Psycholinguistic Correlates of Symbol Grounding in Dictionaries
Philippe Vincent-Lamarre¹, Alexandre Blondin-Massé¹, Mélanie Lord¹, Marcos Lopes² & Stevan Harnad¹,³

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.

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 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&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.

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