Competence in lexical semantics

András Kornai
Institute for Computer Science
Hungarian Academy of Sciences
Kende u. 13-17
1111 Budapest, Hungary
andras@kornai.com

Judit Ács
Dept of Automation and
Applied Informatics, BUTE
Magyar Tudósok krt. 2
1117 Budapest, Hungary
judit@aut.bme.hu

Márton Makrai
Institute for Linguistics
Hungarian Academy of Sciences
Benczúr u. 33
1068 Budapest, Hungary
makrai@nytud.hu

Dávid Nemeskey
Faculty of Informatics
Eötvös Loránd University
Pázmány Péter sétány 1/C
1117 Budapest, Hungary
nemeskeyd@gmail.com

Katalin Pajkossy
Department of Algebra
BUTE
Egry J. u. 1
1111 Budapest, Hungary
pajkossy@math.bme.hu

Gábor Recski
Institute for Linguistics
Hungarian Academy of Sciences
Benczúr u. 33
1068 Budapest, Hungary
recski@mokk.bme.hu

Abstract

We investigate from the competence standpoint two recent models of lexical semantics, algebraic conceptual representations and continuous vector models.

Characterizing what it means for a speaker to be competent in lexical semantics remains perhaps the most significant stumbling block in reconciling the two main threads of semantics, Chomsky's cognitivism and Montague's formalism. As Partee (1979) already notes (see also Partee 2013), linguists assume that people know their language and that their brain is finite, while Montague assumed that words are characterized by intensions, formal objects that require an infinite amount of information to specify.

In this paper we investigate two recent models of lexical semantics that rely exclusively on finite information objects: algebraic conceptual representations (ACR) (Wierzbicka, 1985; Kornai, 2010; Gordon et al., 2011), and continuous vector space (CVS) models which assign to each word a point in finite-dimensional Euclidean space (Bengio et al., 2003; Turian et al., 2010; Pennington et al., 2014). After a brief introduction to the philosophical background of these and similar models, we address the hard questions of competence, starting with learnability in Section 2; the ability of finite networks or vectors to replicate traditional notions of lexical relatedness such as synonymy, antonymy, ambiguity, polysemy, etc. in Section 3; the interface to compositional semantics in Section 4; and language-specificity and universality in Section 5. Our survey of the literature is far from exhaustive: both ACR and CVS have deep roots, with significant precursors going back at least to Quillian (1968) and Osgood et al. (1975) respectively, but we put the emphasis on the computational experiments we ran (source code and lexica available at github.com/kornai/4lang).

1 Background

In the eyes of many, Quine (1951) has demolished the traditional analytic/synthetic distinction, relegating nearly all pre-Fregean accounts of word meaning from Aristotle to Locke to the dustbin of history. The opposing view, articulated clearly in Grice and Strawson (1956), is based on the empirical observation that people make the call rather uniformly over novel examples, an argument whose import is evident from the (at the time, still nascent) cognitive perspective. Today, we may agree with Putnam (1976):

> 'Bachelor' may be synonymous with 'unmarried man' but that cuts no philosophic ice. 'Chair' may be synonymous with 'moveable seat for one with back' but that bakes no philosophic bread and washes no philosophic windows. It is the belief that there are synonymies and analyticities of a deeper nature - synonymies and analyticities that cannot be discovered by the lexicographer or the linguist but only by the philosopher - that is incorrect.

Fortunately, one philosopher's trash may just turn out to be another linguist's treasure. What Putnam has demonstrated is that "a speaker can, by all reasonable standards, be in command of a word like *water* without being able to command the intension that would represent the word in possible worlds semantics" (Partee, 1979). Computational systems of Knowledge Representation, starting with the Teachable Word Comprehender of Quillian (1968), and culminating in the Deep Lexical Semantics of Hobbs (2008), carried on this tradition of analyzing word meaning in terms of 'essential' or 'analytic' components.

A particularly important step in this direction is the emergence of modern, computationally oriented lexicographic work beginning with Collins-COBUILD (Sinclair, 1987), the Longman Dictionary of Contemporary English (LDOCE) (Boguraev and Briscoe, 1989), WordNet (Miller, 1995), FrameNet (Fillmore and Atkins, 1998), and VerbNet (Kipper et al., 2000). Both the network- and the vector-based approach build on these efforts, but through very different routes.

Traditional network theories of Knowledge Representation tend to concentrate on nominal features such as the IS_A links (called hypernyms in WordNet) and treat the representation of verbs somewhat haphazardly. The first systems with a well-defined model of predication are the Conceptual Dependency model of Schank (1972), the Natural Syntax Metalanguage (NSM) of Wierzbicka (1985), and a more elaborate deep lexical semantics system that is still under construction by Hobbs and his coworkers (Hobbs, 2008; Gordon et al., 2011). What we call algebraic conceptual representation (ACR) is any such theory encoded with colored directed edges between the basic conceptual units. The algebraic approach provides a better fit with functional programming than the more declarative, automata-theoretic approach (Huet and Razet, 2008), and makes it possible to encode verbal subcategorization (case frame) information that is at the heart of FrameNet and VerbNet in addition to the standardly used nominal features (Kornai, 2010).

Continuous vector space (CVS) is also not a single model but a rich family of models, generally based on what Baroni (2013) calls the *distributional hypothesis*, that semantically similar items have similar distribution. This idea, going back at least to Firth (1957) is not at all trivial to defend, and not just because defining 'semantically similar' is a challenging task: as we shall see, there are significant design choices involved in defining similarity of vectors as well. To the extent CVS representations are primarily used in artificial neural net models, it may be helpful to consider the state of a network being described by the vector whose nth coordinate gives the activation level of the nth neuron. Under this conception, the meaning of a word is simply the activation pattern of the brain when the word is produced or perceived. Such vectors have very large (10^{10}) dimension so dimension reduction is called for, but direct correlation between brain activation patterns and the distribution of words has actually been detected (Mitchell et al., 2008).

2 Learnability

The key distinguishing feature between 'explanatory' or competence models and 'descriptive' or performance models is that the former, but not the latter, come complete with a learning algorithm (Chomsky, 1965). Although there is a wealth of data on children's acquisition of lexical entries (McKeown and Curtis, 1987), neither cognitive nor formal semantics have come close to formulating a robust theory of acquisition, and for intensions, infinite information objects encoding the meaning in the formal theory, it is not at all clear whether such a learning algorithm is even possible.

2.1 The basic vocabulary

The idea that there is a small set of conceptual primitives for building semantic representations has a long history both in linguistics and AI as well as in language teaching. The more theory-oriented systems, such as Conceptual Dependency and NSM assume only a few dozen primitives, but have a disquieting tendency to add new elements as time goes by (Andrews, 2015). In contrast, the systems intended for teaching and communication, such as Basic English (Ogden, 1944) start with at least a thousand primitives, and assume that these need to be further supplemented by technical terms from various domains. Since the obvious learning algorithm based on any such reductive system is one where the primi-

tives are assumed universal (and possibly innate, see Section 5), and the rest is learned by reduction to the primitives, we performed a series of 'ceiling' experiments aiming at a determination of how big the universal/innate component of the lexicon must be. A trivial lower bound is given by the current size of the NSM inventory, 65 (Andrews, 2015), but as long as we don't have the complete lexicon of at least one language defined in NSM terms the reductivity of the system remains in doubt.

For English, a Germanic language, the first provably reductive system is the Longman Defining Vocabulary (LDV), some 2,200 items, which provide a sufficient basis for defining all entries in LDOCE (using English syntax in the definitions). Our work started with a superset of the LDV that was obtained by adding the most frequent words according to the Google unigram count (Brants and Franz, 2006) and the BNC, as well as the most frequent words from a Slavic, a Finnougric, and a Romance language (Polish, Hungarian, and Latin), and Whitney (1885) to form the `4lang` conceptual dictionary, with the long-term design goal of eventually providing reductive definitions for the vocabularies of all Old World languages. Ács et al. (2013) describes how bindings in other languages can be created automatically and compares the reductive method to the familiar term- and document-frequency based searches for core vocabulary.

This superset of LDV, called '4lang' in Table 1 below, can be considered a directed graph whose nodes are the disambiguated concepts (with exponents in four languages) and whose edges run from each definiendum to every concept that appears in its definition. Such a graph can have many cycles. Our main interest is with selecting a defining set which has the property that each word, including those that appear in the definitions, can be defined in terms of members of this set. Every word that is a true primitive (has no definition, e.g. the basic terms of the Schank and NSM systems) must be included in the defining set, and to these we must add at least one vertex from every directed cycle. Thus, the problem of finding a defining set is equivalent to finding a *feedback vertex set,* (FVS) a problem already proven NP-complete in Karp (1972). Since we cannot run an exhaustive search, we use a heuristic algorithm which searches for a defining set by gradually eliminating low-frequency nodes whose outgoing arcs lead to not yet eliminated nodes, and make no claim that the results in Table 1 are optimal, just that they are typical of the reduction that can be obtained by modest computation. We defer discussion of the last line to Section 4, but note that the first line already implies that a defining set of 1,008 concepts will cover all senses of the high frequency items in the major Western branches of IE, and to cover the first (primary) sense of each word in LDOCE 361 words suffice.

Dictionary	**#words**	**FVS**
4lang (all senses)	31,192	1,008
4lang (first senses)	3,127	361
LDOCE (all senses)	79,414	1,061
LDOCE (first senses)	34,284	376
CED (all senses)	154,061	6,490
CED (first senses)	80,495	3,435
`en.wiktionary` (all senses)	369,281	2,504
`en.wiktionary` (first senses)	304,029	1,845
formal	2,754	129

Table 1: Properties of four different dictionaries

While a feedback vertex set is guaranteed to exist for any digraph (if all else fails, the entire set of vertices will do), it is not guaranteed that there exists one that is considerably smaller than the entire graph. (For random digraphs in general see Dutta and Subramanian 2010, for highly symmetrical lattices see Zhou 2013 ms.) In random digraphs under relatively mild conditions on the proportion of edges relative to nodes, Łuczak and Seierstad (2009) show that a strong component essentially the size of the entire graph will exist. Fortunately, digraphs built on definitions are not at all behaving in a random fashion, the strongly connected components are relatively small, as Table 1 makes evident. For example, in the English Wiktionary, 369,281 definitions can be reduced to a core set of 2,504 defining words, and in CED we can find a defining set of 6,490 words, even though these dictionaries, unlike LDOCE, were not built using an explicit defining set. Since LDOCE pioneered the idea of actively limiting the defining vocabulary, it is no great surprise that it has a small feedback vertex set, though everyday users of the LDV may be somewhat sur-

prised that less than half (1,061 items) of the full defining set (over 2,200 items) are needed.

We also experimented with an early (pre-COBUILD) version of the Collins English Dictionary (CED), as this is more representative of the traditional type of dictionaries which didn't rely on a defining vocabulary. In 154,061 definitions, 65,891 words are used, but only 15,464 of these are not headwords in LDOCE. These words appear in less than 10% of Collins definitions, meaning that using LDOCE as an intermediary the LDV is already sufficient for defining over 90% of the CED word senses. An example of a CED defining word missing not just from LDV but the entire LDOCE would be *aigrette* 'a long plume worn on hats or as a headdress, esp. one of long egret feathers'.

This number could be improved to about 93% by detail parsing of the CED definitions. For example, *aigrette* actually appears as crossreference in the definition of *egret*, and deleting the cross-reference would not alter the sense of *egret* being defined. The remaining cases would require better morphological parsing of latinate terms than we currently have access to: for now, many definitions cannot be automatically simplified because the system is unaware that e.g. *nitrobacterium* is the singular of *nitrobacteria*. Manually spot-checking 2% of the remaining CED words used in definitions found over 75% latinate technical terms, but no instances of undefinable non-technical senses that would require extending the LDV. This is not that every sense of every nontechnical word of English is listed in LDOCE, but inspecting even more comprehensive dictionaries such as the Concise Oxford Dictionary or Webster's 3rd makes it clear that their definitions use largely words which are themselves covered by LDOCE. Thus, if we see a definition such as *naphtha* 'kinds of inflammable oil got by dry distillation of organic substances as coal, shale, or petroleum' we can be nearly certain that words like *inflammable* which are not part of the LDV will nevertheless be definable in terms of it, in this case as 'materials or substances that will start to burn very easily'.

The reduction itself is not a trivial task, in that a simplified definition of *naphtha* such as 'kinds of oils that will start to burn very easily and are produced by dry distillation ...' can eliminate *inflammable* only if we notice that the 'oil' in the definition of *naphtha* is the 'material or substance' in the definition of *inflammable*. Similarly, we have to understand that 'got' was used in the sense *obtained* or *produced*, that *dry distillation* is a single concept 'the heating of solid materials to produce gaseous products' that is not built compositionally from *dry* and *distillation* in spite of being written as two separate words, and so forth. Automated detection and resolution of these and similar issues remain challenging NLP tasks, but from a competence perspective it is sufficient to note that manual substitution is performed effortlessly and near-uniformly by native speakers.

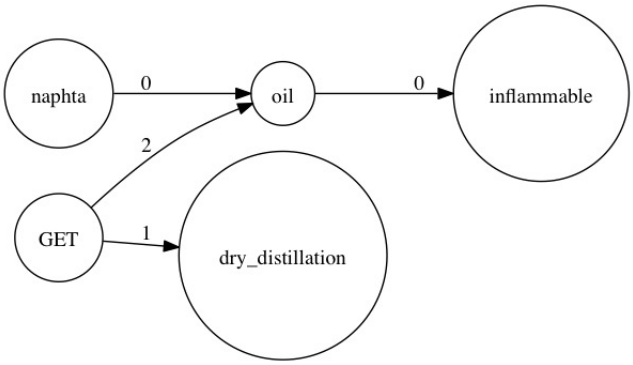

Figure 1: Original definition of `naphtha`

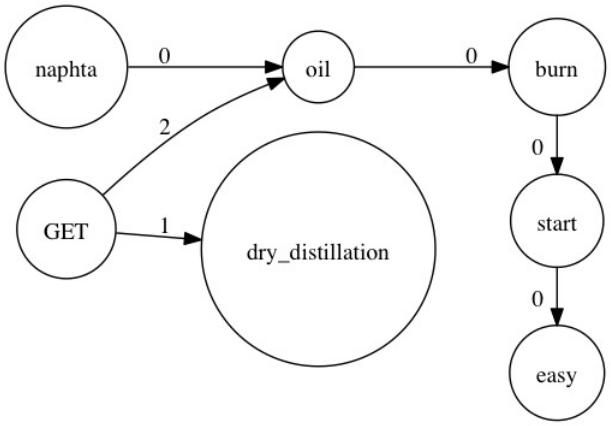

Figure 2: Reduced definition of `naphtha`

2.2 Learnability in CVS semantics

The reductive theory of vocabulary acquisition is a highly idealized one, for surely children don't learn the meaning of *sharp* by their parents telling them it means 'having a thin cutting edge or point'. Yet it is clear that computers that lack a sensory system that would deliver intense signals upon encoun-

tering sharp objects can nevertheless acquire something of the meaning by pure deduction (assuming also that they are programmed to know that cutting one's body will CAUSE PAIN) and further, the dominant portion of the vocabulary is not connected to direct sensory signals but is learned from context (see Chapter 6 of McKeown and Curtis 1987).

This brings us to CVS semantics, where learning theory is idealized in a very different way, by assuming that the learner has access to very large corpora, gigaword and beyond. We must agree with Miller and Chomsky (1963) that in real life a child exposed to a word every second would require over 30 years to hear gigaword amounts, but we take this to be a reflection of the weak inferencing ability of current statistical models, for there is nothing in the argument that says that models that are more efficient in extracting regularities can't learn these from orders of magnitude less data, especially as children are known to acquire words based on a single exposure. For now, such *one shot learning* remains something of an ideal, in that CVS systems prune infrequent words (Collobert et al., 2011; Mikolov et al., 2013a; Luong et al., 2013), but it is clear that both CVS and ACR have the beginnings of a feasible theory of learning, while the classical theory of meaning postulates offers nothing of the sort, not even for the handful of lexical items (tense and aspect markers in particular, see Dowty 1979) where the underlying logic has the resources to express these.

3 Lexical relatedness

Ordinary dictionary definitions can be mined to recover the conceptual entailments that are at the heart of lexical semantic competence. Whatever naphtha is, knowing that it is inflammable is sufficient for knowing that it will start to burn easily. It is a major NLP challenge to make this deduction (Dagan et al. 2006), but ACR can store the information trivially and make the inference by spreading activation.

We implemented one variant of the ACR theory of word meaning by a network of Eilenberg machines (Eilenberg, 1974) corresponding to elements of the reduced vocabulary. Eilenberg machines are a simple generalization of the better known finite state automata (FSA) and transducers (FSTs) that have become standard since Koskenniemi (1983) in describing the rule-governed aspects of the lexicon, morphotactics and morphophonology (Huet and Razet, 2008; Kornai, 2010). The methods we use for defining word senses (concepts) are long familiar from Knowledge Representation. We assume the reader is familiar with the knowledge representation literature (for a summary, see Brachman and Levesque 2004), and describe only those parts of the system that differ from the mainstream assumptions. In particular, we collapse attribution, unary predication, and `IS_A` links in a single link type '0' (as in Figs. 1-2 above) and have only two other kinds of links to distinguish the arguments of transitive verbs, '1' corresponding to subject/agent; and '2' to object/patient. The treatment of other link types, be they construed as grammatical functions or as deep cases or even thematic slots, is deferred to Section 4.

By creating graphs for all LDOCE headwords based on dependency parses of their definitions (the 'literal' network of Table 1) using the unlexicalized version of the Stanford Dependency Parser (Klein and Manning, 2003), we obtained measures of lexical relatedness by defining various similarity metrics over pairs of such graphs. The intuition underlying all these metrics is that two words are semantically similar if their definitions overlap in (i) the concepts present in their definitions (e.g. the definition of both *train* and *car* will make reference to the concept *vehicle*) and (ii) the binary relations they take part in (e.g. both *street* and *park* are `IN` *town*). While such a measure of semantic similarity builds more on manual labor (already performed by the lexicographers) than those gained from state-of-the-art CVS systems, recently the results from the 'literal' network have been used in a competitive system for measuring semantic textual similarity (Recski and Ács, 2015). In Section 4 we discuss the 'formal' network of Table 1 built directly on the concept formulae. By spectral dimension reduction of the incidence matrix of this network we can create an embedding that yields results on world similarity tasks comparable to those obtained from corpus-based embeddings (Makrai et al., 2013).

CVS models can be explicitly tested on their ability to recover synonymy by searching for the nearest word in the sample (Mikolov et al., 2013b); antonymy by reversing the sign of the vector (Zweig, 2014); and in general for all kinds of analogical

statements such as *king is to queen as man is to woman* by vector addition and subtraction (Mikolov et al., 2013c); not to speak of cross-language paraphrase/translation (Schwenk et al., 2012), long viewed a key intermediary step toward explaining competence in a foreign language.

Currently, CVS systems are clearly in the lead on such tasks, and it is not clear what, if anything, can be salvaged from the truth-conditional approach to these matters. At the same time, the CVS approach to quantifiers is not mature, and ACR theories support generics only. These may look like backward steps, but keep in mind that our goal in competence modeling is to characterize everyday knowledge, shared by all competent speakers of the language, while quantifier and modal scope ambiguities are something that ordinary speakers begin to appreciate only after considerable schooling in these matters, with significant differences between the naive (preschool) and the learned adult systems (É. Kiss et al., 2013). On the traditional account, only subsumption (IS_A or '0') links can be easily recovered from the meaning postulates, the cognitively central similarity (as opposed to exact synonymy) relations receive no treatment whatsoever, since similarity of meaning postulates is undefined.

4 Lexical lookup

The interaction with compositional semantics is a key issue for any competence theory of lexical semantics. In the classical formal system, this is handled by a mechanism of *lexical lookup* that substitutes the meaning postulates at the terminal nodes of the derivation tree, at the price of introducing some lexical redundancy rule that creates the intensional meaning of each word, including the evidently non-intensional ones, based on the meaning postulates that encode the extensional meaning. (Ch. 19.2 of Jacobson (2014) sketches an alternative treatment, which keeps intensionality for the intended set of cases.) While there are considerable technical difficulties of formula manipulation involved, this is really one area where the classical theory shines as a competence theory – we cannot even imagine to create a learning algorithm that would cover the meaning of infinitely many complex expressions unless we had some means of combining the meanings of the lexical entries.

CVS semantics offers several ways of combining lexical entries, the simplest being simply adding the vectors together (Mitchell and Lapata, 2008), but the use of linear transformations (Lazaridou et al., 2013) and tensor products (Smolensky, 1990) has also been contemplated. Currently, an approach that combines the vectors of the parts to form the vector of the whole by recurrent neural nets appears to work best (Socher et al., 2013), but this is still an area of intense research and it would be premature to declare this method the winner. Here we concentrate on ACR, investigating the issue of the inventory of graph edge colors on the same core vocabulary as discussed above. The key technical problem is to bring the variety of links between verbs and their arguments under control: as Woods (1975) already notes, the naive ACR theories are characterized by a profusion of link types (graph edge colors).

We created a version of ACR that is limited to three link types. Both the usual network representations (digraphs, as in Figs. 1 and 2 above) and a more algebraic model composed of extended finite state automata (Eilenberg machines) are produced by parsing formulas defined by a formal grammar summarized in Figure 3. For ease of reading, in unary predication (e.g. `mouse` $\xrightarrow{0}$ `rodent`) we permit both prefix and suffix order, but with different kinds of parens `mouse[rodent]` and `rodent(mouse)`; and we use infix notation (`cow MAKE milk`) for transitives (`cow` $\xleftarrow{1}$ `MAKE` $\xrightarrow{2}$ `milk`, link types '1' and '2').

The right column of Figure 3 shows the digraph obtained from parsing the formula on the right hand hand side of the grammar rules. There are no '3' or higher links, as ditransitives like *x give y to z* are decomposed at the semantic level into unary and binary atoms, in this case `CAUSE` and `HAVE`, 'x cause (z have y)', see Kornai (2012) for further details. A digraph representing the whole lexicon was built in two steps: first, every clause in definitions was manually translated to a formula (which in turns is automatically translated into a digraph), then the digraphs were connected by unifying nodes that have the same label and no outgoing edges.

The amount of manual labor involved was considerably lessened by the method of Section 3 that

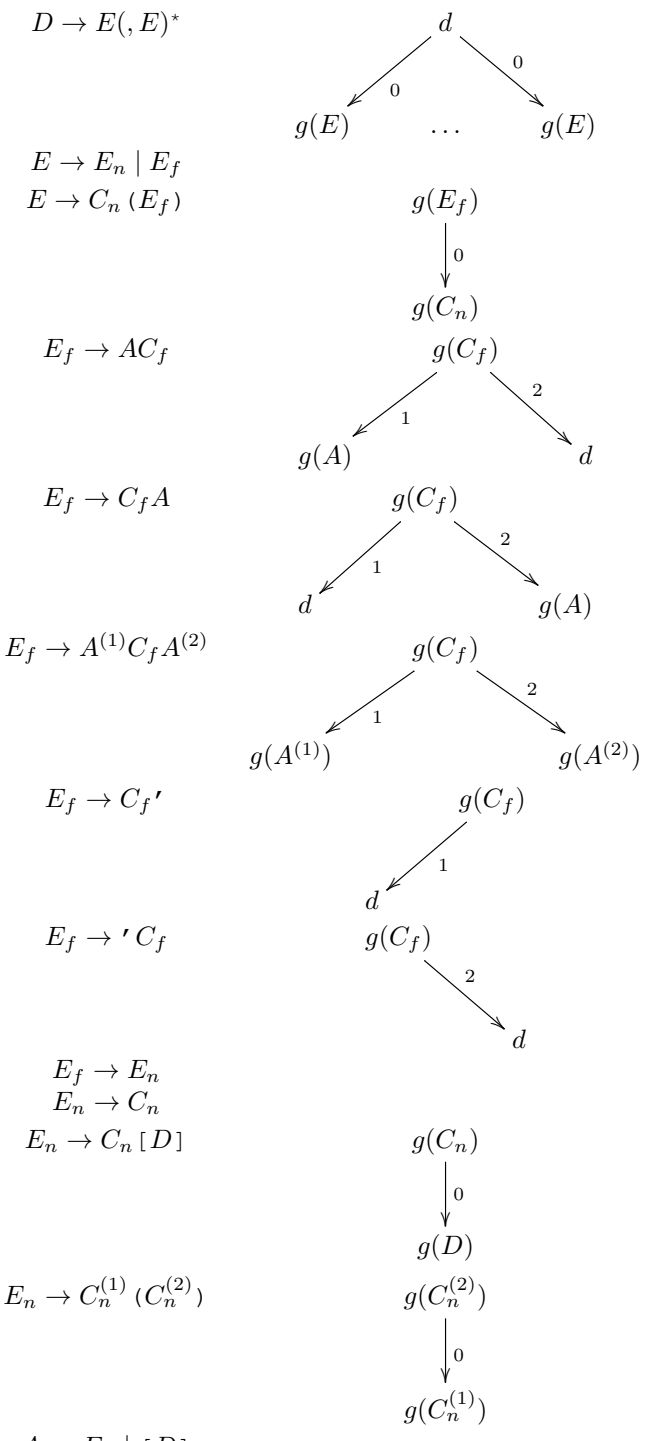

Figure 3: The syntax of the definitions

finds the feedback vertex set, in that once such a set is given, the rest could be built automatically. This gives us a means of investigating the prevalence of what would become different deep cases (colors, link types) in other KR systems. Deep cases are distinguishers that mediate between the purely semantic (theta) link types and the surface case/adposition system. We have kept our system of deep cases rather standard, both in the sense of representing a common core among the many proposals starting with Gruber (1965) and Fillmore (1968) and in the sense of aiming at universality, a subject we defer to the next section. The names and frequency of use in the core vocabulary are given in Table 2. The results are indicative of a primary (agent/patient, what we denote '1'/'2'), a secondary (DAT/REL/POSS), and a tertiary (locative) layer in deep cases – how these are mapped on language-specific (surface) cases will be discussed in Section 5.

freq	**abbreviation**	**comment**
487	AGT	agent
388	PAT	patient
34	DAT	dative
82	REL	root or adpositional object
70	POSS	default for relational nouns
20	TO	target of action
15	FROM	source of action
3	AT	location of action

Table 2: Deep cases

To avoid problems with multiple word senses and with constructional meaning (as in *dry distillation* or *dry martini*) we defined each entry in this formal language (keeping different word senses such as light/739 'the opposite of *dark*' and light/1381 'the opposite of *heavy*' distinct by disambiguation indexes) and built a graph directly on the resulting conceptual network rather than the original LDOCE definitions. The feedback vertex set algorithm uroboros.py determined that a core set of 129 concepts are sufficient to define the others in the entire concept dictionary, and thus for the entire LDOCE or similar dictionaries such as CED or Webster's 3rd. This upper bound is so close to the NSM lower bound of 65 that a blow-by-blow comparison would be justified.

5 Universality

The final issue one needs to investigate in assessing the potential of any purported competence theory is that of universality versus language particularity. For CVS theories, this is rather easy: we have one system of representation, finite dimensional vector spaces, which admits no typological variation, let alone language-specific mechanisms – one size fits all. As linguists, we see considerable variation among the surface, and possibly even among the deeper aspects of case linking (Smith, 1996), but as computational modelers we lack, as of yet, a better understanding of what corresponds to such mechanisms within CVS semantics.

ACR systems are considerably more transparent in this regard, and the kind of questions that we would want to pose as linguists have direct reflexes in the formal system. Many of the original theories of conceptual representation were English-particular, sometimes to the point of being as naive as the medieval theories of universal language (Eco, 1995). The most notable exception is NSM, clearly developed with the native languages of Australia in mind, and often exercised on Russian, Polish, and other IE examples as well. Here we follow the spirit of GFRG (Ranta, 2011) in assuming a common abstract syntax for all languages. For case grammar this requires some abstraction, for example English NPs must also get case marked (an idea also present in the 'Case Theory' of Government-Binding and related theories of transformational grammar). The main difference between English and the overtly case-marking languages such as Russian or Latin is that in English we compute the cases from prepositions and word order (position relative to the verb) rather than from overt morphological marking as standard. This way, the lexical entries can be kept highly abstract, and for the most part, universal. Thus the verb *go* will have a source and a goal. For every language there is a `langspec` component of the lexicon which stores e.g. for English the information that source is expressed by the preposition *from* and destination by *to*. For Hungarian the `langspec` file stores the information that source can be linked by delative, elative, and ablative; goal by illative, sublative, or terminative. Once this kind of language-specific variation is factored out, the `go` entry becomes `before AT src, after AT goal`. The same technique is used to encode both lexical entries and constructions in the sense of Berkeley Construction Grammar (CxG, see Goldberg 1995).

Whether two constructions (in the same language or two different languages) have to be coded by different deep cases is measured very badly, if at all, by the standard test suits used e.g. in paraphrase detection or question answering, and we would need to invest serious effort in building new test suites. For example, the system sketched above uses the same deep case, `REL`, for linking objects that are surface marked by quirky case and for arguments of predicate nominals. Another example is the dative/experiencer/beneficent family. Whether the experiencer cases familiar from Korean and elsewhere can be subsumed under the standard dative role (Fillmore, 1968) is an open question, but one that can at least be formulated in ACR. Currently we distinguish the dative `DAT` from possessive marking `POSS`, generally not considered a true case but quite prevalent in this function language after language: consider English *(the) root of a tree,* or Polish *korzen drzewa*. This is in contrast to the less frequent cases like *(an excellent) occasion for martyrdom* marked by obliques (here the preposition *for*). What these nouns (*occasion, condition, reason, need*) have in common is that the related word is goal of the definiendum in some sense. In these cases we use `TO` rather than `POSS`, a decision with interesting ramifications elsewhere in the system, but currently below the sensitivity of the standard test sets.

6 Conclusion

It is not particularly surprising that both CVS and ACR, originally designed as performance theories, fare considerably better in the performance realm than Montagovian semantics, especially as detailed intensional lexica have never been crafted, and Dowty (1979) remains, to this day, the path not taken in formal semantics. It is only on the subdomain of the logic puzzles involving Booleans and quantification that Montagovian solutions showed any promise, and these, with the exception of elementary negation, do not even appear in more down to

earth evaluation sets such as (Weston et al., 2015). The surprising conclusion of our work is that standard Montagovian semantics also falls short in the competence realm, where the formal theory has long been promoted as offering psychological reality.

We have compared CVS and ACR theories of lexical semantics to the classical approach based on meaning postulates by the usual criteria for competence theories. In Section 2 we have seen that both ACR and CVS are better in terms of learnability than the standard formal theory, and it is worth noting that the number of ACR primitives, 129 in the version implemented here, is less than the dimensions of the best performing CVS embeddings, 150-300 after data compression by PCA or similar methods. In Section 3 we have seen that lexical relatedness tasks also favor ACR and CVS over the meaning postulate approach (for a critical overview of meaning postulates in model-theoretic semantics see Zimmermann 1999), and in Section 4 we have seen that compositionality poses no problems for ACR. How compositional semantics is handled in CVS semantics remains to be seen, but the problem is not a dearth of plausible mechanisms, but rather an overabundance of these.

Acknowledgments

The `4lang` conceptual dictionary is the work of many people over the years. The name is no longer quite justified, in that natural language bindings, automatically generated and thus not entirely free of errors and omissions, now exist for 50 languages (Ács et al., 2013), many of them outside the Indoeuropean and Finnougric families originally targeted. More important, formal definitions of the concepts that rely on less than a dozen theoretical primitives (including the three link types) and only 129 core concepts, are now available for all the concepts. So far, only the theoretical underpinnings of this formal system have been fully described in English (Kornai, 2010; Kornai, 2012), with many details presented only in Hungarian (Kornai and Makrai, 2013), but the formal definitions, and the parser that can build both graphs and Eilenberg machines from these, are now available as part of the `kornai/4lang` github repository. These formal definitions were written primarily by Makrai, with notable contributions by Recski and Nemeskey.

The system has been used as an experimental platform for a variety of purposes, including quantitative analysis of deep cases by Makrai, who developed the current version of the deep case system with Nemeskey (Makrai, 2015); for defining lexical relatedness (Makrai et al., 2013; Recski and Ács, 2015); and in this paper, for finding the definitional core, the feedback vertex set.

Ács wrote the first version of the feedback vertex set finder which was adapted to our data by Ács and Pajkossy, who also took part in the computational experiments, including preprocessing the data, adapting the vertex set finder, and running the experiments. Recski created the pipeline in the `http://github/kornai/4lang` repository that builds formal definitions from English dictionary entries. Kornai advised and wrote the paper.

We are grateful to Attila Zséder (HAS Linguistics Institute) for writing the original parser for the formal language of definitions and to András Gyárfás (HAS Rényi Institute) for help with feedback vertex sets. Work supported by OTKA grant #82333.

References

Judit Ács, Katalin Pajkossy, and András Kornai. 2013. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria, August. ACL.

Avery Andrews. 2015. Reconciling NSM and formal semantics. *ms*, pages v2, jan 2015.

Marco Baroni. 2013. Composition in distributional semantics. *Language and Linguistics Compass*, 7(10):511–522.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Branimir K. Boguraev and Edward J. Briscoe. 1989. *Computational Lexicography for Natural Language Processing*. Longman.

R.J. Brachman and H. Levesque. 2004. *Knowledge Representation and reasoning*. Morgan Kaufman Elsevier, Los Altos, CA.

Thorsten Brants and Alex Franz. 2006. *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, volume 3944 of *LNCS*, pages 177–190. Springer.

David Dowty. 1979. *Word Meaning and Montague Grammar*. Reidel, Dordrecht.

Kunal Dutta and C. R. Subramanian. 2010. Induced acyclic subgraphs in random digraphs: Improved bounds. In *Discrete Mathematics and Theoretical Computer Science*, pages 159–174.

Katalin É. Kiss, Mátyás Gerőcs, and Tamás Zétényi. 2013. Preschoolers interpretation of doubly quantified sentences. *Acta Linguistica Hungarica*, 60:143–171.

Umberto Eco. 1995. *The Search for the Perfect Language*. Blackwell, Oxford.

Samuel Eilenberg. 1974. *Automata, Languages, and Machines*, volume A. Academic Press.

Charles Fillmore and Sue Atkins. 1998. Framenet and lexicographic relevance. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain.

Charles Fillmore. 1968. The case for case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*, pages 1–90. Holt and Rinehart, New York.

John R. Firth. 1957. A synopsis of linguistic theory. In *Studies in linguistic analysis*, pages 1–32. Blackwell.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Andrew S. Gordon, Jerry R. Hobbs, and Michael T. Cox. 2011. Anthropomorphic self-models for metareasoning agents. In Michael T. Cox and Anita Raja, editors, *Metareasoning: Thinking about Thinking*, pages 295–305. MIT Press.

Paul Grice and Peter Strawson. 1956. In defense of a dogma. *The Philosophical Review*, 65:148–152.

Jeffrey Steven Gruber. 1965. *Studies in lexical relations*. Ph.D. thesis, Massachusetts Institute of Technology.

J.R. Hobbs. 2008. Deep lexical semantics. *Lecture Notes in Computer Science*, 4919:183.

Gérard Huet and Benoît Razet. 2008. Computing with relational machines. In *Tutorial at ICON, Dec 2008*.

Pauline Jacobson. 2014. *Compositional Semantics*. Oxford University Press.

Richard M. Karp. 1972. Reducibility among combinatorial problems. In R. Miller and J.W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–104. Plenum Press, New York.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class based construction of a verb lexicon. In *AAAI-2000 Seventeenth National Conference on Artificial Intelligence*, Austin, TX.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430.

András Kornai and Márton Makrai. 2013. A 4lang fogalmi szótár. In Attila Tanács and Veronika Vincze, editors, *IX. Magyar Számitógépes Nyelvészeti Konferencia*, pages 62–70.

András Kornai. 2010. The algebra of lexical semantics. In Christian Ebert, Gerhard Jäger, and Jens Michaelis, editors, *Proceedings of the 11th Mathematics of Language Workshop*, LNAI 6149, pages 174–199. Springer.

András Kornai. 2012. Eliminating ditransitives. In Ph. de Groote and M-J Nederhof, editors, *Revised and Selected Papers from the 15th and 16th Formal Grammar Conferences*, LNCS 7395, pages 243–261. Springer.

Kimmo Koskenniemi. 1983. Two-level model for morphological analysis. In *Proceedings of IJCAI-83*, pages 683–685.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *ACL (1)*, pages 1517–1526.

Tomasz Łuczak and Taral Guldahl Seierstad. 2009. The critical behavior of random digraphs. *Random Structures and Algorithms*, 35:271–293.

Thang Luong, Richard Socher, and Christopher D. Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*, pages 104–113.

Márton Makrai, Dávid Márk Nemeskey, and András Kornai. 2013. Applicative structure in vector space models. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 59–63, Sofia, Bulgaria, August. ACL.

Márton Makrai. 2015. Deep cases in the 4lang conceptlexicon. In Attila Tancs, Viktor Varga, and Veronika Vincze, editors, *X. Magyar Szmtgpes Nyelvszeti Konferencia (MSZNY 2014)*, pages 50–57 (in Hungarian), 387 (English abstract).

Margaret G. McKeown and Mary E. Curtis. 1987. *The nature of vocabulary acquisition*. Lawrence Erlbaum Associates.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, editors, *Proc. ICLR 2013*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Tomas Mikolov, Wen-tau Yih, and Zweig Geoffrey. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT-2013*, pages 746–751.

George A. Miller and Noam Chomsky. 1963. Finitary models of language users. In R.D. Luce, R.R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*, pages 419–491. Wiley.

George A. Miller. 1995. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

T. M. Mitchell, S.V. Shinkareva, A. Carlson, K.M. Chang, V.L. Malave, R.A. Mason, and M.A. Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191.

C.K. Ogden. 1944. *Basic English: A General Introduction with Rules and Grammar*. Psyche miniatures: General Series. Kegan Paul, Trench, Trubner.

Charles E. Osgood, William S. May, and Murray S. Miron. 1975. *Cross Cultural Universals of Affective Meaning*. University of Illinois Press.

Barbara H. Partee. 1979. Semantics - mathematics or psychology? In R. Bäuerl, U. Egli, and A. von Stechow, editors, *Semantics from Different Points of View*, pages 1–14. Springer-Verlag, Berlin.

Barbara Partee. 2013. Changing perspectives on the 'mathematics or psychology' question. In *Philosophy Wkshp on "Semantics Mathematics or Psychology?"*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

H. Putnam. 1976. Two dogmas revisited. *printed in his (1983) Realism and Reason, Philosophical Papers*, 3.

M. Ross Quillian. 1968. Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12:410–430.

Willard van Orman Quine. 1951. Two dogmas of empiricism. *The Philosophical Review*, 60:20–43.

Aarne Ranta. 2011. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford.

Gábor Recski and Judit Ács. 2015. Mathlingbudapest: Concept networks for semantic similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 543–547, Denver, Colorado, June. Association for Computational Linguistics.

Roger C. Schank. 1972. Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, 3(4):552–631.

Holger Schwenk, Anthony Rousseau, and Mohammed Attik. 2012. Large, pruned or continuous space language models on a gpu for statistical machine translation. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 11–19. Association for Computational Linguistics.

John M. Sinclair. 1987. *Looking up: an account of the COBUILD project in lexical computing*. Collins ELT.

Henry Smith. 1996. *Restrictiveness in Case Theory*. Cambridge University Press.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1):159–216.

R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. D. Manning, and A. Y. Ng. 2013. Zero-shot learning through cross-modal transfer. In *International Conference on Learning Representations (ICLR)*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv:1502.05698*.

William Dwight Whitney. 1885. The roots of the Sanskrit language. *Transactions of the American Philological Association (1869-1896)*, 16:5–29.

Anna Wierzbicka. 1985. *Lexicography and conceptual analysis*. Karoma, Ann Arbor.

William A. Woods. 1975. What's in a link: Foundations for semantic networks. *Representation and Understanding: Studies in Cognitive Science*, pages 35–82.

Hai-Jun Zhou. 2013. Spin glass approach to the feedback vertex set problem. *ms, arxiv.org/pdf/1307.6948v2.pdf*.

Thomas E. Zimmermann. 1999. Meaning postulates and the model-theoretic approach to natural language semantics. *Linguistics and Philosophy*, 22:529–561.

Geoffrey Zweig. 2014. Explicit representation of antonymy in language modeling. Technical report, Microsoft Research.