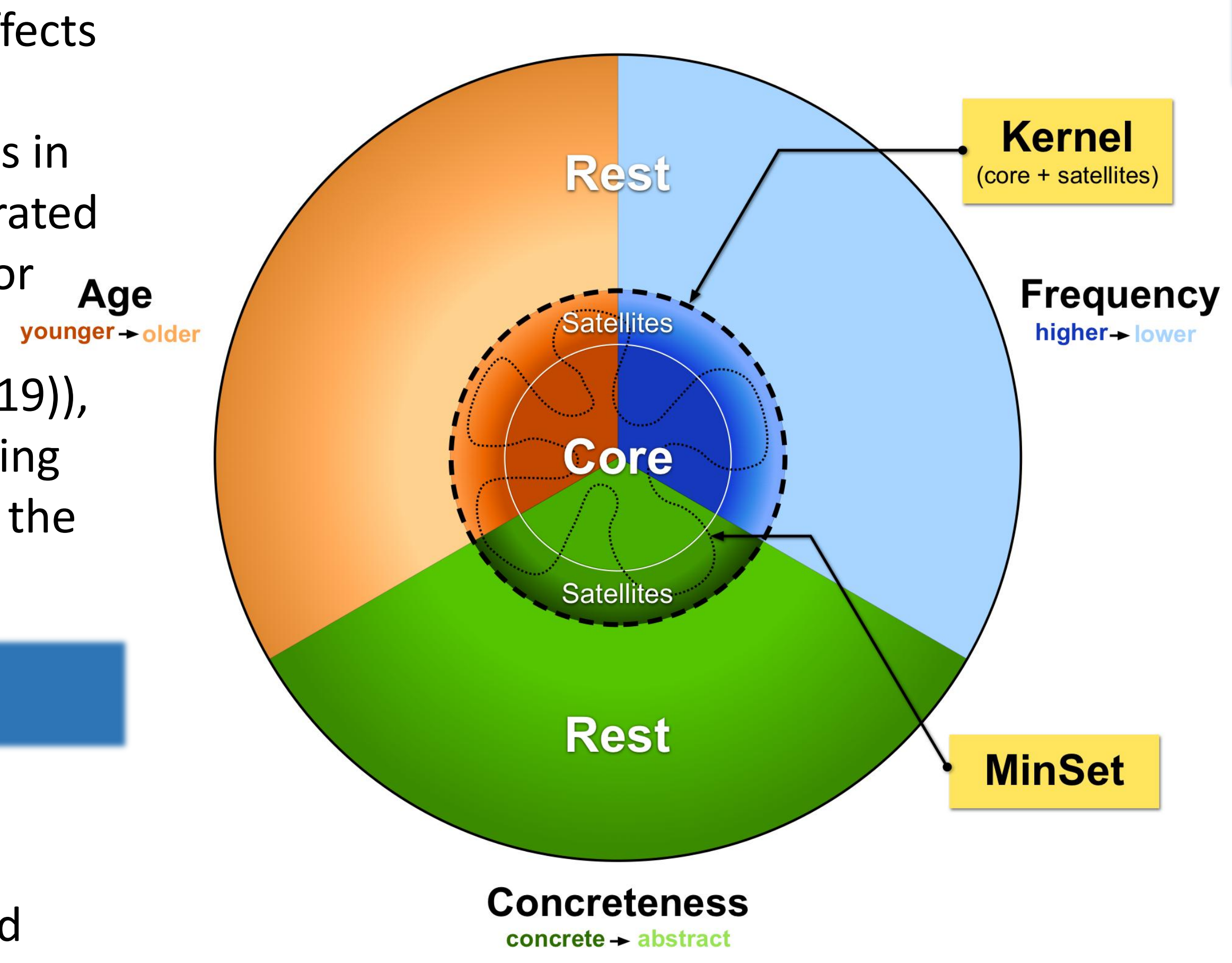


Definitional Distance from Kernel



Definitional Distance from Core within Kernel





Discussion

The words in the dictionary components revealed by our graph-theoretic analysis differ in their psycholinguistic correlates. Every MinSet has a C-part that is younger and more frequent and an S-part, that is more concrete. To understand the functional role of these components will require a close study of the words themselves, and how they are combined into definitions. (For this we will need to analyze even smaller dictionaries, which are generated through an online dictionary game in which participants must define a word, then define the words in the definition, etc. The game ends when all words are defined.) We can already conclude that the closer a word is to the MinSets that can define all other words, the more concrete and frequent the word is likely to be, and the earlier it is likely to have been learned. This is what one would expect if the words in the MinSets were the ones that had been acquired through direct sensorimotor grounding.

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., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 1-8.

Harnad, S. (1990) *The Symbol Grounding Problem* Physica D 42:335-346

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

Picard, O., Lord, M., Blondin-Massé, A., Marcotte, O., Lopes, M., & Harnad, S. (2013) Hidden Structure and Function in the Lexicon, *NLPCS 2013 : 10th International Workshop on Natural Language Processing and Cognitive Science*: 65-77