Glossa
a journal of general linguistics

Varying Abstractions: a conceptual vs. distributional view on prepositional polysemy

RESEARCH

LAUREN FONTEYN

]u[ubiquity press

Abstract

The term 'meaning', as it is presently employed in Linguistics, is a polysemous concept, covering a broad range of operational definitions. Focussing on two of these definitions, meaning as 'concept' and meaning as 'context' (also known as 'distributional semantics'), this paper explores to what extent these operational definitions lead to converging conclusions regarding the number and nature of distinct senses a polysemous form covers. More specifically, it investigates whether the sense network that emerges from the principled polysemy model of *over* as proposed by Tyler & Evans (2003; 2001) can be reconstructed by the neural language model BERT. The study assesses whether the contextual information encoded in BERT embeddings can be employed to succesfully (i) recognize the abstract sense categories and (ii) replicate the relative distances between the senses of *over* proposed in the principled polysemy model. The results suggest that, while there is partial convergence, the two models ultimately lead to different global abstractions because the imagistic information that plays a key role in conceptual approaches to prepositional meaning may not be encoded in contextualized word embeddings.

CORRESPONDING AUTHOR:
Lauren Fonteyn
Leiden University,
Arsenaalstraat 1, 2311 CT
Leiden, NL
l.fonteyn@hum.leidenuniv.nl

KEYWORDS:
Prepositions; Polysemy; Image Schema; Metaphor; Distributional Semantics; BERT

TO CITE THIS ARTICLE:
Fonteyn, Lauren. 2021. Varying Abstractions: a conceptual vs. distributional view on prepositional polysemy. *Glossa: a journal of general linguistics* 6(1): 90. 1–28. DOI: *https://doi.org/10.5334/gjgl.1323*

DOI: 10.5334/gjgl.1323

1 Introduction

The aim of the present study is to empirically investigate whether there is any correspondence between the generalizations that emerge from different operational models of meaning representation. More specifically, this study focuses on the operational definition of meaning as 'concept', as commonly employed in Cognitive Linguistics, and meaning defined as (or derived from) 'context', also known as distributional semantics, which has rapidly gained popularity in Corpus Linguistics and Computational Linguistics/NLP. As a case study, it will home in on the semantics of the English preposition *over*.

Following Brugman (1988)'s and Lakoff (1987)'s extensive treatment of *over*, prepositions have become default examples in descriptions of the core tenets 'the cognitive approach' to meaning (see, e.g. Lemmens 2016). In the cognitive-conceptual approach to semantics, meanings are defined as 'concepts' which are connected and grounded in complex knowledge structures (e.g. Clausner & Croft 1999), often described as complex chains or 'networks' of connected senses. Central to the conceptual approach is that these concepts (and their network of senses) are to a large extent experiential – that is, they are grounded in the physical or cultural experience of the language user.

When it comes to their discussion of polysemy, such cognitive-conceptual accounts have faced substantial criticism, predominantly aimed at their apparent lack of principled and objective methods to determine how many senses can be distinguished, and how the global design of the polysemy networks is construed. One part of the problem appeared to be that the identification of the core node (or 'prototype') of the network seemed to rely solely on subjective, introspective judgements (Sandra & Rice 1995; Rice 1996: 137). These criticisms triggered a search for concrete criteria and data-driven tests for prototypicality (e.g. Gilquin & McMichael 2018; Newman 2011). A similar discussion also arose regarding the position of derived nodes, with many scholars acknowledging that more objective, non-introspective discussions regarding 'distances' between senses will remain all but impossible as long as we are not "able to *measure* the degree of similarity between senses" (Gries & Divjak 2009: 57; emphasis mine).

In response to the need for more well-defined and data-driven definitions of word senses (e.g Kilgarriff 2003), and more objective, measurable ways of determining distances between senses, a number of proposals have been devised that could be subsumed under the header of 'contextual' or 'distributional' approaches to semantics. Underlying these approaches is the premise that the meaning of a word can be derived from the context in which it occurs (an idea which dates back to at least Harris (1954) and Firth (1957)). While not necessarily equating context to meaning, distributional approaches to semantics are based on the assumption that co-occurrence patterns and other distributional frequencies are indicative of the meaning of a linguistic item (i.e., they serve as "proxies" for meaning representation; see, for instance Baroni et al. (2014: 238)). Hence, distributional similarities between linguistic items in a corpus can be used to approximate a measurement of their functional or semantic similarity, thus placing more prominence on the geometrical relationship between linguistic items (Boleda & Erk 2015). Particularly in recent years, the approach has been met with great enthusiasm in Computational Linguistics and Machine Learning, as recent incarnations of such distributional semantic models seem to perform astonishingly well on a wide range of NLP, production, and machine translation tasks (Young et al. 2018; Radford et al. 2018).

A similar (yet overall more cautious) enthusiasm has been expressed in Linguistics: being almost exclusively corpus-based, distributional approaches have provides a welcome bridge between the "well-established corpus-linguistic research tradition and Langacker's idea that linguistic representations emerge from linguistic usage" (Stefanowitsch 2010: 370). Furthermore, because some of the more traditional conceptual models of polysemy already depended on distributional criteria to some extent (Gries & Divjak 2009: 58; Geeraerts 2016: 241), the path to a full-fledged distributional approach had already been cleared. Such an approach involves virtually no introspective manual annotation (but see, e.g., comments in Gries & Divjak (2009) and Heylen et al. (2015: 154)), and hence it allows the elusive concept of meaning to be studied in a more rigorous, large-scale, and ultimately quantifiable way. The confidence that distributional methods provide an appropriate, objective and fully data-driven alternative to more introspective models appears to be growing steadily, as there have been some successes in replicating experimental, survey-based accounts of polysemy by means of distributional

methods (e.g. Berez & Gries 2008; Gries & Divjak 2010). In particular in historical linguistics, where native-speaker intuitions regarding the semantics of linguistic items is inevitably inaccessible, distributional methods are now considered a welcome methodological innovation (e.g. Sagi et al. 2011; Hilpert & Correia Saavedra 2017; Budts 2020).

Yet, while the distributional approach and the more traditional, cognitive-conceptual approach essentially share the same goal (that is, to capture the complex internal semantic structure of linguistic items in a rigorous, theoretically motivated, principled manner), it is also clear that there may be a non-trivial epistemological difference between the two approaches. The distributional approach to meaning and sense distinctions is based on a premise that essentially conflicts with one of the core criteria of, for instance, the principled polysemy approach (Tyler & Evans 2001; 2003), which treats a linguistic item's meaning as precisely that which cannot be directly inferred from contextual cues. Naturally, then, the question arises to what extent the two approaches are related to one another, and whether they can still lead to comparable results. A positive answer to this question (i.e. the results of the two approaches converge) suggests that distributional models are able to use contextual cues to extract and encode the conceptual information that lies at the core of the cognitive-conceptual model. This would mean that we can indeed further fine-tune proposals within the conceptual approach by enabling a quantitative discussion on sense distinctions and possible derivational paths. A negative answer, by contrast, would raise questions about the compatibility of the two approaches, and may even have larger implications. It may seem fair to operate under the "a priori" assumption that different operational definitions of meaning are designed to capture the same, unitary phenomenon (Stefanowitsch 2010: 371), but if the study of meaning is operationalized in two different ways, and we find that the results of a distributional and cognitive-conceptual model lead to different generalizations and abstractions, one may start to question whether they are in fact capturing the same phenomenon (Geeraerts 2016).

Prioritizing depth over width, this study is set up as a detailed empirical comparison between two operational models of meaning representation, with one serving as a representative of the cognitive semantic ('concept') approach, and one representing the distributional ('context') approach. More specifically, this study focuses on one of the most well-developed cognitive-conceptual proposals – the principled polysemy model of *over* as set out by Tyler & Evans (2001; 2003) – and aims to investigate whether the polysemy network that emerges from this theoretical model of meaning representation can be reconstructed by means of a recent neural distributional language model called BERT ('Bidirectional encoder Representations from Transformers'). To this end, a stratified sample of 808 contextualized instances of the preposition *over* has been annotated following the criteria outlined in Tyler & Evans (2001; 2003). This annotated data set serves as a 'theory-specific standard' against which the output of BERT will be assessed. This assessment is targeted at determining whether one can distinguish the same sense categories as proposed in the principled polysemy model by means of BERT embeddings, but it also addresses the question whether there is overlap between the operational models in terms of the suggested similarities and relations between these senses.

What emerges from the analysis is that BERT clearly captures fine-grained, local semantic similarities between tokens. Even with an entirely unsupervised application of BERT, discrete, coherent token groupings can be discerned that correspond relatively well with the sense categories proposed by means of the principled polysemy model. Furthermore, embeddings of *over* also clearly encode information about conceptual domains, as concrete, spatial uses of *over* are neatly distinguished from more abstract, metaphorical extensions (into the conceptual domain of time, or other non-spatial domains). However, there are no indications that BERT embeddings also encode information about the abstract configurational resemblances between tokens across those domains. As such, the global picture of resemblance between sense categories that emerges from the unsupervised application of BERT differs substantially from the theoretical proposal by Tyler & Evans (2003; 2001), which heavily relies on the language user's ability to recognize schematic, imagistic similarities within and across conceptual domains. These findings highlight the fact that such imagistic similarities are not captured by the embeddings of *over*, which provides further insight into the kind of semantic information that can be encoded by means of (unsupervised) BERT embeddings. This can provide an interesting basis for further experimental research (e.g. testing to what extent these different operational models of meaning representation are complementary when assessed against

ellicited behavioural data), as well as a discussion on how we can bring about a "greater cross-fertilization of theoretical and computational approaches" to the study of meaning (Boleda 2020: 2; also see, e.g., Baroni & Lenci 2011; Pater 2019).

2 Background

2.1 Cognitive-conceptual approaches to prepositional semantics

The interest in prepositional semantics in Cognitive Linguistics stems from the observation that language users are able to use a relatively small set of prepositions refer to an indefinitely large number of relations and scenes because of their cognitive ability to categorize concepts schematically (Kreitzer 1997). The cognitive-conceptual approach to prepositional semantics relies on two important constructs: the so-called "embodiment" of meaning, i.e. "[t]he idea that the properties of certain categories are a consequence of the nature of human biological capacities and of the experience of functioning in a physical and social environment" (Lakoff 1987: 12), and the notion of image schemas, which can be considered as condensed, schematic, recurring patterns of perceptual experience (e.g. Oakley 2010; Gibbs et al. 1994).

Some key publications in developing the notion of image schemas and embodiment, and integrating those notions into the discussion of meaning representation, are Brugman (1988) and Lakoff (1987). In their analyses of *over*, Brugman and Lakoff distinguish a vast number of distinct image schemas, all of which map onto a distinct 'sense' of the preposition. The image schema underlying an example such as *Devi lives over the hill*, for instance, conveys a static horizontal spatial configuration, whereby the focal point or "trajector" (TR), Devi, is positioned on the other side of the "landmark" (LM), the hill. This schema is different from the one underlying an example such as *The helicopter hung over the hill*, which conveys a vertical spatial configuration in which the helicopter (TR) is positioned above the hill (LM). The schema furthermore differs from those underlying examples such as *Devi walks over the hill* or *The helicopter flies over (the hill)*, which involve a (horizontal) path, and so on.

Yet, while they evoke different spatial configurations, the image schemas underlying these examples are still connected to one another, as humans are able to recognize general similarities between abstract image schemas (as demonstrated experimentally by, for instance, Gibbs et al. (1994)). Thus, a complex yet structured 'network' of linked senses is formed. Such networks, often termed 'lexical networks' or 'polysemy networks', comprise of nodes which are situated at varying distances from one another, and are centered around a primary sense or prototype (Lakoff 1987; Rice 1996). At their core, prepositions are spatial expressions, but the general human ability to apply metaphorical and analogical reasoning allows them to extend the use of prepositions to embody the non-physical domain ubiquitously (Kreitzer 1997: 317; Lee 1998: 334; Rice 1996: 135; Rice 1999: 227), as demonstrated by examples such as *Devi works over the weekend* (embodiment of time) and *Devi is over her ex-boyfriend* (embodiment of mental state).

As each small modification to an image schema is mapped onto a discrete sense category, the meticulous and comprehensive accounts set out by Brugman and Lakoff are sometimes called the "full-specification" approach. In the case of Lakoff (1987), the full-specification approach led to a fine-grained overview of 24 senses of *over*, which are connected in a sizable polysemy network. In later work, Lakoff's proposal was criticized amply for the fact that it leads to a virtually unconstrained number of sense categories, and for lacking methodological rigour (e.g. Rice 1996; Kreitzer 1997; Sandra & Rice 1995; Tyler & Evans 2001; 2003). This led Sandra (1998: 361) to coin the term "polysemy fallacy" in reference to "the tendency to look for polysemy even when there is no evidence for it".

Subsequent proposals, then, set out ways to tackle the apparent lack of a principled procedure to determine the number of distinct (sub)senses. Two notable examples are the proposal of Kreitzer (1997), and the "principled polysemy" approach advocated by Tyler & Evans (2001; 2003). Drawing strongly on the spatial information encoded in linguistic expression, Kreitzer (1997: 308) defines a prepositional sense as "a class of uses sharing a unique relational level image schema". In the case of *over*, Kreitzer argues that only three such schemata can be distinguished: (1) a static relation between two points on a vertical axis ($over_1$), (2) a dynamic relation involving a path schema ($over_2$), and (3) a static relation where one point occludes the other ($over_3$):

(1) The painting hung over the fireplace.

(2) The cat jumped over the fence.

(3) The mask is over my face.

These three relational schemata are also applicable to non-spatial domains, which, Kreitzer explains, are consistently conceptualized in terms of spatial image schemata: the use of *over* in *I finally got over that relationship* (indicating a path obstructed by an obstacle), for instance, can be motivated by the dynamic schema underlying $over_2$, whereas *over* in *The box is over six feet tall* (indicating excess) is motivated by $over_1$. Yet, as pointed out by Tyler & Evans (2001: 729), Kreitzer does not motivate the existence of those three relational image schemas in light of each other, as he "makes no attempt to account for how $over_1$ could give rise to $over_2$ and $over_3$ respectively". Additionally, many senses touched on by Lakoff (1987) are simply ignored in Kreitzer's account.

Addressing these issues, Tyler & Evans (2001; 2003) devised a more encompassing proposal based on slightly different principles. More specifically, they argue that senses can be considered distinct if (and only if) under the following criteria: (i) First, assuming that the primary sense of the preposition involves "a particular spatial relation between a TR and an LM", the distinct sense "must involve a meaning that is not purely spatial in nature", or "the spatial configuration between the TR and LM is changed vis-a-vis the other senses associated with a particular preposition" (Tyler & Evans 2001: 731). (ii) Second, the sense must exist in examples where it cannot be inferred from the combination of another sense and encyclopedic or contextual knowledge (i.e., they must be instantiated in semantic memory; Evans (2005)). Following these criteria, there is no reason to assume that the use of *over* involves two distinct and separately stored senses in *The helicopter hovered over the hill* and *The helicopter flew over the hill*, as both examples involve a spatial relation in which the TR (*the helicopter*) is located above the LM (*the hill*). Furthermore, the difference in stativity/dynamicity of the scene can simply be inferred from the lexical verb (*hover* vs. *fly*). As such, Tyler and Evans effectively constrain the full-specification network to a more digestible size.

An issue that remains, however, is that there is still no objective, measurable means of determining the global structure of the polysemy network. Besides further attempts to establish which sense constitutes the core node or prototype (Sandra & Rice 1995; Rice 1996; Gilquin & McMichael 2018; Newman 2011: 137), there is still much room for discussion regarding the position of derived nodes. To illustrate the issue, we can consider the multitude of possible derivation pathways of the repetitive sense of *over*, as in *She sang the same song over (and over)* (Tyler & Evans 2003: 105–106). First, based on their comparable, cyclical image schemas, the repetitive sense can be connected to reflexive uses of *over* (e.g. *She turned the page over/ The vase tipped over*). Second, it is possible that repetitive *over* marks an iterative trajectory, in which case the sense could be derived from cases where *over* marks the end of a linear temporal trajectory or process (e.g. *The race is over*). A third possibility is that the repetitive sense constitutes a conceptual blend of reflexivity and trajectory completion, a notion which may equally apply to many other derived senses.

In their accounts, Tyler & Evans choose to remain agnostic on the matter, explaining that "language does not function like a logical calculus which would allow us to … establish absolutely a single, precise derivation for each sense" (Tyler & Evans 2003: 62). This is not to say that 'anything goes', but rather that there is a delimited set of general principles or paths of derivation which may individually or simultaneously give rise to derived senses, and different individuals may draw different connections between senses, if they draw any such connections at all (Langacker 2010: Ch.10). Yet, even so, the agnostic position is somewhat unsatisfactory if one is interested in, for instance, comparing the general probability of multiple derivational paths across individuals, or even across time. Such queries will remain difficult to address in absence of methods that enable researchers to quantify and measure the degree of similarity between sense categories. The further integration of distributional semantic models into Cognitive Linguistics, then, can at least partially be linked to the research community's growing desire to approach the study of polysemy (and synonymy) in a more rigidly corpus-driven and measurable way (Gries & Divjak 2009; Newman 2011).

2.2 Advances in Distributional Semantic Models

At its core, the distributional approach conceptualizes the meaning of a word (or, more generally, of constructions) as a function of its lexical and grammatical context, and as such, meaning can be approached statistically (Turney & Pantel 2010; Boleda 2020). Statistical approaches to meaning have a long tradition in corpus linguistics, with functional-semantic classifications into distinct usages being increasingly based on explicit, automatically detectable contextual cues as corpora grew increasingly large.

An interesting observation made by Heylen et al. (2015) concerns the statistical-manual hybridity of the corpus linguistic tradition. As an example, they take the "British tradition in corpus linguistics", in which lexical collocations and syntactic patterns are employed to capture or approximate word meaning, while the classification of these meanings into categories is conducted manually. By contrast, the more recent application of "Behavioral Profiles" (Gries 2006; Gries & Divjak 2009), for instance, presents a means of statistically automating the classification by means of hierarchical cluster analysis (or correspondence analysis in Glynn (2010)). In such cases, a set of tokens can be annotated along a number of variables or dimensions, such as the type of trajector (TR), landmark (LM), as illustrated in ***Table 1*** from Newman (2011).

Table 1 Example data set adapted from Newman (2011: Table 4).

context variables of 'over'	dynamicity	*TR*	TR_concrete	TR_animate	LM	...
over_1	dynamic	PERSON	concrete	animate	PLACE	...
over_2	stative	THING	concrete	non-animate	PLACE	...
over_3	dynamic	EVENT	abstract	non-animate	TIME	...
...	...	...	...	...	...	...

Such data frames can subsequently be converted into a table with numeric infomation (e.g. the relative frequency of each example with each label), which can then be used as input for statistical analysis (Gries & Otani 2010). Focusing only on Trajector-Landmark combinations found in the ICE-GB corpus, Newman (2011) ultimately identifies seven statistically distinct uses of *over* by means of a Hierarchical Configural Frequency Analysis (HCFA):

(4) [AMOUNT] *over* AMOUNT: …over 700 farms still cannot sell their meat for human consumption

(5) EVENT *over* TIME: the blood pressure <unclear-words> at such a level after repeat measurements over a considerable period of time sometimes as long as six months

(6) THING *over* PLACE: a minute on each side on high and then 5 minutes over a low flame will do it

(7) PSYCH-STATE *over* STIMULUS: In view of the furore over the transmission of news from the Falklands

(8) STATE *over* DEPENDENT ENTITY: Abortion is the right of a woman over her own body

(9) COMMUNICATION *over* INSTRUMENT: When digital data are transmitted over a single parallel interface there is no crosstalk between the codes

(10) ATTRIBUTE *over* STANDARD ITEM: It offered many advantages over other systems including rapid action

The appeal of this approach, according to Newman (2011: 541–542), is that it "offers a systematic corpus-based procedure", which is "strongly grounded in facts of usage, complementing any other (intuition-based or experimentally based) methods the researcher might employ".

Still, the selection and annotation of the variables (which have been selected and defined by the analyst) is predominantly manual. As a "logical extension of the statistical state-of-art"

(Heylen et al. 2015: 154), then, Semantic Vector Space Models were introduced, in which all aspects of semantic analysis are approached statistically. In such models, the contextual properties that are fed into statistical classification models are no longer manually annotated features, but automatically generated numeric representations of syntactic and lexical co-occurence patterns.

Because the distributional approach to meaning is based on a relatively simple, and concrete premise, one may be tempted to assume that studies adopting this approach are highly comparable, if not identical in how they operationalize and model meaning. This would, however, be a mistaken assumption. It would be far beyond the scope of the present paper to survey the many different ways in which 'meaning as context' has been operationalized (for such a survey, one may consult Turney & Pantel (2010), Lenci (2018), Boleda (2020), or specifically for deep learning based models, Young et al. (2018)). However, to clarify the model choice in the present study, a brief discussion of two relatively recent developments is warranted. This concerns (i) the rise of models operating with contextualized (or, rather, token-based) semantic vectors, and (ii) the rise of context-predicting models (also known as 'neural language models') that create semantic vectors often referred to as 'embeddings'.

2.2.1 Semantic vectors: type vs. token

A first development of note is the gradual turn from models that produce vectors of word types, to models that are able to create token-based (or 'contextualized') vectors. The distinction between type-based and token-based is not so much one of whether or not the resulting vector representations include contextual information – this is the case for both type-level as well as token-level vectors – but whether or not all contextual occurences of a single word are conflated into a single vector representation.

Type-based models work from the assumption that a word has a single, constant, 'core' meaning (which can be understood as a prototype, cf. Erk & Padó (2010: 92)), thus representing a 'lumped' approach to meaning representation. Given a number of examples involving the words *cat*, *mouse*, a type-based model will provide a single numeric representation for all of the context in which these words occur (see ***Table 2***).

(11) The **cat** ate some food and purred.

(12) Do not pet the paws of a **cat** unless it purrs.

(13) The **mouse** held some food between its paws.

(14) I bought an external **mouse** and keyboard for my computer.

TYPE representation	*food*	*purr*	*paws*	*keyboard*	*computer*	...
cat	1	2	1	0	0	...
mouse	1	0	1	1	1	...

Table 2 Example of contextual input in (based on lexical co-occurrence frequencies) for cat and mouse using a word type representation.

For a word such as *mouse*, for instance, the contextual information that suggests it is an animal would therefore be conflated with contextual information typical of the object. The problem with such aggregated vector representations is that they may render unsatisfactory or problematic vector representations in cases of polysemy, and, unarguably even more so, in cases of homonymy (Erk & Padó 2010; Desagulier 2019; De Pascale 2019; 'meaning conflation deficiency' in Camacho-Collados & Pilehvar).

In response to this issue, models that generate token-specific vector representations were developed. These token-based distributional models – in which individual vectors are assigned to, for instance, the two different examples of *mouse* as in ***Table 3*** – are better equipped to handle the complex internal semantic structure of words, and, hence, are naturally better suited for specific NLP tasks such as word sense disambiguation (see, e.g. ELMo (Peters et al. 2018a), as well as the model described in, e.g., Heylen et al. (2015)). Because the aim of the present study

is precisely to home in on the differences and similarities between different uses of a single preposition, it evidently employs a distributional model that produces vector presentations at the token level.

TOKEN representation	*food*	*purr*	*paws*	*keyboard*	*computer*	...
cat_1	1	1	0	0	0	...
cat_2	0	1	1	0	0	...
mouse_1	1	0	1	0	0	...
mouse_2	0	0	0	1	1	

Table 3 Example of contextual input in (based on lexical co-occurrence frequencies) for *cat* and *mouse* using a word token representation.

2.2.2 Semantic vectors: count vs. predict

Using the terminology employed in Baroni et al. (2014), I wish to point out that a distinction can be made between 'count models', and 'predict(ive) models' (also see 'explicit' and 'implicit' models in Dubossarsky et al. (2017: 1136)). Count models represent, in a sense, the most straightforward way of operationalizing the distributional hypothesis, in that they make use of numerical vectors that are essentially based on co-occurrence counts (for an accessible explanation of how such vectors are constructed for word types and word tokens, see, for instance, Heylen et al. (2015) and Hilpert & Correia Saavedra (2017)). Still, describing count models as such is a severe simplification, as more than often the vectors are optimized in some way (e.g. by changing context window sizes, reweighting function words, leaving out function words, applying dimensionality reduction, etc.).

By contrast, context-predicting models (yet again a cover-term for an extremely varied group of models, including weighted bag-of-words, to more syntactically informed variations, with new types of model architectures being added continuously) are designed to approach the construction of semantic vectors from a training-based angle: "Instead of first collecting context vectors and then reweighting these vectors based on various criteria, the vector weights are directly set to optimally predict the contexts in which the corresponding words tend to appear" (Baroni et al. 2014: 238). In other words, predictive models construct vectors as part of a learning task, which, to some degree, eliminates the vector transformation and optimization process. This, in addition to the performance improvements observed in a range of NLP tasks compared to count models, is why the relatively recent emergence predictive models is often portrayed as an attractive advancement (Baroni et al. 2014). This is, however, not to say that predictive models involve absolutely no parameter tuning – and it has been suggested that, given comparable settings and tuning, the vectors created with count models are as effective as the embeddings yielded by predictive models (Levy et al. 2015). Yet, what does make predictive neural language models particularly appealing for the present study, which focuses on prepositional semantics, is that count models are generally less successful in providing useful representations of function words (e.g. Bullinaria & Levy 2012: 7), whereas "recent neural network models do provide usable representations for them" (Boleda 2020: 7).

2.3 BERT

In the present study, the distributional approach is represented by a single model architecture: Devlin et al. (2019)'s Bidirectional Encoder Representations from Transformers (BERT). First launched in November 2018, BERT quickly became the model to beat due to its impressive performance on a wide range of NLP tasks. Soon after, it also grabbed the attention of computational linguists, who were interested in determining precisely what kind of linguistic information such models acquire and capture (Clark et al. 2019; Jawahar et al. 2019; Alishahi et al. 2019).

In a nutshell, BERT is a deep contextualized model based on a particular type of neural architecture, called "the Transformer", which is entirely based on so-called "attention mechanisms" (Vaswani et al. 2017). A context-predicting model, BERT has been pre-trained on approximately 3.3 billion words (800 million words taken from the BooksCorpus, and 2.5 billion words from English Wikipedia) of unlabelled data over a masked word prediction task

(in which the objective is to predict randomly masked input tokens based only on the context in which they occur) and a next sentence prediction task (so that the model will also understand sentence relationships).

Like other Transformers, BERT consists of multiple layers (or 'transformer blocks'), all of which contain multiple self-attention heads which behave similarly within their layer. The smallest pre-trained model, called $\text{BERT}_{\text{base}}$, consists of 12 layers with 12 attention heads, whereas the larger model, called $\text{BERT}_{\text{large}}$, consists of 24 layers with 16 attention heads. Each of these layers captures the *n* tokens in the input sentence (or rather 'sequence', as the input need not correspond with what linguists have traditionally defined as a sentence) in compressed numerical vector representations or 'embeddings'.

The attention heads within BERT's layers have been probed for the linguistic phenomena they capture. This revealed that particular heads capture syntactic relations (e.g. valency patterns and dependency relations), while others perform well at coreference resolution (Clark et al. 2019) – which is remarkable given that the model has not received any explicit input about syntax or coreference. This "syntax-aware attention" (Clark et al. 2019) may be why BERT is succesful the downstream NLP tasks it has been employed in (cf. Peters et al. 2018b). Finally, it is important to note that the different layers (and accompanying attention heads) perform slightly differently on different tasks. In various sources, the second-to-last layer (or a concatenation of the last four layers) is suggested to perform best on token-level tasks such as word sense disambiguation (e.g. Devlin et al. 2019; Wiedemann et al. 2019), but many applications also operate with the final hidden layer (e.g. Huang et al. 2019; Blevins & Zettlemoyer 2020).

With respect to linguistic investigation into polysemy and sense disambiguation, the contextualized embeddings produced by BERT have thus far not been explored. One reason may be that neural models have grown into increasingly intransparant systems (Linzen et al. 2019: iii), making linguists more reluctant to rely on them for lingusitic analysis. Still, it is worth investigating to what extent they could be employed as analytic tools in linguistic research, as neural language models like (but consistently outperformed by) BERT have already been shown to capture very nuanced aspects of meaning, and they even seem to provide usable representations for function words (see Boleda (2020: 18), in reference to Peters et al. (2018a)). Furthermore, a model such as BERT also unites the strengths of different types of token-based distributional methods. First, the fact that the model is syntax-aware agrees with the cognitive-linguistic (and constructionist) view that differences in syntactic structures reflect differences in meaning (Langacker 1991; Goldberg 1995). As such, its syntax-awareness sets BERT apart from bag-of-words approaches to contextualized vectors (e.g. Heylen et al. 2015), and thus makes it more akin to, for example, the Behavioural Profiles approach. Second, the application of BERT to the study of polysemy does not involve any manual annotation, and neither does it involve making an a priori selection of syntactic features to be included, making it a fully data-driven approach to the question at hand.

3 Data and Methodology

In the present study, $\text{BERT}_{\text{base}}$ has been used to create contextualized embeddings for all occurrences of *over* in the final decade of the Corpus of Historical American English (COHA, 2000–2010). In total, embeddings were created for 39,834 tokens of *over* using the Spacy implemenation of $\text{BERT}_{\text{base}}$ (which, at the time the analysis was conducted, only offered access to the final hidden layer). The embeddings of the target tokens were created with a context window set to 20 words preceding and 20 words following *over*. In principle, the performance of the model in the task at hand could still be improved by experimenting with different hyperparameter settings or by fine-tuning the model to specific tasks or corpus data, but no such operations were undertaken.

Of the 39,834 tokens, 808 examples were manually annotated by two human annotators, following the sense description in Tyler & Evans (2001; 2003). Note that the proposal by Tyler & Evans does not consitute a 'gold standard', as its cognitive reality remains to be tested against ellicited, experimental data. Yet, their proposal was chosen as a point of comparison because (i) it is well-documented, (ii) is firmly grounded in and motivated by linguistic theory, (iii) and presents the most comprehensive assessment of all possible senses of *over* since Lakoff (1987). Furthermore, because their proposal focusses not only on motivating the number of distinct

senses, but also on motivating the connections between those senses by foregrounding the importance of image schemas, it lends itself well to an assessment of whether such imagistic information is encoded in corpus data and captured by word embeddings.

In total, 16 different sense categories were distinguished. Following the example of Tyler & Evans (2001), the categories are given a label that corresponds with their status as a discrete sense (i.e., 1, 3, etc.) and subsense (i.e., A, B, etc.). Note that the 808 examples constitute a stratified sample: first, a random sample of 300 tokens was manually annotated. Subsequently, the sample was expanded with further examples until each sense category was represented by at least 10 tokens. The token frequencies per sense category are listed in ***Table 4***.

Table 4 Token Frequencies per category.

Category	Tokens	Category	Tokens
1. Protoscene 'above'	152	*4A. Focus-of-attention*	62
2A. On-the-other-side	122	*5A. More*	39
2B. Excess 'beyond'	24	*5B. Control*	50
2C. Completion	37	*5C. Preference*	33
2D. Transfer	35	*6. Reflexive*	31
2E. Time span	50	*6A. Repetition*	32
3. Covering	49	*7. Communication line*	23
4. Examining	32	*8. Hangover*	11
		Indeterminate	26

Between the two human annotators, inter-rater agreement was found to be very good (Fleiss' Kappa = 0.867). In the statistical analyses presented below, 26 were examples were excluded because they were considered indeterminate between multiple categories. Further information on the sense categories is provided in Section 3.1.

Ultimately, the sense categorization proposed by Tyler and Evans was created in response to models that are too fine-grained, and hence lack what Tyler & Evans consider to be meaningful, principled abstractions. Thus, the question we are in fact asking is to what extent these abstractions are also 'meaningful' to models such as BERT, which approach prepositional meaning by compressing contextual data. To address this question, I adopt a procedure based on the Varying Abstraction Model (Vanpaemel & Storms 2008). Originally, the Varying Abstraction Model (VAM) was designed in response to the debate in psychology on how the classification accuracy of exemplar-based models (involving no abstraction) compares to that of prototype models (involving complete abstraction) as well as models involving intermediate levels of abstraction. The procedure adopted here is based on the *k*-means variant of the VAM (Verbeemen et al. 2005), where a particular level of abstraction is operationalized by the degree to which category members are clustered.

The VAM conducts a series of evaluation tasks, where it predicts the category label of unseen test tokens against a manually assigned label. The series start with the prediction of the category label of an unseen test token based on its nearest neighbour embedding in a labelled training set. At this level, none of the token embeddings in the training set have been clustered, which, one could argue, means that the number of 'clusters' k equals the number of tokens n in the training set ($k = n$). This is also called the 'exemplar level'. In other word sense disambiguation studies, the performance of models such as BERT is commonly assessed solely at this level (see, e.g. Wiedemann et al. 2019). In the present study, however, the assessment is also taken beyond the exemplar level: the VAM will subsequently attempt the same classification task again, but instead of using all the embeddings of the concrete tokens in the training set as a reference set, it will create a slightly higher level of 'abstraction' by merging the embeddings of some concrete tokens of the same sense category into a slightly more schematic, averaged representation. At every step of the VAM procedure, an increasing number of token embeddings are merged, until all embeddings of all training tokens that belong the same sense category are merged into a single averaged embedding. A schematic representation of the different steps of abstraction is presented in ***Figure 1***.

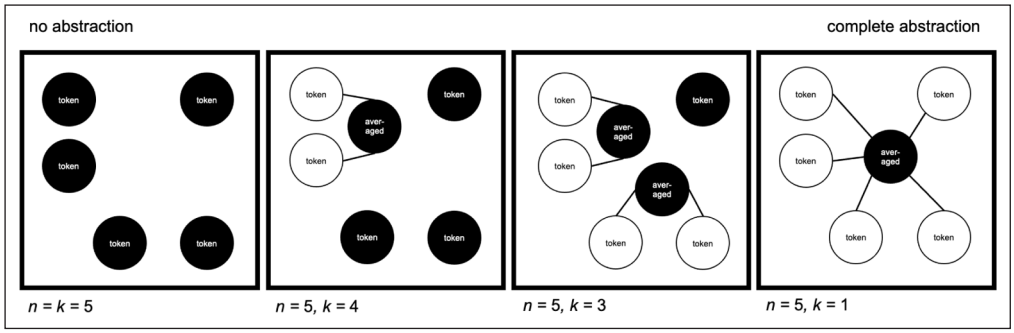

Figure 1 Schematic representation of abstraction continuum. At the lowest level of abstraction, the items in the training set that can be used for classifiying an unseen token are the embeddings of concrete tokens in that set. At intermediate levels, the number of items that can be used to classify an unseen token is gradually reduced, as an increasing number of token embeddings are merged into an averaged embedding. When complete abstraction is reached, all items of the same category are merged into a single, averaged 'sense embedding'.

At the highest level of abstraction, then, the classification task of the unseen test tokens is attempted by means of a clustered or averaged representation of all embeddings assigned to that category. This averaged embedding can hence be thought of as a 'contextualized sense embedding'. Because I consider the 16 sense categories of the principled polysemy model (described in Section 3.1) to represent the highest level of abstraction, the number of clustered representations by means of which classification is attempted at this level is 16 ($k = 16$). Note that the level of abstraction could be increased further by reducing the number of clusters to 8 (to attempt the classification task using clustered representations of, for instance all tokens labelled as sense 5A, 5B and 5C), but no such actions were undertaken in this study.

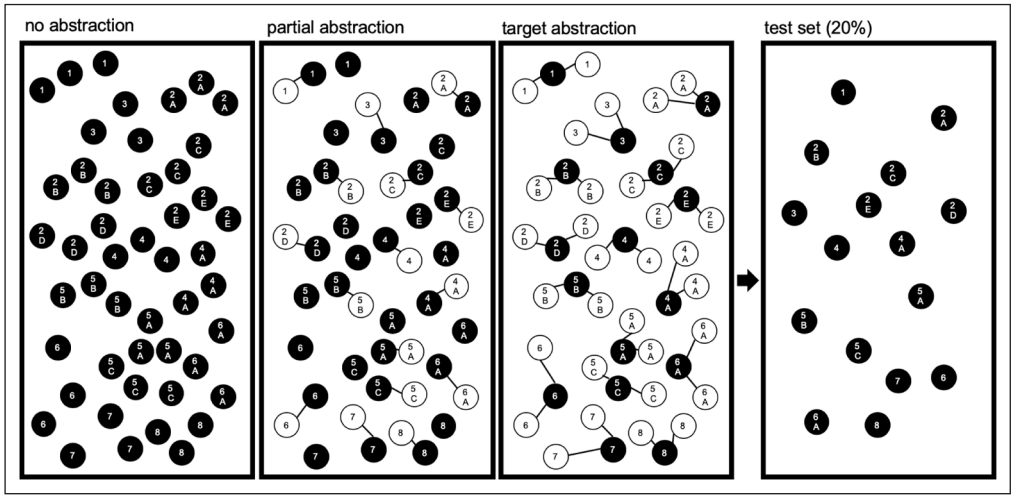

Figure 2 Schematic representation of VAM procedure. Starting from no clustered items, the VAM uses different configurations of clustered or averaged embeddings (which represent different levels of abstraction) in a training set (80% of the data) of the data to classify an unseen test token set (20% of the data).

The series of classification tasks (from $k = n$ to $k = 16$) is evaluated against a test set (20% of the data; 100 iterations per level; see ***Figure 2***), and will be expressed in an accuracy score (F_1-score, between 0 and 1, with 1 representing perfect accuracy). The resulting series of classification accuracy scores allows us to assess the following: if the sense classification task goes well at the lowest level of abstraction (the exemplar level), we find that the contextual information encoded in the BERT embeddings of *over* encodes and captures local similarities between concrete tokens of the same sense category. As the level of abstraction increases, the classification task will involve classifiying unseen tokens not by means of other, concrete tokens, but by means of averaged contextual representations of multiple tokens that have been assigned the same label. In other words, the model will attempt the classification of unseen tokens by means of contextual representations that are decreasingly concrete and increasingly schematic (that is, representing abstractions over multiple tokens in the same sense category). If classification accuracy of the unseen test tokens remains high when all training tokens of the same category are averaged into a 'sense embedding', we find that these abstract contextual representations are helpful tools to categorize new, unseen tokens. In that case, we could say that the 'meaningful abstractions' or sense categories proposed by Tyler and Evans also make sense in terms of the contextual information encoded in BERT embeddings.

Besides assessing to what extent BERT embeddings can be used to distinguish the sense categories proposed in the principled polysemy model, I will also discuss the global structure of

the network proposed by Tyler & Evans (2001). To discuss distances between sense categories, I use the cosine similarity between the embeddings (see, e.g., Bullinaria & Levy (2007); Heylen et al. (2015); Peters et al. (2018b)).

3.1 Sense categories

In what follows, I will describe the categories distinguished in Tyler & Evans, illustrating them with examples from the data set. For an in-depth description of the sense categories and a full argumentation as to why these (and only these) categories have been distinguished, I refer to Tyler & Evans (2001; 2003).

3.1.1 Sense 1: 'above'

The first category contains all examples in which *over* signals that the TR is located above the LM. This category is considered to be the primary sense or 'protoscene' from which all other senses can be derived (Tyler & Evans 2001: 735–737). The relation expressed is an atemporal, spatial relation, where the TR is typically in close proximity to the LM. In many cases, TR is typically movable and smaller than the LM (as in (15)), but immovable (e.g. (16)) and larger (e.g. (17)) TRs occur as well.

(15) I noticed a painting hanging **over** the piano (COHA, 2006)

(16) He was bleeding from a cut **over** his eye (COHA, 2003)

(17) And, so that's how I got into the apartment **over** the garage. (COHA, 2005)

3.1.2 Sense (group) 2: A-B-C trajectory

Besides the protoscene, Tyler & Evans also distinguish a number of derived senses. Four of these can be conceived of as a 'cluster' of senses where *over* marks a trajectory from a starting point (A), a midpoint (B), and an endpoint (C). While not all senses in this cluster put equal focus on all points in the trajectory, the uniting factor seems to be that there is a certain linearity to the expressed relation.

ON-THE-OTHER-SIDE-OF (2A) In examples (18) and (19), the TR is portrayed as being not above, but on the other side of the LM:

(18) I'd grown up only a few hours away, **over** the Kentucky line. (COHA, 2007)

(19) God, let this be the peak. Let us be **over** the mountain (COHA, 2007)

Note that the verb itself does not trigger the trajectory reading (when it does, the example will be assigned to the protoscene).

While the examples in (18) and (19) both function as prepositions, a large group of tokens in this category function as an adprep. In some cases, such as (20), the verb is combined with an adverbial phrase that indicates the endpoint of the trajectory. Thus, it could be argued that the combination of the verb and the endpoint adverbial already implies movement along a trajectory. The addition of *over*, then, seems to have a mere emphatic function. However, in other examples, such as (21), the adprep is non-optional if a trajectory is to be evoked. All adprep trajectory uses are assigned to category 2A.

(20) After he left us, he drove **over** to my brother Jacob 's apartment (COHA, 2006)

(21) Smiling, she hurries **over** (COHA, 2004)

Examples such as (22), where a look is thrown at an explicit endpoint, were also assigned to category 2A:

(22) He looked **over** at the computer. (2007, COHA)

Finally, a number of examples assigned to this category do not refer to a spatial relation. If we consider examples such as (23), where the LM represents an obstacle or hurdle, one

can metaphorically extend the use of *over* to non-physical obstacles (often relating to past relationships), as in (24) and (25):

(23) … that old and painful relationship. But Mike had seemed okay with it, as if he was completely **over** Lindsey (COHA, 2009).

(24) I had a thing with her a bunch of years ago, and I guess I never got **over** the attraction. (COHA, 2007)

(25) Memphis had gotten **over** her steering problems. (COHA, 2004)

'EXCESS': ABOVE-AND-BEYOND (2B) In this category, we find cases where the TR moves above the LM, and thus misses or exceeds a point it should not have crossed. What is key about Sense 2B, and what distinguishes it from Sense 1 and Sense 2A, is the implicature that "the LM represents an intended goal or target and that the TR moved beyond the intended or desired point" (Tyler & Evans 2001: 749), as in (26):

(26) Your article is **over** the page limit. (Tyler & Evans 2001: 749)

With such an example, it is difficult to say whether interpretation of excess is entirely 'context-free' and not evoked or supported by the lexeme *limit*. Similarly, the 'excess' implicature in examples (27) and (28) may be triggered by *fault* and *illegally*. Still, the decision to include a separate category of Sense 2B was maintained.

(27) … seldom-called violations in tennis – the foot fault. It occurs when a player's foot brushes or goes **over** the baseline when serving. (COHA, 2006)

(28) He's illegally parked. His ass is **over** the white line. (2003, COHA)

Furthermore included in this category are cases such as (29), which portray a situation where the 'missed target' is a person in line for a reward (usually in the form of a job offer or promotion, as in (30)). The implicature here is that the reward was expected or deserved, but those expectations were not fulfilled.

(29) … his monumental 1957 paper on the origins of elements, for which – to his annoyance – he was passed **over** for a nobel prize. (COHA, 2001)

(30) … a lot of times he'll pass **over** the most talented and put someone in with the biggest heart. (COHA, 2003)

COMPLETION (2C) This category contains examples such as (31) and (32), where *over* indicates that something is finished or completed.

(31) My school days are finally officially **over**. (2007, COHA)

(32) All the decisions had been made, the story was **over**. (2006, COHA)

This category solely contains examples where *over* functions as an adprep, and consistently combines with the verb *be*.

TRANSFER (2D) Another category that solely includes adpreps is category 2D, which includes examples such as (33):

(33) The woodsman reached in his pocket, pulled out the thirty euros, and handed the two bills **over** to the new man. (2003, COHA)

The only difference between these examples and examples of 2A is that some sort of transfer has taken place. However, one could argue that such a transaction is encoded by the verb used in the same construction. Consider, for instance, the example in (34):

(34) The clerk lifted the bill from Peterson's hand and took it **over** to the second clerk sitting at the desk (COHA, 2003).

Here, an object is indeed moved by a clerk to another clerk, but there is no explicit indication that the object was given to the second clerk. While different in lexical material, the example in (34) is structurally identical Tyler & Evans' examples of transfer (e.g. *The teller handed the money over to the investigating officer*). The key element that triggers the meaning of transfer is therefore perhaps not *over*, but the verb *hand*, which complicates the suggestion that 'transfer' constitutes an encoded sense somewhat. Still, examples such as (33) were placed in a separate category 2D.

Finally, Non-physical transfers, as illustrated in (35) are also considered as instances of 2D:

(35) … the London office had grown considerably in the last eight years. Boyd wouldn't half mind taking **over** the running of it. (2008, COHA)

As non-physical transfers almost exclusively involve a transfer of control or authority, these examples are, at times, difficult to distinguish from examples of 5B (see below).

TIME SPAN (2E) In a fairly large number of examples, *over* "mediates a temporal relation of concurrence between a process or activity and the times during which the process or activity elapses" (Tyler & Evans 2001: 748–749), as illustrated in (36) and (37):

(36) The war on witchcraft intensified **over** the next 200 years, sending millions of cats, not to mention humans, to their deaths. (2001, COHA)

(37) Geologists and biologists before Darwin noted that the Earth and its inhabitants change **over** time. (2004, COHA)

Note that the question may be raised whether these examples do in fact constitute a distinct sense of *over*, as the temporal reading is inferable from the fact that the LM consistently involves a noun that refers to a time-related concept. The choice was made to create a separate label for these examples, but if the inferrability criterion is adhered to more strictly, these examples could perhaps be classified as instances of Sense 2A.

3.1.3 Sense 3: Covering

Like Lakoff, Tyler & Evans also distinguish a category of examples such as (38), in which there is "an understood viewpoint from which the TR is blocking accessibility of vision to at least some part of the LM" (Lakoff 1987: 429). In these cases, the TR is not located above the LM from the vantage point of the viewer (Tyler & Evans 2001: 752):

(38) A ratty leather jacket gaped open to reveal a white button-front shirt **over** an ample but not outrageous bosom (2005, COHA)

The 'covering' sense also includes a examples there is a "multiplex trajector" (Lakoff 1987: 428) that is scattered over the LM, as in (39), or where the TR has covered a path consisting of multiple points over the LM, as in (40):

(39) He searches through the papers scattered **over** the desk, but finds nothing (COHA, 2005)

(40) I can get into the crawlspace from my closet and climb all **over** the house. (COHA, 2005)

3.1.4 Sense (group) 4: Proximity

As in sense 3, the TR is no longer necessarily positioned above the LM from the perspective of the viewer. Instead, *over* conveys that there is close proximity between the TR and LM. This proximity goes beyond the spatial realm, and is manifested in the attention paid by the TR to the LM.

EXAMINING (4) The first category in Sense group 4 includes examples where the TR is examining the LM. The majority of cases involve the verb *look* (or near-synonyms such as *glance*), as in (41), but other verbs (e.g. *read, go*) occur as well:

(41) Chad got out and walked around the truck, looking it **over**. (2008, COHA)

(42) After studying my folder and going **over** the exact sequence of what to speak on, I allow myself the pleasure of flipping on the news (2004, COHA)

FOCUS-OF-ATTENTION (4A) In the majority of examples included in this category, the LM is the focus of the TRs attention. In these examples, *over* is equivalent to *about*, and in some cases, the LM can be considered the cause of the TRs actions:

(43) That debate, **over** how fast and how far to cut emissions, was the right battle to have (2004, COHA)

(44) Ben was pictured displaying great emotion by crying **over** her loss. (2001, COHA)

In discussing sense 4A, Tyler & Evans also mention examples such as (45):

(45) John Stewart presides **over** Comedy Central's The Daily Show, a blessed wedding of performer and format. (2001, COHA)

Note that, because the verb *preside* is used, it is also implied that the TR controls the LM. The same could also be said for examples such as (46), where the the notion of control or authority is not encoded in the lexical verb:

(46) Francis watched **over** the boy's education (2004, COHA)

Examples such as (45) and (46) were classified as instances of sense 4A, but it must be noted here that the distinction between these examples and examples of sense 5B, which are discussed below, is difficult to maintain.

3.1.5 Sense (group) 5: Up

Four further senses fall under "the *up* cluster", which are suggested to derive "from construing a TR located physically higher than the LM as being vertically elevated or up relative to the LM" (Tyler & Evans 2001: 755).

MORE (5A) The first (and most frequently occurring) sense is 5A. In all examples in this category, *over* indicates that a quantity is higher than the quantity expressed in the LM:

(47) Jerry and I were parents to **over** fifty foster kids in our thirty years of marriage. (2004, COHA)

Tyler & Evans distinguish one further sense, sense 5A.1, where the TR is understood as something that is contained by, but exceeds the capacity of, the LM. The only clear example discussed is *overtired*, in *The child was overtired and thus had difficulty falling asleep*. The data set in the present paper does not include compounds with *over*. In a footnote, Tyler & Evans (2001: 757) explain that it is often possible to "construct a 'more' conceptualization" alongside "an 'excess' interpretation". In practice, this seemed to apply to nearly all examples in the data set. As such, no distinction was made between sense 5A and 5A.1.

It is, in many cases, also extremely difficult to distinguish cases of 'excess' as crossing a target point, or excess as exceeding an amount or capacity. Consider, for instance, the example in (48):

(48) But for kids **over** age 5, as the portion size got larger, so did the amount they ate. (2003, COHA)

While suitable for the 'up' conceptualization, it is not inconceivable that examples such as (34) could also be classified under 2B (as time is a linear concept rather than a container). Tyler & Evans (2001: 758) also address this issue, stating that their network of senses "should be thought of as a semantic continuum, in which complex conceptualizations can draw on meanings from distinct nodes as well as the range of points between nodes, which provide nuanced semantic values". For simplicity's sake, the choice was made to assign all cases where a numeric threshold was exceeded to sense 5A.

CONTROL (5B) The *up*-cluster further includes a category for all cases where *over* is used to mark that the "TR exerts influence, or control over the LM" Tyler & Evans (2001: 758), as in (49) and (50):

(49) She was moved by her power **over** me. I would have fallen down for her any day. (2002, COHA)

(50) But Nolan has final say **over** all personnel decisions, including the draft. (2005, COHA)

As noted earlier, there is also a sense of 'control' in examples classified as 2D, where control is transferred from one party to another, and examples classified under 5B.

PREFERENCE (5C) The final group of examples in the *up*-cluster convey a preference of one option, the TR, over another, the LM. In many examples, the notion of preference is encoded by the verb (e.g. *prefer*), but this need not be the case, as shown in (51) and (52):

(51) We haven't switched to a local pediatrician, believing irrationally in Manhattan doctors **over** Brooklyn doctors. (2001, COHA)

(52) His name was Miguel Santiago, and he insisted on being called Miguel **over** Mike. (2005, COHA)

3.1.6 Sense (group) 6: Reflexivity and Repetition

Two more senses distinguished by Tyler & Evans are the reflexive use of **over**, as in (53), and the repetitive use, as in (54) and (55):

(53) Elaine pulls her leg back and kicks the grill. The coals fly up and out, the grill tips **over**. (2000, COHA)

(54) Sometimes even horrible memories play **over** and *over* in my mind (2003, COHA)

(55) I hope someday soon we can begin again … start **over**. (2002, COHA)

Examples such as (53) are categorized as Sense 6, whereas (54) and (55) are classified as Sense 6A. All instances of reflexive and repetitive *over* function as adpreps.

3.1.7 Other Senses

Besides the senses discerned by Tyler & Evans, a few further categories were distinguished.

INDETERMINATE First, when the precise sense of *over* in a given example was considered vague or ambiguous between multiple readings, the example was classified as 'indeterminate'. Consider, for instance, example (56), in which it is unclear whether the interpretation is reflexive ('she bent/tilted forward'), or whether a trajectory is implied ('she leaned over (to us)').

(56) She leaned **over** and talked with excitement. (2009, COHA)

Finally, 49 examples were not classifiable along the categories set out above. These examples seem to fall within two categories: 'means of communication' and 'hangover'.

MEANS OF COMMUNICATION The first category concerns examples such as (57) and (58), in which the TR (if expressed) is a person who uses a particular channel or means of communication (the LM). In these examples, *over* seems to be paraphrasable as 'by means of':

(57) You could break the news **over** the phone (COHA, 2005)

(58) In Napster's case the transfers took place **over** the internet (COHA, 2006)

HANGOVER The second category concerns cases where *over* is used in an idiomatic expression with *hung*, indicating the (unpleasant) after-effects of excessive substance abuse, as in (59). In these cases, *over* is an adprep.

(59) … that he had gotten drunk the night before and that he was still horribly hung **over**. (COHA, 2003)

4 Results

4.1 Sense Distinctions

As a first point of enquiry, it is investigated whether BERT indeed recognises the sense categories proposed in the principled polysemy model in a relatively distinct and coherent manner. Following Kilgarriff (2003: 108), I define 'senses' as "abstractions over clusters of word usages". In other words, if the abstract, conceptual sense categories proposed in the principled polysemy model are recognized by BERT, we would expect to find that the geometrical distance (operationalized as the cosine distance) between the embeddings of all tokens labelled as Sense 1, for instance, is shorter than the distance between those tokens and tokens with a different category label, thus forming a cluster.

To visualize the local embedding clusters and the global positioning of those clusters relative to one another, a two-dimensional representation of the token embeddings was created based on the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm for dimensionality reduction of high-dimensional data (van der Maaten & Hinton 2008). In ***Figure 3***, each token is represented by a dot, which has been coloured according to its manually assigned label. Note that the embeddings were created solely based on the contextual information surrounding *over*, and that the manual labels were assigned separately. The distributional model was hence not fed any human-defined knowledge about the number or nature of labelled sense categories. The two-dimensional plot below therefore visualizes the overall correspondence between clustered token embeddings, which can be conceptualized as 'distributionally defined senses', and the proposed 'conceptually defined' sense labels.

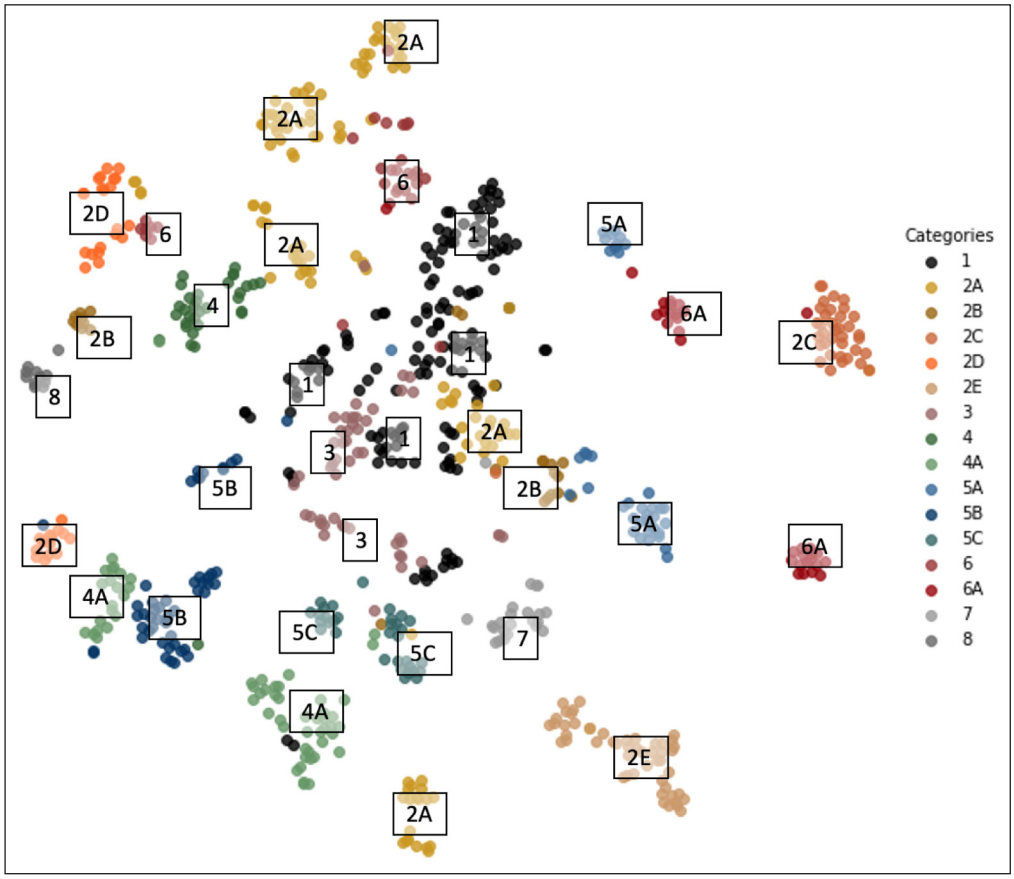

Figure 3 t-SNE embeddings of *over*, perplexity = 20, KL divergence after 1,000 iterations: 0.457.

From eyeballing ***Figure 3***, it looks like the distributional model proposes a fair number of distinct token-clusters that correspond relatively well with the suggested sense categories: in the majority of cases, the local clusters (or cluster areas) consist of tokens that were assigned the same label. At the bottom of ***Figure 3***, for instance, a local cluster area was marked, which comprises entirely of examples of Sense 2A (more specifically, those cases where the TR has mentally overcome an obstacle or past relationship). Yet, at the same time, the correspondence between the two models seems to become weaker at the global level. In some cases, such as Sense 2C ('completion'), 2E ('time span'), 4 ('examining'), 7 ('means of communication') and 8 ('hangover'), it appears that all tokens with the same label are assigned to the same local cluster

area, but in others, such as 2D ('transfer'), tokens are grouped in separate, relatively distant areas. Thus, the question arises how we can assess the degree to which there is correspondence between the distributionally defined and conceptually defined senses.

To address this question in a way that goes beyond eyeballing a visualization, this study uses a series of classification tasks, which help quantify the extent to which there is correspondence between the sense categories emerging from the two models (VAM). ***Figure 4*** presents the results of the VAM when applied to all categories (over 100 iterations). The x-axis represents the abstraction continuum, which starts at no abstraction (the exemplar level, where classification of unseen tokens (20% of the data) is attempted by means of the nearest neighbour embeddings of concrete tokens) and reaches up to the target level (the highest level of abstraction, where all items within the 16 labelled categories are clustered into averaged 'sense embeddings'). The y-axis represents the classification accuracy of the distributional model (F_1-score, i.e. the harmonic mean of precision and recall when applied to the test set).

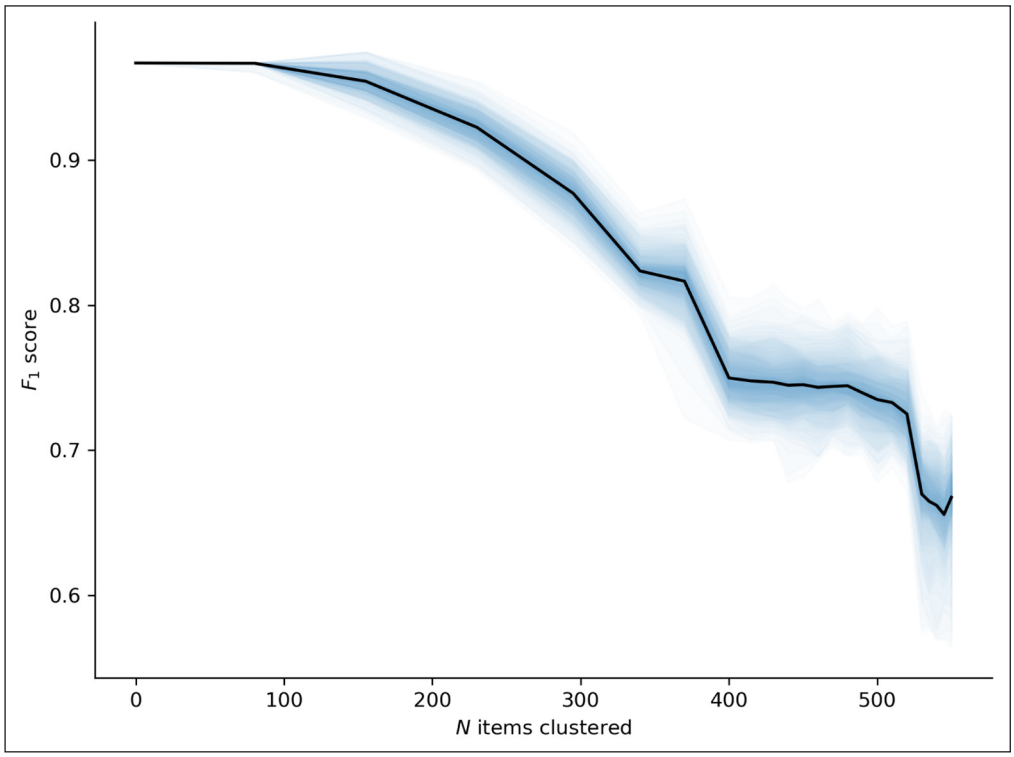

Figure 4 VAM output over all data.

At the lowest level of abstraction, the model's classification accuracy is quite high at 0.95, remaining relatively stable at this level until approximately 200 tokens have been clustered. Subsequently, its accuracy gradually drops to 0.8 (at approx. 300 tokens), after which it drops below 0.7 at the highest levels of abstraction. These findings imply that the BERT embeddings do encode the similarities between members of the categories proposed in the principled polysemy model, but only up to a certain point. Beyond that point, the proposed abstractions no longer optimally fit the output of the distributional model.

When we assess the models classification accuracy per sense category (***Figure 5***), we also find that the model is more successful in 'recognizing' some sense abstractions than others. Overall, the embeddings of *over* clearly encode the similarities between concrete tokens, and the performance of the model remains high at lower-intermediate levels of abstraction where only some of the concrete token embeddings are merged into slightly more schematic representations. Yet, whether higher levels of abstraction are still meaningfully encoded in the BERT embeddings of *over* seems to depend on the sense category under scrutiny. Unsurprisingly, perhaps, it is precisely the sense categories that occur in fixed syntactic configurations and have clear collocational preferences (e.g. Sense 2C 'completion', in which *over* consistently functions as an adprep in combination with a form of *BE*), are easier to group than more 'schematic' senses.

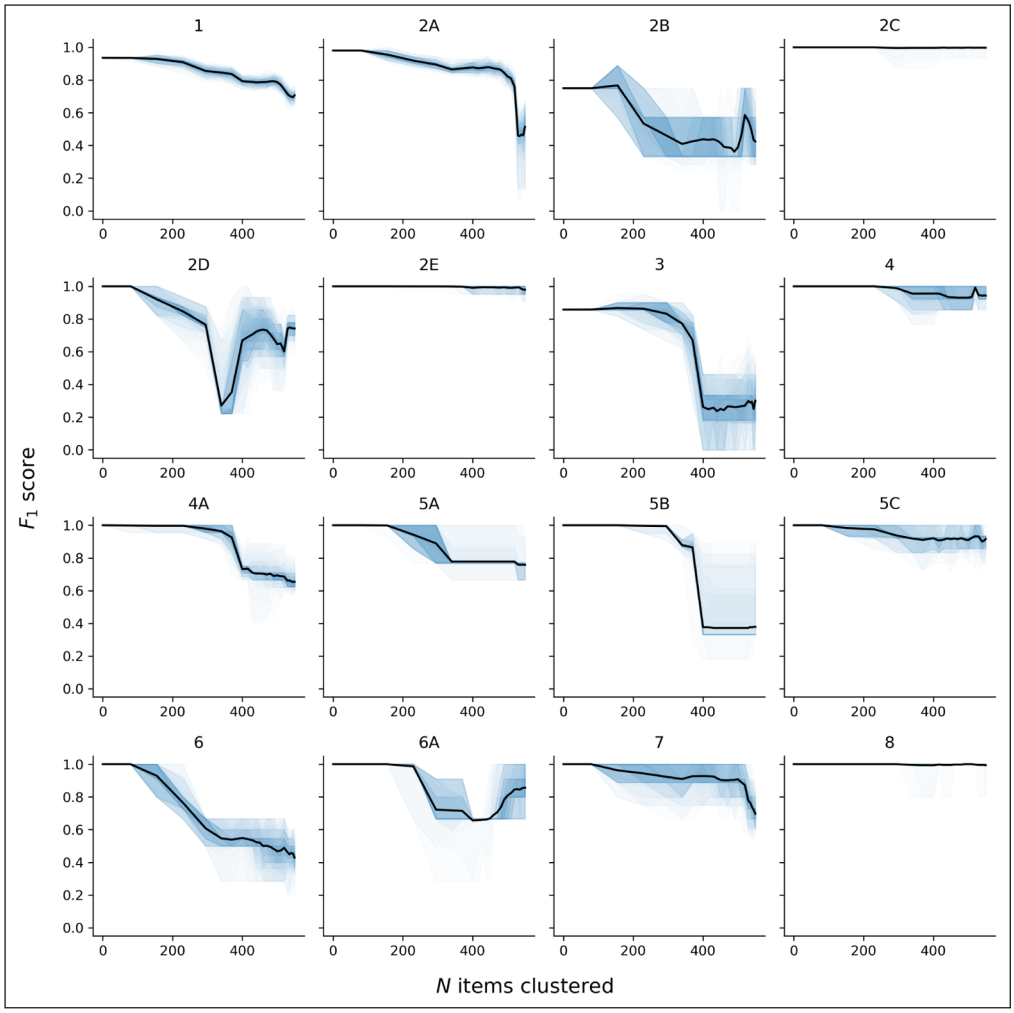

Figure 5 VAM output per sense category.

There are, however, a number of notable cases where the model's classification accuracy drops sharply when more more items of the same category are merged. This pertains to Sense 3, where the model does not recognize similarities between, for instance, *Spread a tablecloth over the table* and *He received votes from all over the floor*. Reassuringly, this ties in with other analyses that have argued that 'covering' does not adequately capture the interpretation of *all over* (Queller 2001; Taylor 2006; Pawelec 2010: 98–101). Furthermore, Sense 4A ('focus of attention') and 5B ('control') suffer from the high number of false positives of the other category the model wishes to assign to them (see Section 3.3.4–3.3.5 on the difficulty of distinguishing 4A and 5B). Finally, drops in performance can also be witnessed for category 2A and 2D, as the model seems to have difficulties in relating examples describing physical and non-physical scenes.

4.2 Relations between senses

Having established that there is some correspondence between clustered BERT embeddings and the proposed sense categories (up to a certain point), we can now turn to the question whether the geometrical distances between the various senses of *over* (as emergent from the distributional semantic model) correspond with the semantic relationships proposed in the prepositional polysemy network of *over* proposed by Tyler & Evans (2001: 746), reproduced in ***Figure 6***.

The dark, full nodes in the suggested network representation in ***Figure 6*** constitute what are considered to be separate senses, whereas the empty nodes are included as abstractions over a proposed cluster of related, derived senses. At the centre, we find the protoscene (Sense 1). Because the representation in Tyler & Evans has been constructed based on theoretical principles, it makes little sense to assess the representation based on the absolute geometrical distances between embeddings derived from the distributional model. It does make sense, however, to operationalize the directness of node linkage in the proposed network to relative distances between embeddings. More specifically, we could hypothesize that, if Sense 4A is not

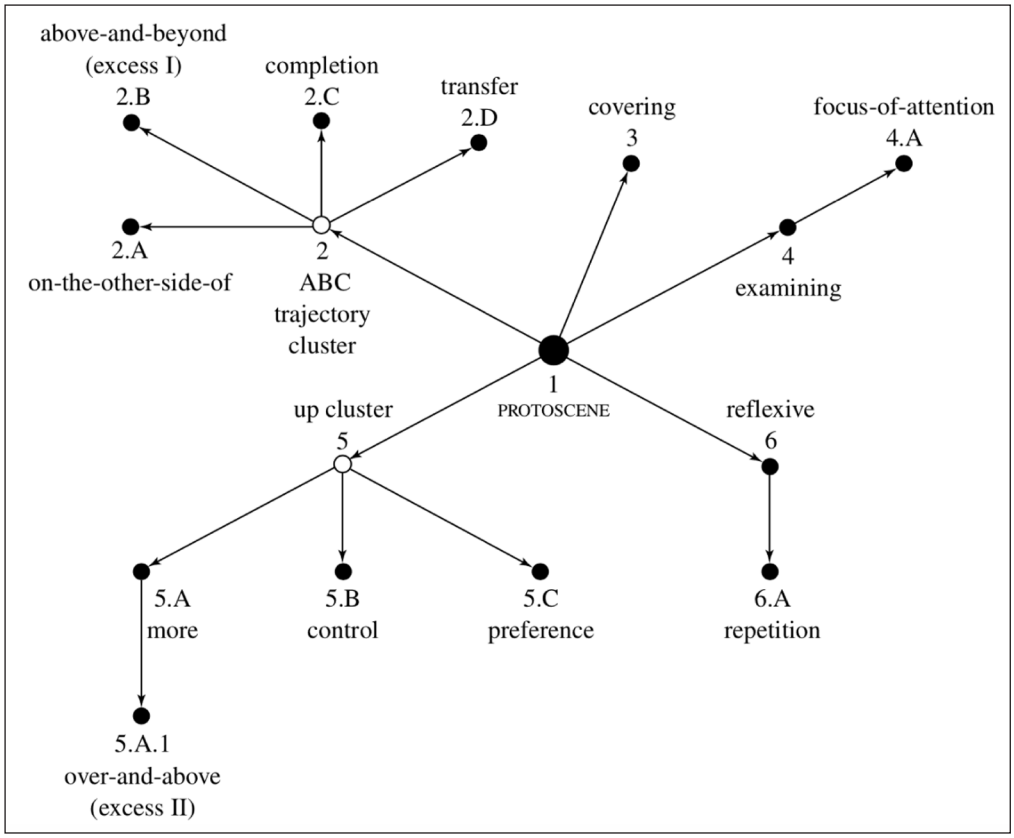

Figure 6 Polysemy Network of *over* as presented in Tyler & Evans (2001: 746).

directly derived from the protoscene (Sense 1), but emerged as a further extension derived from Sense 4, we would expect that the relative distance between Sense 4A and Sense 1 is bigger than the relative distance between Sense 4A and Sense 4. Similarly, if Sense 2A, 2B, 2C and 2D form a cluster of related senses, we would expect the relative distance between, for instance, Sense 2A and 2B or 2A and 2D, to be shorter than the relative distance between Sense 2A and Sense 3.

To compare the principled polysemy network to the output of the distributional model by means of cosine distances, two approaches can be taken. First, we may approach the comparison by taking the senses proposed by Tyler & Evans (2001; 2003) as given, and rely on manually assigned labels to create an averaged sense embedding – that is, a summary embedding similar to the clusters created at the highest level of abstraction in the VAM. Subsequently, we can calculate cosine similarities between these sense embeddings. If the result of this assessment turns out to be that the relative distance between the sense embeddings maps onto the suggested relative distances in the network in ***Figure 6***, we could conclude that both models arrive at the same network representation in a relatively straightforward manner.

However, as explained in Section 4.1, the proposed sense categories do not always correspond with the way in which the tokens of a manually labelled category cluster, with categories such as 2A falling apart into multiple rather distinct groupings. A second approach, then, would be to adhere less strictly to the sense categories proposed by Tyler & Evans, and determine the geometrical distance between token clusters proposed by the distributional model. In what follows, I restrict myself to this second approach.

Figure 7 presents a hierarchical cluster tree of the annotated tokens. Of the 808 annotated examples, 26 were examples were excluded because they were considered unclear or ambiguous between multiple readings (cf. example (42)). The clustering presented in ***Figure 7*** is based on the cosine distance between the embeddings of the remaining 782 examples. The coloured areas represent clusters of neighbouring tokens that were assigned to the same category. In order for a group of tokens to be considered a cluster, it was decided that there should be at least 5 neighbouring tokens of the same sense category. As such, the smallest clusters represented in ***Figure 7*** are based on at least 5 examples, and the largest (i.e. cluster 2E, 'time span') contains 51 examples. In total, 47 examples did not have at least 4 neighbours of the same type.

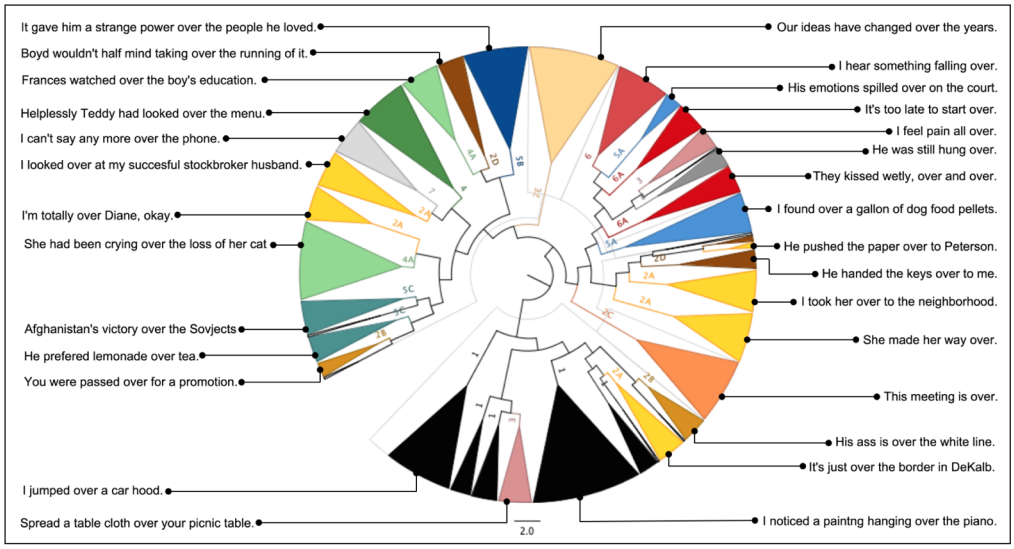

Figure 7 Cluster tree (distance = cosine) with representative examples.

As was briefly pointed out in Section 4.1, seven of the conceptual categories are 'consistent', with all tokens clustering together. This includes 2E ('time span'), 2C ('completion'), 4 ('examining'), 5B ('control'), 6 ('reflexive'), 7 ('communication channel') and 8 ('hangover'). For the remaining nine categories, the model suggests that there are at least two different clusters. In some cases, the separate clusters are still part of the same higher-order branch (e.g. 1 'above', 4A 'focus-of-attention', 6A 'repetition'), whereas for others, the embeddings are less closely related (e.g. 2A 'other-side', 2D 'transfer', 5A 'excess'). All in all, the suggested distances between the sense groupings differ substantially from the proposal put forward by Tyler & Evans: not only does the distributional distributional model suggest fairly large distances between token groupings that Tyler & Evans would have assigned to the same category (based on their shared underlying image schema), the proposed relative distances in the sense network are also not reflected in the geometrical distances between the (groupings) of embeddings.

Given that BERT does not group tokens of *over* according to abstract similarities in spatial configurations between them, the question that remains is what kind of groupings the model does suggest, and whether any (other) meaningful abstractions can be made. To address this question, we could examine the token groupings illustrated in ***Figure 7***.

GROUP 1 – SPACE The first group that appears to be recognized appears to be a collection of spatial uses of *over*. Tokens grouped in Group 1 include all tokens assigned to Sense 1 ('above'), and some tokens of the spatial 'excess' sense, 2B, in which a spatial border or threshold is exceeded. Note, however, that this cluster does not correspond with Kreitzer (1997)'s static $over_1$: Group 1 also includes a subgroup of tokens of Sense 3 ('covering', $over_3$), (stative) uses of Sense 2A ('other-side-of', $over_2$), and examples involving dynamic verbs (e.g. *the cat jumped over the fence*, $over_2$).

GROUP 2 – TIME SPAN and GROUP 3 – COMPLETION The two temporal uses of 'over', 2C ('completion') and 2E ('time span'), constitute fairly separate categories. The coherent cluster of 2E tokens will henceforth be referred to as Group 2. The coherent cluster of 2C will henceforth be referred to as Group 3. In all examples in Group 2, *over* functions as a preposition, whereas in all examples in Group 3 function as an adprep.

GROUP 4 – MIND AND PERCEPTION As the closest neighbour of Group 2, we can discern a very large grouping of tokens where *over* consistently marks a non-spatial relation (Group 4). In Group 4, we find all examples categorized as 4A ('focus of attention'). One cluster of examples of Sense 4A involves verbs such as *preside* and *watch*. These are closely associated with examples of 2D, in which there is a transfer of power or authority (e.g. *He took over the business*). These are, in turn, closely related to 5B ('control'). The other token cluster of 4A ('focus-of-attention'), which includes all examples where *over* can be paraphrased with 'about' or 'because of' (e.g. *He agonized over it*), is most closely linked to the subgroup of tokens in category 2A, where the TR has mentally overcome or lost interest in the LM (e.g. *I am over the drama*). We furthermore find all tokens of 5C ('preference') in Group 4, closely positioned next to a subgroup of 2B, where the LM is omitted or skipped in a selection procedure (possibly implying absence of preference,

e.g. *She was passed over for the job*). Group 4 also contains all tokens of 4 ('examining'). Given that the latter group of examples frequently (but not exclusively) involves the phrasal verb combination *look over*, it is not surprising that these tokens are positioned relatively close to (but, notably, are not confused with) cases where a glance is cast (classified as 2A). Finally, we also find all tokens of Sense 7 ('means of communication') in Group 4. While perhaps more loosely related to the 'mind' and 'perception' relations, Sense 7 also involves animate TRs and a non-spatial interpretation (i.e. in a sentence like *they spoke over the phone*, the preposition does not capture a physical, spatial positioning of the TR and LM).

GROUP 5 – PATH Related to Group 3, we find a cluster of tokens where *over* again has a spatial interpretation. Yet, unlike the tokens in Group 1, *over* functions as an adprep, and involves movement along a path (and hence partially overlaps with Kreitzer (1997)'s dynamic $over_2$). These include examples of 2D ('transfer') as well as examples of 2A where the TR moves to a different location (e.g. *I made my way over (to the computer)*).

GROUP 6 – EXCESS (IN NUMBERS) In the remaining set of tokens, then, a clustering of 5A tokens can be distinguished, where the LM is an amount or quantitative threshold (e.g. She mentioned that I donated *over* $100,000 to Katrina victims (2006, COHA)) or an amount (this cluster is also identified by Newman 2011).

GROUP 7 – NON-SPATIAL ADPREPS AND FIXED PHRASES While it is not easy to make sense of Group 7, it is interesting to note that all tokens classified as Sense 6 ('reflexive') and 6A ('repetition') are part of this group. Yet, rather than clustering together, they are positioned closely to token groupings classified as Sense 3 (in the fixed combination *all over*, e.g. *I feel pain all over*), Sense 5A (non-literal adprep uses, as in *their personal involvement will spill over into their workplace interaction*), and Sense 8 ('hangover'). The relation between these groupings seems formal rather than semantic, as the majority of groupings present cases where *over* functions as an adprep, and occurs in relatively fixed phrases.

If we approach the complex internal semantic structure of *over* by means of BERT embeddings, then, it appears that global clusters are formed based on in intersection of similarities in, on the one hand, conceptual domain (i.e., spatial, temporal, mental, etc.), and syntactic resemblance on the other. Overall, the lack of correspondence between the suggested network configuration in ***Figure 6*** and the global distances between the grouped embeddings does not necessarily indicate that the global groupings are not interpretable, or that no abstractions can be made – rather, it suggests that the use of an embedding-based approach leads to different abstractions.

5 Discussion and Conclusion

Within Cognitive Linguistics, there has been no shortage of proposals for modelling polysemy networks, in which the syntactic configurations, collocations, and the notion of underlying image schemas are of central concern. To minimize the arguably subjective nature of further proposals, linguists are increasingly turning to the use of distributional, statistical methods, and, most recently, to deep contextualized neural language models. In the present study, I investigated to what extent the output of a fully unsupervised application of BERT (a 'meaning as context' model) corresponds with the sense network of *over* proposed by Tyler & Evans (2003; 2001) (a 'meaning as concept' model). The analyses reveal that, while there are interesting correspondences the two approaches, they ultimately lead to different abstractions. Which of these abstractions most closely approximate the abstractions that emerge from ellicited, experimental data (or the extent to which the 'context' and 'concept' models are complementary) remains an open question that needs to be addressed by means of behavioural studies. However, because Tyler & Evans' proposal foregrounds the importance of sense connections via image schemas, the analysis presented in this study does provide some insight into the extent to which such imagistic information may be encoded in BERT embeddings.

What emerges from the preceding analyses is that the extent to which the models appear to converge varies depending on whether we rely on BERT embeddings to detect local (concrete, token-based) or global (schematic) similarities between examples. At the local level, BERT's focus on collocational and syntactic patterns helps it in 'recognizing' similarities between tokens of the same category, resulting in relatively coherent local clusters. When the consistency of

these local clusters does deviate from what is proposed by the principled polysemy model, it is furthermore reassuring that we can come up with a reasonable explanation for the divergence. For instance, when the embeddings of the tokens of the same principled polysemy category are split in separate clusters, the split can be motivated semantically: examples of Sense 2A ('other-side') that are used in a literal, physical sense (e.g. *I got over the hill*) are distinguished from more metaphorical uses (e.g. *I got over my puberty weirdness*). Notably, BERT's 'recognition' of such metaphorical uses extends into cases where surrounding context words are themselves used metaphorically (e.g. *We will* move on *and get over these* rough patches), indicating an unexpected aptitude for coping with elaborated metaphors (cf. De Pascale 2019: 157).

Yet, the fact that BERT is apt at recognizing metaphorical uses of *over* does not necessarily imply that it also recognizes that they are, in fact, metaphorical extensions of a literal, spatial source. This becomes evident when we consider the model's output at the global level. When we examine the relations between the token groupings emergent from the cluster analysis presented in Section 4.2, we find that the geometrical distance between the embedding of a particular spatial use of *over*, which may have given rise to a particular non-spatial use via metaphorical extension, is not shorter than the geometrical distance between the embeddings of two literal spatial uses (e.g. Group 1) or two metaphorical uses (e.g. Group 4) with different underlying image schemas. As such, if there is indeed a close connection between a spatial configuration and a non-physical scene it embodies, there may be a discrepancy between the geometric distances between the token embeddings of *over* and the actual conceptual similarity between those tokens.

The observation BERT does not immediately capture similarities in terms of image-schema resemblances can be understood in light of the fact that the model has been trained on linguistic data alone, and has no experience with (physical) non-linguistic, perceptual information (such as spatial configurations, but also, for example, visual properties such as colours: Sommerauer & Fokkens 2018). Hence, BERT embeddings pick up fine-grained semantic distinctions based on collocational and morpho-syntactic cues, and can be employed to successfully group senses in distinct domains. However, BERT embeddings seem less equipped to flag abstract configurational resemblances in image schemas across those domains, which helps highlight what sort of semantic information is (and is not) encoded in contextualized embeddings.

Note that, if the perceptual, imagistic information that motivates the abstractions and sense connections made in the cognitive-conceptual model is still somehow encoded in contextual information (as appears to be suggested by Gromann & Hedblom (2017)), such information could be brought to the fore by further experimentation with the model's hyperparameters (e.g. different context window sizes, different (combinations of) layers), or fine-tuning the model to a sense classification task by exposing it to manually labelled examples in training. Yet, if such perceptual information is not represented in context embeddings and requires extra-linguistic knowledge, an interesting avenue to pursue is, for instance, to train language models based on coupled textual and visual input (Chrupała et al. 2015). Of course, whether such additional training and supervision is desirable depends entirely on the question the researcher wishes to address, and which facets of meaning they deem relevant within their study or theoretical framework. In Cognitive Linguistics, researchers may be inclined to say that a model of meaning representation should capture the global resemblance between the underlying image schemas of prepositions, as image schemas (and embodiment) are part of the core tenets of the framework (e.g. Oakley 2010; Gibbs & Matlock 2001: 233), and play an important role in, for instance, studies of semantic change and grammaticalization (e.g. Rhee 2002).

As a final concluding remark, I wish to add that the findings presented in this study have important implications for the integration of neural language models – and perhaps, more generally, the application of Semantic Vector Space Models – in theoretical linguistic research, and in particular, to research on semantic change. In a recent publication, Boleda (2020) surveys a number of studies that have applied either count or predictive models to historical and diachronic corpus data. Such studies, which involve examination of nearest neighbours and cosine similarities between type- and/or token-vectors, have provided the key to detecting, as well as describing the diachronic trajectory of lexical and, albeit less commonly, grammatical semantic changes (e.g. Hilpert & Perek 2015; Hamilton et al. 2016; Dubossarsky 2018; Sagi et al. 2011; Hilpert & Correia Saavedra 2017; Budts & Petré 2020; Giulianelli et al. 2020). In some of these studies, it is argued that distributional semantic models could also be employed to detect

different types of semantic change (and, by extension, I could add that they may also help assess competing hypotheses regarding the mechanisms of change at play in a particular diachronic development). Some steps have already been taken in this direction (e.g. the automated detection of semantic broadening and narrowing in Sagi et al. (2011); Giulianelli et al. (2020)), and indeed, BERT could be an excellent tool for detecting metaphorical extensions of linguistic items in diachronic corpora (Giulianelli et al. 2020). However, it should be clear that, at least when left entirely unsupervised, BERT does not seem to pick up that there may be abstract, imagistic similarities between domains. As such, researchers interested in studying metaphorical extensions (of prepositions or otherwise) should take into consideration that unsupervised BERT will be great at indicating *that* a metaphorical extension has occurred from one domain to another, but they do not reveal which perceptual similarity pattern is the most likely source of the extension. It could be possible, however, to tackle these issues by experimenting with additional supervision and different model architectures, and, crucially, by accelerating the dialogue on how to integrate these models in theoretical linguistic research, and vice versa.

Data accessibility statement

All data and code can be found at *https://github.com/LFonteyn/Glossa_over*.

Acknowledgements

I wish to thank Charlotte Maekelberghe for acting as the second annotator. I furthermore thank all anonymous reviewers as well as Folgert Karsdorp and Stefano De Pascale for their insightful feedback on earlier versions of this manuscript.

Competing interests

The author has no competing interests to declare.

Author affiliation


Lauren Fonteyn *orcid.org/0000-0001-5706-8418*
Leiden University, Arsenaalstraat 1, 2311 CT Leiden, NL

References

Alishahi, Afra, Grzegorz Chrupała & Tal Linzen. 2019. Analyzing and interpreting neural networks for nlp: A report on the first blackboxnlp workshop. *Natural Language Engineering* 25(4). 543–557. DOI: *https://doi.org/10.1017/S135132491900024X*

Baroni, Marco & Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, 1–10. Edinburgh, UK: Association for Computational Linguistics.

Baroni, Marco, Georgiana Dinu & Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 238–247. Association for Computational Linguistics. DOI: *https://doi.org/10.3115/v1/P14-1023*

Berez, Andrea L. & Stefan Th. Gries. 2008. In defense of corpus-based methods: A behavioral profile analysis of polysemous get in English. In *Proceedings of the 24th NWLC*, 157–166. Seattle, WA.

Blevins, Terra & Luke Zettlemoyer. 2020. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1006–1017. Online: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/2020.acl-main.95*

Boleda, Gemma. 2020. Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics* 6(1). 213–234. DOI: *https://doi.org/10.1146/annurev-linguistics-011619-030303*

Boleda, Gemma & Katrin Erk. 2015. Distributional Semantic Features as Semantic Primitives – Or Not. In *Aaai spring symposium on knowledge representation and reasoning*, 2–5. USA: Stanford University.

Brugman, Claudia M. 1988. *The Story of Over: Polysemy, Semantics, and the Structure of the Lexicon*. New York: Garland.

Budts, Sara. 2020. *On periphrastic do and the modal auxiliaries: a connectionist approach to language change*. Antwerp: Universiteit Antwerpen PhD dissertation.

Budts, Sara & Peter Petré. 2020. Putting connections centre stage in diachronic construction grammar. In Lotte Sommerer & Elena Smirnova (eds.), *Nodes and Networks in Diachronic Construction Grammar*, 317–352. Amsterdam: John Benjamins. DOI: *https://doi.org/10.1075/cal.27.09bud*

Bullinaria, John A. & Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods* 39(3). 510–526. DOI: *https://doi.org/10.3758/BF03193020*

Bullinaria, John A. & Joseph P. Levy. 2012. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods* 44(3). 890–907. DOI: *https://doi.org/10.3758/s13428-011-0183-8*

Chrupała, Grzegorz, Ákos Kádár & Afra Alishahi. 2015. Learning language through pictures. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 112–118. Beijing, China: Association for Computational Linguistics. DOI: *https://doi.org/10.3115/v1/P15-2019*

Clark, Kevin, Urvashi Khandelwal, Omer Levy & Christopher D. Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. *arXiv:1906.04341*. DOI: *https://doi.org/10.18653/v1/W19-4828*

Clausner, Timothy C. & William Croft. 1999. Domains and image schemas. *Cognitive Linguistics* 10(1). 1–31. DOI: *https://doi.org/10.1515/cogl.1999.001*

De Pascale, Stefano. 2019. *Token-based vector space models as semantic control in lexical sociolectometry*. Leuven: KU Leuven PhD dissertation.

Desagulier, Guillaume. 2019. Can word vectors help corpus linguists? *Studia Neophilologica* 91(2). 219–240. DOI: *https://doi.org/10.1080/00393274.2019.1616220*

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, 4171–4186. Minneapolis, Minnesota.

Dubossarsky, Haim. 2018. *Semantic change at large: A computational approach for semantic change*. Jerusalem: the Senate of the Hebrew University of Jerusalem PhD dissertation.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2017. Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1136–1145. Copenhagen, Denmark: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/D17-1118*

Erk, Katrin & Sebastian Padó. 2010. Exemplar-Based Models for Word Meaning in Context. In *Proceedings of the ACL 2010 Conference Short Papers*, 92–97. Uppsala, Sweden: Association for Computational Linguistics. DOI: *https://doi.org/10.1017/S0022226704003056*

Evans, Vyvyan. 2005. The meaning of time: polysemy, the lexicon and conceptual structure. *Journal of Linguistics* 41(1). 33–75.

Firth, John Rupert. 1957. *Papers in Linguistics 1934–1951*. London: Oxford University Press.

Geeraerts, Dirk. 2016. Sense individuation. In Nick Riemer (ed.), *The Routledge Handbook of Semantics*, 233–247. London: Routledge.

Gibbs, Raymond W., Dinara A. Beitel, Michael Harrington & Paul E. Sanders. 1994. Taking a Stand on the Meanings of *Stand*: Bodily Experience as Motivation for Polysemy. *Journal of Semantics* 11(4). 231–251. DOI: *https://doi.org/10.1093/jos/11.4.231*

Gibbs, Raymond W. & Teenie Matlock. 2001. Psycholinguistic perspectives on polysemy. In Hubert Cuyckens & Britta Zawada (eds.), *Polysemy in Cognitive Linguistics*, 213–239. Amsterdam: John Benjamins. DOI: *https://doi.org/10.1075/cilt.177.10gib*

Gilquin, Gaëtanelle & Andrew McMichael. 2018. Through the prototypes of through: A corpus-based cognitive analysis. *Yearbook of the German Cognitive Linguistics Association* 6(1). 43–70. DOI: *https://doi.org/10.1515/gcla-2018-0003*

Giulianelli, Mario, Marco Del Tredici & Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 3960–3973. Online: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/2020.acl-main.365*

Glynn, Dylan. 2010. Testing the hypothesis. Objectivity and verification in usage-based Cognitive Semantics. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*. Berlin, New York: De Gruyter Mouton. DOI: *https://doi.org/10.1515/9783110226423.239*

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

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: The many senses of to run. In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Trends in Linguistics. Studies and Monographs [TiLSM]*. Berlin, New York: Mouton de Gruyter.

Gries, Stefan Th. & Dagmar Divjak. 2009. Behavioral profiles: A corpus-based approach to cognitive semantic analysis. In Vyvyan Evans & Stéphanie Pourcel (eds.), *Human Cognitive Processing* 24. 57–75. Amsterdam: John Benjamins Publishing Company. DOI: *https://doi.org/10.1075/hcp.24.07gri*

Gries, Stefan Th. & Dagmar Divjak. 2010. Quantitative approaches in usagebased Cognitive Semantics: Myths, erroneous assumptions, and a proposal. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*. Berlin, New York: De Gruyter Mouton. DOI: *https://doi.org/10.1515/9783110226423.331*

Gries, Stefan Th & Naoki Otani. 2010. Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME journal* 34. 30.

Gromann, Dagmar & Maria M. Hedblom. 2017. Kinesthetic Mind Reader: A Method to Identify Image Schemas in Natural Language. *Advances in Cognitive Systems* 5. 14.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501. Berlin, Germany: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/P16-1141*

Harris, Zellig S. 1954. Distributional Structure. *WORD* 10(2–3). 146–162. DOI: *https://doi.org/10.1080/00437956.1954.11659520*

Heylen, K., T. Wielfaert, D. Speelman & D. Geeraerts. 2015. Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua* 157. 153–172. DOI: *https://doi.org/10.1016/j.lingua.2014.12.001*

Hilpert, Martin & David Correia Saavedra. 2017. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory* 16(2). 393–424. DOI: *https://doi.org/10.1515/cllt-2017-0009*

Hilpert, Martin & Florent Perek. 2015. Meaning change in a petri dish: constructions, semantic vector spaces, and motion charts. *Linguistics Vanguard* 1(1). 339–350. DOI: *https://doi.org/10.1515/lingvan-2015-0013*

Huang, Luyao, Chi Sun, Xipeng Qiu & Xuanjing Huang. 2019. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3509–3514. Hong Kong, China: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/D19-1355*

Jawahar, Ganesh, Benoît Sagot & Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. Florence, Italy: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/P19-1356*

Kilgarriff, Adam. 2003. "I don't believe in word senses". In Brigitte Nerlich, Zazie Todd, Vimala Herman & David D. Clarke (eds.), *Polysemy*. Berlin, New York: De Gruyter Mouton.

Kreitzer, Anatol. 1997. Multiple levels of schematization: A study in the conceptualization of space. *Cognitive Linguistics* 8(4). 291–326. DOI: *https://doi.org/10.1515/cogl.1997.8.4.291*

Lakoff, George. 1987. *Women, fire, and dangerous things: what categories reveal about the mind*. Chicago: The University of Chicago Press. DOI: *https://doi.org/10.7208/chicago/9780226471013.001.0001*

Langacker, Ronald W. 1991. *Foundations of Cognitive Grammar 2: Descriptive Application*. Stanford, CA: Stanford University Press.

Langacker, Ronald W. 2010. *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Walter de Gruyter.

Lee, David. 1998. A Tour through through. *Journal of English Linguistics* 26(4). 333–351. DOI: *https://doi.org/10.1177/007542429802600404*

Lemmens, Maarten. 2016. Cognitive semantics. In Nick Riemer (ed.), *The Routledge Handbook of Semantics*, 90–105. London: Routledge.

Lenci, Alessandro. 2018. Distributional Models of Word Meaning. *Annual Review of Linguistics* 4(1). 151–171. DOI: *https://doi.org/10.1146/annurev-linguistics-030514-125254*

Levy, Omer, Yoav Goldberg & Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics* 3. 211–225. DOI: *https://doi.org/10.1162/tacl_a_00134*

Linzen, Tal, Grzegorz Chrupala, Yonatan Belinkov & Dieuwke Hupkes (eds.). 2019. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics.

Newman, John. 2011. Corpora and cognitive linguistics. *Revista Brasileira de Linguística Aplicada* 11(2). 521–559. DOI: *https://doi.org/10.1590/S1984-63982011000200010*

Oakley, Todd. 2010. Image Schemas. In Dirk Geeraerts & Hubert Cuyckens (eds.), *Handbook of Cognitive Linguistics*, 214–235. Oxford: Oxford Univeristy Press.

Pater, Joe. 2019. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language* 95(1). e41–e74. DOI: *https://doi.org/10.1353/lan.2019.0009*

Pawelec, Andrzej. 2010. *Prepositional network models a hermeneutical case study*. Krakow: Jagiellonian University Press.

Peters, Matthew, Mark Neumann, Luke Zettlemoyer & Wen-tau Yih. 2018b. Dissecting Contextual Word Embeddings: Architecture and Representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1499–1509. Brussels, Belgium: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/D18-1179*

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee & Luke Zettlemoyer. 2018a. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/N18-1202*

Queller, Kurt. 2001. A usage-based approach to modeling and teaching the phrasal lexicon. In Dirk Geeraerts, René Dirven, John R. Taylor & Ronald W. Langacker (eds.), *Applied Cognitive Linguistics, II, Language Pedagogy*. Berlin, Boston: De Gruyter.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei & Ilya Sutskever. 2018. Language Models are Unsupervised Multitask Learners. San Francisco, California, United States: OpenAI.

Rhee, Seongha. 2002. Semantic Changes of English Preposition against A Grammaticalization Perspective. *Language Research* 38(2). 563–583.

Rice, Sally. 1996. Prepositional prototypes. In Martin Pütz & René Dirven (eds.), *The Construal of Space in Language and Thought*. Berlin, New York: De Gruyter Mouton.

Rice, Sally A. 1999. Aspects of prepositions and prepositional aspect. In Leon de Stadler & Christoph Eyrich (eds.), *Issues in Cognitive Linguistics*. Berlin, New York: De Gruyter Mouton.

Sagi, Eyal, Stefan Kaufmann & Brady Clark. 2011. Tracing semantic change with Latent Semantic Analysis. In Kathryn Allan & Justyna A. Robinson (eds.), *Current Methods in Historical Semantics*. Berlin, Boston: De Gruyter. DOI: *https://doi.org/10.1515/9783110252903.161*

Sandra, Dominiek. 1998. What linguists can and can't tell you about the human mind: A reply to Croft. *Cognitive Linguistics* 9(4). 361–378. DOI: *https://doi.org/10.1515/cogl.1998.9.4.361*

Sandra, Dominiek & Sally Rice. 1995. Network analyses of prepositional meaning: Mirroring whose mind—the linguist's or the language user's? *Cognitive Linguistics* 6(1). 89–130. DOI: *https://doi.org/10.1515/cogl.1995.6.1.89*

Sommerauer, Pia & Antske Fokkens. 2018. Firearms and Tigers are Dangerous, Kitchen Knives and Zebras are Not: Testing whether Word Embeddings Can Tell. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276–286. Brussels, Belgium: Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/W18-5430*

Stefanowitsch, Anatol. 2010. Empirical cognitive semantics: Some thoughts. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*. Berlin, New York: De Gruyter Mouton. DOI: *https://doi.org/10.1515/9783110226423.355*

Taylor, John R. 2006. Polysemy and the lexicon. In Gitte Kristiansen, Michel Achard, René Dirven & Francisco J. Ruiz Mendoza Ibáñez (eds.), *Cognitive Linguistics: Current Applications and Future Perspectives*, 51–80. Berlin: Mouton De Gruyter.

Turney, Peter D. & Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37. 141–188. DOI: *https://doi.org/10.1613/jair.2934*

Tyler, Andrea & Vyvyan Evans. 2001. Reconsidering Prepositional Polysemy Networks: The Case of Over. *Language* 77(4). 724–765. DOI: *https://doi.org/10.1353/lan.2001.0250*

Tyler, Andrea & Vyvyan Evans. 2003. *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning, and Cognition*. Cambridge University Press 1st edn. DOI: *https://doi.org/10.1017/CBO9780511486517*

van der Maaten, Laurens & Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9. 2579–2605.

Vanpaemel, Wolf & Gert Storms. 2008. In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review* 15(4). 732–749. DOI: *https://doi.org/10.3758/PBR.15.4.732*

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. Attention Is All You Need. In *31st conference on neural information processing systems (nips 2017)*. USA: Long Beach, CA.

Verbeemen, Timothy, Gert Storms & Tom Verguts. 2005. Varying Abstraction in Categorization: a K-means Approach. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, 2301–2306. Mahwah, NJ: Erlbaum.

Wiedemann, Gregor, Steffen Remus, Avi Chawla & Chris Biemann. 2019. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. *arXiv:1909.10430*.

Young, Tom, Devamanyu Hazarika, Soujanya Poria & Erik Cambria. 2018. Recent Trends in Deep Learning Based Natural Language Processing. *arXiv:1708.02709*. DOI: *https://doi.org/10.1109/MCI.2018.2840738*

TO CITE THIS ARTICLE:
Fonteyn, Lauren. 2021. Varying Abstractions: a conceptual vs. distributional view on prepositional polysemy. *Glossa: a journal of general linguistics* 6(1): 90. 1–28. DOI: *https://doi.org/10.5334/gjgl.1323*

Submitted: 23 May 2020
Accepted: 16 February 2021
Published: 06 July 2021

Glossa: a journal of general linguistics is a peer-reviewed open access journal published by Ubiquity Press.