Organon F 23 (3) 2016: 299-323

Distribution and Inference: What Philosophical and Computational Semantics can Learn from Each Other

RADEK OCELÁK

Institute of Philosophy. Academy of Sciences of the Czech Republic
Jilská 1. 110 00 Praha 1. Czech Republic
radioc@seznam.cz

RECEIVED: 07-03-2016 ▪ ACCEPTED: 26-04-2016

ABSTRACT: Distribution of a word across contexts has proved to be a very useful approximation of the word's meaning. This paper reflects on the recent attempts to enhance distributional (or vector space) semantics of words with meaning composition, in particular with Fregean compositionality. I discuss the nature and performance of distributional semantic representations and argue against the thesis that semantics is in some sense identical with distribution (which seems to be a strong assumption of the compositional efforts). I propose instead that distribution is merely a reflection of semantics, and a substantially imperfect one. That raises some doubts regarding the very idea of obtaining semantic representations for larger wholes (phrases, sentences) by combining the distributional representations of particular items. In any case, I reject the generally unquestioned assumption that formal semantics provides a good theory of semantic composition, which it would be desirable to combine with distributional semantics (as a theory that is highly successful on the lexical field). I suggest that a positive alternative to the strong reading of the distributional hypothesis can be seen in the philosophy of inferentialism with respect to language meaning. I argue that the spirit of inferentialism is reasonably compatible with the current practice of distributional semantics, and I discuss the motivations for as well as the obstacles in the way of implementing the philosophical position in a computational framework.

KEYWORDS: Lexical semantics – distribution – compositionality – inferentialism.

1. Introduction

One of the most crucial insights of the present-day computational, application-oriented approach to the semantics of natural language is this: we can usefully capture the meaning of a word by characterizing its distribution, or the contexts in which the word appears. As one famous aphorism goes, "you shall know a word by the company it keeps" (Firth 1957, 11). This proclamation may sound odd, and surely there are many ways of reading it. But it has been made clear by now that at least in some readings, the "distributional hypothesis" lends itself to remarkably successful computational applications. Models based on this insight have been applied to a variety of semantic tasks. Even if the results are still far from perfection, they generally seem to be far above anything achieved, first, in the other paradigms of semantic thinking, such as formal or cognitive semantics, and second, in the computational semantic branches that draw their inspiration from them.

Neither of these two points is quite surprising. As concerns the latter point, the distributional formulation of the natural language meaning problem is the key that enables us to treat the problem based on large amounts of actual language data, using the mechanical efficiency of a computer, or many computers at a time. It thus offers an interesting alternative to relying on our creative (see Schneider 1992) but relatively inefficient minds operating with language intuitions (which are, moreover, sometimes not too reliable). In the simplest case, word meanings as mysterious objects exclusively accessed by human minds are replaced by word meanings as patterns of textual co-occurrence of the target words with other words. Textual words being nothing but sequences of characters, that provides for efficient processing of the language material collected in extremely large corpora of written text. State-of-the-art models in distributional computational semantics are nowadays standardly built upon corpora containing billions of lexical tokens.

As concerns the former point, we might argue that the superior results in applications follow from the very nature of computational semantics, and computational linguistics in general. Computational linguistics differs from the theoretical approaches to language rather substantially in its orientation. At least as much as for theoretical *understanding* of language phenomena, the struggle here is for efficient "engineering" solutions to well-defined applied problems (such as machine translation or automatic summarization). Providing a good engineering solution nonetheless does not imply knowing *why* the solution approximates the related phenomenon of natural language, or actually understanding that phenomenon. One might therefore claim that theoretical semanticists need not be unsettled by the success in application achieved by their computational colleagues.

In this paper, however, I would like to reverse that perspective in the following way. The tasks considered in the computational paradigm, and the distributional branch in particular, are not aimed to capture any sort of detached mechanical processes unrelated to human language use. Instead, they closely resemble some of the tasks that any competent speaker is likely to perform on a daily basis. Every now and then, we are expected to paraphrase, summarize, distinguish between two senses of a word, choose an appropriate synonym, sometimes even translate, etc. Suppose that a machine achieves human-like mastery in the *whole* spectrum of semantic tasks of which our everyday struggle with language consists. Then from a pragmatic point of view there will be little reason to claim that what the machine does has nothing to do with "real" semantics. Given the psychological and neurological aspects of our semantic competence, this machine will obviously not embody all there is to such a competence (or to the *implementation* thereof in our minds and brains). But it is also clear that the position that such an intelligent device has nothing whatsoever to teach us about "real" semantics would be absurd. Since the rise of automatic dishwashers, there have not been many complaints to the effect that what they actually do has nothing in common with *true* dish-washing as performed by humans.

Yet we are still nowhere near that ultimate stage in distributional computational semantics, and in the following I will try to argue that with a purely engineering approach we are not on our way there either. This is where theoretical understanding comes in. We should not claim that a machine ideally performing in semantic tasks would provide no such understanding to us. But it seems equally clear that without theoretical understanding of language we will not be able to bring a machine to such an ideal performance level (or any close to it) in the first place. Even if you occupy yourself with fairly practical tasks, you should not systematically ignore what appears as a good theory. Otherwise you might find yourself in the position of someone who keeps driving nails with a screwdriver, refusing all theoretical lessons in the mechanics of a hammer.

For this reason, I find it appropriate in this paper to combine two perspectives that are seemingly quite disparate in their assumptions and goals. First is that of distributional semantics, as a very fruitful—although by no means exclusive—branch of computational semantics. Second is that of inferentialism, as a position in the philosophy of semantics. Interbreeding two remote perspectives, the paper of course runs the risk of not being digestible for either party. But I think that both approaches to natural language meaning can be mutually enhanced as to their *own* respective goals: success in semantic applications on one hand, understanding of meaning in language (and crucially, *validation* of such understanding) on the other. My hope is that the reader, having successfully navigated between the computational Scylla and the philosophical Charybdis, will be in a position to judge whether this claim is a correct one.

The paper is further structured as follows. In section 2, the stage is set by characterizing distributional semantics as to its basic ideas, methods, results and their broader significance. Further, some recent ideas regarding the possible enrichment of distributional semantics with semantic composition are discussed. (Except for the second part of 2.3, which is more critical in character, the whole section 2 is meant to be fairly consensual, and a reader who is familiar with distributional semantics and its recent development should feel free to just scan through it.) In section 3, I reflect on the theoretical status of distributional semantics, more specifically the question of the relation between distribution and meaning. I argue in favor of a weak, rather than strong, reading of the distributional hypothesis. In section 4, I return to the performance of distributional semantics from a more critical angle. With the observations made, I try to support the position that there is a serious gap between distribution and meaning, and I draw some consequences for the project of compositional distributional semantics. Finally, in section 5, I work towards presenting the inferentialist approach to semantics as a positive and viable alternative to the strong version of distributionalism.

2. Distributional semantics

2.1. Distributional semantic models

In this section, I outline the most important features of the distributional program in computational semantics. Note that this is just a very basic sketch. A much more thorough picture of the framework, its origins, assumptions, methods, goals and results can be found in works such as Lenci (2008), Turney and Pantel (2010), or Erk (2012).

Let me start the presentation with a toy example. Assume that the following table expresses how often each of the target words *dog*, *cat*, *tortoise*, *comb* occurred in the proximity of the words *hair* and *run* in a toy corpus.

	hair	*run*
dog	6	7
cat	8	6
tortoise	0	2
comb	5	0

Each row of the table determines a vector in a two-dimensional space, where each dimension corresponds to one of the context words; so, e.g., the vector for *cat* begins in the point [0,0] and ends in [8,6].

The distributional hypothesis generally states that the meaning of a word can be approximated by its pattern of occurrence in various contexts. Now, since the vector of each of the four target words is defined to (partly) capture just such a distributional pattern, we may decide to treat it as a *semantic representation* of the word in question. An important feature of vector semantic representations is that they are graded: a set of such representations is not merely a list of items (such as, for instance, the set of entries in a dictionary). We have a graded measure of similarity for any two of them: the angle formed by the two vectors in question, or more conveniently, the cosine of that angle. The smaller the angle (higher the cosine), the more semantic similarity we should expect between the words represented by these vectors. Thus in our toy example, at least some of the predictions will appear quite intuitive. (That is how the example is made up, of course.) *Cat* will come out as fairly similar in meaning to *dog*; *tortoise* not so much; *comb* will come out as particularly dissimilar from *tortoise*. One should note that no semantic information in any traditional sense went into these representations. All the table contains are (hypothetical but arguably plausible) co-occurrence counts of particular *words*.

There are literally dozens of reasons why the above does not constitute an adequate semantic analysis of the target words *dog*, *cat*, *tortoise*, and *comb*. However, a more interesting question is, which of the problems are—or can be—overcome by scaling the approach up with the available computing power, and by considering the many variants of the model that have been explored in distributional semantics up to now?

It is just for the sake of illustration that the previous example works with a small number of vectors in a two-dimensional space constituted by two context words, reflecting co-occurrence counts in a very small (hypothetical) corpus. In fact, the simple mathematics employed is easily generalized to multidimensional spaces with an arbitrary number of context dimensions. Thus the state-of-the-art distributional semantic models typically contain vectors for many thousands target words, vectors that "live" in several hundreds of dimensions. (Usually these are secondary dimensions which are gained from the original dimensions, given by many thousands of context words, by means of dimensionality reduction techniques.) As has been mentioned already, it is nowadays possible to build the vectors based on the co-occurrence counts in corpora of several billion textual words. That is, current distributional semantic models try to approximate lexical meaning using amounts of distributional information that are utterly incomparable to the toy example above.

Further, there are many alternatives to using the raw word co-occurrence counts as the basis of semantic representation. Some sort of automatic reweighting of these counts is usual, or even necessary, so as to ensure that the more informative co-occurrences (such as that between *dog* and *bark*) will count more than those which are frequent but rather uninformative (e.g., *dog* and *the*). Also the notion of occurrence in a context can be made precise in various ways. Sometimes, it is defined as occurrence within a textual "window" of *n* word positions to the left and to the right from a particular token of the context word. Another option is to look for *any* co-occurrences within a single web-based document. It is possible to define the occurrence contexts in terms of lemmas[1] rather than plain word forms; or we can define the contexts with the use of syntactic characteristics (such as *dog* in the syntactic function of a direct object). The last two options depend on there being a method of automatic lemmatization or syntactic parsing applicable in the whole extent of the primary corpus, which is supposed to be as large as possible.

In theory (much less in practice, so far), *extralinguistic* contexts are considered as well. The fact that current models almost exclusively work with *textual* distribution seems to be a matter of contingent limitations rather than of

[1] Lemma is a representative form standing for the plurality of forms a lexical item can take, such as *bark* for *bark*, *barks*, *barked*, *barking*.

a theoretical commitment (cf. Lenci 2008, 10). Apparently, distributionalists are prepared to include as contexts whatever is technically manageable in a sufficiently large scale. For instance, some models derive their sets of contexts from large databases of labeled images. That seems important for the philosophical assessment of the program, for in this, distributionalism arguably diverges from the narrow, intralinguistic distributional analysis once practiced by the American linguistic descriptivism.[2] At the same time, it comes closer to the use-theoretic view of meaning originating from later Wittgenstein. After all, the hypothesis that the meaning of an expression is a matter of *where* it is used differs from the famous Wittgensteinian dictum solely by replacing *how* with *where*. That seems to open some room for a use-theoretic reappraisal of the distributional program, attempted in section 5.[3]

[2] Zellig Harris, the main descriptivist figure, is seen as a precursor of distributional semantics by Lenci (2008, 3ff.).

[3] It should be noted, finally, that there is also what Baroni et al. (2014a) call a new generation of distributional semantic models, represented notably by Mikolov et al. (2013). They are models that grew in the natural language processing field and the now dramatically developing area of neural network research, quite independently of the distributional tradition outlined above, which has more connections to theoretical linguistics. These models, referred to as *neural network language models* or *context-predicting models*, also semantically represent words with vectors in a multidimensional space. Instead of counting co-occurrences and applying heuristic transformations, however, the vectors are estimated by means of automatic learning, optimizing the success in the prediction of missing words in a known context. The evaluation by Baroni et al. (2014a) indicates, to the authors' own surprise, that these models perform consistently better than the traditional distributional models. In the following, context-predicting models are not systematically addressed. While I originally thought most of the critical considerations in this paper would apply to these models as well, Tomáš Musil (personal communication) pointed out to me an important difference which might prevent this from being the case. Namely, in context predicting models, the change in the semantic representation of an expression permeates further into the system by bearing on the representations of other expressions. That is not true in the traditional distributional models, where an expression's semantic representation is given by its co-occurence with other expressions but not by the representations of those expressions (which are, again, defined in terms of their own co-occurences).

2.2. The performance of distributional models

The previous technical characterization of distributional semantic models might appear omissible from the standpoint of some philosophical preconceptions about meaning which we may hold. But it is useful to see some details of the techniques that achieve as much in practical terms as distributional semantic models do. These models have been applied, with non-negligible success, to a variety of semantic tasks. From the theoretical perspective, many of these tasks are, in some form, part and parcel of our everyday operating with language. From the perspective of computational linguistics, methods successful in dealing with the tasks are likely to contribute to final language processing applications such as machine translation or question answering systems.

For instance, the performance of distributional models on the task of synonym detection is rather impressive, at least at first glance. The well-known TOEFL test consists of 80 multiple-choice questions where the subject is asked to pick one synonym for the target word out of four candidates (e.g., to choose the synonym *imposed* for the target *levied* from the candidate set *believed*, *imposed*, *correlated*, *requested*). In this test, the most successful distributional models, relying exclusively on the similarity of the vector representations of the words in question, are able to match in performance and even outperform the average college-educated native speaker of English (cf. Landauer and Dumais 1997; Baroni and Lenci 2010; Baroni et al. 2014a). Other tasks in which distributional models enjoy highly non-trivial success include, among others, prediction of human judgments of semantic similarity and relatedness, categorization of concepts into natural categories, detection of relational analogies (such as, *brother* is to *sister* as *grandson* is to *granddaughter*), even prediction of the psycholinguistic effect of semantic priming; (see, e.g., Erk 2016; Baroni and Lenci 2010; Baroni et al. 2014a; Baroni et al. 2014b; and their references.)

This is not to say that the current distributional models are able to solve all the semantic tasks that an average human speaker can, and with comparable accuracy. In fact, there is much that they *cannot* do in any satisfactory manner. (I will go into some detail in section 4.) But it is very much worth attention that they achieve relative success, and even approximate human performance, in *some*—undeniably semantic—tasks. This is especially manifest in comparison with the situation in formal semantics. In that field, there exists very little transparent evaluation in terms of what the proposed models can actually *do*, which can be probably linked to the fact that they cannot do much in practical terms. (That seems to be agreed upon by the critics and the outsiders as well as the insiders of formal semantics, even if the other opinions regarding the value of formal semantic work differ; cf. Maddirala 2014.) By contrast, in computational semantics a lot of attention is traditionally paid to evaluation against independent data, and a substantial part of work goes into devising new evaluation methods, sets of testing data, etc.

Another difference from the more theoretical approaches to semantics, which is however closely related to the previous, is that distributional models require little[4] or no human "supervision", little or no semantic information brought in manually by semantically competent humans. They can thus be automatically trained for tens of thousands of target and context words on huge amounts of actual language data. This is not the case with formal semantic representations, which are typically crafted manually, as if one by one, by a semanticist, based on a small sample of actual language instances. (Here, I gloss over the fact that formal semantics hardly ever deals with problems of lexical meaning, whereas distributional semantics is, to a large extent, lexical semantics.) This is clearly an important part of the relative practical success of distributional semantics: with the limited descriptive capacities of individual humans, it is hard, or extremely expensive, to cover the vastness of human language use.

One more fact can be noted in favor of distributional vectors as genuine semantic representations in some sense, rather than as mere *ad hoc* engineering constructions. Although different parameter settings are often optimal for capturing different aspects of lexical meaning, one and the same distributional model can be used, with moderate success, for a plurality of purposes or semantic tasks. This thought is elaborated, e.g., in Baroni and Lenci (2010).

2.3. Composition in distributional semantics

An obvious drawback of the distributional approach to semantics as presented so far is the limitation to lexical meaning, or, in the best case, to the meaning of short and common phrases (such as *fall apart* or *kick the bucket*). Larger phrases and whole sentences will generally not occur in an arbitrarily

[4] Baroni et al. (2014a, 1): "Occasionally, some kind of indirect supervision is used: Several parameter settings are tried, and the best setting is chosen based on performance on a semantic task that has been selected for tuning."

large corpus with a frequency that could make the distributional information any informative in the semantic respect. (On the level of phrases and sentences, the number both of possible target vectors and of possible context dimensions grows tremendously, as presumably does the number of semantic distinctions that must be made. But there are not more tokens of phrases or sentences in a corpus than there are tokens of words, so the distributional information in the table of co-occurrence counts will be extremely sparse.) And indeed, semantic composition has recently been a hot topic in distributional semantics.

The question is: Can you combine the vector representations of particular words in a phrase (such as *black dog*) so as to obtain a useful semantic representation of that phrase, without having to rely on the distributional properties of the phrase as a whole? The most rudimentary attempts in this respect involve some very basic mathematical operations with the vectors, the resulting "phrasal" vector being obtained by simple addition or multiplication of the basic vectors. Some sort of linear weighting is possible, e.g., in order to stress the semantic role of nouns as compared to adjectives (Mitchell and Lapata 2010). These all are clearly very *ad hoc* solutions, with hardly any motivation other than mathematical simplicity.

A more ambitious program in compositional distributional semantics is formulated by Baroni et al. (2014b). Here, the idea of meaning composition as functional application, a fundamental notion from formal (model-theoretic) semantics, is adopted. Some words, nouns in particular, are represented in the familiar fashion, with their basic distributional vectors. Other words, such as adjectives, are semantically conceived as functions turning vectors into vectors; thus e.g. the vector for *black dog* can be obtained by the application of the functional meaning of *black* to the basic vector of *dog*. Yet other words are conceived as binary functions, etc., roughly in correspondence with the matching between grammatical categories and semantic types in Montague grammar (see e.g. Gamut 1991).

Despite the inspiration, this approach to semantic composition also differs from the formal semantic treatment in some important respects. First, unlike in formal semantics, the lexical functions are given concretely and informatively, not only defined as to their type and otherwise left unspecified (or specified just informally using disquotation, such as, "black" refers to the function that assigns truth value 1 to all black objects and only them). Namely, they are estimated based on the short phrases that still occur in the corpus often enough for their distributional representation to be semantically informative.

Basically, the functional representation of *black* is automatically estimated based on how the distributional vector of *black dog* differs from that of *dog*, that of *black book* from that of *book*, etc.[5] Once it is learned in this way, it can be used to derive the representations of longer phrases for which representation by the basic distributional vector cannot be assumed.

Second, the correspondence to the Montagovian matching between grammatical categories and semantic types is only partial, as attested by the treatment of common nouns such as *dog* (cf. Baroni et al. 2014b, 59). In formal semantics, common nouns, just like intransitive verbs or adjectives, are standardly conceived as logical predicates; that is, words with a functional meaning. The reason why Baroni and colleagues do not preserve this choice, in which the semantic types of nouns, adjectives and intransitive verbs are unified,[6] is clearly pragmatic. Representing common nouns with basic distributional vectors works remarkably well, and it would be unwise to force the distributionalist program into the scheme of formal semantics, a discipline whose outcomes are not nearly as efficient in practical terms.

But then, why should we bother incorporating *any* of the formal semantic tenets into the distributionalist program? It makes sense if we believe that formal semantics provides a good theory of semantic composition nevertheless. In any case, this is in accordance with how formal semanticists themselves tend to present the discipline (facing the lack of practical applications), and Baroni et al. (2014b) seem to share that belief. I do not, and I think there are serious reasons to believe the contrary. In Ocelák (manuscript), I attempt to elaborate these. Just briefly, my argument regarding formal semantics is that the lack of interest in lexical meaning, combined with the lack of empirical evaluation of the proposed semantic formulas, leads to the construction of chimerical compositional structures whose "adequacy" is a purely formal matter.

[5] That is, the semantic representation of short phrases like *black dog* can be, in principle, either obtained by composing the representations of their parts, or specified directly as their basic distributional vectors. Given the method of estimating the functional representations, the outcome will typically be different in these two cases. The choice between the two options is upon the theorist. There is however also an argument for keeping both, pointing out the difference between the compositional and the idiomatic reading of, e.g., *kick the bucket* (Baroni et al. 2014b, 7).

[6] That, in any case, is an option much more intuitive to logicians than to linguists.

For instance, the quantifier *all men* is in the most basic (extensional) case translated as $\lambda X \forall x (Man(x) \rightarrow X(x))$, which is supposed to be interpreted with a function that assigns truth values to functions from individuals to truth values (that is, to logical predicates). This function, however, is never given in full. It is only informally specified as *that function which* assigns the *appropriate* values to all relevant predicates (such as, 1 to *mortal* and 0 to *dark-haired*: for all men are mortal but not all of them are dark-haired). But that actually amounts to little more than saying that the meaning of a part is *whatever gives the right meaning* for the whole when applied to what we regard as the meaning of another part. It is then hard to see where such a quasi-analysis could possibly go wrong. At the same time, this can be found in the core of most formal semantic analyses. I therefore suspect that the existing body of work in compositional, lambda-phrased formal semantics can largely be seen as aprioristic elaboration of the Fregean idea of functional application. Whether the resulting theory of semantic composition is any good in empirical terms is highly questionable.

It moreover seems to me as a sort of wishful thinking to suggest that distributional and formal (or "denotational") semantics cover "complementary aspects of meaning" (Baroni 2014, 24; cf. also Erk 2016). The authors support this suggestion with the observation (in itself right) of the different focus in both approaches: generic knowledge in the former, episodic knowledge in the latter (Baroni 2014b, 22ff.). But at the same time, these approaches have been often pronounced complementary in dealing with the *lexical* and the *compositional* (or structural) aspects of meaning, respectively. How are these two divisions of labor supposed to square with one another? Surely, the distinction of the lexical and the compositional does not run parallel to that of the generic and the episodic. Lexical semantic competence, for instance, has both generic and episodic aspects to it. Thus the position that distributional semantics aims at the lexical *and* the generic, whereas formal semantics aims at the structural *and* the episodic, *and yet* they fully complement each other in the examination of language meaning seems problematic, even incoherent. For me, that as well constitutes a reason for being suspicious about the proposed boosting of distributional semantics with Fregean compositionality.

Altogether, I suggest we drop the assumption that formal semantics is a successful program in a domain that is complementary to the core domain of distributional semantics. And clearly, that would reduce the alleged need of encompassing both approaches in one framework.

As to distributional semantics alone, I have so far presented the framework in a more or less uncontroversial way, basically describing what people have done in the field. At this point, the very idea of enriching distributionalism with semantic composition invites a more philosophical discussion of the approach: an inspection of what it is that has been done, and what hopes we can (or cannot) derive from that.

3. What is distributional semantics, really?

Despite the general orientation on the performance in semantic tasks, the literature also contains explicit concerns about the philosophical interpretation of the distributionalist framework. In particular, people have made a distinction between a weak and a strong reading of the distributional hypothesis (see Lenci 2008, 14ff.; cf. also Baroni et al. 2014b, 20ff.).

Roughly speaking, distribution in the weak reading *reflects* the meaning of words (and perhaps also of some larger expressions), but does not *constitute* it. Words are generally used in accordance with what they mean (thus *dog* often appears in the context of *bark*, *bone*, *leash*, much less in the context of *fuel* or *oligarchy*). That makes distribution (which can be captured mechanically and efficiently) a useful guide in the exploration of meaning (which cannot), without however making it into a court of appeal as regards semantics. This conception leaves room for divergences of meaning and distribution, since it assumes that distribution is shaped also by factors other than meaning.

In the strong reading, distributionalism amounts to a cognitive hypothesis about the character of our semantic knowledge, or some parts of it. Here, vector space representations acquire the more binding character of cognitive or mental representations, rather than mere theoretical instruments. Sure, there is little reason to believe that the vectors we actually draw from a particular corpus, with a particular choice of target expressions, context dimensions, weighting techniques etc., capture the knowledge of any particular speaker very precisely. Thus distribution, at least as observable practically and in a large scale, can still somewhat diverge from meaning. But something like computing vectors based on the input and using them is (a part of) what is going on in our minds/brains when we acquire and use semantic knowledge—or so the thesis goes.

Baroni et al. (2014b), in their attempt to inject distributional semantics with compositionality, go for the strong reading of the distributional hypothesis. In opposition to them, I would like to defend the weak version of distributionalism here. By philosophical means, it is hard to disprove a cognitive hypothesis directly, stating facts by which it is contradicted. But I believe distributionalism can be presented in a way which will simply make the strong hypothesis not appear worth too much consideration.[7]

To me, it seems rather obvious that distribution is merely a reflection of semantics, and a substantially imperfect one. Apart from meaning, there are other important factors bearing on how words are put to use in a text; that is to say, factors that are also reflected in distribution. What the world is like is one of such factors. What we prefer to communicate about is another. (All these factors are interrelated and there are borderline phenomena: indeed much of the 20th century philosophy of language can be viewed as a struggle with the idea that they can be neatly separated and subsequently interlinked in a controlled fashion. But there are all sorts of clear cases which justify making the distinction nonetheless.)

Years ago, there was a fierce war in Bosnia, which made *Bosnia* co-occur with *war*, *tank* and *suffering* particularly often. Later, the situation stabilized, but people kept talking and writing about the past war. Yet neither of these periods added to the meaning of *Bosnia* a substantial something that we do not find in the meaning of *Switzerland*; neither made *Bosnia* markedly more related in meaning, e.g., to *war* than *Switzerland* is. I do not deny that many semantic changes do indeed proceed this way. But it is crucial to note that a semantic change is incomparably *slower* than the change in distribution to which it is linked. First, a massive change in distribution seems to be followed by hardly anything in the semantic respect. Slowly, something we call *connotation* may arise. It is only much later that a full-fledged semantic change can sometimes be recognized. Over past two centuries, *Waterloo* may have evolved into a synonym of *utter loss*, but very little of that change seems to have taken place in the first days or years after the co-occurrences of *Waterloo* in speech or writing rapidly changed in 1815. I believe this issue is overlooked

[7] That, incidentally, is a philosophical method of later Ludwig Wittgenstein, whom Lenci (2008) or Baroni et al. (2014b) mention among the historical sources of the distributionalist thinking.

when meanings are equated with distributional patterns, as seems to be more or less the case with the strong version of distributionalism.

Now, one can object that this, rather than being an objection to the strong reading of the distributional hypothesis, simply expresses a conservative view of meaning to which strong distributionalism provides a fresh alternative. Let me leave it at that for the moment: I hope to justify this conservatism later when a more positive program is finally outlined.

Provided that distribution is shaped also by factors other than meaning, its utility in the exploration of semantics may still be considerable, but is limited on principle. Consider an analogy: The ripples on Loch Ness may give us a clue about where underwater Nessie is at the moment. Yet the evidence is imperfect, since rippling is, besides the timid monster, also caused by the wind, by other creatures in the lake, etc. It would certainly be naive to insist that our methods of counting and measuring the ripples, and they alone, should make Nessie perfectly traceable, let alone to insist that the pattern of rippling is in some sense *identical* with her. To be sure, Nessie *can* be traced perfectly based on that pattern, but for that we would need to know the other factors and subtract their effects. By contrast, distributional semantics does *not* attempt to study the impact of factors other than semantics on distribution, and therefore is not in a position to subtract that impact.

4. Performance, nature and composition of distributional representations (again)

In section 2.2., I emphasized what distributional models are capable of doing in practical terms, in order to contrast them with other, more theoretical approaches to linguistic semantics. At this point, it seems convenient to mention what they have as yet *failed* to achieve.

Lenci (2008, 19ff.) identifies three main issuess with distributional semantics: semantic composition, reference or grounding, and inference. Of these, the first is discussed separately in this paper, and the second can perhaps be laid aside as a matter of technical limitations (see the discussion of extralinguistic contexts in section 2.1.). But the third problem, accounting for inferences, deserves some attention.

Inference, or entailment, plays a central role in a number of semantic approaches, including formal semantics and the inferentialist view of meaning which is to be outlined in the next section. Correct inference, in the simplest case, is a transition between two sentences or utterances that is in a specific (namely, the *semantic*) sense appropriate.

It might seem that lexical semantics, the primary domain of distributionalism, does not concern sentential meaning at all, and therefore that we cannot expect this branch of semantics to provide an account of inference. That is however not quite true: the lexical semantic relations which are traditionally a crucial interest of lexical semantics are characteristic by licensing particular classes of inferences. Knowing that A is a *synonym* of B, we know that (by way of example and under certain additional conditions) we can infer "this is a B" from "this is an A" and the other way round. The information that A is a *hyponym* of B allows us to draw the inference from "this is an A" to "this is a B", but not the other way round. If A is an *antonym*, *meronym*, *co-hyponym* of B, that again seems to license at least some specific inferences in each case. Note that the same does not hold for the broad semantic similarity, which is supposed to be the relation primarily captured by distributional models. The information that A is semantically *similar* to B is not sufficient to license particular inferences from sentences containing A to sentences containing B.

Assuming there is a connection (to say the least) between understanding a sentence and knowing the appropriate inferences in which it is involved, it seems not unreasonable to expect of lexical semantics that it will do its part in accounting for inferences—that is, it will reliably detect lexical semantic relations. But for distributional semantics, with its basic notion of underspecified semantic similarity, this is a chronic problem.

It was mentioned above that the best of the current distributional models perform admirably on the standard TOEFL synonym detection task, easily reaching the performance of native human speakers. That is, however, a very specific task: it requires detecting exactly one synonym for a given term among three non-synonyms which also stand in no other particular semantic relation to the target. It is remarkable that this can be done very successfully on a distributional basis, but it is clearly not enough. In order to account for inferences, you need to be able to tell for arbitrary two terms whether or not they stand in the relation of synonymy, in the relation of hyper-/hyponymy, etc. A model's good performance in the TOEFL task does not guarantee this for synonymy. The vector representations of synonyms can be generally more similar to one another than those of semantically unrelated words, without the former being on the whole more similar than the vectors of antonyms, co-hyponyms etc.

And indeed, experimental results suggest that distributional models are too weak to tell apart cases of particular lexical relations reliably (Lin et al. 2003; Baroni et al. 2011.) Generally, the vectors most similar to the vector representation of a given word tend represent synonyms, co-hyponyms, and antonyms of the target word, without clear order. At the same time, not all synonyms, co-hyponyms etc. reach higher similarity than all words semantically less related to the target. That of course further complicates the classification task.

Admittedly, it is possible to construct the model or redefine the similarity measure so as to favor instances of a particular lexical relation; e.g., to enhance the "similarity" of co-hyponyms and suppress that of synonyms, antonyms etc. That seems to be the case at least for synonymy, co-hyponymy, and hyper-/hyponymy (cf. Baroni et al. 2011; Erk 2016). But the sorting success achieved is moderate in each case. For instance, one can find a specific similarity measure which, unlike the standard cosine measure, is likely to assign higher "similarity" on average to the instances of hyper-/hyponymy than to the instances of co-hyponymy (cf. Erk 2016, 21-22). That however does not imply that the measure is capable of sorting out hyper-/hyponymical pairs very efficiently. To give a parallel, men are no doubt taller than women on average; yet the utility of height alone in telling apart men from women is limited. The clue is better than random, but far from perfect. In accounting for inference, arguably, better than random is not good enough. You won't entrust a robot with making pancakes if its knowledge of appropriate inferences between sentences containing *egg*, *milk*, *food*, *poison*, *hot*, *cold* etc. is merely better than random.

As a side note, this approach also makes distributionalism as a cognitive hypothesis more problematic than it already seems to be. Namely, it is one thing to assume that what we do in our minds/brains when acquiring and using meanings is something like constructing and comparing distributional vectors. It is another thing, arguably a more involved one, to defend that we should actually need a whole bunch of vector spaces and/or similarity measures in order to cope with *various* lexical relations.

Above, the efficiency of distributional models in detecting lexical semantic relations is deliberately discussed in rather vague terms, despite there being many experimental results phrased in concrete numbers. I do not go into the evaluation numbers here, for that would make little sense in the absence of a detailed discussion of the respective semantic tasks, and of their relevance with respect to the problem in question. I nonetheless take it for given that the current distributional semantic models, in spite of their achievements that are highly non-trivial from the point of view of theoretical semantics, are still far from giving a satisfactory account of lexical semantic relations (as an important part of natural language inference).

To this, we may react with the standard *more research is necessary* statement and keep trying to wring out what we can from distributional models. And no doubt, some improvement *can* be reached, in particular by exploiting ever bigger corpora and ever higher dimensionality, made possible by more efficient implementation and by using ever more computing power.[8] But my impression is that these improvements in performance are not promising enough to validate the position that in the limit, distribution *is* semantics.

Instead, I suggest that we bite the bullet of admitting that it is not. In my opinion, the problems with accounting for lexical relations are inherent to the approach as such. I believe that at the moment, the performance of distributional models is somewhere near the ceiling, and that is simply because distribution is a useful, yet imperfect reflection of semantics.

The hunt of Baroni and colleagues for composition in distributional semantics seems somewhat questionable from this perspective. In this view, composing distributional representations of particular words (even the advanced, functional representations) necessarily amounts to adding up the considerable imprecision that arises already on the lexical level. Very likely, there will still be some tasks on which the compositional representations (in particular those of relatively short phrases) will achieve a non-trivial performance. But if the claim that non-negligible amounts of error are being added up in composition is correct, then it is unclear whether such achievements can be of theoretical or practical consequence.

Let us go back to the Loch Ness parallel. If using a word's distributional pattern to explore its meaning is like tracing Nessie based on the momentary pattern of rippling, then the struggle for compositional distributional representations seems to be like trying to write up her biography based on a series of snapshots of the lake's surface. The former is limited in precision; in the latter, shortcomings are being piled up.

[8] Cf. Mikolov et al. (2013), who report on models which it took days on hundreds of processing cores to build up.

5. Distributionalism and inferentialism

I am aware that the previous critical considerations, pertinent as they may be, can hardly have much impact in lack of a positive alternative, one that would be viable from the point of view of computational linguistics. Also, one might want to bypass their theoretical relevance by insisting that the strong, cognitive distributional hypothesis gives rise to a radically new conception of meaning, whereby my assumptions regarding distribution, meaning etc. are simply not shared. But I think there *is* an alternative way to go, other than in the direction of contemporary compositional semantics. The alternative inspiration source is well-founded theoretically and I believe it can be stated precisely enough so as to invite computational implementation. Being use-theoretic in character, it seems to better fit the distributional reliance on language corpora, as documents of actual language *use*. There are moreover reasons to think that the implementation need not be quite disconnected from the current practice of distributional semantics.

I see such an alternative in the inferentialist philosophy of meaning, elaborated in particular by Brandom (1998); for a more accessible introduction, see Part I of Peregrin's (2014). Inferentialism draws on the idea that the meaning of a sentence is basically a matter of the appropriate inferences in which the sentence is involved. The meaning of a word, or generally of a subsentential expression, is then seen as its contribution to the inferential properties of the sentences in which it is contained. Here, the notion of inference is very broad, covering *language-language* transitions (that is, from sentences or sets thereof to sentences), as well as *world-language* and *language-world* transitions (that is, from worldly *circumstances* to sentences; and from sentences or sets thereof to worldly *actions*).

The inferentialist view is a specific elaboration of the Wittgensteinian idea that the meaning of a word consists in how the word is put to use, plus the aged observation that the primary use of a word is in the context of a sentence. It is specific, first, in that it emphasizes the normative character of our language use (the meaning of a sentence is identified not with its *actual* use, but with its *appropriate* use), and second, in that it narrows down the general notion of *use* to the transitions to which our sentences are subject. So, the meaning of a sentence (and closely related, the content of a belief) is given by what we *should* infer it from and by what we *should* infer from it in the context of other sentences (beliefs) which we are committed to assert (hold). Brandom's crucial idea is that normative *statuses* of agents (i.e., what agents should do) can be reduced to factual normative *attitudes* (i.e., how the agents treat one another, as well as themselves, in relation to what they do). In this way, semantics is underlain by pragmatics. What people believe, or what their sentences mean, is explained—in a rather sophisticated way—in terms of what people do non-linguistically.

Argumentation for why Brandomian inferentialism is a fruitful and highly adequate philosophical approach to the semantics of natural language is far beyond the scope of this paper. Here, let me simply assume it is. On this assumption, I would like to make some comments towards bridging the gap between inferentialism as a philosophical project and distributionalism as a program in computational semantics, as I believe that enhancing a practical application with adequate philosophy is something desirable in principle.

The practical problem of inferentialism (which distributional semantics might be in a position to solve) is the following. Brandom's inferentialism is a holistic philosophy of meaning. What he draws is a picture of an overwhelmingly complex network in which any node standing for a sentence or a belief is deeply integrated. Any ordinary sentence is involved in myriads of appropriate inferences.[9] Little wonder that inferentialism as concerns natural language has not made it far beyond a mere philosophical idea until now: no *content* expression has ever been explicitly analyzed in inferential terms. Virtually the only inferentialist semantic analyses of natural expressions that seem plausible to some extent are the natural-deduction-style characterizations of sentential connectives such as *and*, *or*. ("A *and* B" can be appropriately inferred if A as well as B are given; from "A *and* B" we can appropriately infer A as well as B. That is all one needs to characterize the meaning of *and*, at least as traditionally employed in logic. The analysis is tempting in that we in this way completely avoid the need to postulate an object, typically a truth function, as the *meaning* to be mysteriously connected to the expression in question.) But

[9] Take, e.g., the belief/sentence stating that the cat is in the garden. It can be appropriately drawn from seeing the cat in the garden; or from hearing familiar noise from the garden; or from the belief that the cat was in the garden five minutes before plus the belief that it is an extremely lazy creature, etc. And given various "collateral commitments", it may be appropriate to infer that the cat is safe from the street traffic, or that there will soon be no mice in the garden, or that the cat will make a mess when it's back in the house; or it can lead to a *lemme-drive-the-cat-out-of-the-garden* practical commitment, etc.

application to this restricted vocabulary can hardly provide sufficient validation for such a general philosophical theory.

Think as we may that inferentialism is the right way of thinking about meaning, it cannot be considered an option by computational semanticists unless it is presented as viable by their methods. Preferably, it should be made feasible using the valuable resources that are available and that make computational linguistics successful as it practically is: large corpora of actual language use in the first place. In my opinion, inferentialists should side with the idea of computational implementation of their program. At least, the philosophical ambition of inferentialism is to reduce the mysterious notion of meaning to something more transparent, something that we *do*: something that computers, therefore, might be also capable of doing one day.

Here is why I think inferentialism is fundamentally compatible with the distributional perspective. Recall that distributional semantics attempts to capture the meaning of an expression in terms of its occurrence *in various contexts*. Usually, these are lexical contexts, so what is typically counted are lexical co-occurrences. But the distributional project does not set any *a priori* bounds to what we can regard as contexts. Various options have been considered: among others, lexico-syntactic contexts, web-based documents, extralinguistic contexts (such as labeled images)—and crucially, we may think of *inferential contexts* as well. We may want to count a sentence's occurrences in the context of sentences inferred from it, and in the context of sentences from which it is inferred.

There is a number of problems with this proposal. The first is that what primarily features in an inference are *sentences.* As mentioned in section 2.3, the actual co-occurrence information in the co-occurrence space of sentences (unlike the space of words) is extremely sparse for corpora of all available sizes—sparse beyond imagination. We could count co-occurrences of *words* in inferential contexts, but it seems to be of little use to know, e.g., that *freezing* and *green* co-occured within the inference "It is freezing outside. – I'd better take the green cap, the wooly one, since the red is really thin." (The co-occurrence of *freezing* and *wooly*, or *outside* and *cap* is perhaps more informative, but it occurs to me that counting word co-ocurrences would open the door for the same kind of imprecision that has been observed in the standard distributional representations.) Somehow, we need to treat a sentence as a whole, nevertheless. I do not have a solution for this, I only hope one can be given. Perhaps, a clever engineering solution can exploit the idea that the meaning of a word is the way it contributes to the inferential properties of the sentences in which it is involved, and perhaps, the process of inferential characterizing can be bootstrapped from minimal inferences such as "this is a banana: this is yellow". Syntactic information will be surely indispensable in such a scheme.

Second, the issue with *world-language* and *language-world* inferences. Given the technical difficulties limiting the utilization of extralinguistic contexts, I suggest that we follow current distributional semantics in focusing on linguistic contexts, at least for the time being. That is, we may focus on *language-language* transitions. (Existing distributional models have shown clearly enough that non-trivial practical success can be hoped for even in the absence of extralinguistic information.) One more thing that needs to be technically overcome is that often, *language-language* inferences are inferences not from individual sentences, but from *sets* of sentences.

Third, the problem of normativity. Bradomian inferentialism explains the meaning of a sentence in terms of appropriate inferences, not in terms of actual inferences. Contrariwise, what we can (at best) gather from a corpus of actual language use are the inferences people make, possibly the inferences they make *regularly*, but not the inferences they *should* make. Here again, I suggest we take a pragmatic stance. The practical success of distributional models (that is, on tasks that are unequivocally semantic in nature) indicates that the cleft between actual and appropriate use is narrow enough for at least some practical purposes. One may here also consider Davidsonian arguments to the effect that it is incoherent to assume a massive amount of factual or semantic error among speakers (cf. Davidson 1974).

The fourth problem is likely the most serious one in practical terms. Actual inferences occurring in a corpus are not very reliably marked with formal means such as *therefore*, *thus*, *so*, etc. Yet worse, rudimentary inferences such as "this is a banana: this is yellow" scarcely make it to the communication of competent speakers. Usually, such inferences are assumed rather than pronounced. What gets explicitly communicated instead are complex inferences relying on a number of collateral commitments or shared assumptions: "People still remember the Denver incident. *Therefore*, Smith won't get more than 15 percent of the votes."

An option that comes to mind in this context is utilizing language *acquisition* corpora, rather than corpora of grown-up communication. Unfortunately, the corpora of the former type are several orders of magnitude smaller in size, which may be hard to bite for a distributionalist, and the data is very expensive to gather. Its quality could nonetheless make up for that. It is first and foremost with children that we explicitly state what is otherwise obvious, talking about the color of bananas, etc.

This approach also seems to constitute an additional answer to the normativity issue. In talking to children, we are generally engaged not only in communication, but also in tuition and training that are relevant for the child's future communication. Thus stating "This is a banana. (*So*) it is yellow" in this situation is not merely an actual inference. Much more it amounts to the formulation of an inferential rule, to stating what inferences should be drawn. [10] Even so, there remains the problem that not all inferences are formally marked, and an amount of manual annotation may be necessary.

6. Conclusion

No doubt, the difficulties involved are considerable, and the "inferentialized distributionalism" just proposed may not reach the practical performance of the current distributional models any time soon. Still, I believe something in these lines is worth elaborating. Distributionalism in computational semantics has had highly non-trivial achievements, but in the end that all comes down to clever exploitation of the fact that meaning is reflected in distribution. If that is not *all* there is to meaning, the prospects of exploiting the idea further are of course limited.

Ultimately, the goals of computational and philosophical semantics cannot be as divergent as they possibly appear to be at the moment. Computational semantics is supposed to come up with something that can do what natural language meaning does, or what humans do using their semantic knowledge. Fair enough, but why would we think this can be achieved without paying attention to our best opinions about what natural language meaning *is*?

What there is in the project for inferentialism as a philosophical program seems also quite clear. Boosting computational semantics with inferentialist insights would constitute important empirical validation for the philosophical

[10] Note that there would be no point in stating such rules incorrectly. Joking or lying about bananas makes sense only after the child has mastered some basic inferential properties of banana-related sentences.

theory. A theory of an empirical phenomenon, as human language altogether is, has surely no right to spurn such a prospect.

Acknowledgments

I am grateful to Jarda Peregrin and Tomáš Musil as well as to several anonymous reviewers, for their relevant comments on previous versions. All remaining flaws and omissions are my own responsibility. Work on this paper was supported by the Research Grant No. 13-21076S of the Czech Science Foundation.

References

BARONI, M. and LENCI, A. (2010): Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics* 36, No. 4, 673-721.

BARONI, M. and LENCI, A. (2011): How We BLESSed Distributional Semantic Evaluation. *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, 1-10.

BARONI, M., DINU, G. and KRUSZEWSKI, G. (2014a): Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 238-247.

BARONI, M., BERNARDI, R. and ZAMPARELLI, R. (2014b): Frege in Space: A Program of Compositional Distributional Semantics. *Linguistic Issues in Language Technology* 9, 5-110.

BRANDOM, R. (1998): *Making It Explicit: Reasoning, Representing and Discursive Commitment*. Cambridge, (Mass.): Harvard University Press.

DAVIDSON, D. (1974): Belief and the Basis of Meaning. *Synthese* 27, 309-323.

ERK, K. (2012): Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass* 6, No. 10, 635-653.

ERK, K. (2016): What Do You Know About an Alligator When You Know the Company It Keeps? *Semantics and Pragmatics* 9.

FIRTH, J. R. (1957): *Papers in Linguistics*. London: Oxford University Press.

GAMUT, L. T. F. (1991): *Logic, Language, and Meaning. Vol. 2: Intensional Logic and Logical Grammar*. University of Chicago Press.

LANDAUER, T. K. and DUMAIS, S. T. (1997): A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review* 104, No. 2, 211-240.

LENCI, A. (2008): Distributional Semantics in Linguistic and Cognitive Research. *Italian Journal of Linguistics* 20, No. 1, 1-31.

LIN, D., ZHAO, S., QIN, L. and ZHOU, M. (2003): Identifying Synonyms Among Distributionally Similar Words. *Proceedings of the 18th International Joint Conference On Artificial Intelligence*, 1492-1493.

MADDIRALA, N. (2014): *Philosophy of Logical Practice: A Case Study in Formal Semantics*. Master thesis, ILLC, University of Amsterdam.

MIKOLOV, T., CHEN, K., CORRADO, G. and DEAN, J. (2013): Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.

MITCHELL, J. and LAPATA, M. (2010): Composition in Distributional Models of Semantics. *Cognitive Science* 34, 1388-1429.

OCELÁK, R. (manuscript): Besieging Model-Theoretic Semantics. Available at: http://ocelak.cz.

PEREGRIN, J. (2014): *Inferentialism: Why Rules Matter*. Basingstoke, UK: Palgrave Macmillan.

SCHNEIDER, H. J. (1992): *Phantasie und Kalkül: Über die Polarität von Handlung und Struktur in der Sprache*. Berlin: Suhrkamp.

TURNEY, P. D. and PANTEL, P. (2010): From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37, 141-188.