The Latent Structure of Dictionaries

Philippe Vincent-Lamarre[1,2], Alexandre Blondin Massé[1], Marcos Lopes[3], Mélanie Lord[1], Odile Marcotte[1], Stevan Harnad[1,4]

1 Université du Québec à Montréal, **2** Université d'Ottawa, **3** University of São Paulo (USP), **4** University of Southampton

ABSTRACT: How many words – and which ones – are sufficient to define all other words? When dictionaries are analyzed as directed graphs with links from defining words to defined words, they turn out to have latent structure that has not previously been noticed. Recursively removing all those words that are reachable by definition but do not define any further words reduces the dictionary to a *Kernel* of 10%, but this is still not the smallest number of words that can define all the rest. About 75% of the Kernel is its *Core*, a strongly connected subset (with a definitional path to and from any word and any other word within it), but the Core cannot define all the rest of the dictionary. The 25% surrounding the Core are *Satellites*, small strongly connected subsets. The size of the smallest set of words that can define all the rest – a graph's "minimum feedback vertex set" or *MinSet* – is about 1% of the dictionary, about 15% of the Kernel, about half-Core and half-Satellite, but every dictionary has a huge number of MinSets. The words in the Core turn out to be learned earlier, more frequent, and less concrete than the Satellites, which are learned earlier and more frequent but more concrete than the rest of the Dictionary. The findings are related to the symbol grounding problem and the mental lexicon.

The Representation of Meaning. One can argue that the set of all the written words of a language constitutes the biggest and richest digital database of all. Numbers and algorithms are just special cases of words and sentences, so they are all part of that same global verbal database. Analog images are not words, but even their digitized versions only become tractable once they are sufficiently tagged with verbal descriptions. So in the end it all comes down to words. But how are the meanings of words represented? There are two prominent representations of word meaning: one is in our external dictionaries and the other is in our brains: our "mental lexicon." How are the two related?

The Symbol Grounding Problem. We consult a dictionary in order to learn the meaning of a word whose meaning we do not yet already know. Its meaning is not yet in our mental lexicon. The dictionary conveys that meaning to us through a definition consisting of further words, whose meanings we already know. If a definition contains words whose meanings we do not yet know, we can look up their definitions too. But it is clear that meaning cannot be dictionary look-up all the way down. *The meanings of some words, at least, have to be learned by some means other than dictionary look-up, otherwise word meaning is ungrounded*: It is just strings of meaningless symbols (defining words) pointing to meaningless symbols (defined words). This is the "symbol grounding problem" (Harnad 1990).

This paper addresses the question of *how many words* – and *which words* – have to be learned (grounded) by means other than dictionary look-up so that all the rest of the words in the dictionary can be defined either directly, using solely those grounded words, or,

recursively, using further words that can themselves be defined using solely those grounded words. Let us call those grounded words in our mental lexicon -- the ones sufficient to define all the others -- a "Grounding Set."

Category Learning. The process of word grounding itself is the subject of other ongoing work on the sensorimotor learning of categories (Harnad 2005; Blondin Massé et al 2013). Here we just note that almost all the words in any dictionary (nouns, verbs, adjectives and adverbs) are "content" words[1], meaning that they are the names of *categories* (objects, individuals, kinds, states, actions, events, properties, relations) of various degrees of abstractness. Many of those categories, and hence the words that name them, can be learned directly through trial-and-error sensorimotor experience, guided by feedback that indicates whether an attempted categorization was correct or incorrect. A grounding set composed of such experientially grounded words would then be enough to allow the meaning of all further words to be learned through verbal definition alone.

Expressive Power. Perhaps the most remarkable and powerful feature of natural language is the fact that *it can say anything and everything that can be said* (Katz 1978). There exists no language in which you can say this, but not that. (Pick a pair of languages and try it out.) Word-for-word translation may not work: you may not be able to say everything in the same number of words, equally succinctly, or equally elegantly, in the same *form*. But you will always be able to translate in paraphrase the *propositional content* of anything and everything said in any one language into any other language. (If you think that may still leave out anything that can be said, just say it in any language at all and it will prove to be sayable in all the others too; Steklis & Harnad 1976.)

One counter-intuition about this is that the language may lack the words: its vocabulary may be insufficient: How can you explain quantum mechanics in the language of isolated Amazonian hunter-gatherers? But one can ask the very same question about how you can explain it to an American 6-year-old – or, for that matter, to an eighteenth century physicist. And the banal answer is that it takes time, and a lot of words, to explain -- but you can always do it, in any language. Where do all those missing words come from, if not from the same language? We coin (i.e., lexicalize) words all the time, as they are needed, but we are coining them within the same language. Nor are most of the new words we coin labels for unique new experiences, like names for new colors (e.g., "ochre") or new odors ("acetic") that you have to see or smell directly at first hand in order to know what the words refer to.

Consider the German word "*Schadenfreude*" for example. There happens to be no single word for this in English. It means "feeling glee at another's misfortune." English is highly assimilative, so instead of bothering to coin a new English word (say, "misfortune-glee," or, more latinately, "malfelicity") whose definition is "glee at another's misfortune," English has simply adopted *Schadenfreude* as part of its own lexicon. All it needed was to be defined,

---

[1] Content or Open Class words are growing in all spoken languages all the time. In contrast, Function or Closed Class words like *if, off, is, or his* are few and fixed, with mostly a formal or syntactic function: Our study considers only content words. Definitions are treated as unordered strings of content words, ignoring function words, syntax and polysemy (i.e, multiple meanings, of which we use only the first and most common meaning for each word-form).

and then it could be added to the English dictionary. The shapes of words themselves are arbitrary, after all, as Saussure (1911/1972) noted: words do not resemble the things they refer to.

So what gives English or any language its limitless expressive power is *its capacity to define anything with words*. But is this defining power really limitless? First, we have already skipped over one special case that eludes language's grasp, and that is new sensations that you have to experience at first hand in order to know what they are – hence to understand what any word referring to them means. But even if we set aside words for new sensations, what about other words, like *Schadenfreude*? That does not refer to a new sensory experience. We understand what it refers to because we understand what the words "glee at another's misfortune" refer to. That definition is itself a combination of words, and we have to understand those words in order to understand the definition. If we don't understand some of the words, we can of course look up their definitions too – but as we have noted, it cannot be dictionary look-ups all the way down! The meanings of some words, at least (e.g., "glee") need to have been grounded in direct experience, whereas others (e.g., "another" or "misfortune") may be grounded in the meaning of words that are grounded in the meaning of words… that are grounded in direct experience.

Direct Sensorimotor Grounding. How the meaning of a word referring to a sensation like "glee" can be grounded in direct experience is fairly straightforward: It's much the same as teaching the meaning of "ochre" or "acetic": "Look (sniff): that's ochre (acetic) and look (sniff) that's not." "Glee" is likewise a category of perceptual experience. To teach someone which experience "glee" is, you need to point to examples that are members of the category "glee" and examples that are not: "Look, that's glee" – pointing to someone who looks and acts and has reason to feel gleeful - and "Look, that's not glee" – pointing[2] to someone who looks and acts and has reason to feel ungleeful (Harnad 2005).

What about the categories denoted by the words "another" and "misfortune"? These are not direct, concrete sensory categories, but they still have examples in our direct sensorimotor experience: "That's you" and "that's another" (i.e., someone else). "That's good fortune" and "that's misfortune." But it is more likely that higher-order, more abstract categories like these would be grounded in verbal definitions composed of words that each

---

[2] Wittgenstein had some cautions about the possibility of grounding words for private experiences because there would be no basis for correcting errors. But thanks to our "mirror neurons" and our "mind-reading" capacity we are adept at inferring most private experiences from their accompanying public behavior, and reasonable agreement on word meaning can be reached on the basis of common experience together with these observable behavioral correlates (Apperley 2010). Because of the "other-minds problem" -- i.e., because the only experiences you can have are your own -- there is no way to know for sure whether private experiences accompanied by the same public behavior are indeed identical experiences. These subtleties do not enter into the analyses being done in this paper. Word meaning is in any case not exact but approximate in all fields other than formal mathematics and logic. Even observable, empirical categories can only be defined or described provisionally and approximately. Like a picture or an object, an experience is always worth more than a thousand (or any number) of words (Harnad 1987).

name already grounded categories, rather than being grounded in direct sensorimotor experience.

Dictionary Grounding. This brings us to the question that is being addressed in this paper: A dictionary provides an (approximate) definition for every word in the language. Apart from a small, fixed set of words whose role is mainly syntactic ("function words," e.g. articles, particles, conjunctions), all the rest of the words in the dictionary are the names of categories ("content words," i.e. nouns, verbs, adjectives, adverbs). *How many content words (i) -- and which ones (ii) -- need to be grounded already so that all the rest can be learned from definitions composed only out of those grounded words?* The answer may cast light on how the meaning of words is represented – externally, in dictionaries, and internally, in our mental lexicon -- as well as on the evolutionary origin and adaptive function of language for our species (Blondin Massé et al. 2013).

Synopsis of Findings. Before we describe in detail what we did, and how, here is a synopsis of what we found: Dictionaries turn out to have a latent structure that has not previously been reported by others, as far as we know. Dictionaries have a special subset of words – about 10% (**Table 1**; **Figure 1**) -- that we have called their "Kernel". The Kernel is unique, and its words can define the remaining 90% of the dictionary. The Kernel is hence a grounding set. But it is not the *smallest* grounding set. That smallest subset – which we have called the "Minimal Grounding Set" (MinSet) -- is much smaller than the Kernel (about 1% of the dictionary and about 15% of the Kernel), but it is not unique: The Kernel contains a huge number of different MinSets; each of them is of that same minimum size and each is able to define all the other words in the dictionary. The Kernel also turns out to have further latent structure: About 75% of the Kernel turns out to be one huge "strongly connected subset" (i.e., one within which there is a definitional path between any two words in both directions), which we call the Kernel's "Core." The remaining 25% of the Kernel surrounding the Core consists of many tiny strongly connected subsets, which we call the Core's "Satellites." It turns out that each MinSet is part-Core and part Satellite. The words in these different latent structures also turn out to differ in their psycholinguistic properties: As we go deeper into the dictionary, from the 90% Rest to the 10% Kernel, to the Satellites (1-4%) surrounding the Kernel's Core, to the Core itself (6-9%), the words turn out on average to be more frequently used (orally and in writing) and to have been learned at a younger age. This is reflected in a gradient within the Satellite layer: the shorter a word's definitional distance (the number of definitional steps to reach it) from the Core, the more frequently it is used and the earlier it was learned. The average concreteness of the words in the Core and the Rest outside the Kernel is about the same. Within the Satellite layer, however, words become more concrete the greater their definitional distance from the Core. There is also a (much weaker) definitional distance gradient from the Kernel outward into the Rest of the dictionary for age and concreteness, but not for frequency. We will now describe how this latent structure was discovered.

Control Vocabularies. Our investigation began with two small, special dictionaries – the *Cambridge International Dictionary of English* (47,147 words; Procter 1995; henceforth *Cambridge*) and the *Longman Dictionary of Contemporary English* (69,223 words; Procter 1978; henceforth *Longman*) (Table 1). These two dictionaries were created especially for people with limited English vocabularies, such as non-native speakers; all words are defined using only a "control" vocabulary of 2000 words that users are likely to know already. Our objective was to analyze each dictionary as a directed graph (*digraph*) in

which there is a directional link from each defining word to each defined word. Each word in the dictionary should be reachable, either directly or indirectly, via definitions composed of the 2000-word control vocabulary.

A direct analysis of the graphs of *Longman* and *Cambridge*, however, revealed that their underlying control-vocabulary principle was not faithfully followed: There turned out to be words in each dictionary that were not defined using only the 2000-word "control" vocabulary, and there were also words that were not defined at all. So we decided to use each dictionary's digraph (a directed graph with arrows pointing from the words in each definition to the word they define) to work backward in order to see if we could generate a genuine control vocabulary out of which all the other words could be defined. (We first removed all undefined words.)

Dictionaries as Graphs. Dictionaries can be represented as directed graphs $D = (V, A)$ (*digraphs)*. The *vertices* are words and the *arcs* connect defining words to defined words, i.e. there is an arc from word *u* to word *v* if u is a word in the definition of v. Moreover, in a complete dictionary, every word is defined by at least one word, so we assume that there is no word without an incoming arc. A *path* is a sequence $(v_1, v_2, \dots, v_k)$ of vertices such that (v_i, v_{i+1}) is an arc for $i = 1,2, \dots, k-1$. A *circuit* is a path starting and ending at the same vertex. A graph is called *acyclic* if it does not contain any circuits.

Grounding Sets. Let $U \subseteq V$ be any subset of words and let *u* be some given word. We are interested in computing all words that can be learned through definitions composed only of words in U. This can be stated recursively as follows: We say that u is *learnable* from *U* if either *u* belongs to *U* or all predecessors of *u* are learnable from *U*. The set of words that can be learned from U is denoted by *L(U)*. In particular, if $L(U) = V$, then U is called a *grounding set* of D. Intuitively, a set U is a grounding set if, provided we already know the meaning of all the words in U, we can learn the meaning of all the remaining words just by looking up the definitions of the unknown words (in the right order).

Grounding sets are equivalent to well-known sets in graph theory called *feedback vertex sets*. These are sets of vertices *U* that *cover* all circuits, i.e. for any circuit *c*, there is at least one of *c* belonging to *U*. It is rather easy to see this. On the one hand, if there exists a circuit of unknown words, then there is no way to learn any of them by definition alone. On the other hand, if every circuit is covered, then the graph of unknown words is acyclic, which means that the meaning of at least one word can be learned – a word having no unknown predecessor (Blondin Massé et al 2008)).

Clearly, every dictionary D has many grounding sets. For example, the set of all words in D is itself a grounding set. But how small can grounding sets be? In other words, what is the smallest number of words you need to know already in order to be able to learn the meaning of all the remaining words in D through definition alone? These are the Minimal Grounding Sets (*MinSets)* mentioned earlier. It is already known that finding a minimum feedback vertex set in a general digraph is NP-hard (Karp, 1972), which implies that finding Minsets is also NP-hard. Hence, it is highly unlikely that one will ever find an algorithm that solves the problem without taking an exponentially long time. However, since some real dictionary graphs are relatively small and also seem to be structured in a favorable way, there are ways to compute their MinSets.

Kernel. As a first step, we observed that, in all dictionaries analyzed so far, there exist many words that are never used in any definition. These words can be removed without changing the MinSets. This reduction can be repeated iteratively until no further word can be removed without leaving any word unlearnable from the rest. The resulting subgraph is what we called the dictionary's *(grounding) Kernel*. Each dictionary's Kernel is unique, in the sense that every dictionary has one and only one Kernel. The Kernels of our two small dictionaries, *Longman* and *Cambridge* turned out to amount to 8% and 7% of the dictionary as a whole, respectively. We have since extended the analysis to two larger dictionaries, *Merriam-Webser* (248,466 words; Webster 2006; henceforth *Webster*) and *Wordnet* (132,477 words; Fellbaum 2010) whose Kernels are both 12% of the dictionary as a whole (**Table 1**).

Core and Satellites. Next, since we are dealing with directed graphs, we can divide the words according to their *strongly connected components*. Two words *u* and *v* are *strongly connected* if there exists a path from *u* to *v* as well as a path from *v* to *u*. There is a well-known algorithm in graph theory that computes the *strongly connected components* (SCCs) very efficiently (Tarjan 1972). It turns out that in the Kernel of all four dictionaries we have analyzed so far there is one SCC that is much bigger than all the rest: it is about 75% of the Kernel and about 7-8% of the dictionary as a whole. It is this largest strongly connected component of a dictionary's Kernel that we call the Kernel's *Core*. All the remaining strongly connected components in the Kernel (about 25% of the Kernel and about 1-2% of the dictionary as a whole) constitute the Kernel's *Satellites*.

Definitional Distance from the Kernel: the K-Hierarchy. Another potentially informative graph-theoretic property is the "definitional distance" of any given word from the Kernel or from the Core in terms of the number of arcs separating them. We define these two distance hierarchies as follows. First, for the Kernel hierarchy, suppose *K* is the Kernel of a dictionary graph *D*. Then, for any word *u*, we define its distance recursively as follows:

1. $dist(u) = 0$, if *u* is in *K*;
2. $dist(u) = 1 + max\{dist(v) : v \text{ is a predecessor of } u\}$, otherwise.

In other words, to compute the distance between K, as origin, and any word *u* in the rest of D, we compute the distance of all words defining *u* and add one. This distance is well defined, because *K* is a grounding set of *D* and hence the procedure cannot cycle because every circuit is covered. The mapping that relates every word to its distance from the K is called the *K-hierarchy*.

Definitional Distance from the Core: the C-Hierarchy. The second metric is slightly more complicated but based on the same idea. Let *D* be the directed graph of a dictionary, and *D'* be the graph obtained from *D* by merging each strongly connected component (SCC) into a single vertex. The resulting graph is acyclic. We can then compute the distance of any word from the Core (the vertex corresponding to the biggest of the merged strongly connected components of the Kernel) as follows:

1. $dist(u) = 0$, if *u* is in a source vertex of *D'*;
2. $dist(u) = 1 + ma\square\{dist(v) : v \text{ is in a predecessor merged vertex of } u\}$, otherwise.

The words in the merged vertices of the Core have no predecessor and constitute the origin of the C-hierarchy for two of our four dictionaries, and level 1 for the other two (because of a tiny, probably artefactual predecessor). The distance from C of the words at each succeeding level is computed by taking the minimum among the predecessors plus one. Like the K-hierarchy, the C-hierarchy is well defined because *G'* is acyclic.

MinSets. We have computed the Kernel K, Core C, and Satellites S as well as the K-hierarchy and the C-hierarchy for four English dictionaries: two smaller ones -- (1) Longman's Dictionary of Contemporary English (Longman, 47147 words), (2) Cambridge's International Dictionary of English (Cambridge, 69, 223 words) – and two larger ones - (3) Merriam-Webster (Webster, 248,466 words), (4) WordNet (132,477 words). Because of polysemy (multiple meanings)[3], there can be more than one word with the same word-form (lexeme). As an approximation, for each stemmatized word-form we used only the first (and most frequent) meaning for each part of speech of that word-form (noun, verb, adjective, adverb). This reduced the total number of words by 50% for the smaller dictionaries and by 65% for the larger dictionaries. The size of their respective Kernels turned out to be between 8% of the whole dictionary for the smaller dictionaries and 12% for the larger dictionaries. The Kernel itself varied from 10% Satellite and 90% Core for the two small dictionaries to 35% Satellite and 65% Core for the two large dictionaries (Table 1).

As noted earlier, computing the MinSets is much more demanding than computing K, C and S, because the problem is NP-hard. As a first step, we represented the problem as a linear integer program in which the constraints are given by the circuits. In a general digraph, the number of these constraints grows exponentially, but in the special case of dictionaries and their structure, it was possible to concentrate on the shortest circuits. Next, we tried to solve the linear program using the powerful CPLEX solver. For the smaller dictionaries Longman and Cambridge we were able to compute a few MinSets (though not all of them, because there are a very large number). For Webster and WordNet, we could only compute some almost-minimal sets, by using the best solution found by CPLEX after many days of computation.[4]

These analyses answered our first question about the *size* of the MinSet for these four dictionaries (373 and 452 words for the small dictionaries; 1396 and 1094 for the larger ones; about 1% for each dictionary). But because, unlike a dictionary's unique Kernel, its MinSets are not unique, a dictionary has a vast number of MinSets, all within the Kernel, all the same minimal size, but each one different in terms of which combination of Core and Satellite words it is composed of. The natural question to ask now is whether the words contained in these latent components of the dictionary, identified via their graph-theoretic properties -- the MinSets, Core, Satellites, Kernel and the rest of the dictionary – differ from

---

[3] Once the problem of polysemy is solved for both defined and defining words, the analysis described in this paper can be applied to each unique word/meaning pair instead of just to the first meaning of each defined word.

[4] This is yet another approximation in a study that necessitated many approximations: ignoring syntax and word order, using only the first meaning, and finding only something close to the MinSet for the biggest dictionaries. Despite all these approximations and potential sources of error, systematic and interpretable effects emerged from the data.

one another in any systematic way that might give a clue as to the function (if any) of any of the different latent structures identified by our analysis.

	Cambridge	Longman	Webster	WordNet	Game dictionaries (average)
Total words meanings	47147	69223	248466	132477	182
First sense meanings	25132	31026	91388	85195	-
Rest	22891 (91%)	28700 (93%)	80433 (88%)	75393 (88%)	10.14(7%)
Kernel	2241 (9%)	2326 (8%)	10955 (12%)	9802 (12%)	171.68 (93%)
Satellites	232 (1%)	540 (2%)	2978 (3%)	3410 (4%)	54.47 (29%)
Core	2009 (8%)	1786 (6%)	7977 (9%)	6392 (8%)	117.21 (64%)
MinSets	373 (1%)	452 (1%)	1396 (2%)	1094 (1%)	32.81 (18%)
Satellites-MinSets	59 (16%)	167 (37%)	596 (43%)	532 (49%)	20.59 (63%)
Core-MinSets	314 (84%)	285 (63%)	800 (57%)	562 (51%)	12.22 (37%)

Table 1. Number and percentage of word-meanings for each latent structure in each of the four dictionaries used (plus averages for game-generated dictionaries). Based on using only the first word-meaning for each stemmatized part of speech wherever there are multiple meanings (hence multiple words).

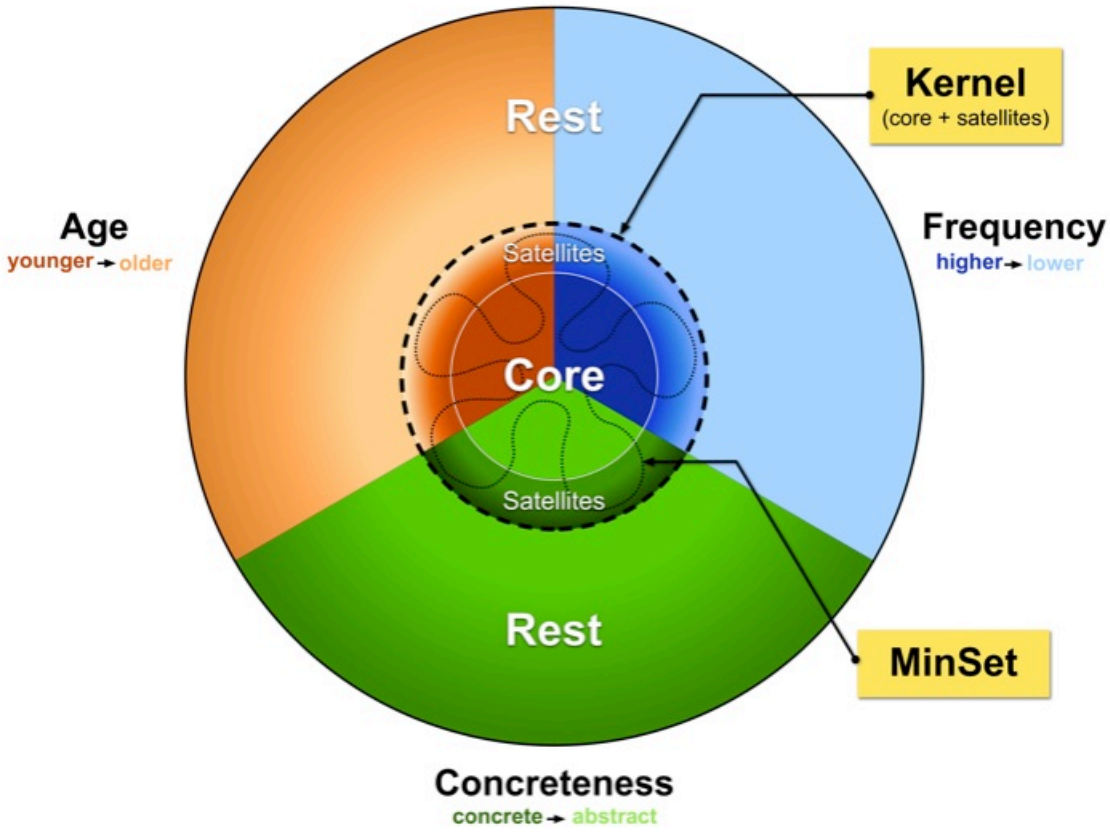

Figure 1. Overall pattern of average psycholinguistic differences (age of acquisition, concreteness, frequency) among the words in the latent structures revealed by the analysis of the dictionary digraph. Pattern is the same for all four dictionaries analyzed but image is not drawn to scale: for exact numbers and percentages see **Table 1** and **Figures 2 & 6**). (MinSets are part Core and part Satellite. Core + Satellites = Kernel [~10%]. Outside the Kernel is the Rest [~90%]). Core words are more frequent (blue) and learned younger (orange) than the Rest of the dictionary. Within the Kernel's Satellite layer, this difference increases gradually as definitional distance from the Core increases. Outside the Kernel, for age, the difference decreases gradually (but weakly) as definitional distance from the Kernel increases; frequency remains uniform. For concreteness (green), it is the Satellite layer that is more concrete than the Core. This difference increases gradually as definitional distance from the Core increases within the Satellite layer of the Kernel. Outside the Kernel, concreteness is at first equal to the Core and then increases gradually (but weakly) as definitional distance from the Kernel increases.

Psycholinguistic Correlates of Dictionary Latent Structure. A number of databases have been compiled that index various psycholinguistic properties of words (e.g., Wilson 1988). We used three of them: For word frequency, we used the SUBTLEX¬US Corpus, which has been found to be more reliable than the widely used Kučera and Francis (1967) word frequency norms (Brysbaert & New, 2009). Raw frequencies range from 1 to over 2 million, with an average of 669 and with about 1% of the values over 5000. For our goal of determining the average frequency for different sets of words, instead of using raw frequency, we used the Lg10WF metric (log10(FREQcount+1)) to reduce the effect of extreme values. For concreteness, the Brysbaert et al. (2013) concreteness ratings for 40,000 commonly known English word lemmas were used. For age of acquisition, we used the Kuperman et al. (2012) age-of-acquisition ratings for 30,000 English words.

We tested whether the words in the latent components we identified in dictionary graphs differ systematically in frequency, concreteness or age of acquisition. Our overall pattern of findings (for all four dictionaries) is illustrated in **Figure 1**, which shows the latent structures of the dictionary: the 90% Rest and the 10% Kernel, and within it the Core surrounded by its Satellites. Shown also is one MinSet (just one of many); all MinSets are part Core and part Satellite.

Based on the data for word frequency (blue), concreteness (green) and age of acquisition (orange) from the psycholinguistic databases, the words in the Core for all four dictionaries are more frequent and learned younger than the Satellite words, which are in turn more frequent and younger than the Rest of the dictionary. The Satellites are more concrete than the Core or the Rest. The average values for each of the psycholinguistic variables in each of the latent substructures are shown in **Figure 2**. The pattern is the same for all four dictionaries. Because the results are based on the entire population of each dictionary graph, no statistical tests were done. All differences would be highly significant because the number of words in each dictionary is so big. The effects themselves, however, are not very big; there are clearly many other factors underlying these variables apart from the dictionary latent structures.

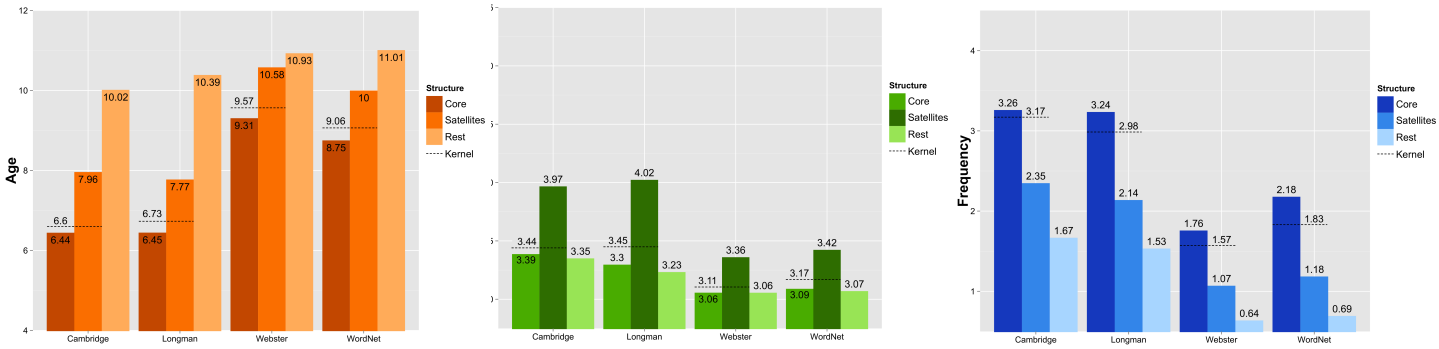

Figure 2: Average age, concreteness and frequency of words in Core, Satellites, Kernel and Rest. The pattern is the same for all four dictionaries: The Core is youngest and most frequent, then the Satellites, then the Rest. The Satellites are more concrete than the Core and the Rest, which are about equal (but see the gradients in Figure 6.)

The effect size for each of the pairwise differences in **Figure 2** is shown in **Figure 3**. Note that the biggest effect size tends to be for frequency. This may be because the psycholinguistic database coverage for frequency is close to 100% complete[5] for all the words in all three latent structures, Core, Satellites and Rest, whereas the coverage for age and concreteness declines with frequency, especially for the two larger dictionaries **(Figure 4)**. It is possible that the effect sizes for age and concreteness would have been larger, especially for the larger dictionaries, if the database coverage had been more complete. It is likely that the incompleteness of the data for age and concreteness is itself an indirect effect of word frequency: Age and concreteness data are lacking for the less frequent words.

All three variables – age, concreteness, and frequency – are intercorrelated (frequency/age: -0.5915; frequency/concreteness: 0.1583; age/concreteness: -0.3773). Decorrelating frequency from age and concreteness by recalculating effect sizes for only the residual variance left after removing the frequency variance reduces the effect sizes for age and concreteness (**Figure 5**). Age and concreteness data, which are much harder to gather than frequency data, are less available for less frequent words:

> "*From a list of English words that one of the authors (M.B.) is currently compiling, we selected all of the base words (lemmas) that are used most frequently as nouns, verbs, or adjectives*" (Kuperman et al 2012).
>
> "*Because ratings are only useful for well known words, we used a cut-off score of 85% known. In practice, this meant that not more than 4 participants out of the average of 25 raters indicated they did not know the word well enough to rate it. This left us with a list of 37,058 words and 2,896 two-word expressions (i.e., a total of 39,954 stimuli)*" (Brysbaert 2013).

This introduces a frequency bias into our analysis, because of missing age and concreteness data for less frequent words. This frequency bias could either be (1) helping to reveal valid effects, (2) spuriously inflating them or (3) spuriously reducing them (**Figure 3**); or (4) removing the frequency bias by decorrelating frequency could be masking valid effects (**Figure 5**). We think it is unlikely that word frequency *causes* concreteness or age effects. It is more likely that age of acquisition and concreteness are part of the cause of frequency effects. But the direction of causality cannot be resolved by the available data.

[5] Words in our four dictionaries that had no values for SUBTELXus' frequencies were assigned frequency value zero. The SUBTELXus frequency data were collected on a corpus used as the reference database; zero means the word never occurred in that corpus.

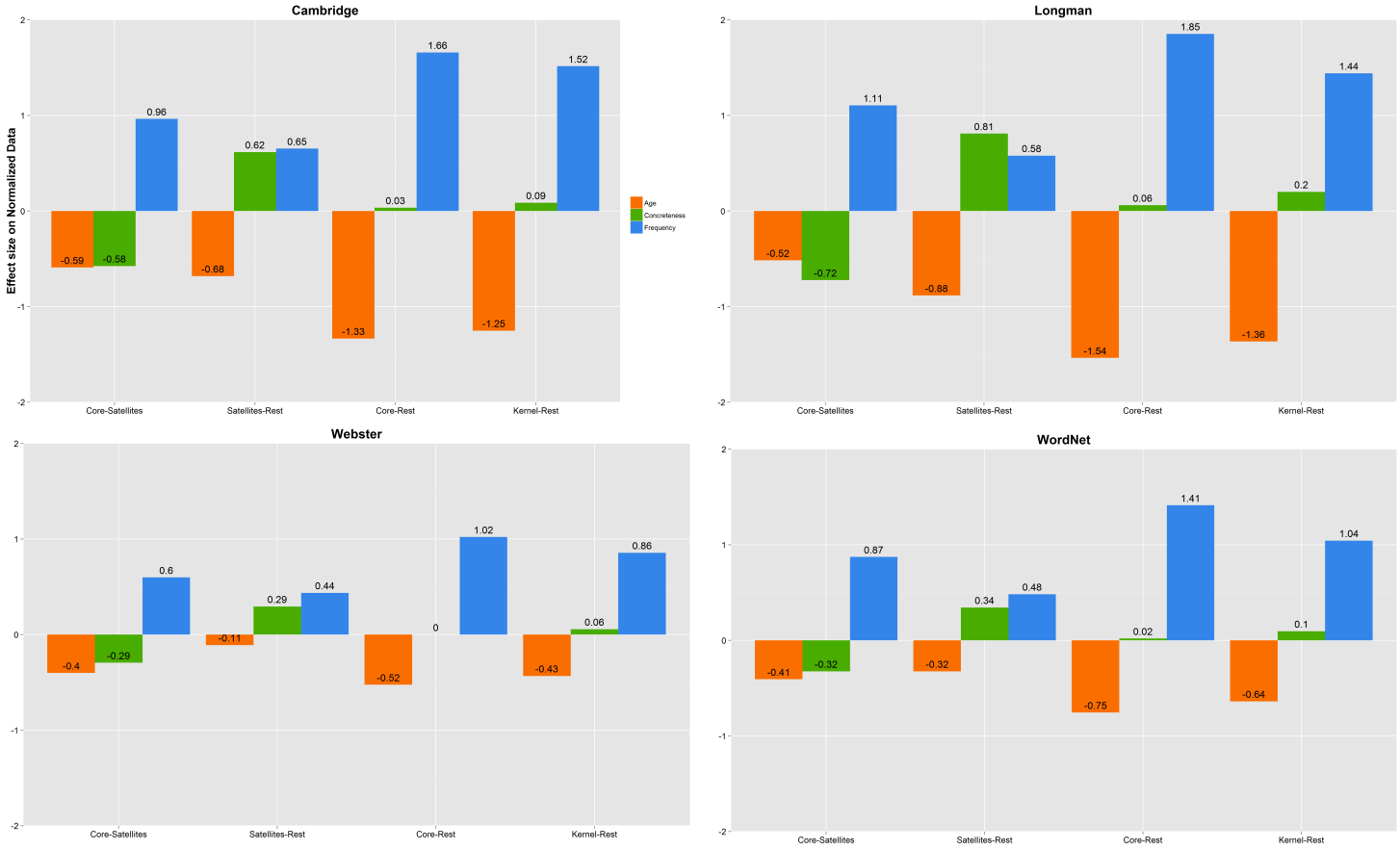

Figure 3. Effect size and direction for the principal comparisons among Core, Satellites, Kernel and Rest for age, concreteness and frequency, for each of the four dictionaries. Note that the effect size for frequency tends to be the biggest, then age, then concreteness.

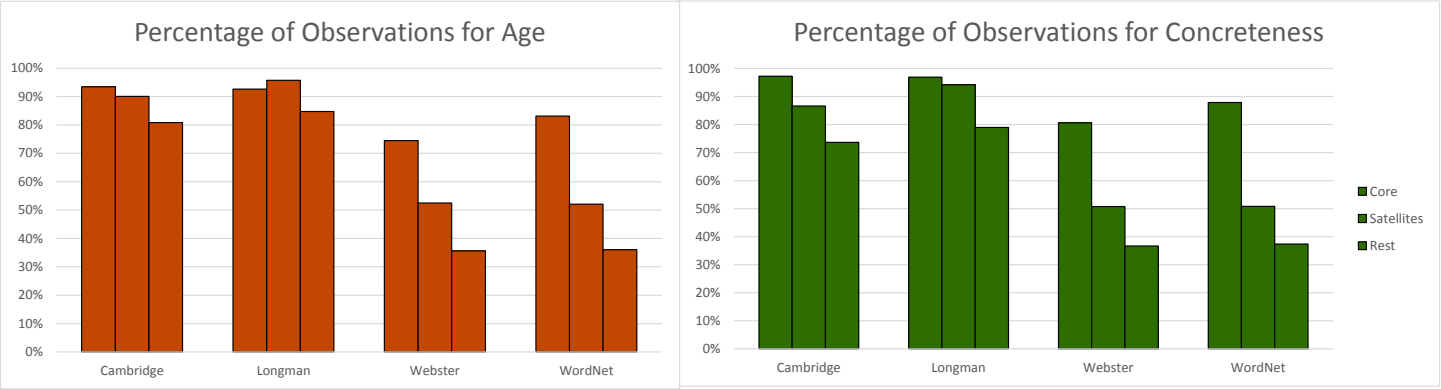

Figure 4. Percentage of words in Core, Satellites and Rest for which psycholinguistic data were available for age and concreteness for each of the four dictionaries. (Frequency data not shown because they are at 100% for all dictionaries.) Note that the percentage of available data is lower for the two bigger dictionaries, and decreases from the Core to the Satellites to the Rest.

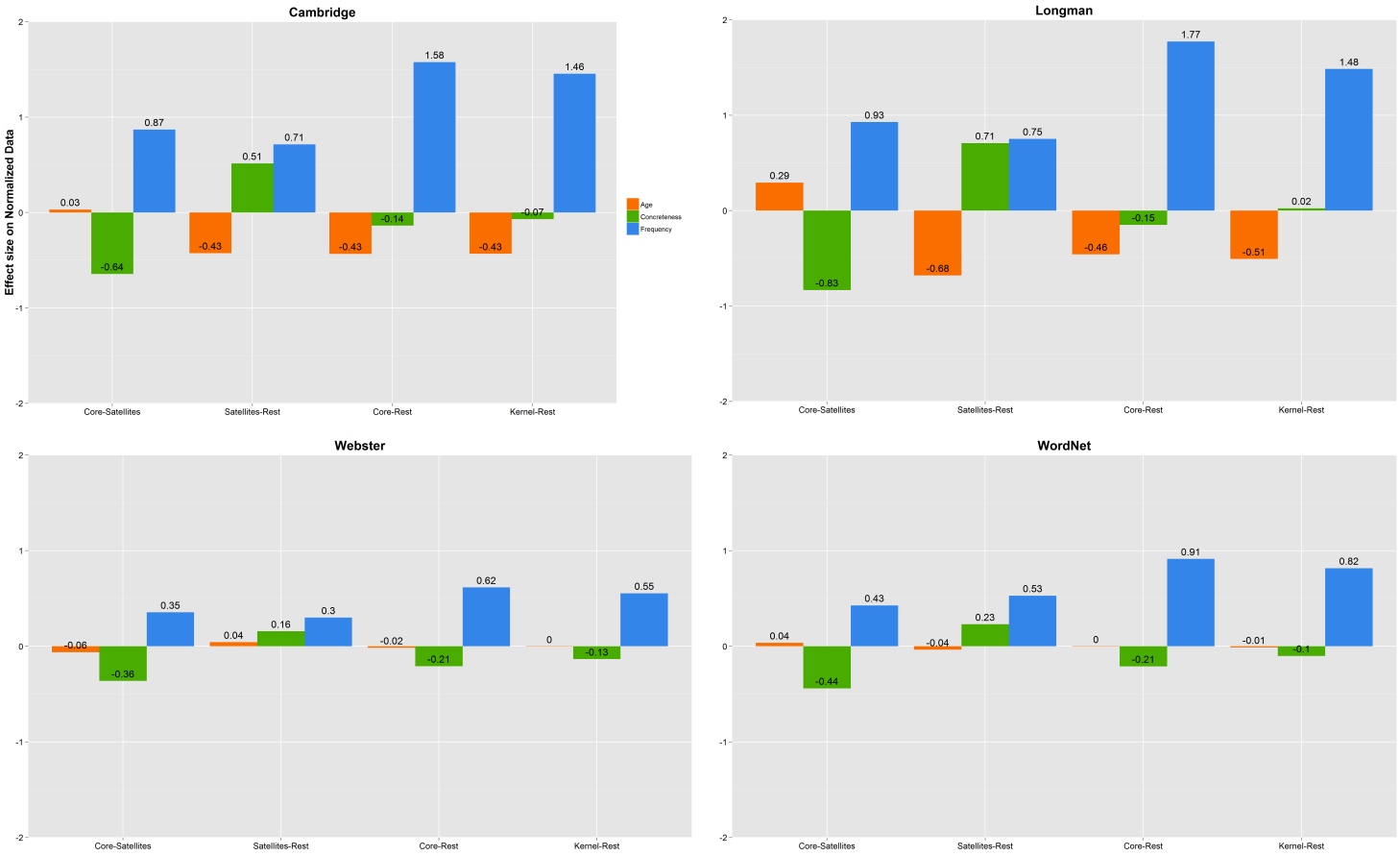

Figure 5. Effect size and direction for the principal comparisons among Core, Satellites, Kernel and Rest for age, concreteness and frequency, for each of the four dictionaries when correlation with frequency is removed. Age and concreteness effects are reduced considerably (cf. **Figure 3**), especially for the bigger dictionaries, for which the psycholinguistic database coverage for age and frequency was lower for the less frequent words (cf. **Figure 6**).

Definitional Distance Gradients. Alongside the main effects – the average differences in frequency, age and concreteness between the Core, Satellites and the Rest -- our analysis also revealed two kinds of graded effects:

The upper part of **Figure 6** shows the gradient for the K-Hierarchy, which is the definitional distance of words in the Rest of the dictionary from the Kernel (i.e. the number of definitional steps to reach a word starting from the Kernel). The first step in this gradient, from distance level 0 (the Kernel) to level 1 corresponds roughly to the main effects in **Figure 2**: For frequency there is a decrease from level 0 to 1 for all four dictionaries; then frequency is flat for all but Cambridge. For age there is an increase from level 0 to 1 (i.e., level 1 words are "older" -- i.e., learned later -- than the Kernel) for all four dictionaries, then descending slightly for all but WordNet. For concreteness there is a decrease (i.e., becoming more abstract) from 0 to 1, and then a gradual increase. Apart from the first step, from 0 to 1, the K-Hierarchy curves are hard to interpret because not only do the words at each succeeding distance level become fewer (**Table 1**) and less frequent, but the psycholinguistic database coverage for age (orange) and concreteness (green) is incomplete, especially for the two bigger dictionaries (**Figure 7**, left).

The lower part of **Figure 6** shows the gradient for the C-Hierarchy, which is the definitional distance from the Core for words in the Satellite layer (i.e. the number of definitional steps to reach a Satellite word starting from the Core). Here the gradients are consistent for all four dictionaries and all three psycholinguistic variables: they are descending (less frequent) for frequency, rising (getting older) for age, and rising (getting more concrete) for concreteness. Here too the number of words diminishes at each distance level (**Table 2**), but, for the two larger dictionaries there is a particularly marked decrease in database coverage for age (orange) and frequency (**Figure 7**, left). (This very visible negative correlation between definitional distance from the Core within the Satellite layer and psycholinguistic database coverage is probably due to the decline of word frequency with definitional distance from the Core within the Satellite layer (**Figure 6**, lower, blue). The red lines show the same effects when we analyze words that are present at the same level (intersection) in both large dictionaries (thick red line) and (separately) words that are present in both smaller dictionaries (thick red line).

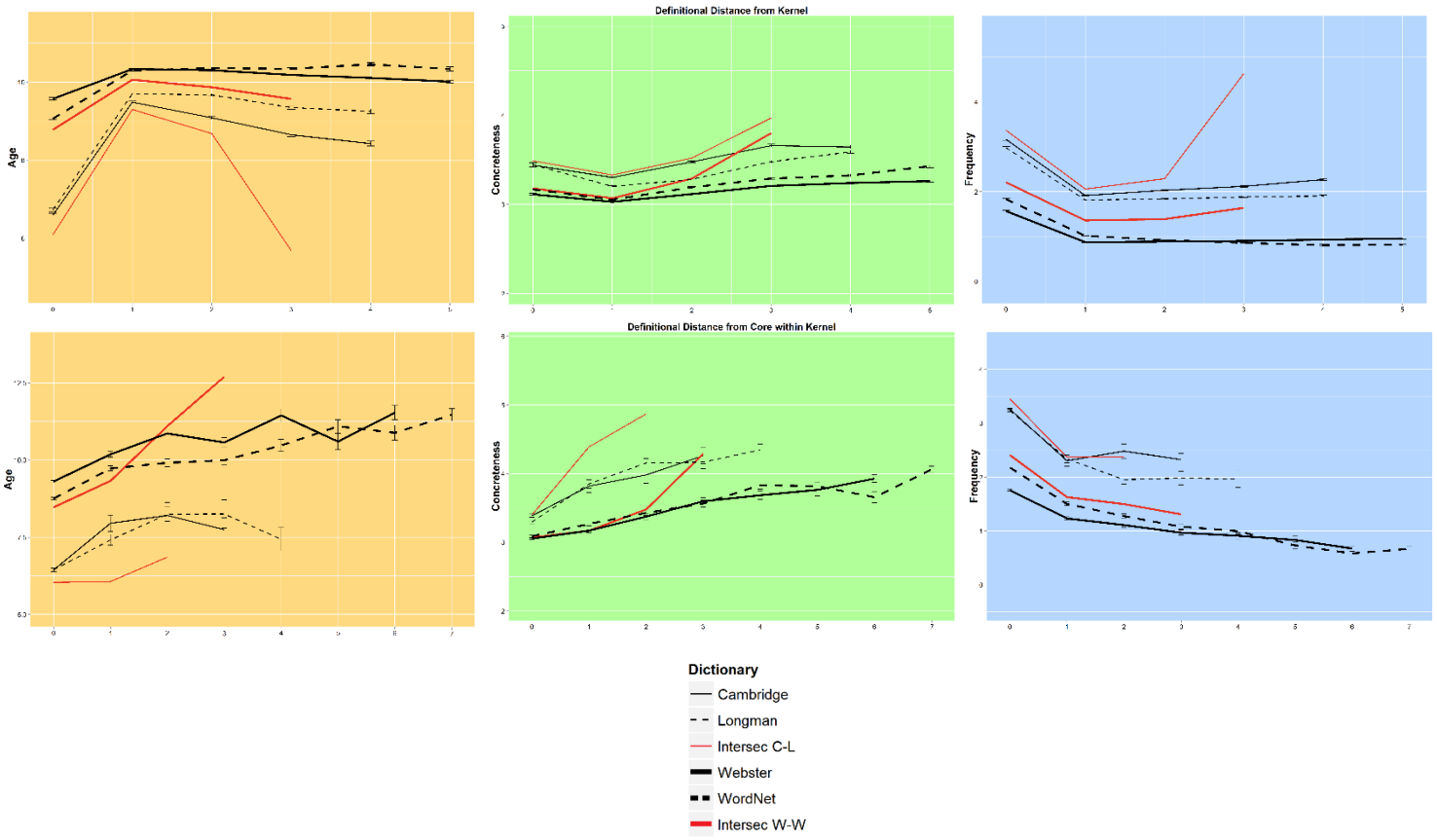

Figure 6: Average age, concreteness and frequency at each level of the definitional distance hierarchy starting from the Kernel through the Rest of the dictionary (K-Hierarchy, above), and within the Kernel, starting from the Core through the Satellites (C-Hierarchy, below), for each of the four dictionaries. K-Hierarchy: for age there is a big increase from the Kernel to level 1 and then a slight decrease at higher levels; for concreteness a slight decrease from K to 1, then slight increase; for frequency a big decrease from K to 1, then mostly flat. C-Hierarchy: increases for age and concreteness and decreases for frequency. All effects are stronger in the smaller dictionaries. The thick red lines show that the pattern is the same when considering only those words that occur at the same level (intersection) in both bigger dictionaries. The thin red lines show the pattern for words that occur at the same level in both smaller dictionaries.

	K-Hierarchy														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Cambridge	2241	15935	9483	4122	1663 (2796)	730	276	105	20	2	-	-	-	-	-
Longman	2326	20555	12231	4966	1906 (3025)	611	259	121	81	39	6	2	-	-	-
Webster	10955	59160	38111	19860	9577	4396 (7615)	1904	666	292	201	89	38	25	2	2
WordNet	9802	48186	26186	12642	5379	2248 (4304)	989	609	293	125	32	7	1	-	-

	C-Hierarchy												
	0	1	2	3	4	5	6	7	8	9	10	11	12
Cambridge	2008	127	51	26 (54)	20	4	4	-	-	-	-	-	-
Longman	1786*	248	159	66	34 (64)	14	6	4	4	2	-	-	-
Webster	7976*	1220	640	425	245	153	106 (287)	68	49	39	19	6	-
WordNet	6391	1270	683	443	308	252	179	117 (275)	77	56	17	6	2

Table 2. Number of words at each level of the definitional distance hierarchy starting from the Kernel through the Rest of dictionary (K-Hierarchy, above), and, within the Kernel, starting from the Core through the Satellites (C-Hierarchy below), for each of the four dictionaries. Note that Figure 6 was truncated at the blue level past which frequencies became too low to be representative. Words past the truncation point were added to the blue value (total number of words for blue level shown in parentheses).

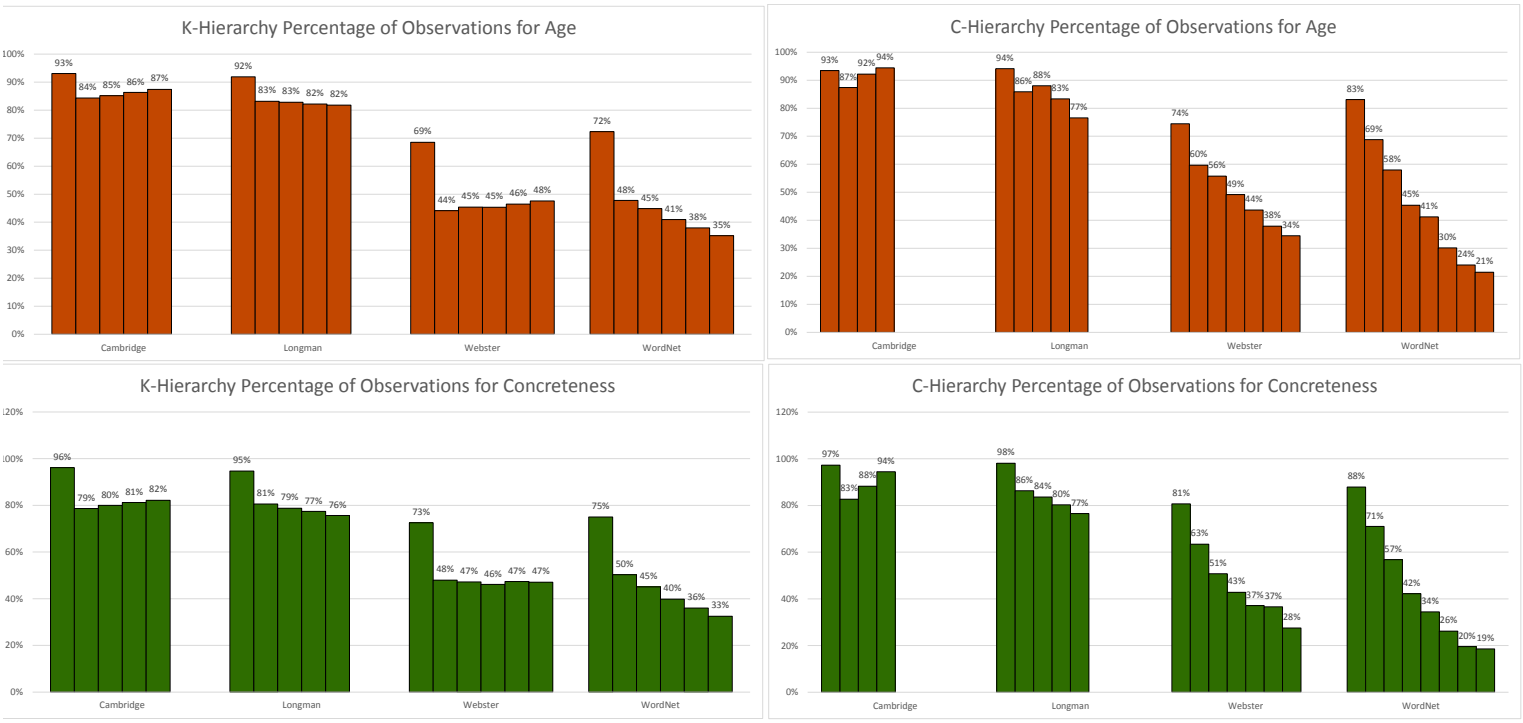

Figure 7. Percentage of words at each level of the definitional distance hierarchy starting from the Kernel through the Rest of dictionary (K-Hierarchy, left), and, only within the Kernel, starting from the Core through the Satellites (C-Hierarchy right), for which psycholinguistic data were available for age and concreteness for each of the four dictionaries. (Frequency data not shown because 100% for all dictionaries.) Note that the percentage of available data is lower for the two bigger dictionaries, and that within the Satellite layer it decreases with increasing definitional distance from the Core in the C-Hierarchy.

Core and Satellite Components of the MinSets. Because it takes so long to compute MinSets, even for the two small dictionaries, we do not have many of them yet; and for the two large dictionaries we so far only have one approximate MinSet each. Every MinSet is part-Core and part-Satellites. A natural question to ask is: What is the difference between the words in these two subcomponents of every MinSet? In the Kernel, the Core is more frequent, younger and less concrete than the Satellites. Comparing the words in the Core component of each MinSet with equal-sized random sets of Core words, and comparing the words in the Satellite component of each MinSet with equal-sized random sets of Satellite words also shows this ratio: For all four dictionaries, the Core component of the Minset is more frequent, younger and less concrete than its random counterparts, and the Satellite component is less frequent, older and more concrete (**Figure 8**). (This effect was confirmed by t-tests ($p<0.001$) for the two smaller dictionaries, for which we had enough Minsets ($n=20$ and $n=19$ for Cambridge and Longman respectively). Because we were only able to compute one MinSet each for the two larger dictionaries, we could not do t-tests, but their pattern of results was the same as for the small dictionaries.) The Core/Satelite effect is hence even more pronounced within the MinSets than within the Kernel.

Comparing the Core, Satellites and Rest in terms of parts of speech again points to the Satellite layer, which has more nouns and fewer adjectives, adverbs and verbs than the Core or the Rest in all four dictionaries **(Figure 9)**. This may be a hint of some sort of functional complementarity between Core and Satellites. Our digraphs and computations treat definitions as if they were just unordered strings of stemmatized content words' first meanings, ignoring syntax and even part of speech – yet definitions themselves are all subject/predicate propositions. It is time to look into this formal black box, at the words themselves. There are very many words in the Core and the Satellites, and very many potential MinSets within each Kernel. But to get a better idea of what the functional role of Core and Satellite words might be in making up a MinSet, we are beginning in ongoing work to examine the respective words themselves, as well as the actual definitions of which they are each a part.

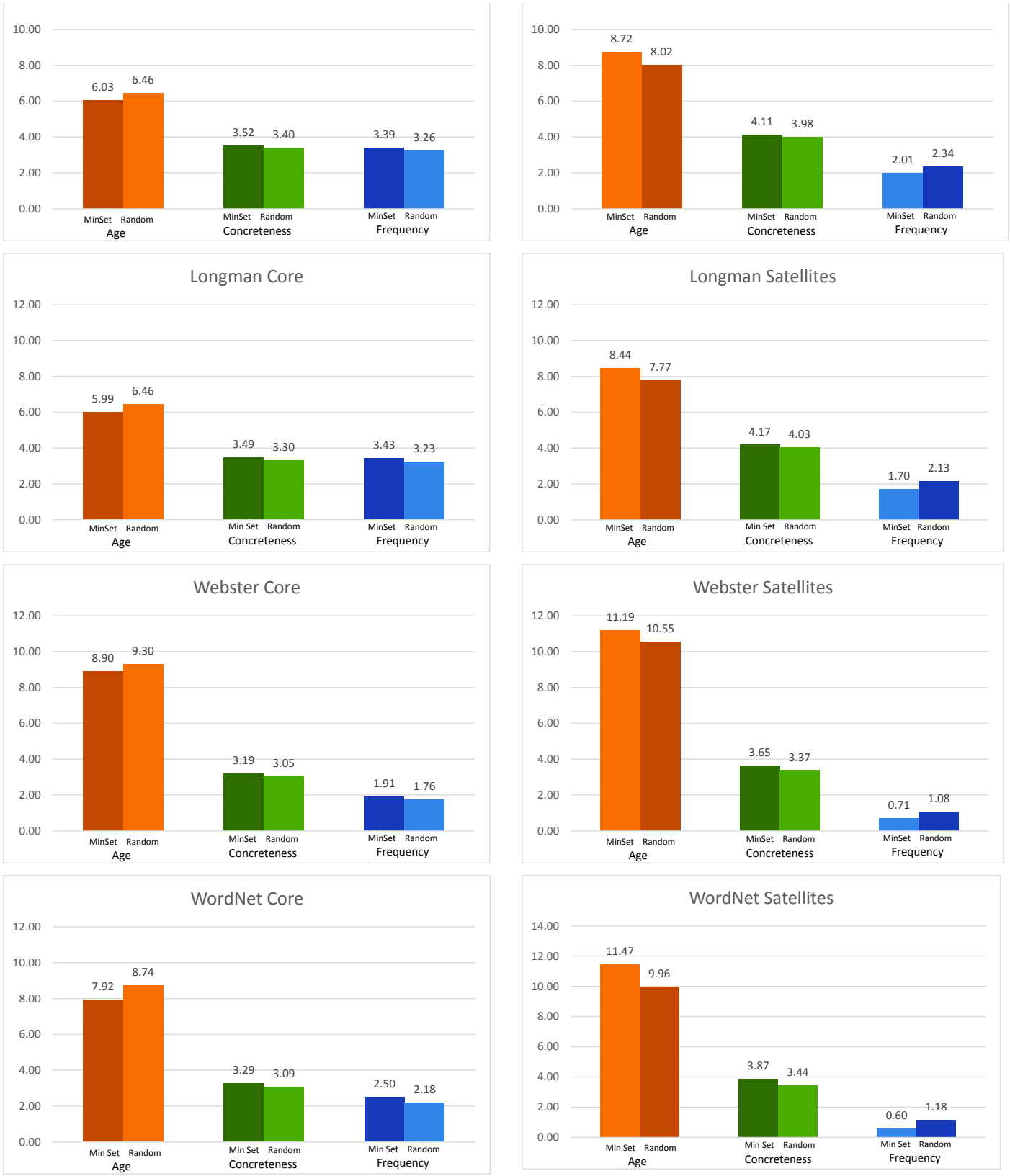

Figure 8. Comparing average age, concreteness and frequency of words in Minsets and equal-sized random subsets of the Core (left) and the Satellites (right) for each of the four dictionaries. In all four dictionaries the average Minsets are younger, more concrete and more frequent than random Core words and older, more abstract and less frequent than random Satellite words.

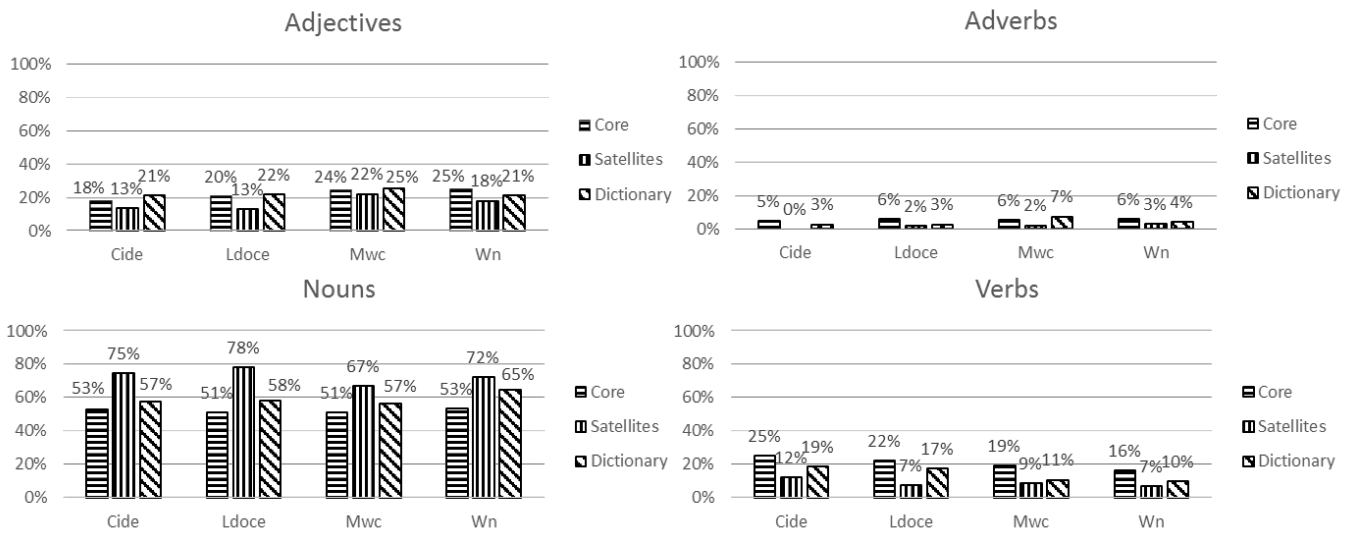

Figure 9. Percentage of parts of speech in the Core, Satellites and Rest for each of the four dictionaries. Note that the percentage of both nouns and adjectives is higher in the Satellite layer, whereas the percentage of verbs and adverbs is lower.

Discussion and Conclusions

What we have learned from the graph-theoretic analysis of dictionaries so far is that knowing the meaning of a grounding set of as few as 373 words for a small dictionary of 25,132 words (1st meanings only) or 1396 words for a larger dictionary of 91,388 words (1st meanings) we could in principle learn the (1st) meanings of all the rest of the words through definition alone. It does not follow, of course, that that is the way we actually do learn the meanings of all the rest of the words. If the grounding set was learned through direct sensorimotor experience, it is probable that a lot of later words are learned in a hybrid way, through a combination of direct experience and verbal definition (or description, or instruction or explanation). Most of our categories are not lexicalized at all, and are described (rather than defined) by ad hoc verbal descriptions: there is no dictionary entry for "things that are bigger than a breadbox," for example, nor for "things that I saw last Tuesday" – nor even, in most people's vocabularies, for "feeling glee at another's misfortune." But even the words in ad hoc verbal descriptions of unlexicalized categories have to be grounded, just as dictionary definitions have to be. So that's back to the grounding set.

Language and Propositions. One can equally well ask: "Why couldn't the meanings of *all* words be learned through direct sensorimotor grounding?" First of all, if it really were possible to learn the meaning of every word through direct sensorimotor experience, then why bother to have words at all? Presumably it is to transmit what one of us has learned (say, via direct experience) to another who has not. Here we cannot avoid considering the question of the nature of language itself, and its adaptive value for our species. No other species speaks (in any modality, including gesture). What do other species lack, and what has ours gained, for having evolved the capacity for language? It is the ability to say anything and everything that can be said: the ability to express every possible proposition.

Most of what we say consists of subject/predicate propositions (even questions and commands). Some propositions are "deictic" which means that they point to the sensorimotor here and now: "She is here." Those are all function words, which were excluded from our dictionary analysis. We were only interested in content words, which, as noted, are the names of categories, and make up almost all the words in the dictionary. The kind of proposition that corresponds to most of what we say (when it is not deictic) is "Apples are red." This could be the reply to someone who does not know, asking "What color are apples?" This is much like someone consulting a dictionary to find out what "apple" means, and learning (to a first approximation), that "An apple is a round, red fruit."

Age, Experience and Abstraction. So far, all these categories could have been learned either from a verbal definition or from direct sensorimotor experience: They are all pretty concrete, and they could all be learned early, fairly quickly, and without any particular risk. "Goodness, truth, and beauty" are becoming more abstract -- although "that's good (true, beautiful)" and "that's not good (true, beautiful)" could be learned from experience too. Learning what "quiddity" or "quark" mean nonverbally, from direct experience, would be quite a bit harder, and the meaning of "peekaboo-unicorn" ("A one-horned horse that vanishes without a trace whenever either senses or an instrument are aimed at it") would be impossible to learn directly via the senses, whereas its verbal definition is as well grounded as the definition of apple.

Now suppose the category that someone lacks is not apples but toadstools, and that the person is starving, and the only thing available to eat is edible mushrooms or poisonous toadstools that look very much like the edible mushrooms. Being told, by someone who knows, that "The striped gray mushrooms are poisonous toadstools" could save someone a lot of time (and possibly their life) by making it unnecessary to find out through direct trial-and-error experience which kind is which.

Category Learning: the Hard and Easy Way. And that, in a nutshell, is our hypothesis about the nature and adaptive value of language (Blondin Massé et al 2013): Language makes it possible to learn new categories by word of mouth, by recombining already grounded category names into propositions, instead of having to do it the hard way, from direct experience. But to do so, some words, at least, still have to be grounded in direct experience. The grounding would need to occur earlier, before the grounded words could be used to define and transmit further categories. And because grounding is sensorimotor, the grounding words would tend to be more concrete.[6]

There is no reason to expect the grounding words to be unique and identical for everyone. The minimal grounding set of any individual's mental lexicon might be like the basis set of an N-dimensional vector space: the basis can generate every point in the vector space, but it is not unique: just a set of N linearly independent points with the property that linear combinations of them can generate any and every point in the vector space. But, because people share a lot of common experiences, and this is in turn reflected in the vocabulary of their language, there is nevertheless reason to expect that some words will be part of many people's grounding vocabularies, so those words would be spoken and written more frequently (see the frequency curves as well as the red curves for intersections in **Figure 6**).

This hypothesis is certainly not *entailed* by our findings on the greater frequency of Kernel words, the earlier age of acquisition of Core words, the greater concreteness of Satellite words, or the multiplicity of MinSets. But if the hypothesis were correct, it would help make sense of some of these findings: Not all. It remains a puzzle why Kernel words in the Satellite layer become increasingly concrete but also older and less frequent, the greater their definitional distance from the Core. We will not understand that until we get a better idea of the complementary role of Core words and Satellite words in making up a MinSet. But even for our two smallest dictionaries there are still very many words in their Kernels (over 2000), and they have very many different MinSets (each of c. 400 words each). So we are currently also generating tiny dictionaries by means of an online dictionary game: The participant is given a word, asked to define it, and then to define the words used to define it, and so on, until all the words used have been defined. This yields dictionaries with an average size of about 200 words, 90% of them in the Kernel, and with MinSets of about 30 words, 2/3 of them Satellite words and 1/3 Core words (which is a reversal of our observed ratio for the full-size dictionaries) (**Table 1**). We hope that these much smaller dictionaries generated by individuals may reflect the way meanings are represented in the

[6] There are some conceptual problems, however, with the notion of concreteness (Borghu & Binkofski 2014), and hence also with judgments of concreteness: To name a kind, like "apple," rather than just a unique individual on a unique occasion, is already to abstract.

mental lexicon and will allow us to get a better idea of the complementary roles played by Core and Satellite words in jointly making up a MinSet.

The question of the causal role of frequency is also an open one. There is no doubt that frequency is correlated with grounding -- the Core words are the most frequent ones, then the Satellites, then the Rest. The frequency gradient within the Satellite layer also follows this pattern, and there is no detectable frequency gradient in the rest of the dictionary, even though the frequency database is 100% complete. It may well be that some words are learned earlier because they are more frequent in the language: But why are they more frequent in the language? Frequency is undeniably the strongest of the psycholinguistic correlates of the latent structures of the dictionary. But its causal role must be explained by something other than frequency: It cannot be frequency all the way down, any more than it can be definitions all the way down. We hope this article will encourage “crowd-sourcing” the analysis of dictionary digraphs for further psycholinguistic variables as well as in further languages.

Apperly, I. (2010) Mindreaders: the cognitive basis of" theory of mind". Psychology Press.

Blondin Masse, A, Chicoisne, G, Gargouri, Y, Harnad, S, Picard, O & Marcotte (2008). How Is Meaning Grounded in Dictionary Definitions? In *TextGraphs-3 Workshop - 22nd International Conference on Computational Linguistics* http://www.archipel.uqam.ca/657/

Blondin Massé, A., Harnad, S., Picard, O. & St-Louis, B. (2013) Symbol Grounding and the Origin of Language: From Show to Tell. In: Lefebvre C, Comrie B & Cohen H (Eds.) *Current Perspective on the Origins of Language*, Benjamin

Borghi, A. M., & Binkofski, F. (2014). The Problem of Definition. In *Words as Social Tools: An Embodied View on Abstract Concepts* (pp. 1-17). Springer New York.

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.

Fellbaum, C. (2010). WordNet. Springer: Netherlands. http://wordnet.princeton.edu

Harnad, S. (1987) Category Induction and Representation. In: Harnad, S. (ed.) (1987) *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

Harnad, S. (2005) To Cognize is to Categorize: Cognition is Categorization. in Lefebvre, C. and Cohen, H., Eds*. Handbook of Categorization*. Elsevier.

Karp, R.M. (1972) Reducibility Among Combinatorial Problems. In R.E. Miller and J.W. Thatcher (editors) *Complexity of Computer Computations*. New York: Plenum, 85-103.

Katz, J. J. (1978). Effability and translation. *Meaning and Translation: Philosophical and Linguistic Approaches*, London: Duckworth, 191-234.

Kucera, H. & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

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, Olivier; Lord, Mélanie; Blondin Massé, Alexandre; Marcotte, Odile; Lopes, Marcos; & Harnad, Stevan (2013) Hidden Structure and Function in the Lexicon, *NLPCS 2013 : 10th International Workshop on Natural Language Processing and Cognitive Science*: 65-77

Procter, P. (1978) *Longman Dictionary of Contemporary English (LDOCE)*. Essex: Longman

Procter, P. (1995) *Cambridge International Dictionary of English (CIDE)*. Cambridge University Press.

Saussure, Ferdinand de (1911/1972) Cours de Linguistique Générale. Paris: Payot.

Steklis, H D and Harnad, S (1976) From hand to mouth: Some critical stages in the evolution of language, In: *Origins and Evolution of Language and Speech* (Harnad, S, Steklis , HD & Lancaster, JB., Eds.), 445-455. *Annals of the New York Academy of Sciences 280*.

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2), 146-160.

Webster, M. (2006). Merriam-Webster online dictionary. Chicago http://www.merriam-webster.com/dictionary.htm

Wilson, M.D. (1988) The MRC Psycholinguistic Database: Machine Readable Dictionary. *Behavioral Research Methods, Instruments and Computers*, 20(1), 6-11.