Lexical Semantics and Model Theory: Together at Last?

András Kornai
HAS Institute of Computer Science and Automation
11-13 Kende utca
H-1111 Budapest, Hungary
andras@kornai.com

Marcus Kracht
Bielefeld University
Postfach 10 01 31
33501 Bielefeld, Germany
marcus.kracht@uni-bielefeld.de

Abstract

We discuss the model theory of two popular approaches to lexical semantics and their relation to transcendental logic.

1 Introduction

Recent advances in formal and computational linguistics have brought forth two classes of theories, algebraic conceptual representation (ACR) and continuous vector space (CVS) models. Together with Montague grammar (MG) and its lineal descendants (Discourse Representation Theory, Dynamic Predicate Logic, etc.) we now have three broad families of semantic theories competing in the same space. MG and related theories fit well with most versions of transformational and post-transformational grammar and retain a strong presence in theoretical linguistics, but have long been abandoned in computational work as too brittle (Landsbergen, 1982). As we have argued elsewhere, MG-like theories fail not just as performance grammar but, perhaps more surprisingly, on competence grounds as well (Kornai et al., 2015). Nevertheless, MG will be our starting point, as it is familiar to virtually all linguists.

From an abstract point of view we should distinguish between a framework for compositionality and a commitment to a particular brand of semantics. While we still want to uphold the idea of compositionality, we are less enthusiastic about the dominance of standard first order models, even if suitably intensionalized, in explaining or representing meanings. Luckily, other choices can be made, though they come with a different conception of meaning. The main difference between ACR, CVS, and the standard MG treatment is in fact the choice of model structures: both ACR and CVS aim at modeling 'concepts in the head' rather than 'things in the world', and thus clash strongly with the ostensive anti-psychologism of MG. How can we make sense of such theories after Lewis (1970) without being attacked for promulgating yet another version of markerese? The answer proposed in this paper is that we divest model theory from the narrow meaning it has acquired in linguistics, as being about formulas in some first- or higher-order calculus, and interpret natural language expressions either directly in the models, the original approach of Montague (1970a), or through some convenient knowledge representation language, still composed of formulas, but without the standard logical baggage. The main novelty is that the formulas themselves will be very close to the models, though not quite like in Herbrand models for reasons that will become clear as we develop the theory.

Section 2 provides a brief justification for the enterprise, and sketches as much of ACR and CVS as we will need for Section 3, where essential properties of their models are discussed. Our focus will be on CVS, and we shall discuss the challenge of compositionality, which appears to be nontrivial for CVSs. ACR graphs are simple discrete structures, very attractive for representing meaning (indeed, they have a long history in Knowledge Representation), but more clumsy for syntax. CVS representations, finite dimensional vectors over $\mathbb{R}$, are primarily about distribution (syntactic cooccurrence), and

Proceedings of the 14th Meeting on the Mathematics of Language (MoL 14), pages 51–61,
Chicago, USA, July 25–26, 2015.

meaning, especially the linear structures that encode analogy such as *king:queen* = *man:woman* will arise in them chiefly as a result of probabilistic regularities (Arora et al., 2015). We take the view that CVS models 'concepts in the head' and to understand how these can be similar across speakers we need to invoke 'concepts in the world' as described by ACR. Section 4 discusses the challenge posed by changing to mentalist semantics. If meanings are in the head, we are losing, or so it appears, the objectivity of meanings. However, we think that this is not so. Instead, our working hypothesis of this paper is what we call 'One Reality': meanings describe a common reality so that anything that is true of the world must be compatible with anything else that is true. The section explores some immediate consequences of this hypothesis. We close with some speculative remarks in Section 5.

2 Out with the old, in with the new

Classical MG (Montague, 1970b; Montague, 1973) provides a translation from expressions of natural language into (higher order) predicate logic. Predicate logic itself is just a technical device, a language, to represent the actual meanings, which are thought to reside in models. Thus, already at the inception, formal semantics differentiated two kinds of "semantics": the abstract level, consisting of linguistic objects (here: expressions of simple type theory), and the concrete level, represented by a model. In what is to follow we shall investigate the effects of making two changes. One is to replace the simple type theory by radically different kinds of semantics, and the second to uphold the idea that the semantics is not just about some model, but about reality, and as such cannot be arbitrarily fixed.

Let us briefly recall how a Montague-style semantics looks like. Following (Kracht, 2011), a grammar consists of a finite signature (F, Ω) of function symbols ($\Omega : F \to \mathbb{N}$ assigns an arity to the symbols), together with an interpretation that interprets each function symbol f as an $\Omega(f)$-ary function on the space of signs, (see also Hodges 2001). Further down we shall meet only two kinds of functions: constants, where $\Omega(f) = 0$, representing the lexicon, and binary functions ($\Omega(f) = 2$), representing syntax proper.) We may take as signs either pairs (w, m), where w is a word over the alphabet and m its meaning; or we may take them as triples (w, c, m), where c is an additional component, the *category* (Kracht, 2003). In the best of all cases, the action of f on the signs is independent in each of the components. The independence of the string action from the meanings is exactly Chomsky's famous principle of the *autonomy of syntax* while the independence of the meaning action from the words is the principle of *compositionality*. If these are granted, each function symbol f then gives rise to a *pair* of functions (f^ε, f^μ), where f^ε is an $\Omega(f)$-ary function on strings and f^μ an $\Omega(f)$-ary function on meanings. Further, given any constant term t over this signature, "unfolding" (the homomorphic operation denoted by $\spadesuit$) it into a sign means

$$(f(t_1, t_2))^\spadesuit = (f^\varepsilon(t_1^\spadesuit, t_2^\spadesuit), f^\mu(t_1^\spadesuit, f_2^\spadesuit))$$

and for 0-ary f, simply $f^\spadesuit() = (f^\varepsilon(), f^\mu())$. Omitting obvious brackets this is simply $f^\spadesuit = (f^\varepsilon, f^\mu)$. A constant therefore is defined by its two components, the string (f^ε) and the meaning (f^μ).

Effectively, we can now view not only the terms as elements of an algebra (called the *term algebra*), but also the strings together with the functions f^ε ($f \in F$), and the meanings together with the functions f^μ ($f \in F$). Expressions and meanings thus become algebras, and there are then two homomorphisms from the algebra of terms: one to the algebra of strings and another to the algebra of meanings. Both algebras may have additional functions, of course. This will play a role in the case of CVS models which have the structure of a vector space, whose natural operations enter into the definition of the functions.

When we say 'out with the old' we will not dwell much on the inadequacies of the standard MG treatment, except to summarize some of the well known issues. Technical inadequacies, ranging from narrow issues of proposing invalid readings and missing valid ones to more far-reaching problems as provided e.g. by hyperintensionals (Pollard, 2008) are not viewed as fatal – to the contrary, these provide the impetus for further developments. A more general, systemic issue however is the chronic *lack of coverage*. The problem is not so much that the pioneering examples from *Every man loves a woman such that she loves him* to *John seeks a unicorn*

and Mary seeks it could hardly be regarded examples of ordinary language as the alarming lack of progress in this regard – forty years have passed, and best of breed implementations such as CatLog (Morrill, 2011) and grammar fragments such as Jacobson (2014) still cover only a few dozen constructions. An equally deep, and perhaps even more critical, problem is the continuing disregard for *information*. No matter how we look at it, well over 80% of the information carried by sentences comes from the lexicon, with only 10-15% coming from compositional structure (Kornai, 2010). By putting lexical semantics front and center, we will address both these issues.

One of the biggest challenges for MG is the disambiguation. In the standard picture, readings correspond to parse terms. Thus, a string w has as many readings as there are parse terms t such that $t^\varepsilon = w$. Unfortunately, scholars in the MG tradition have spent little effort on building grammatical models of natural language that could serve as a starting point of disambiguation in the sense Montague urged, and the disambiguation in terms of parse terms is more a promissory note than an actual algorithm. This is particularly clear as we come to effects of *contextuality*, restated from Frege by Janssen (2001) as follows: 'Never ask for the meaning of a word in isolation, but only in the context of a sentence'.

In other words, what a particular word means in a sentence can be determined only by looking at the context, since the context selects a particular reading. The standard MG picture handles lexical ambiguity by invoking separate lexical entries (that is, 0-ary function symbols) for each sense a word may have, e.g. for pen_1 'writing instrument' and pen_2 'enclosed area for children or cattle'. When we say *The box is in the pen* we clearly have pen_2 in mind, and when we say *The pen is in the box* it is pen_1. Strict adherence to MG orthodoxy demands that we bite the bullet and claim that the false readings are actually wonderful to have, since a smaller playpen could really be delivered in a box, and even for a large cattle pen an artist like Christo could always come by and box up the entire thing. Yet somehow the claim rings false, both from a cognitive standpoint, since the odd readings do not even enter our mind when we hear the sentence unless we are specifically primed, and from the computational standpoint, since it is common knowledge (at least since Bar-Hillel (1960) where the box/pen example originates) that the bulk of the effort e.g. in machine translation is to disambiguate the word meanings. According to Bar-Hillel, the average English word is 3-way ambiguous, so a sentence of length 15 will require over 14 million disambiguated options. However much our computational resources have grown since 1960, and they have actually grown more than 14 million-fold, this is still unrealistic.

Another part of the theory that remained, for the past forty years, largely unspecified, is the mapping g that would *ground* elements of the mathematical model structure in reality (as opposed to the 'valuation' that is built into the model structure). For a mathematical theory, such as the theory of groups, there is no need for g as such in that there are no groups "in the world". All objects in mathematics that have group structure (e.g. the symmetries of some geometrical figure) can be built directly from sets (since a symmetry is a function, and functions are sets), so restricting attention to model structures that are sets is entirely sufficient for doing mathematics. Here we must give some thought to what we consider 'ground truth', a notion that is already problematic for proper names without referents such as *Zeus*.

The abstract structure outlined above does not require the meanings to be anything in particular. All that is required is that we come up with an algebra of meanings into which the terms can be mapped so that certain equations, the meaning postulates, come out true. Our exercise consists therefore in throwing out the old semantics and bringing in the new, here CVS and ACR, and see where this leads us. When we say 'in with the new' this is something of an exaggeration – both ACR and CVS theories go back to the late 1960s and early 1970s, and are thus as old as MG, except both suffered a long hiatus during the 'AI Winter'. Algebraic conceptual representation (ACR) begins with Quillian (1969) and Schank (1972), who put the emphasis on associations (graph edges) between concepts (graph nodes). Quillian only used one kind of (directed) edge, while Schank used several – the ensuing proliferation of link types is famously criticized in Woods (1975). For a summary of the early work see Findler (1979), for modern treatments see Sowa (2000), Banarescu

et al. (2013).

Continuous vector space (CVS) representation was first developed by (Osgood et al., 1975), whose interest is also with association between concepts, which they directly measured by asking informants to rate the strength of the association on a 7-point scale. From such data, Osgood and his coworkers proceeded by data reduction via principal component analysis (PCA), obtaining vectors that were viewed as directions in semantic space. In the modern version, which has taken computational linguistics by storm in the past five years, the associations are mined from cooccurrence data in large corpora (Schütze, 1993), but data reduction by PCA or similar techniques is still a central part of establishing the mapping from the vocabulary V to $\mathbb{R}^n$.

Importantly, both ACR and CVS are essentially type free. They assume that the representation of the whole utterance is not any different from the representation of the constituents, down to the lexical entries: in ACR every meaning is a graph, and in CVS a vector. As said above, there are two types of function symbols. Those of arity 0 constitute the *conceptual dictionary*. The remaining function symbols are of arity 2. On the string side they are interpreted as concatenation, giving rise to a CFG. For ACR, the meanings of the parts are combined by ordinary substitution operations, graph rewriting and adjunction. For CVS, several combination operations have been proposed, including vector addition (Mitchell and Lapata, 2008), coordinatewise (weighted) multiplication (Dinu and Lapata, 2010), function application (Coecke et al., 2010) and substitution into recurrent neural nets (Socher et al., 2013). For a summary, see Baroni (2013). Here we will use $\otimes$ to denote any composition operation, as tensorial products have long been suggested in this area (Smolensky, 1990).

A key point is that $\otimes$ itself may be parametrized, more similar to the 'type-driven' versions of MG (Klein and Sag, 1985) than to the classic variant which has a single composition operation, function application. Berkeley Construction Grammar (CxG, see Goldberg 1995) has long urged a full theory of constructional meanings, and Kracht (2011) makes clear that languages must employ many, many simple constructions, if as above compositionality and autonomy of syntax are assumed.

3 The structure of CVS and ACR model structures

Recall that the functions f^ε and f^μ ($f \in F$) impose an algebraic structure both on the set of exponents and the set of meanings, respectively. There may be additional structure on the meanings, which we may take advantage of. For example, if meanings are vectors, we additionally have scalar multiplication and addition, which can be used in calculations, but which also have their own semantic relevance. Indeed, it has been observed by Mikolov et al. (2013) that in an analogy $a : b = c : d$ we can calculate v_d approximately as $v_a - v_b + v_c$. Or, what is the same, we expect $v_a - v_b = v_c - v_d$.

The currently best performing Context Vector Grammar (CVG, see Socher et al. 2013) uses what looks like a single binary function $\otimes$, however it is parametrized by part of speech. CVGs work on ordered pairs $(\vec{v}, X)$ where $\vec{v}$ contributes the semantics, and X is some part of speech category (including nonterminals such as NP). In our notation $(\vec{v}, X)$ combines with $(\vec{w}, Y)$ by two square matrices L_{XY}, R_{XY} and a bias $\vec{b}_{XY}$ that depend on X and Y (but not on $\vec{v}$ or $\vec{w}$) to yield $\vec{v} \otimes \vec{w} = \tanh(L\vec{v} + R\vec{w} + \vec{b})$ (dropping the parts of speech) where the squishing function $\tanh$ is applied coordinatewise.

Since $\tanh$ is strictly monotonic, we have $x = y$ iff $\tanh(x) = \tanh(y)$, so the last step of squishing can be ignored in the kind of equational deduction that we will deal with. As an example, consider the `gram3-comparative` task. It is an accident of English that comparative is sometimes denoted by the suffix *-er* and sometimes by the prefix *more* written as a separate word. Ideally, the semantics should support equations such as

$$\vec{big} \otimes \vec{er} - \vec{nice} \otimes \vec{er} = \vec{big} - \vec{nice} \qquad (1)$$

or, equivalently,

$$\tanh(L\vec{big}+R\vec{er}+\vec{b})-\tanh(L\vec{nice}+R\vec{er}+\vec{b}) = \vec{big} - \vec{nice}$$

In reality both the matrix and the vector coefficients are small enough for $\tanh(x) = x$ to be a reasonable approximation, so we have

$$L\vec{big} - L\vec{nice} = \vec{big} - \vec{nice} \qquad (2)$$

or, what is the same, $(L-I)(\vec{big}-\vec{nice})=0$ not just for *big* and *nice* but for every pair of adjective vectors $\vec{u},\vec{v}$. This is possible only if $\langle A\rangle$, the subspace generated by the adjectives, is contained in $\mathrm{Ker}(L-I)$. Since L does not even need to be defined outside $\langle A\rangle$, and must coincide with I within $\langle A\rangle$, the simplest assumption is $L=I$ everywhere. Now, R and b are fixed for the comparative task, so $R\vec{er}+\vec{b}$ is some constant vector $\vec{c}$ on $\langle A\rangle$, so that we finally get

$$\forall \vec{x}\in\langle A\rangle : \vec{x}\otimes\vec{er}=\vec{x}+\vec{c} \tag{3}$$

and obviously if (3) holds the analogical requirement in (1) is satisfied. The same argument can be made (with different constant $\vec{c}$) for every derivational and inflexional suffix such as the *-ly* of the `gram1-adjective-to-adverb` or the *-ing* of the `gram5-present-participle` Google task. Further, the same must hold for every case where a fixed formative is used to derive a higher constituent, such as PP[from] from a base NP and a prefix *from*, or NP from a base N and the prefix *the*. Remarkably, just as PP[from] can differ from PP[by] only by a fixed offset, the difference between the constant for *from* and that for *by*, NP[every] and NP[some] can also differ only in a fixed offset irrespective of what the base N was.

This shows how analogies can help in identifying the functions for certain derivations. However, more can be achieved. Consider the case of two synonymous expressions e and e'. Retracing their respective parses, assuming that the result vectors are the same we derive further constraints. Consider *the mayor's hat* and *the hat of the mayor* which should get the same vector assigned compositionally through two different routes. If $\vec{m}$ and $\vec{h}$ are the vectors for *mayor* and *hat*, we have some $\vec{m}+\vec{c_1}$ for *the mayor* and $\vec{h}+\vec{c_1}$ for *the hat*. If the *'s* possessive construction is defined by matrices L_1, R_1 and bias $\vec{b_1}$, and the *of*-possessive by $L_2, R_2, \vec{b_2}$, the fact that these mean the same will be expressed, again ignoring the squishing, by

$$\begin{aligned} &L_1(\vec{m}+\vec{c_1})+R_1\vec{h}+\vec{b_1} \\ &\qquad = L_2(\vec{h}+\vec{c_1})+R_2(\vec{m}+\vec{c_1})+\vec{b_2} \end{aligned} \tag{4}$$

By collecting like terms together, this means

$$(L_1-R_2)\vec{m}+(R_1-L_2)\vec{h}+\vec{c_4}=\vec{0} \tag{5}$$

for some constant $\vec{c_4}$ and for all noun vectors $\vec{m},\vec{h}$. This of course requires $L_1=R_2, L_2=R_1$ and $\vec{c_4}=0$, meaning that the two constructions differ only in the order they take the possessor and possessed arguments. Also, if instead of *((the hat) of (the mayor))* we had chosen the structure *(the (hat of (the mayor)))* the matrices would be the same.

To summarize, all productive derivational and inflectional processes will have the output differ from the input by some constant $\vec{c}$ that depends only on the construction in question, and the same goes for all 'syntactic' processes such as forming a PP or NP whose output differs from its input only by the addition of some fixed grammatical formative, including the formation of modal verb complexes (*must go, will eat,* ...) by a fixed auxiliary. Note that such processes crosslinguistically often end up in the morphology, cf. Romanian *-ul* 'the' or Hungarian *-val/vel* 'with'.

An important consequence of what we said so far is that the effects of fixed formatives, be they attached morphologically or by a supporting clitic or full word, are commutative. This explains how even closely related languages like Finnish and Hungarian can have different conventional suffix orders (e.g. between case endings and possessive endings), as it takes no effort to rejigger the semantics with a change of inflection order. Also, a good number of bracketing paradoxes (Williams, 1981; Spencer, 1988) simply disappear: in light of commutative semantics brackets are not at all called for, and the 'paradox' is simply a by-product of an overly detailed (context free) descriptive technique.

The less productive a process, the less compelling the argument we made above, since it depends on some identity holding not just for a handful of vectors but for an entire subspace generated by the part of speech class of the input. For example the morphologically still perceptible relatedness of latinate prefixes and stems (Aronoff, 1976) as in *commit, remit, permit, submit, compel, repel, impel, confer, refer, infer,* ... will hardly allow for computing separate vectors for *con-, re-,* ... on the one hand and *pel, mit, sume, fer, ceive,* ... on the other as we have

too many unknowns for too few equations. Or consider *bath:bathe, sheath:sheathe, wreath:wreathe, teeth:teethe, safe:save, strife:strive, thief:thieve, grief:grieve, half:halve, shelf:shelve, serf:serve, advice:advise, ...* where the relationship between the noun and the verb is quite transparent, yet the set on which the rule applies is almost lost among the much larger set of nouns that can be 'verbed' by zero affixation or stress shift alone.

This is not to say that suppletive forms, such as found in irregular plurals or strong verbs are outside the scope of our finding, for clearly if plural formation is the addition of a single fixed $\vec{c}$ in all regular cases, $\vec{horses} = \vec{horse} + \vec{c}$, we must also have $\vec{oxen} = \vec{ox} + \vec{c}$ since the analogy *horse:horses=ox:oxen* is intact. But given their paucity, derivational forms may still be sensitive to order of affixation, so that something like the Mirror Principle (Baker, 1985) may still make sense.

Looking at the 882 L and R matrices (25 by 25 dimensions) in the CVG instance available as part of the Stanford Dependency Parser, we note that over half (55% for L, 53% for R) of the variance in this set is explained by the first 25 eigenmatrices, so the structure is likely considerably simpler than the full CVG model allows for. We tested this hypothesis by grammars CVG$^{(k)}$ constructed from the Socher et al. (2013) CVG$^{(882)}$ by replacing all 882 L and R matrices by approximations based on the first k eigenmatrices (middle column of Table 1). The case $k = 1$ corresponds to the earlier RNN (Socher et al., 2011) with a single global $\otimes$, and gets only 81.0% on the WSJ task. As we increase the number of coefficients kept, we obtain results closer and closer to the original CVG$^{(882)}$: at $k = 100$ we are already within 1% of the full result.

k	no I	I first
1	81.02	
5	82.85	84.59
25	86.88	89.32
50	88.50	90.07
100	89.47	90.24
200	90.08	90.32
882	90.36	90.36

Table 1 Parsing performance as a function of the number of coefficients kept in $\otimes$ definitions

As Socher et al. (2013) already observe, the diagonal of the L (resp. R) matrix is dominant for left- (resp. right-)headed endocentric constructions, so we also experimented with keeping only $k - 1$ of the eigenmatrices and replacing the kth by I before finding the best approximations (right column of Table 1). With this choice of basis, the phenomenon is even more marked: it is sufficient to keep the top 24 (plus the coefficient for I) to get within 1% of the original result.

By limiting k we can limit the actual information content of $\otimes$, which would otherwise grow quadratically in d. Given that the 882 matrix pairs were already abstracted on the basis of sizeable corpora (63m words from the Reuters newswire, see Turian et al. 2010), direct numerical investigation of the 882 $\otimes$ operators to detect this simpler structure faces stability issues. In fact, it is next to impossible to guess, based strictly on an inspection of the eigenmatrices, that replacing the least one by I would be advantageous – for this we need to have a more model-based strategy, to which we now turn.

We speak about distributions in two main senses: discrete (class-level) and continuous (item-level). The distinction is reflected in the notation of generative grammar as between preterminals and terminals, and in the practice of language modeling as between states and emissions of Hidden Markov Models (HMMs). In generative grammar, the class-level distribution is typically conceived of in 0-1 terms: either a string of preterminals is part of the language or it is not – weighted grammars that make finer distinctions only became popular in the 1980s, decades after the original work on constituency (Wells, 1947; Harris, 1951; Chomsky, 1957). The standard (unweighted) grammar already captures significant generalizations such that A+N (adjective followed by noun) is very likely in English, while N+A is more likely in French. However, as (Harris, 1951) already notes,

> All elements in a language can be grouped into classes whose relative occurrence can be stated exactly. However, for the occurrence of a particular member of one class relative to a particular member of another class, it would be necessary to speak in terms of probability, based on the frequency of that occurrence in a sample.

Retrofitting generative rules such as $N \rightarrow AN$ achieves very little, in that it is not clear which adjective will go with which noun. As (Kornai, 2011) noted, HMM transition probabilities tend to stay in a relatively narrow range of $10^{-4} - 10^{-1}$ (the low values typically coming from smoothing) while emissions can span 8-9 orders of magnitude – this is precisely why n-gram HMMs remain a viable alternative to PCFGs to this day. CVS models capture a great deal of the distributional nuances because the vectors encode not just an estimate of unigram probabilities

$$\log(p(w)) = \frac{1}{2d}||\vec{w}||^2 - \log Z \pm o(1) \quad (6)$$

but also a cooccurrence estimate

$$\log p(w, w') = \frac{1}{2d}||\vec{w} + \vec{w'}||^2 - 2\log Z \pm o(1) \quad (7)$$

for some fixed Z (Arora et al., 2015). For unigrams, the GloVe dictionary (Pennington et al., 2014) actually shows a Pearson correlation of 0.393 with the Google 1T frequencies and 0.395 with the BNC. While these are not bad numbers, (especially considering that G1T and BNC only correlate to 0.882), clearly a lot more need to be done before (7) becomes realistic. Table 2 shows some frequent, rare, and nonexistent A+N combinations together with their Google 1T frequency; the right-hand side of eq. (7); the scalar product of the GloVe word vectors; and their cosine angles.

A-N pair	freq	(6) rhs	$\langle , \rangle$	cos
popular series	95k	-3.153	13.95	0.39
popular guidance	127	-3.158	2.80	0.08
popular extent	0	-3.175	6.78	0.23
rapid development	299k	-3.137	20.40	0.50
rapid place	182	-3.165	7.88	0.25
rapid percent	0	-3.115	11.30	0.24
private student	134k	-3.121	16.79	0.37
rare student	989	-3.133	5.30	0.13
cold student	0	-3.121	4.58	0.10

Table 2 Cooccurrence predictors for frequent, rare, and nonexistent adjective+noun combinations

Evidently, GloVe captures a great deal of the distribution, clearly ranking the frequent above the rare/nonexistent both in unnormalized (scalar product) and normalized (cosine) terms, while (7) largely obscures this. All of these predictors fare badly when it comes to comparing rare to nonexistent forms. (Of course Google 1T 'nonexistence' only means 'below the cutoff' but here this is as good as nonexistence since such pairs don't participate in the training.) It is reasonable to conclude that embeddings model the high- to mid-range of the distribution quite well, but fail on very rare data, which call for a corrective term in the Arora et al. estimate in Eq. (7).

Remarkably, word similarity measures based on definitional similarity do nearly as well on semantic world similarity tasks as those based on distributions (Recski and Ács, 2015). These definitions, common to ACR models, manifest no distributional similarity between definiendum and definiens, compare *rascal* to *a child who behaves badly but whom you still like*. Yet when we compare *rascal* to *imp* 'a child who behaves badly, but in a way that is funny' the similarity becomes evident: both *rascal* and *imp* are defined as 'children behaving badly'. There are many idiosyncratic traits to these words, for example both *little rascal* and *little imp* are plausible, but *??old imp* is not, even though *old rascal* is. More often than not, these differences in distribution have to do with accidents of history rather than any semantic difference to speak of – this is especially clear on the case of exact synonyms like *twelve* and *dozen*.

Here we simply assume that observable distribution is the result of two factors: pure syntax, as expressed by the system of lexical (part of speech) categories such as N, and their projections such as NP, and pure semantics, expressed by their conceptual representations. The manner these two factors combine is not transparent, we hope to address the issue in a follow-on paper.

4 One Reality

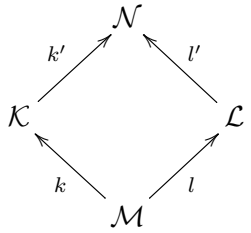
Let us return to the question posed above concerning 'real' meanings: the challenge is not so much to encode meanings into some clever abstract language but to actually account for their successful use in conversation. If we believe in a common reality about which we talk to each other, meanings have to have a property that allows them to be merged in

a particular way: anything that is true of the world must be compatible with anything else that is true. Yet, reality is not given to us in one fell swoop but rather needs to be explored. Despite the fact that we think of the one real model as the justification of our way of talking, we can only hypostatise its existence and take it from there. The constructed models of reality must form a family of models each approaching the single one. This can be explored in two ways. One way is to insist that any language – even an abstract one – is already equipped with a realist interpretation, and that leads to what is known as Robinson Consistency and the so-called Joint Embedding Property. The second approach considers only the constructed models as given and constructs reality out of them. This leads us to inverse systems of models, or dually, direct systems of algebras.

The models of choice, we argue here, are ACR representations, in essence graphs with colored edges. Some additional markup may be necessary on the nodes (to govern the loci of substitution/adjunction operations) and some additional constraints (in particular limiting out-degrees) may hold, but on the whole such structures are well understood. CVS models may stand in various relations to one another, in a manner far more complex than the alternative relations familiar from Kripke-style models. For example, embeddings I_p and I_q created from the same raw data by PCA but keeping a different number of dimensions $p < q$ are in an *extension of* relation which we can state directly on the corresponding models as $\mathcal{M}_p < \mathcal{M}_q$ where $<$ means 'can be embedded in'.

In ACR, there are many cases when one model structure can be embedded in the other, central among these being the case of the smaller structure simply containing fewer *existents* than the larger one. (The term 'existent' is a bit awkward, but helps to avoid non-Meinongian ontological commitments: in a model whose base elements are graphs or vectors corresponding to *mountain* and *gold*, $I(gold) \otimes I(mountain)$ is an 'existent'.) Moreover, if $\mathcal{K}, \mathcal{L}$ are isomorphic substructures of $\mathcal{M}$, the isomorphy between $\mathcal{K}$ and $\mathcal{L}$ can be extended to an automorphism of $\mathcal{M}$, making model structures *homogeneous* in the sense of Fraïssé (1954).

For the graph structures to actually be *models* they must satisfy certain requirements. The requirements are sine qua non because the models are models of something, namely, in first approximation, external reality. If models are about external reality then it follows that there can be only one. As such however it is not to be found in anyone's head. Instead, we picture the acquisition of the model structure as a process that walks through a number of smaller model structures, expanding them as new information comes in. The process of expansion by necessity produces substructures of one bigger structure. Thus, the classes of model structures must satisfy what is known as the *amalgamation property* that for each $\mathcal{K}, \mathcal{L}, \mathcal{M}$ where we have k and l embeddings of $\mathcal{M}$ into $\mathcal{K}$ and $\mathcal{L}$ respectively, we have some $\mathcal{N}$ and embeddings k' and l' of $\mathcal{K}$ and $\mathcal{L}$ into $\mathcal{N}$ such that the following diagram commutes:



On the logical side we expect a joint consistency in the spirit of Robinson' theorem: if T_1 and T_2 are two theories such that the intersection is consistent and there is no formula φ such that $T_1 \vdash \varphi$ while $T_2 \vdash \neg\varphi$, then $T_1 \cup T_2$ is consistent. Assuming that the world is consistent, we expect this behaviour. Let U be the intersection of T_1 and T_2. Suppose our database is U. Then after some steps of learning we may end up in T_1 or in T_2. However, both states cannot be in conflict by deriving one of them a formula and the other its negation. So, they are jointly consistent.

This property makes perfect sense for lexical entries, where extending a model $\mathcal{M}$ with new entries to build $\mathcal{K}$ or $\mathcal{L}$ can be amalgamated to produce $\mathcal{N}$. What this means, in naive terms, is that the lexicon harbors no contradictions. To see that this is already a non-empty requirement, consider the lexical entry for *cancer* which will, under the ACR theory, contain an IS_A link to *incurable*. When (hopefully soon) a cure is found, this means that the lexical entry itself will have to be revised, just as gay marriage forced the revision of 'between a man and a woman'. More significant are the contradictory cases, for in-

stance when in one extension we learn that Colonel Mustard killed Mr. Boddy, and in another we learn that Professor Plum did. Admitting model structures that harbor internal contradictions (as in paraconsistent logic) clashes with the use of a single model; an alternative that suggests itself is to allow for much richer embeddings, e.g. ones that contain propositional attitude clauses: Miss Scarlet believes that Colonel Mustard killed Mr. Boddy, while Mrs. Peacock believes it's Professor Plum.

As the last example shows, there is an additional complicating factor at play. Even if we assume the model to be a model of a single reality, this grounding model may vary from person to person as in the 'lifelong' DRT of (Alberti, 2000). Communication may reveal that this is the case, but the remedy is not simple. Differences may arise about facts of the matter as well as over meanings, hence they may concern either the grounding model itself ('reality') or the map g that grounds the meanings (Kracht, 2011). Thus, the fact that language is shared among a group of individuals in and of itself calls for a different approach in model theory. This must be left for another occasion, however.

If we believe in a single and unique model structure, we will assume that any model we build must be embeddable into the one existing model. Thus, we must have the *joint embedding property* (JEP) for the family of 'candidate' models of reality. This property requires that for any two models $\mathcal{K}, \mathcal{L}$ the existence of an $\mathcal{N}$ in which both can be embedded. Such a consistency requirement must be made if we insist that all semantics is about a single external reality.

However, suppose the real model is unknown, even unknowable. Then, if we want to understand what it means to talk about real objects appeal to external reality is futile if all we have is appearances. This is where Kant saw the need of a logic he called *transcendental*. The transcendental object is so to speak the limit of approximation made by our inquiry. It is our construction of reality, which rationalises our previous models as being about *something*. In this connection it is rather interesting to note the proposal by Achourioti and van Lambalgen (2011) concerning the transcendental logic. The authors propose that what Kant had actually in mind was what is nowadays called an *inverse system*. This is a family of models $\mathcal{M}_s$ indexed by a poset $(S, \leq)$ such that for all s, t there is r such that $s, t \leq r$, together with maps $h_{st} : \mathcal{M}_s \to \mathcal{M}_t$ for $s \geq t$ satisfying $h_{tr} \circ h_{st} = h_{sr}$. Even if there is no unique model structure, the system of model structures itself is *objective* in Kant's sense (that is *about an object*) if it has the structure of an inverse system; and the transcendental object itself can somehow be imagined as a member of the inverse limit of that system. What is interesting to note is that the formulae for which the transition from the inverse system to the inverse limit is what is known as *geometrical formulae*, having the form $\forall \overline{x}(\varphi(\overline{x}) \to \exists \overline{y}\theta(\overline{x}, \overline{y}))$.

5 Conclusions

As usual, model theory does not solve many outstanding problems, but brings a great deal of much needed clarity in organizing the minor variants one could conceive of. As long as we stay with a purely deductive apparatus, we have to figure out whether natural deduction, Beth tableaux, Hilbert systems, sequent calculi, or some new combination of the above is what we use, and this inevitably gets mixed up with other design choices we have within the ACR/CVS world. (Also, for reasons shrouded in the mists of history, proof theory somehow has a very bad reputation within linguistics.)

This paper has taken the first, rather tentative steps towards understanding the structure of the new model structures. We have seen that operations of inflectional morphology, whether realized by actual inflection or by function words, amount to a shift by a constant vector. Second, we have seen that data can be pooled across semantically equivalent but syntactically different constructions such as the *of* and *'s* possessives. Third, we have seen that the numerical limitations of the current model make it impossible to explore the low frequency tail of the distribution where many phenomena of great linguistic interest, such as causativization and other forms of predicate decomposition, are to be found. Even so, our results in Table 1 make clear that the actual complexity of construction operations is considerably less than the POS-pair assumption built into CVGs would suggest.

The use of existents provides, for the first time we

believe, a reasonable framework to approach both standard and Meinongian ontology on equal footing. This is not to say that one has to be committed to some higher plane of ideal existence where the entirety of Meinong's Jungle is present, to the contrary, all one needs is a notion of finitely generated models, and a compositional semantics that is willing to interpret $a \otimes b$ based on the interpretation of a and b. Similarly, the key property of Fraïssé homogeneity is the one at stake in the entire philosophical debate surrounding *inverted qualia* (see Byrne (2014) for a summary). What is clear is that automorphisms mapping one synonym on another can be extended to automorphisms of the whole lexicon, but from here on one may take several paths depending on one's philosophical predilections.

Many questions remain open, and perhaps more importantly, many questions can be meaningfully asked for the first time. The traditional riddle of *class meanings* (how nouns designate 'things', adjectives 'qualities', and verbs 'actions') is now amenable to empirical work relating the vectors of pronouns, proadjectives and other pro-forms to the center of gravity of $\langle N \rangle, \langle A \rangle, \ldots$. On the pure semantics side, we may begin to see how, by finite mechanism, humans are capable of infinite comprehension, learning of (transcendental) objects.

Acknowledgments

We thank Gábor Recski (HAS Research Institute for Linguistics) for performing the PCA on the $\otimes$ operators of the Socher et al. (2013) CVG.

References

Gábor Alberti. 2000. Lifelong discourse representation structure. *Gothenburg Papers in Computational Linguistics*.

Mark Aronoff. 1976. *Word Formation in Generative Grammar*. MIT Press.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2015. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *arXiv:1502.03520v1*.

Mark Baker. 1985. *Incorporation: a theory of grammatical function changing*. MIT.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Yehoshua Bar-Hillel. 1960. A demonstration of the nonfeasibility of fully automatic high quality translation. In *The present status of automatic translation of languages*, volume Advances in Computers I, pages 158–163.

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

Alex Byrne. 2014. Inverted qualia. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2014 edition.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *arXiv:1003.4394v1*.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. pages 1162–1172.

Nicholas V. Findler, editor. 1979. *Associative Networks: Representation and Use of Knowledge by Computers*. Academic Press.

Roland Fraïssé. 1954. Sur l'extension aux relations de quelques propriétés des ordres. *Ann. Sci. Ecole Norm. Sup*, 71:361–388.

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

Zellig Harris. 1951. *Methods in Structural Linguistics*. University of Chicago Press.

Wilfrid Hodges. 2001. Formal features of compositionality. *Journal of Logic, Language and Information*, 10:7–28.

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

T.M.V. Janssen. 2001. Frege, contextuality and compositionality. *Journal of Logic, Language and Information*, 10(1):115–136.

Ewan Klein and Ivan Sag. 1985. Type-driven translation. *Linguistics and Philosophy*, 8:163–201.

András Kornai, Judit Ács, Márton Makrai, Dávid Nemeskey, Katalin Pajkossy, and Gábor Recski. 2015. Competence in lexical semantics. To appear in Proc. *SEM-2015.

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. 2011. Probabilistic grammars and languages. *Journal of Logic, Language, and Information*, 20:317–328.

Marcus Kracht. 2003. *The Mathematics of Language*. Mouton de Gruyter, Berlin.

Marcus Kracht. 2011. *Interpreted Languages and Compositionality*, volume 89 of *Studies in Linguistics and Philosophy*. Springer, Berlin.

Jan Landsbergen. 1982. Machine translation based on logically isomorphic montague grammars. In *Proceedings of the 9th conference on Computational linguistics-Volume 1*, pages 175–181. Academia Praha.

D. Lewis. 1970. General semantics. *Synthese*, 22(1):18–67.

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

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.

Richard Montague. 1970a. English as a formal language. In R. Thomason, editor, *Formal Philosophy*, volume 1974, pages 188–221. Yale University Press.

Richard Montague. 1970b. Universal grammar. *Theoria*, 36:373–398.

Richard Montague. 1973. The proper treatment of quantification in ordinary English. In R. Thomason, editor, *Formal Philosophy*, pages 247–270. Yale University Press.

Glynn Morrill. 2011. CatLog: A categorial parser/theorem-prover. In *Type Dependency, Type Theory with Records, and Natural-Language Flexibility*.

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

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)*.

Carl Pollard. 2008. Hyperintensions. *Journal of Logic and Computation*, 18(2):257–282.

M. Ross Quillian. 1969. The teachable language comprehender. *Communications of the ACM*, 12:459–476.

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.

Hinrich Schütze. 1993. Word space. In SJ Hanson, JD Cowan, and CL Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 895–902. Morgan Kaufmann.

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

Richard Socher, Cliff Chiung-Yu Lin, and Christopher D Manning. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proc. 28th ICML*.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing with compositional vector grammars. In *The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*.

J.F. Sowa. 2000. *Knowledge representation: logical, philosophical, and computational foundations*. MIT Press.

Andrew Spencer. 1988. Bracketing paradoxes and the English lexicon. *Language*, 64:663–682.

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.

Michiel van Lambalgen and Theodora Achourioti. 2011. A Fromalization of Kant's Transcendental Logic. *The Review of Symbolic Logic*, 4:254 – 289.

Roulon S. Wells. 1947. Immediate constituents. *Language*, 23:321–343.

Edwin Williams. 1981. On the notions 'lexically related' and 'head of a word'. *Linguistic Inquiry*, 12:245–274.

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