Content vs. function words: The view from distributional semantics[1]

Márta ABRUSÁN — *IJN, CNRS, ENS, EHESS, PSL, Paris, France*
Nicholas ASHER — *IRIT, CNRS, Toulouse, France*
Tim VAN DE CRUYS — *IRIT, CNRS, Toulouse, France*

Abstract. Counter to the often assumed division of labour between content and function words, we argue that both types of words have lexical content in addition to their logical content. We propose that the difference between the two types of words is a difference in degree. We conducted a preliminary study of quantificational determiners with methods from Distributional Semantics, a computational approach to natural language semantics. Our findings have implications both for distributional and formal semantics. For distributional semantics, they indicate a possible avenue that can be used to tap into the meaning of function words. For formal semantics, they bring into light the context-sensitive, lexical aspects of function words that can be recovered from the data even when these aspects are not overtly marked. Such pervasive context-sensitivity has profound implications for how we think about meaning in natural language.

Keywords: function words, lexical semantics, determiners, distributional semantics.

1. Introduction

Is there a categorical difference between the semantics of content (or open-class/lexical) words and function (or closed-class/logical) words in natural languages? Common wisdom in linguistic research holds that the answer is 'yes'. According to this view, functional items encode the grammatical knowledge of language speakers, while content words are a reflex of world knowledge. In some incarnations of this view, the functional vocabulary is given by the language faculty, and is thus universal and biologically determined (see for example May 1991; Partee 1992; Chierchia 2013). It provides a syntactic skeleton into which lexical content is inserted, a mental coat rack onto which colourful content about the world can be hung.

Despite intuitions about the existence of the two classes, finding a precise semantic difference has proven to be difficult. The most frequently cited idea, borrowed from a tradition in logic aimed at defining logical constants, is that function words have meanings that are invariant across certain mathematical transformations of their domains.[2] Examples of transformations that have been proposed to diagnose logical constants include invariance under permutations (Tarski and Givant 1987; van Benthem 1989; Sher 1991), invariance under surjective functions (Feferman 1999), invariance under potential isomorphisms (Bonnay 2008), etc. What all these have in common is the underlying idea that logical meanings are topic-independent: the validity

[1]We are grateful to the organisers of the Special Session on Semantics and Natural Logic for the invitation, the audience for helpful questions and an anonymous reviewer for copy-editing suggestions. The research reported here was supported by a Marie Curie FP7 Career Integration Grant, Grant Agreement Number PCIG13-GA-2013-618550, a grant overseen by the French National Research Agency ANR (ANR-14-CE24-0014), and by ERC grant number 269427.

[2]Another idea that was advanced is that function words involve higher types than lexical items (cf. Partee 1992). See also MacFarlane (2017) for a review of the philosophical literature on logical constants.

 In: Uli Sauerland and Stephanie Solt (eds.), *Proceedings of Sinn und Bedeutung 22*, vol. 1, ZASPiL 60, pp. 1–21. ZAS, Berlin.

of a logical inference should not be dependent on the particular properties of what one is talking about. The appropriateness of the above ideas for diagnosing logical constants is a subject of lively debate, but they are clearly unsuitable for diagnosing function (logical) words of natural language (see Gajewski 2002; van Benthem 2002). This is because they predict certain lexical items to be logical (e.g. the predicates *self-identical, exist*), and they also predict that certain intuitively logical elements of natural language, e.g. the quantifier *every* or *each*, are not logical since they have a lexical restriction that they need to quantify over countable objects, hence **Every/*Each milk is in the fridge*.

The intuitive distinction between the two classes of words seems at first to be supported by research in Distributional Semantics (DS). This computational approach to natural language semantics is based on the "distributional hypothesis" by Harris (1954), according to which one can infer a meaning of a word by looking at its context. Meanings of words differ in DS, because they will co-occur with different contexts with different probabilities. While the approach has been very successful in capturing the meanings of lexical words and lexical aspects of meaning in general (synonymy, hyponymy, etc.), there is very little evidence that DS can capture semantic properties of function words (though see Baroni et al. 2012; Bernardi et al. 2013; Hermann et al. 2013; Linzen et al. 2016). It is easy to see why: if logical meanings are topic-independent, their logical meanings will not be reflected by their distributions, and all logical words will have the same DS meaning.

However, the actual picture that emerges from DS is not a clear-cut division between the two classes of items. What we show in this paper is that when we approach function words (in particular, determiners) with DS methods, what comes to light is that logical items in natural language have a layer of non-logical meaning in addition to their logical meaning. Function words do not have purely logical content, but are a mixture of logical content and more "worldly" content comprised of lexical and distributional aspects of meaning. This is also one of the reasons why logical methods such as permutation invariance fail to diagnose functional items of natural language correctly. While DS is indeed blind to purely logical meaning, it brings to light the lexical and distributional aspects of functional items in natural language.

Our results suggest the following picture. There are context-invariant, logical aspects of meaning, and lexical/distributional aspects of meaning that tend to be modulated by the context. But the two types of meaning do not map neatly to two different types of words. More often than not, the total conceptual meaning of words is composed of both types of meaning, but to varying degrees. For example, an adjective such as *heavy* has, beside its lexical content relating to heaviness, a logical aspect of being a predicate over degrees. Aspects of the lexical meaning of heaviness can be modulated by context (*heavy elephant* vs. *heavy bleeding*), but not the logical meaning of being a degree predicate. A determiner such as *some* has, besides its logical meaning of being an existential quantifier, context-sensitive lexical aspects, for example an inference of uncertainty about identity on the part of the speaker. This type of lexical content of quantifiers has a high degree of contextual variability, similarly to other types of lexical content (e.g. *Some guy called you* vs. *There is some milk in the fridge*).

Our results connect to a growing body of evidence that challenges the traditional division between lexical and functional words. Firstly, one of the reasons why permutation invariance fails to correctly capture logical words in natural language is because of the sensitivity of these items to the properties of the linguistic and extralinguistic context. Sensitivity of certain quantifiers to the mass/count distinction, indefinites introducing discourse referents, the focus sensitivity of particles and negation, etc. are all examples of such lexical (or pragmatic) dependencies of the context. Such context-sensitivity is the essence of non-logical content.

A second, theoretical argument might come from language variation. The quantifier systems of even very closely related languages can be quite different. However, the variation, at least in the case of well-studied European languages, is not so much in the logical content expressed but more in the non-logical content associated with quantifiers. For example, indefinites in English (*a, some, any*) and German (*ein, irgendein, etwas,* etc.) differ not in their logical meanings but in the non-logical, lexical and distributional aspects associated with them.

A third reason for challenging the idea of a clean cut between the two types of words might come from historical linguistics: recent advances in this field seem to call into question the traditional idea according to which functional items are more stable than lexical items.[3] For example, Greenhill et al. (2017) argue that grammatical features tend to change faster than basic lexical vocabulary. Similarly, a substantial part of the functional vocabulary belongs to the fast-changing items in the lexicon. This shows that functional items and grammatical features are not, generally speaking, the stable pillars in the dynamics of language change that they were often assumed to be. Various subsystems of language show differing patterns of dynamics, but the classification into these subsystems does not follow the lexical/functional division.

What our results add to the above theoretical arguments is that they bring to light the lexical and distributional aspects of the meaning of quantifiers, even when these are not overtly marked by the morphology. Our methods, based on distributional semantics, can associate latent semantic dimensions to these quantifiers. Some of these correspond to well-known aspects of quantifiers with a special distribution (e.g. *any*), and some correspond to semantic distinctions that are unmarked in English but marked in other languages, as in the case of *some*.

Our view of lexical semantics is a mixed model that incorporates elements from both traditional approaches to lexical semantics and distributional semantics. Traditionally, the lexical semantics of a word is the meaning that is associated with it in the lexicon. This meaning is assumed to depend on the circumstances of the evaluation (or contexts in the sense of Kaplan 1989) in the case of many lexical items, for example indexicals, demonstratives and possibly a large number of other items such as adjectives, attitude verbs, etc. Context-sensitivity, in these systems, means that the lexical meaning contains a variable whose value needs to be fixed by some context. For example, the cutoff-point for a degree adjective such as *heavy* might be supplied by the context and will be different for elephants and for mice. The lexical meaning offered by distributional semantics is context-sensitive in a much more radical way: the conceptual structures that we associate with words are gleaned from the contexts of use (dis-

[3]We advance this argument tentatively, since none of the authors is an expert in historical linguistics. Greenhill et al.'s (2017) article seems highly pertinent though, which is why we mention it here.

tributions) and might change with use over time and across different corpora. In the mixed model we assume here (see Asher et al. 2016) words have logical content (which we cannot derive from distributions, for the reasons described above) which plays a role not only in establishing their denotation but also in the composition of meaning as well; the logical content of an adjective, for instance, is that it must modify in some way a noun meaning, and all adjectives have that function, though their modification may proceed in different ways depending on whether they are subsective or non-subsective. However, all words also have lexical and distributional aspects which we can induce from our corpora via DS methods. These include what is traditionally thought of as the conceptual content associated with words (e.g. whatever makes an elephant an elephant) and also distributional (selectional) restrictions.[4] While logical content is by nature context-invariant, lexical content is by nature context-sensitive in the sense that underspecified ('clouds of') meanings get precisified, shifted and modulated in context as in the case of *heavy bleeding* vs. *heavy box*. Our mixed model assumes that the two types of content complement and interact with each other.

Recognising the important aspect that the lexical (and pragmatic) aspects of function words play in their meaning also delineates which avenues are open for distributional semantics when it comes to approaching logical meanings. Lexical aspects of function words open a side-entrance by which it might be possible to approach the meaning of function words in natural language indirectly. One example of such an approach is Kruszewski et al.'s (2016) article, which proposes to tap into the meaning of negation in natural language by exploiting its focus-sensitive nature.

Our view also has consequences for the idea of the 'Logicality of language', proposed recently by Gajewski (2002), Fox and Hackl (2007) and Chierchia (2013), among others. These approaches rely crucially on the idea that there is a fundamental distinction between content and function words in natural language and that grammar is sensitive only to the content of functional vocabulary. If our approach is on the right track, then the presupposition of these accounts is not met in natural languages: the two types of content do not map to two different types of vocabulary. In Abrusán et al. (to appear) we spell out an alternative approach to explain the problems discussed in the 'Logicality of language' tradition that does not need this distinction.

In what follows, we first provide a brief introduction to the DS methods we used in Asher et al. (2016) and outline how these methods can inform us about meaning shifts. In Section 3 we show, based on preliminary results, what these methods give us when applied to determiner-noun combinations. In Section 4 we offer some speculations about what these findings imply for formal semantics.

[4] Although there is a conjunction in the expression 'lexical and distributional', in fact these are the same type of meaning from the point of view of DS.

2. Distributional Semantics and meaning shifts

Distributional Semantics, a computational approach to natural language semantics, can throw new light on meaning shifts in co-composition, as was shown in Asher et al. (2016). This paper outlines a close correspondence between Asher's (2011) Type Composition Logic (TCL) and DS methods that we will describe below. Below we provide a brief description of some of the distributional methods we used in this work. For details concerning how to translate the results of the distributional study into a symbolic system, readers are invited to consult Asher et al. (2016).

2.1. Distributional Semantics

Distributional Semantics is based on the so called "distributional hypothesis" by Harris (1954), according to which one can infer a meaning of a word by looking at its context. Observe the following examples for illustration:

(1) a. tasty *sooluceps*
b. sweet *sooluceps*
c. stale *sooluceps*
d. freshly baked *sooluceps*

The reader, even though they have never heard the word *sooluceps* before, is able to infer from the above examples that it is some sort of food, perhaps a type of cookie. How is this possible? It must be that the adjectives that modify this noun provide a clue as to the meaning of the noun.

In distributional semantics this idea is generalised as follows. The co-occurrence frequencies of two entities are captured by word vectors. Observe first the following toy example in which the co-occurrence frequencies of 4 nouns with 4 adjectives in some corpus are given:

	red	tasty	fast	second-hand
raspberry	728	592	1	0
strawberry	1035	437	0	2
car	392	0	487	370
truck	104	0	393	293

Table 1: A toy example

One way of thinking about word meaning within Distributional Semantics is to assume that it is a vector in some space **V** whose dimensions are contextual features. So in the above toy example, the meaning of *raspberry* is given by the vector that captures its co-occurence frequencies with the adjectives *red, tasty, fast, second-hand*. A graphical representation of such vectors in two-dimensional space (since four-dimensional spaces are hard to draw) is presented in Figure 1, with the two dimensions being *fast* and *red*.

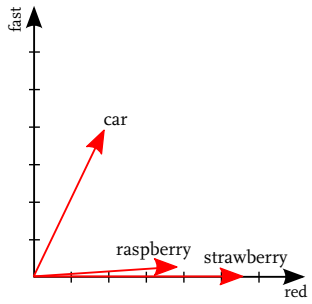

Figure 1: A graphical representation of word vectors in two-dimensional space

The graphical representation suggests a certain intuitive similarity between the words *strawberry* and *raspberry* as opposed to *car*: the vectors of the former two words have similar direction in vector space. This similarity can be captured mathematically by measuring the cosine similarity of the two vectors.[5]

2.2. Dimension reduction and aspects of meaning

When we move from toy examples towards real data, $words \times context$ matrices become very large and very sparse, with thousands if not hundreds of thousands of rows and columns. Contexts can include words and/or grammatical features or dependency relations that appear within a window of any size, where the window might be the sentence that the word appears in, or simply a certain number of words preceding and following a word, or something else still. In order to bring out the 'information content' in such huge matrices, dimensionality reduction techniques are applied. A dimensionality reduction reduces the abundance of overlapping contextual features to a limited number of meaningful, latent semantic dimensions.

Singular value decomposition While rooted in linear algebra, singular value decomposition (SVD) has proven to be a useful tool in statistical applications. It is closely related to statistical methods such as principal components analysis and factor analysis. SVD stems from a well known theorem in linear algebra: a rectangular matrix can be decomposed into three other matrices of specific forms, so that the product of these three matrices is equal to the original matrix:

$$A_{m\times n} = U_{m\times z}\,\Sigma_{z\times z}\,(V_{n\times z})^T \tag{1}$$

where $z = min(m,n)$. Matrix A is the original matrix of size $m \times n$. Matrix U is an $m \times z$ matrix that contains newly derived vectors called left-singular vectors. Matrix V^T denotes the transpose of matrix V, an $n \times z$ matrix of derived vectors called right-singular vectors. The third matrix Σ is a $z \times z$ square diagonal matrix (i.e. a square matrix with non-zero entries only

[5]Cosine similarity is just one of the various similarity measures that can be used, though probably the most popular (Turney and Pantel, 2010).

along the diagonal); Σ contains derived constants called singular values. A key property of the derived vectors is that all dimensions are orthogonal (i.e. linearly independent) to each other, so that each dimension is uncorrelated to the others.

The diagonal matrix Σ contains the singular values in descending order. Each singular value represents the amount of variance that is captured by a particular dimension. The left-singular and right-singular vector linked to the highest singular value represent the most important dimension in the data (i.e. the dimension that explains the most variance of the matrix); the singular vectors linked to the second highest value represent the second most important dimension (orthogonal to the first one), and so on. Typically, one uses only the first $k \ll z$ dimensions, stripping off the remaining singular values and singular vectors.[6] If one or more of the least significant singular values are omitted, then the reconstructed matrix will be the best possible least-squares approximation of the original matrix in the lower dimensional space. Intuitively, SVD is able to transform the original matrix—with an abundance of overlapping dimensions—into a new matrix that is many times smaller and able to describe the data in terms of its principal components. Due to this dimension reduction, a more succinct and more general representation of the data is obtained. Redundancy is filtered out, and data sparseness is reduced.

SVD is the underlying technique of the well-known information retrieval and text analysis method called Latent Semantic Analysis (Landauer and Dumais 1997; Landauer et al. 1998). A key characteristic of the resulting decomposition is that it contains both positive and negative values. Though the decomposition contains usable latent dimensions, it turns out the negative values make the resulting dimensions difficult to interpret. The application of a non-negative constraint, as in the factorization technique described in the following section, remediates this shortcoming.

Non-negative matrix factorization Another dimensionality reduction technique we deem particularly useful for semantic analysis is non-negative matrix factorisation (NMF; Lee and Seung, 1999). There are a number of reasons to prefer NMF over the better known singular value decomposition used in LSA. First of all, NMF allows us to minimize the Kullback-Leibler divergence as an objective function, whereas SVD minimizes the Euclidean distance. The Kullback-Leibler divergence is better suited for language phenomena. Minimizing the Euclidean distance requires normally distributed data, and language phenomena are typically not normally distributed (Baayen 2001). Secondly, the non-negative nature of the factorization ensures that only additive and no subtractive relations are allowed. This proves particularly useful for the extraction of semantic dimensions, so that the NMF model is able to extract much more clear-cut dimensions than an SVD model. And thirdly, the non-negative property allows the resulting model to be interpreted probabilistically, which is not straightforward with an SVD factorization.

Given a non-negative matrix $\mathbf{V}$, NMF finds non-negative matrix factors $\mathbf{W}$ and $\mathbf{H}$ such that when multiplied, they approximately reconstruct $\mathbf{V}$:

[6] A typical choice for k would be 300.

$$\mathbf{V}_{n\times m} \approx \mathbf{W}_{n\times k}\mathbf{H}_{k\times m} \tag{2}$$

A graphical representation of NMF applied to a matrix of nouns by context words is given in Figure 2.

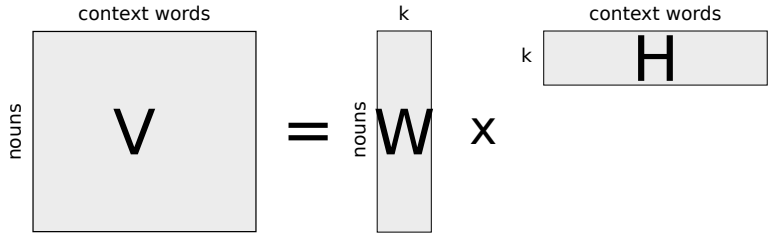

Figure 2: A graphical representation of NMF

As its name indicates, this factorization observes the constraint that all values in the three matrices need to be non-negative (≥ 0). Choosing $k \ll n,m$ reduces data significantly; for word-context matrices, k is typically chosen within the range 100–600.

As it turns out, reducing word-context matrices using NMF is particularly useful for finding topical, thematic information. For many of the k dimensions, the words with the highest value on that dimension seem to belong to the same topical field. Observe for example the nouns with the highest values on a number of example dimensions in Table 2 (computed from a word-context matrix extracted from Wikipedia). The examples indicate that NMF is able to automatically induce topically salient dimensions: dimension 60 has something to do with transport, dimension 88 with publishing, dimension 89 with computing and dimension 120 with living spaces. Although the labels of these dimensions are not given automatically, it is intuitive to think of these dimensions as semantic features, or topics. Factorisation also allows a more abstract way of representing the meaning of a word: we can now say that the meaning of a word is represented by a vector of size k whose dimensions are latent features.

dim 60	**dim 88**	**dim 89**	**dim 120**
rail	journal	filename	bathroom
bus	book	null	lounge
ferry	preface	integer	bedroom
train	anthology	string	kitchen
freight	author	parameter	WC
commuter	monograph	String	ensuite
tram	article	char	fireplace
airport	magazine	boolean	room
Heathrow	publisher	default	patio
Gatwick	pamphlet	int	dining

Table 2: Example dimensions (*k=300*)

Word embeddings Thirdly, we want to briefly touch upon a dimensionality reduction technique known as word embeddings. Though word embeddings are related to the factorization techniques mentioned above, the methods used to induce them operate somewhat differently. Word embeddings became popular with the recent surge of neural network methods for natural language processing applications (Collobert et al., 2011). By word embeddings, one usually denotes the low-level vector representations that serve as input to neural networks; the vector representations are typically automatically induced as parameters within the neural network, training on a particular task at hand. Word embeddings are often pre-trained in an unsupervised fashion by means of context prediction (Mikolov et al., 2013). As such, they are another instantiation of the distributional hypothesis.

As with SVD, word embeddings contain negative values, and therefore are more cumbersome when interpretation is concerned. Moreover, there is research that establishes a connection between SVD and induction methods for word embeddings (Levy and Goldberg, 2014). Still, there is a strong consensus within the NLP community that word embeddings provide adequate semantic representations, and as a result they might be useful for research on logical aspects of lexical items. We have not explored word embeddings in our research so far, but leave this interesting avenue for future work.

2.3. Composition: from aspects to meaning shifts

We have seen above how DS can generate vectors to capture individual word meaning and bring out latent dimensions that might correspond to semantic features. But what sort of semantic features would they be? In a purely denotational theory of meaning in which an expression would denote some sort of an intension, it is unclear how to represent these latent dimensions or indeed the collection of them as represented in a DS vector. In Asher et al. (2016), we took the view that DS vectors correspond to *internal* meanings or *types*, which the composition system uses to construct logical forms. Asher (2011), for instance, uses types to predict semantic anomalies and shifts in the meanings of polysemous words and already makes use of aspects in types, which we can think of as latent dimensions.

But how do we know that these latent dimensions could correspond to semantic features? One way to see this is to examine what happens in composition. If these latent dimensions affect composition and make empirically testable predictions about the semantic values of composing expressions, then that is evidence that these dimensions do correspond to semantic features. DS methods of composition involve the manipulation of vectorial or other algebraic representations of lexical content using various mathematical operations: vector addition, vector multiplication, and more complex forms of combination such as we will see below. The view from the DS approach connects to a growing body of work that assumes that the meaning of lexical words can be shifted or modulated in one way or another: either within the semantics (cf. e.g. Martí 2006; Stanley 2007; Asher 2011; Alxatib and Sauerland 2013) or within the pragmatics (Kamp and Partee 1995; Recanati 2010; Lasersohn 2012). Since we assume that meaning shift diagnosed by DS approaches happens at the compositional level, the view from DS is more in line with semantic approaches.

We have developed models of composition that show how the content of each word is modified during composition. Formally, our method is a DS implementation of the symbolic approach in Asher (2011). Asher's TCL approach provides the basic logical meanings for all expressions, including for instance their basic type information and methods of composition. In addition, it assumes a rich set of subtypes of the type of entities, and this assumption drives TCL's account of meaning shift in coercions and aspect selection in dual aspect nouns (Cruse 1986). However, like other symbolic methods, TCL has little to say about the content of the type associated with those subtypes. DS methods on the other hand tell us what the contents of those types are and how the compositional process modifies those contents. This method, applied for instance to the composition of an adjective with the noun it modifies, looks like this, where *A* is the adjective and *N* is the noun:

(2) *AN*: $\lambda x\,(\mathscr{O}_A(N(x)) \wedge \mathscr{M}_N(A(x)))$

$\mathscr{O}$ and $\mathscr{M}$ are functors intended to capture the shift in meaning induced by the compositional process. For an an expression like *heavy traffic* we would have:

(3) *heavy traffic* : $\lambda x.(\mathscr{O}(\text{heavy})(x) \wedge \mathscr{M}(\text{traffic})(x))$

The meaning of both nouns and adjectives can thus change in this system, according to the words they combine with. However, Asher (2011) does not supply a method for constructing the functors $\mathscr{O}$ and $\mathscr{M}$. This is what we can do with DS automatically. Moreover, as we will see below in our discussion of a previous study on nouns and adjectives, different latent dimensions of meaning of both the adjective and the noun can be reinforced, depending on what these expressions combine with.

2.4. A distributional model for compositionality

In order to capture meaning shift as in the case of *heavy traffic*, the meaning of the adjective needs to be adapted to the context of the particular noun that it co-occurs with. That is, the distributional model needs to provide us with the functors $\mathscr{O}_A$ and $\mathscr{M}_N$ in the TCL approach. In Asher et al. (2016), we chose two different approaches that meet this requirement: one based on *matrix* factorization (Van de Cruys et al., 2011) and one based on *tensor* factorization (Van de Cruys et al., 2013). In what follows, we describe briefly the second method and the results we got with it. Note that the following paragraphs only provide a brief overview of the model; for more details, see Asher et al. (2016).

Tensor factorization The approach based on tensor factorization allows for a rich and flexible modeling of the interaction between adjectives and nouns, in order to provide an adequate representation of each when they appear in each other's context. The key idea is to factorize a three-way tensor[7] that contains the multi-way co-occurrences of nouns, adjectives and other dependency relations (in a direct dependency relationship to the noun) that appear together at

[7] A tensor is the generalization of a matrix to more than two axes or *modes*.

the same time. A number of well-known tensor factorization algorithms exist; we opted for an algorithm called Tucker factorization, which allows for a richer modeling of multi-way interactions using a core tensor. In Tucker factorization, a tensor is decomposed into a core tensor, multiplied by a matrix along each mode. For a three-mode tensor $\mathbf{X} \in \mathbb{R}^{I \times J \times L}$, the model is defined as follows:

$$\begin{aligned} \mathbf{X} &= \mathbf{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} \\ &= \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} \mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r \end{aligned} \tag{3}$$

where $\circ$ represents the outer product of vectors. By setting $P, Q, R \ll I, J, L$, the factorization represents a compressed, latent version of the original tensor $\mathbf{X}$; matrices $\mathbf{A} \in \mathbb{R}^{I \times P}$, $\mathbf{B} \in \mathbb{R}^{J \times Q}$, and $\mathbf{C} \in \mathbb{R}^{L \times R}$ represent the latent factors for each mode, while $\mathbf{G} \in \mathbb{R}^{P \times Q \times R}$ indicates the level of interaction between the different latent factors. Figure 3 shows a graphical representation of Tucker decomposition.[8]

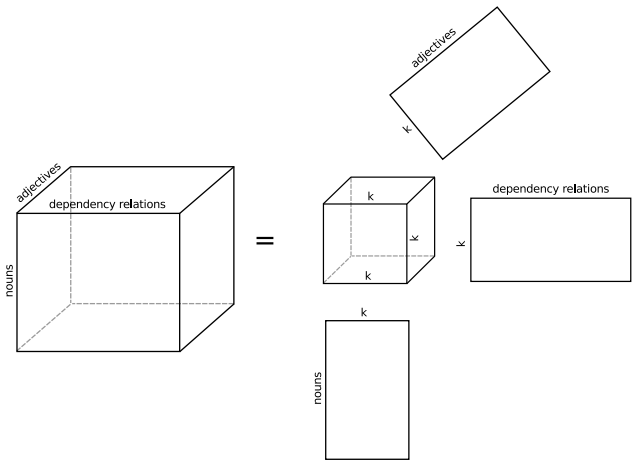

Figure 3: A graphical representation of Tucker decomposition

We carried out the factorization with non-negative constraints, and we found the best possible fit to the original tensor $\mathbf{X}$ using Kullback-Leibler divergence, a standard information-theoretic measure. To ensure that the algorithm for non-negative Tucker decomposition finds a good global optimum, we initialized the three matrices using data that comes from non-negative matrix factorization, cf. Asher et al. (2016).

Computing meaning shifts We can now compute a representation for a particular adjective-noun composition. In order to do so, we first extract the vectors for the noun ($\mathbf{a}^i$) and adjective ($\mathbf{b}^j$) from the corresponding matrices $\mathbf{A}$ and $\mathbf{B}$. We multiply those vectors into the core tensor, in

[8] where $P = Q = R = K$, i.e. the same number of latent factors K is used for each mode

order to get a vector **h** representing the importance of latent dimensions given the composition of noun i and adjective j, i.e.

$$\mathbf{h} = \mathbf{G} \times_1 \mathbf{a}^i \times_2 \mathbf{b}^j \tag{4}$$

By multiplying the vector representing the latent dimension with the transpose of the matrix for the mode with dependency relations ($\mathbf{C}^T$), we are able to compute a vector **d** representing the importance of each dependency feature given the adjective-noun composition, i.e.

$$\mathbf{d} = \mathbf{h}\mathbf{C}^T \tag{5}$$

The vector **d** is in effect the DS version of TCL's functor $\mathscr{O}_A$, which we now have to combine with the original noun meaning. This last step goes as follows in DS: we weight the original noun vector according to the importance of each dependency feature given the adjective-noun composition, by taking the point-wise multiplication of vector **d** and the original noun vector **v**, i.e.

$$\mathbf{v'}_d = \mathbf{d}_d.\mathbf{v}_d \tag{6}$$

Note that we could just keep the representation of our adjective-noun composition in latent space. In practice, the original dependency-based representation provides a much richer semantics, which is why we have chosen to perform an extra step weighting the original vector.

Some implementational details We used the UKWaC corpus (Baroni et al., 2009), an internet corpus of about 1.5 billion words, to construct the algebraic structures for our approaches. We tagged the corpus with part-of-speech tags, lemmatized it with Stanford Part-Of-Speech Tagger (Toutanova and Manning, 2000; Toutanova et al., 2003), and parsed it using MaltParser (Nivre et al., 2006). We extracted our input tensor **X** of 5000 nouns by 2000 adjectives by 80,000 dependency relations from the corpus. The tensor **X** was weighted using a three-way extension of PMI (Van de Cruys, 2011). We set $K = 300$ as our number of latent factors. All similarity computations were performed using cosine as a similarity measure.

An example Finally, observe an example illustrating the unshifted meaning of the adjective *heavy* vs. the shifted meaning of the same adjective in the context of the noun *traffic*:

(4) **heavy**$_A$: *heavy*$_A$ (1.000), *torrential*$_A$ (.149), *light*$_A$ (.140), *thick*$_A$ (.127), *massive*$_A$ (.118), *excessive*$_A$ (.115), *soft*$_A$ (.107), *large*$_A$ (.107), *huge*$_A$ (.104), *big*$_A$ (.103)

(5) **heavy**$_A$, traffic$_N$: *heavy*$_A$ (.293), *motorised*$_A$ (.231), *vehicular*$_A$ (.229), *peak*$_A$ (.181), *one-way*$_A$ (.181), *horse-drawn*$_A$ (.175), *fast-moving*$_A$ (.164), *articulated*$_A$ (.158), *calming*$_A$ (.156), *horrendous*$_A$ (.146)

There is an evident shift in the composed meaning of *heavy* relative to its original meaning; there is no overlap in the lists (4) and (5) above except for *heavy*. We see this also in the quantitative measure of cosine similarity, sim_{cos}, between the original vector for *heavy* $\vec{v}_0$ and the modified vector for *heavy* $\vec{v}_1$ as modified by its predicational context: With the tensor model, on average, $sim_{cos}(\vec{v}_{orig}, \vec{v}_{mod})$ was 0.2 for adjectives and 0.5 for nouns. In addition, these different senses of *heavy* were reflected in the dimensions in which *heavy* occurred, thus confirming that aspects of meaning affect composition and meaning shift. Finally, Asher et al. (2016) validated these meaning shifts in terms of speaker judgments.

3. Determiners, logical meaning and shiftable meaning

We have seen above how distributional semantics can inform us about the nature of meaning shifts. The distributional method that we introduced above for calculating meaning shift adapts the vector of the original predicate to its predicational context using the latent dimensions derived during dimensionality reduction. The way the distributional method calculates meaning shifts implies that meaning shift crucially depends on the latent dimensions that we find during tensor factorisation: it is the semantic features implicitly present in the latent dimension that drive the meaning shift. Distributional semantics thus picks up the aspects of lexical meaning that vary with the context: these are the aspects of the meaning that are affected by changes in the distribution. As a result, DS can tell us about which aspects of meaning of an expression can shift; aspects of the meaning that correspond to (or interact with) semantic dimensions uncovered by distributional semantics methods are in principle shiftable. In contrast, aspects of the meaning that are invisible for DS are unshiftable. In Abrusán et al. (to appear), we argue that clashes in unshiftable content of a predicate and its argument lead to semantic anomaly, and shiftable contents lead to shifts of meaning in composition.

We now apply a distributional approach to determiners. Do determiners have meanings that can shift, or do they have only unshiftable meanings? Logical meaning, the meaning upon which valid inferences rest, must be present in all contexts, and so we expect it to be invisible for DS; in particular, we expect that it will not show up in dimensions of the latent space where certain contexts are operative. So, whether or not we get logical meaning or any other meaning to shift depends on whether we find latent dimensions with our dimensionality reduction methods that correspond to logical meaning.

In recent work we performed a number of preliminary experiments similar to the ones described in the previous section but this time with determiner-noun compositions. Specifically, we looked at four determiners, *a, any, some* and *every* using two different corpora: Wikipedia, and a corpus of unpublished novels collected from the web (Zhu et al., 2015).[9] We extracted an input tensor $\mathbf{X}$ of 5000 nouns by four determiners by 80,000 dependency relations from each of the two corpora. The tensors were weighted using a three-way extension of PMI, cf. Van de Cruys (2011). The tensor was factorized using tensor factorization, with $k = 30$ as our number of latent factors.

[9]The former corpus contains about 1 billion words, the latter about 1.5 billion words. Preprocessing was performed similarly to the approach described in Section 2.4.

In the resulting factorization, determiners and nouns as well as dependency relations are all linked to same latent dimension. We can now go and inspect each of the 30 latent dimensions manually, by looking at the list of highest ranked words in each dimension. In the corpus of novels, we found that out of the 30 dimensions, 5 had *some* as the most important determiner (i.e. the determiner with the highest value), 3 dimensions had *every* as the most important determiner, and in 1 dimension *any* was the most important. The remaining dimensions were dominated by the determiner *a*. We found similar results with the Wikipedia corpus: we found 3 dimensions with *some* being highly ranked, 3 dimensions with *any* being highly ranked and one dimension with *every* highly ranked. The rest of the dimensions were dimensions with the determiner *a* ranked most highly. In the following paragraphs, we try to identify a number of semantic characterizations linked to the latent dimensions.

Dimensions An intuitive evaluation of the semantic coherence of each of the 30 dimensions was conducted by the authors, and we have found that many of these seem to capture interesting semantic features, albeit not logical features. Here we describe the results of the Novels corpus.[10] In the case of the determiner *some*, we found that two of the 5 dimensions seem to capture uncertainty or indifference about the identity of the discourse referent in question. We see this from trying to recompose the highest ranked determiner with the highest ranked nouns and other dependency features on a particular dimension, cf. examples in (6). Another two of the five dimensions capture measure or quantity readings with *some*, the difference between the two dimensions being that in one we found nouns that denote more concrete things (e.g. food), in the other we find more abstract nouns. The fifth dimension arguably captures kind or sort readings with *some*.

(6) *some*
- a. [uncertainty, indifference]: e.g. some people argue; for some reason; on some level
- b. [measure/quantity]: e.g. some food, some protection, some comfort, some help
- c. [kind/sort]: e.g. some kind, some sort

With the determiner *every*, one dimension we got very robustly was a temporal dimension: the highest ranked nouns were all temporal. Another dimension seemed to capture part-whole relations, see (7b). (We could not make sense of the third dimension, which is why we omitted it here.)

(7) *every*
- a. [temporal]: e.g. every day, every year, every minute
- b. [part-whole]: e.g. every inch, every detail, every part

The dimension we got with the determiner *any* seems to correspond to the negative polarity (possibly also free choice?), with negation and the modals *could* and *would* being the highest ranked modifiers among dependency features.

[10] The results of the study on the Wikipedia corpus were similar, but less rich.

(8) *any*
[negative polarity]: e.g. not show any emotion, without any warning, at any moment

In the case of the determiner *a*, we mostly found topical dimensions, e.g. legal, publishing, building construction, political campaigns, people, etc. (especially in the Wikipedia corpus). In some dimensions *a* appeared exclusively within prepositional modifier phrases (*in a chair, with a grin*). The rest of the dimensions were uninterpretable to us. We must hence be careful about making too many claims about these dimensions of determiner meaning. However, we hope to demonstrate in future work that these dimensions recur in different corpora and when choosing different latent spaces. If this is the case, then we think this is good evidence that these semantic principles are part of the determiners' internal meanings.

Interpretation What we have described above is still work in progress, but it is already clear that we are not getting any dimensions via tensor factorisation that correspond to logical meaning. As a consequence, we are not going to get logical meaning to shift. This is not surprising given that logical meanings are supposed to validate logical deductions universally regardless of context. Thus the fact that logical meaning shouldn't shift with content comes with the definition of logical meaning and the universally valid inferences it purports to underwrite.

On the other hand, it seems to us that the dimensions that we do get correspond to some aspects of the *lexical/distributional* meaning of quantifiers. In light of this, one way to interpret our results with the determiner *a* is that this determiner does not have any extra conceptual content beyond its logical meaning. In the case of *any* we get a dimension that captures its peculiar distribution. Most interestingly, perhaps, the dimensions we find with the quantifier *some* correspond to non-logical aspects of its use that have puzzled semanticists for a long time. The first of these is uncertainty about identity (also known as the *epistemic* aspect of indefinites). Indefinites with an epistemic effect can be marked by a special determiner in many languages, e.g. in German or Spanish (cf. Kratzer and Shimoyama 2002; Alonso-Ovalle and Menéndez-Benito 2015). Other aspects of the determiner *some* include measure and kind readings. In some languages, all these different aspects are marked explicitly, e.g. in Hungarian and in many Slavic languages (cf. Haspelmath 1997; Szabolcsi 2015). In particular, the Hungarian determiner *some* incorporates *wh*-words and the relevant aspects of *some* are tied to the *wh*-word:

(9) Indefinite determiners in Hungarian: VALA+WH-word N:
   a. vala+mi N [lit: some+what$_{\text{N}}$]:
   suggests uncertainty about the identity of N, e.g. some guy
   b. vala+milyen N [lit: some+what$_{\text{A}}$]:
   kind or sortal reading: e.g. some kind of drug
   c. vala+melyik N [lit: some+which]:
   partitive reading: one of the Ns
   d. vala+mennyi N [lit: some+how-much]:
   amount reading: some amount of N
   e. vala+hany N [lit: some+how-many]:
   count reading: some number of Ns

Note that *vala-* in itself is not a word and so every occurrence of the determiner *some* incorporates a *wh*-word. As a result, *some* in Hungarian is always classified into one of the readings signaled by the *wh*-word.[11] Below are some typical examples found on the web:

(10) a. A bárban **valami pasas** énekelt Woody Guthrie számokat.
'In the bar *some guy* was singing Woody Guthrie songs.'
b. Próbált emlékezni, de nem ment. Mintha leblokkolt volna az agya, mintha **valamilyen gyógyszer** hatása alatt állt volna.
'He was trying to remember, but couldn't. As if his brain had gone blank, as if he was under the effect of *some drug*.'
c. De mi van olyankor, ha **valamelyik családtag** allergiás?
'But what happens if *some family member* has allergy?'
d. A szeretetben mindig van **valamennyi őrület**. De az őrületben is mindig van **valamennyi ész**. (F. Nietzsche)[12]
'In love there is always *some madness*. But there is also *some reason* in madness.'
e. Volt **valahány gyerek** a kertben.
'There were some children in the garden (I do not know how many).'

Finally, the dimensions we find with the quantifier *every* are somewhat puzzling and not easy to interpret. Probably, the part-whole distinction corresponds to a semantically relevant dimension, since universals (and also existentials) often appear with overt partitive constructions. In contrast, the temporal dimension we found might simply be an artifact of the extremely frequent use of *every* with temporal nouns in context.[13]

4. What does this mean for linguistics?

In the previous section, we showed that determiners, like open class nouns and adjectives, have aspects of meaning that cluster in some but not all latent dimensions. Given our experiments on adjective noun composition, we fully expect that composition of a determiner with a common noun phrase (NP) to form a DP will also exhibit shifts in meaning—not shifts in logical meaning (*every* doesn't suddenly mean *some* or *many*) but shifts in the sort of meanings we have found in the latent dimensions like epistemic indefinites, the negative polarity semantic behavior of *any* and less well known aspects like the dimension of *every* that selects for temporal NPs. In this section, we speculate how our observations relate to semantics as more traditionally construed.

4.1. Different corpora: different meaning aspects?

In our studies we looked at two different corpora to provide us with contexts and finally a set of latent dimensions for our determiners. We also examined spaces with different numbers of latent dimensions. Happily, the aspects of determiner meaning that we reported on above

[11] We find the same pattern in Hungarian with free choice items and also negative existentials: *akármi, akármilyen, akármelyik,* etc.

[12] "Es ist immer etwas Wahnsinn in der Liebe. Es ist aber immer auch etwas Vernunft im Wahnsinn." - Thus Spoke Zarathustra (1885)

[13] To see if this is indeed the case, it would be interesting to compare *every* with *each* and *all*.

showed up across all the latent spaces. Moreover, at least some of the aspects we isolated are grammatically marked in languages other than English, and that in itself is evidence that they correspond to aspects of the semantics of the expressions. On the other hand, it is clear that the analysis of dimensions or the shifts in meaning that these dimensions induce in composition do not exhaust the meaning of expressions. No distributional semanticist, we believe, would conclude that, just because we don't see the logical meaning of *every* showing up in any particular dimension, it does not have the logical meaning that it evidently does. But then how do or should these different aspects of determiner meaning relate to the core logical meaning? In our DS model and the underlying theory of types that it implements, we suppose that since logical meaning is present in every context of use, and hence in every latent dimension, the operations we do to bring out certain aspects of meaning that are more present in some contexts will not affect the logical meaning of the determiner. The logical meaning is a constant component of the type, while the shiftable aspects of meaning are more or less present depending on what the determiner is composing with.

This view of composition already indicates how we might want to formulate an analysis of epistemic indefinite uses of *some*. If we follow our DS and TCL model, the epistemic use should come from a compositional account in which elements of the context of use of *some* reinforce this interpretation. Since our results are very much dependent on the kinds of context we choose, what context we use to analyze this epistemic use is an important question we need answer. With the right notion of context, the DS model of composition could then in principle tell us which contextual elements reinforce this interpretation.

There is also the question about what the various latent dimensions represent. Does each one of them in fact represent an aspect of the semantics of a determiner? Even the ones we can't interpret? If they don't represent semantic aspects, what do they represent? We don't know the answers to this question, but we feel that these are important questions to ask and to resolve for those who are using DS methods and believe that DS can offer an explanatory, theoretically satisfying model of meaning.[14] A related question is, what about the differences we noticed across our two corpora? Some dimensions of meaning were more widespread in one space than another; in some spaces a dimension could be more amenable to interpretation than in others. Do these indicate a difference in semantics too?

4.2. On the cherry picking argument

The questions in the preceding subsection highlight a difficulty in studying latent dimensions of meaning using DS methods. If we can't interpret some dimensions or don't see any semantic relevance in them, then our selection of certain dimensions as being semantically informative can seem suspect. Looking through latent dimensions and "cherry picking" the ones we find interesting doesn't seem like the right way to do semantics as a science. However, once we see that differences in content in dimensions lead to shifting in composition and we can empirically

[14]These questions become all the more pressing once researchers start to exploit nonlinear and neural net methods for representing word meaning, as such architectures are intrinsically much more difficult to interpret than the linear algebraic techniques we use here.

test the effects of composition, the cherry picking argument loses its force. In addition, the fact that some of our dimensions are grammaticalized in some languages attests to their semantic relevance. The cherry picking argument, however, still points out a potential embarrassment about the dimensions that we can't interpret. Our inability to explain aspects of the model hampers its theoretical power.

4.3. The content vs. functional distinction

Logical and lexical aspects of meaning do not map neatly to two different types of words. There is no purely logical vocabulary nor purely lexical/conceptual vocabulary. Instead, this distinction cross-cuts word boundaries. The meanings of lexical as well as logical words have both logical aspects (their model-theoretic meaning) and lexical/conceptual aspects. We cannot neatly separate grammar and conceptual knowledge because they are packaged together within lexical entries. Similarly, the boundary between shiftable and unshiftable content does not map neatly to lexical vs. functional vocabulary: both closed and open class words can have shiftable and unshiftable aspects to their meaning.

Though our study described above is still work in progress, it suggests that there are aspects of the meaning of quantificational determiners that might shift, namely the conceptual content that they have on top of their logical meaning. We suspect that this is the case at least for the quantifiers that have such meanings, e.g. *some*, though probably not the determiner *a*. Assuming that the conceptual meaning of determiners also includes at least some reflex of their logical meaning (Szymanik and Zajenkowski 2010), we can say that there are parts of the conceptual content of quantificational determiners that might shift, and there is also a part of their conceptual content that is invariant with context. The first type of conceptual meaning corresponds to the semantic dimensions that distributional semantics can uncover. The second type of conceptual meaning is the conceptual reflex of the logical meaning of these items.

For example, in the case of the determiner *some*, the conceptual reflex of the existential quantification is non-shiftable. However, the other conceptual effects associated with it, e.g. uncertainty about identity, measure readings, kind readings, partitive readings, etc. might be shiftable.

5. Conclusion

In this paper we have argued, based on findings from Distributional Semantics, that both function and content words have lexical/distributional content in addition to their logical content. Our results, based on a preliminary study of determiners, indicate that the difference between the two types of words is a difference in degree rather than being categorical. These findings, if correct, have implications both for distributional and formal semantics. For distributional semantics, they indicate a possible avenue that can be used to tap into the meaning of function words. For formal semantics, they bring into light the context-sensitive, lexical aspects of function words that can be recovered from the data even when these are not overtly marked. Such pervasive context-sensitivity has profound implications for how we think about meaning in natural language.

References

Abrusán, M., N. Asher, and T. V. de Cruys (to appear). Grammaticality and meaning shift. In G. Sagi and J. Woods (Eds.), *The Semantic Conception of Logic*. Cambridge University Press, forthcoming.

Alonso-Ovalle, L. and P. Menéndez-Benito (2015). *Epistemic indefinites: Exploring modality beyond the verbal domain*. Oxford University Press, USA.

Alxatib, S., P. P. and U. Sauerland (2013). Acceptable contradictions: Pragmatics or semantics? *Journal of Philosophical Logic 42*, 619–634.

Asher, N. (2011). *Lexical Meaning in Context: A web of words*. Cambridge University Press.

Asher, N., T. van de Cruys, A. Bride, and M. Abrusán (2016). Integrating type theory and distributional semantics: a case study on adjective-noun compositions. *Computational Linguistics 42*(4), 703–725.

Baayen, R. H. (2001). *Word frequency distributions*. Kluwer.

Baroni, M., R. Bernardi, N.-Q. Do, and C.-c. Shan (2012). Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 23–32. Association for Computational Linguistics.

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta (2009). The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation 43*(3), 209–226.

van Benthem, J. (1989). Logical constants across varying types. *Notre Dame Journal of Formal Logic 30*(3), 315–342.

van Benthem, J. (2002). Invariance and definability: two faces of logical constants. essays in honor of sol feferman. In W. Sieg, R. Sommer, and C. Talcott (Eds.), *Reflections on the Foundations of Mathematics.*, ASL Lecture Notes in Logic 15., pp. 426–446.

Bernardi, R., G. Dinu, M. Marelli, and M. Baroni (2013). A relatedness benchmark to test the role of determiners in compositional distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 53–57.

Bonnay, D. (2008). Logicality and invariance. *Bulletin of Symbolic Logic 14*(1), 29–68.

Chierchia, G. (2013). *Logic in grammar: Polarity, free choice, and intervention*. OUP Oxford.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research 12*, 2493–2537.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Feferman, S. (1999). Logic, logics, and logicism. *Notre Dame Journal of Formal Logic 40*(1), 31–54.

Fox, D. and M. Hackl (2007). The Universal Density of Measurement. *Linguistics and Philosophy 29*, 537–586.

Gajewski, J. (2002). On Analyticity in Natural Language. Ms., MIT.

Greenhill, S. J., C.-H. Wu, X. Hua, M. Dunn, S. C. Levinson, and R. D. Gray (2017). Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences 114*(42), E8822–E8829.

Harris, Z. S. (1954). Distributional structure. *Word 10*(2-3), 146–162.

Haspelmath, M. (1997). *Indefinite pronouns*. Clarendon Press Oxford.

Hermann, K. M., E. Grefenstette, and P. Blunsom (2013, August). "not not bad" is not "bad": A distributional account of negation. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, Sofia, Bulgaria, pp. 74–82. Association for Computational Linguistics.

Kamp, H. and B. Partee (1995). Prototype theory and compositionality. *Cognition 57*(2), 129–191.

Kaplan, D. (1989). Demonstratives. In J. Almog, J. Perry, and H. Wettstein (Eds.), *Themes from Kaplan*. Oxford.

Kratzer, A. and J. Shimoyama (2002). Indeterminate pronouns: The view from Japanese. *The Proceedings of the Third Tokyo Conference on Psycholinguistics*, 1–25.

Kruszewski, G., D. Paperno, R. Bernardi, and M. Baroni (2016). There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics. *Computational Linguistics 42*(4), 637–660.

Landauer, T. and S. Dumais (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychology Review 104*, 211–240.

Landauer, T., P. Foltz, and D. Laham (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes 25*, 295–284.

Lasersohn, P. (2012). Contextualism and compositionality. *Linguistics and Philosophy 35*, 171–189.

Lee, D. D. and H. S. Seung (1999). Learning the parts of objects by non-negative matrix factorization. *Nature 401*(6755), 788–791.

Levy, O. and Y. Goldberg (2014). Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27*, pp. 2177–2185. Curran Associates, Inc.

Linzen, T., E. Dupoux, and B. Spector (2016). Quantificational features in distributional word representations. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pp. 1–11.

MacFarlane, J. (2017). Logical constants. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.). Metaphysics Research Lab, Stanford University.

Martí, L. (2006). Unarticulated constituents revisited. *Linguistics and Philosophy 29*.

May, R. (1991). Syntax, semantics, and logical form. *The Chomskyan Turn, Blackwell, Oxford*, 334–359.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pp. 3111–3119.

Nivre, J., J. Hall, and J. Nilsson (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC-2006*, pp. 2216–2219.

Partee, B. (1992). Syntactic categories and semantic type. *Computational linguistics and formal semantics*, 97–126.

Recanati, F. (2010). *Truth-conditional pragmatics*. Clarendon Press Oxford.

Sher, G. (1991). *The Bounds of Logic*. Cambridge, Mass: MIT Press.

Stanley, J. (2007). *Language in context: Selected essays.* Oxford: Clarendon Press.

Szabolcsi, A. (2015). What do quantifier particles do? *Linguistics and Philosophy 38*(2), 159.

Szymanik, J. and M. Zajenkowski (2010). Comprehension of simple quantifiers: Empirical evaluation of a computational model. *Cognitive Science 34*(3), 521–532.

Tarski, A. and S. R. Givant (1987). *A formalization of set theory without variables*, Volume 41. American Mathematical Soc.

Toutanova, K., D. Klein, C. Manning, and Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, pp. 252–259.

Toutanova, K. and C. D. Manning (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63–70.

Turney, P. and P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research 37*(1), 141–188.

Van de Cruys, T. (2011, June). Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, Portland, Oregon, USA, pp. 16–20. Association for Computational Linguistics.

Van de Cruys, T., T. Poibeau, and A. Korhonen (2011, July). Latent vector weighting for word meaning in context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., pp. 1012–1022.

Van de Cruys, T., T. Poibeau, and A. Korhonen (2013). A tensor-based factorization model of semantic compositionality. In *Conference of the North American Chapter of the Association of Computational Linguistics (HTL-NAACL)*, pp. 1142–1151.

Zhu, Y., R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR abs/1506.06724*.