Word Embeddings are Word Story Embeddings (and That's Fine)

Katrin Erk
The University of Texas at Austin

Gabriella Chronis
The University of Texas at Austin

CONTENTS

ABSTRACT

Many words are polysemous, and most of them in a manner that is entirely idiosyncratic, not regular. Word embeddings give us a window into idiosyncratic polysemy: They are computed as an aggregate of many observed word uses, from different speakers. Interestingly, they also pick up what could be called traces of stories: text topics, judgements and sentiment, and cultural trends. This holds for all types of

DOI: 10.1201/9781003205388-9

word embeddings, from count-based vectors to word type embeddings and word token embeddings. In this chapter, we trace the many ways in which stories show up in word vectors, and we argue that this is actually an interesting signal and not a bug. We also perform an in-depth analysis of clusters of contextualised word embeddings, again finding traces of stories, along with an interesting pattern of clustering that could be described equally well as driven by lexico-syntactic patterns and story traces.

9.1 INTRODUCTION: WORD EMBEDDINGS AND STORIES

Deep learning models in machine learning are universal function approximators. Given enough pairs of inputs and outputs of some function, they can learn to approximate it. They project the inputs to latent representations, combine and re-combine them through multiple layers, and thereby learn internal features to best match the data they have seen. This is interesting from a linguistic point of view, not just for the performance of these models on natural language processing tasks but for the latent features that the models learn. Baroni (Chapter 1 of this volume) contends that language models should be treated as linguistic theories in their own right. What linguistic representations, if any, do these models build when they solve natural language tasks? Deep learning models that are language models are particularly intriguing because they are trained on very general and basic tasks such as predicting a held-out word in context, so that we can observe what the models learn when they are simply pushed to observe statistical regularities in texts.

There is a lively ongoing discussion about the syntactic information that deep learning models encode, for example, Gulordava et al. (2018), Lake and Baroni (2018), Lappin (2021), and Linzen et al. (2016), about what syntactic knowledge these models notice, how systematic they are in observing syntactic regularities, and whether they encode syntactic knowledge in a way that is different from standard linguistic formalisms. In the same way, we can ask what machine learning models learn about lexical semantics – and this is what we want to do in this chapter. The questions that we want to look at are analogous to those that are being asked about syntax: What differences in word meaning do machine learning models observe, and encode in word embeddings? And do their observations of lexical regularities differ from how we usually describe word senses?

Word embeddings to some extent reflect categories of things and events in the world, in that words that belong to the same categories tend to be more distributionally similar than words that do not share a category (Burgess and Lund, 1997). Padó and Lapata (2003) and Sahlgren (2008) remark that distributional similarity encompasses a wide array of semantic relations, including (but not limited to) taxonomic relations such as synonymy, hyponymy, and co-hyponymy. There is even work which sets inferring taxonomic structure as an objective for distributional modelling. For example, Kruszewski et al. (2015) and Nickel and Kiela (2017) construct hierarchical taxonomic spaces, Mrkšić et al. (2017) aim to separate antonyms, and Roller et al. (2014) tackle lexical entailment and hypernymy. Webson et al. (2020) remove

connotative associations. When one does so, one is selecting meanings related to categorisation as the particular kind of meaning to represent faithfully in the model.

Sahlgren notes that these relations are neither axiomatic nor all-encompassing, and cautions against a prescriptive perspective that would construe any particular kind of meaning as the *a priori* goal of distributional modelling. It is certainly useful to have a lexical model with taxonomic knowledge, but it is a mistake to think that distributional similarity is merely a noisy measure of taxonomic relatedness. That "noise" can contain information about other kinds of meaning, which we are calling stories.

Word embeddings do not just learn about categories of things and events in the world, they also pick up on stories connected to words. This is not an entirely new insight, it has been mentioned over and over again, in bits and pieces, and not just with recent word embeddings but also with old-fashioned count-based vector spaces. For example, word sense induction models seem to pick up not only on what we would call dictionary senses but also more fine-grained usage context (Reisinger and Mooney, 2010). Such usage context can be used to study changes in word connotations (Kozlowski et al., 2019; Kutuzov et al., 2017), but it has also been found to be shot through with harmful social biases that reflect stereotypes. We believe it is useful to view all these observations as facets of the same larger phenomenon: that word embeddings record how humans view the world, what they find important, useful, and useless, what judgements they make, and what stories they like to tell using the words.

What we mean by *stories that people tell with words* is an amalgam of several things, including but not limited to affordances (ways that humans typically use an object), judgements and emotions, and the topical contexts in which words frequently appear. In this chapter, we will explore how these aspects of meaning are reflected in the organisation of contextual distributional space.

When putting a deep learning model into production, there are very real concerns about unforeseen negative impacts lurking inside the black box: cultural bias, prejudice, and misinformation potential. NLP research oriented towards applications often asks how models trained on text can be shaped to more accurately reflect the shape of the world, for example through de-biasing. The same characteristics that pose an obstacle for NLP practitioners present an opportunity to linguists and social scientists: a chance to study the interaction between many different factors which contribute to the meaning and interpretation of language. The traditional distinction between semantics and pragmatics leads to the desire for a transparent model in which we think of the factors which contribute to meaning as separate or separable; there are extensional ('actual world') factors and social factors. On the other hand, Potts (2019) suggests that if we follow the cue that word embeddings give us, we are likely to view these categories as connected and influenced by one another.

In addition to providing a framework for studying distributional semantic similarity, the holistic idea of story meanings is useful for discussions about the mental lexicon. McRae and Matsuki (2009) show that during sentence processing, listeners make use of story knowledge in forming expectations about upcoming words. This could mean that story knowledge is part of the lexicon, or that it is used in sentence

processing in addition to the lexicon, or even, as Elman (2009) points out, it could mean that there is no mental lexicon, just mental story knowledge. We think that if there is a mental lexicon, it must contain some story knowledge, for prominent affordances, emotions, and topical contexts. Listeners can draw on this knowledge in sentence understanding and in metaphorical uses of a word that may foreground its social and emotional facets. It may not be possible to draw a clear, principled line between story knowledge in the lexicon and story knowledge that is used in addition to the lexicon, and we do not see it as necessary to draw such a line. Our aim in this paper is to point to work in lexical semantics that is compatible with story knowledge in the lexicon and to the many ways in which embeddings exhibit traces of such knowledge.

The debate about the nature of the lexicon is an old one. In Section 9.2, we look at some theories in linguistics and psychology that consider scenes, events, and background knowledge to be important to word meaning, both out of context and in context. We want to see to what extent the information that we find in word embeddings matches what these theories say about the human mental lexicon and human sentence processing, and whether therefore word embeddings can be used in linguistic analyses informed by those theories. In Section 9.3 we take a tour through research about word embeddings, from count-based word vectors to very recent models, to point out all the different observations that are glimpses of a general "story bias."

The most recent addition to word embedding frameworks are contextualised word embeddings (CWEs), embeddings for individual word tokens instead of word types. These embeddings constitute a fascinating new opportunity for studying lexical meaning in context. It was possible to some extent to do this with word type embeddings that were modified based on their sentence context (Erk and Padó, 2008; Schütze, 1998; Thater et al., 2011) but recent contextualised embeddings have the potential to observe higher-order co-occurrence regularities and therefore much more subtle meaning differences. Do we indeed observe more subtle meaning differences, and can we actually interpret them? This is not a question that we can settle in this chapter, but to get some insight into contextualised word embeddings as lexical representations, we perform an in-depth qualitative analysis of contextualised word embeddings for a small sample of nouns and verbs in Section 9.4. We close the text by speculating on how word embeddings can help us understand how humans understand sentences: how the stories connected to individual words combine to form a story for the overall sentence, and how the sentence story in turn modulates the meanings of the words from which it is constructed.

9.2 FRAMES, AFFORDANCES, EVENTS, AND CONTEXTUAL MODULATION IN WORD MEANING

When you know the meaning of a word, what do you know? When you know the word *cat*, you certainly know properties of cats, such as: that they are furry, whiskered, and carnivorous. You may also know what things in the world you can label as cats. But that is not all. Words are also connected to connotations, emotional affects, larger scenes, pieces of background knowledge, typical events, circumstances, even

perspectives. When a speaker (or signer, or reader) encounters a word in context, the word might evoke for them a particular referent. It might also evoke for them any and all of these other dimensions of meaning. To refer to these facets in a theory-neutral way, we use the term *stories*. Story meaning is related to referential meaning, but it is not the same. The idea of story meaning is not new. It had been translated into theoretical frameworks and motivated by psycholinguistic evidence. In this section, we trace these different notions through the literature in linguistics and psychology. What arises, in spite of all differences, is a clear picture of story-like knowledge as an important part of word meaning, both lexicalised and in context.

Fillmore (1982) describes words as connected to *frames*. A frame is a "categorisation of experience" (p. 112), a piece of background knowledge connected to a word. Fillmore's frames are prototypes in the sense of Rosch (1973). They are part of a word's memorised meaning; when the word is used, it evokes its frame. A frame can be a scene with typical participants. For example, the scene for the word *criticise* involves a Judge and a Defendant. In an utterance, these scene participants are verbalised as semantic roles of the verb *criticise*. Frames also include cultural knowledge. Fillmore uses the example of *breakfast*, writing (p. 118): "to understand this word is to understand the practice in our culture of having three meals a day, at more or less conventionally established times of the day, and for one of those meals to be the one which is eaten early in the day, after a period of sleep."

The frame that forms the background of a word can be highly specific. Fillmore's example is the word *decedent*. A decedent is a dead person, but the word is only used in the context of inheritance. The frame concept is later developed into a "semantics of understanding", or U-semantics which contrasts with T-semantics, a semantics of truth (Fillmore, 1985). Fillmore argues that truth values are not enough to describe what it means to understand a sentence, that it is important to consider the frames evoked by the words. Multiple words can connect to the same prototypical scene but take different perspectives on it. This is the case with *buy* and *sell*, which both describe a commerce scene, but differ in who is the profiled agent, and which other participants are typically named. Emotions and judgements, too, are important parts of frames. Fillmore (1985) uses the example *My dad wasted most of the morning on the bus* (p. 230), where the choice of the word *dad* instead of *father* lets us draw conclusions about the speaker's relationship with his father, and the word *wasted* "brings into play a judgement that the time was not used profitably [...] and this depends on a framing of time as a limited resource (Lakoff and Johnson, 1980)."

Prototypical scenes can have participants that are heroes and villains, so evoking a frame can mean passing a judgement. This connects frames to the notion of *framing* in the social sciences (Kuypers, 2009), where a framing is a schema of interpretation. Linguistically, framing can take the form of a frame in the sense of Fillmore, or a conceptual metaphor (Lakoff and Johnson, 1980). Lakoff (2004, p. 56) uses the example of *tax relief*. The word *relief* evokes a frame with clear heroes and villains. The term *tax relief* casts taxes as a scourge, taxpayers as suffering victims, and politicians who lower taxes are the rescuers. So the notion of *framing*, like Fillmore's frames, involves stored, prototypical scenes but focuses on the judgements, in some cases the biases, imparted by the use of a particular frame.

In contextualism, the focus is on the contextual *modulation* of word meaning: the influence of context on word meaning, including what could be called story context. In a classical example by Travis (1997, p. 89), Pia has a red-leaf shrub but does not like the red, so she paints the leaves green. If, upon completion of her task, she says: "This is better. The leaves of my tree are now green," she has spoken the truth. But imagine that now a botanist asks around for some green leaves for a study on photosynthesis, and Pia says, "The leaves of my tree are green, you can use them." This time, so the argument goes, she has not spoken the truth – the point being that context can interfere with word meanings to an extent that the truth conditions of a sentence are changed. The context influence in the *green leaves* example is a story influence, but one that comes from the discourse context, imposed on the word *green* from the outside, if you will, rather than evoked lexically.[1] Recanati (2017) connects these dynamic context influences back to the lexicon, building on the theory of Meaning Eliminativism from Recanati (2003, ch. 9). In that earlier work, Recanati offers several accounts of word meaning that allow for context influences. One of them is Meaning Eliminativism, where the idea is that listeners do not generalise over observed word uses to form abstract word senses, they just remember observed occurrences in all their story context. In a new situation, the speaker chooses to use the word if the situation is sufficiently (contextually) similar to a previous use of the word. Recanati (2017) sketches how polysemy can arise in a Meaning Eliminativism setting: The listener remembers the observed usages, so if a usage context is sufficiently frequent, it will be remembered strongly enough to function as a separate sense.

In psychology, a number of studies have shown the influence of event knowledge on human sentence processing, measured through priming effects, and effects on reading times (McRae and Matsuki, 2009). Event verbs cue typical arguments, but conversely, nouns cue event verbs for which they are typical agents, patients, or instruments (McRae, Hare, et al., 2005): *Chef* cues *cooking*, *tax* cues *paid*, and *chainsaw* cues *cutting*. This could be interpreted, at least in part, as objects cueing their affordances (Gibson, 1966), that is, the actions that humans typically do with them. The event knowledge that affects human sentence processing can be very fine-grained, such that different patients are cued depending on different agents of the action. *Journalists* are expected to *check the spelling*, while *mechanics* are expected to *check the brakes* (Bicknell et al., 2010). Bicknell et al. (2010) use words where this expectation cannot be explained by a direct association between the agent and patient, for example, *journalist* does not cue *spelling*. The effect is only found in the presence of the verb – which means that it is hard to describe this effect in any way as lexical. In fact, Elman (2009) speculates about "lexical knowledge without a lexicon:" Given that event knowledge affects sentence processing strongly and early on, and given that this knowledge does not seem to be lexical, is it possible that there is no lexicon at all? Intriguingly, the computational mechanism that Elman uses to demonstrate lexical knowledge without a lexicon is a language model (an RNN). We

[1] Del Pinal (2018) affirms the effect of context on word meaning but seeks to rein in context effects on a lexical basis: Context can only modify specific lexically given dimensions, which coincide with the qualia used in Pustejovsky (1991).

TABLE 9.1 Some annotator choices for lexical substitutes for the noun *charge* in the sense of a person in one's care, from Kremer et al. (2014).

Now, how can I help the elegantly mannered friend of my Nepthys and his surprising young charge?	dependent, companion, person, lass, protégé
The distinctive whuffle of pleasure rippled through the betas on the bridge, and Rakal let loose a small growl, as if to caution his charges against false hope.	dependent, companion, private, underling, prisoner, troop

will return to the question of "lexical knowledge without a lexicon" below, when we look at recent contextualised language models.

Summing up, there are many different notions of story-like influences connection with word meaning. There are affordances, prototypical scenes and events, and pieces of background knowledge, sometimes with a focus on the facts of the situation, sometimes with explicit inclusion of cultural contexts as well as emotions and judgements connected to a scene. (Fillmore's frames encompass all of these notions of stories when they are prototypical; we use the term *story* to encompass both frames and dynamic, non-prototypical story contexts.)[2] In the next section, we check which of these different types of lexicalised story influences we can observe in word embeddings.

Words lexically evoke stories, but they also jointly cue stories that they do not evoke lexically. And their meanings are dynamically modulated by the story told by the sentence, and the discourse, as a whole. This interaction between frequent, lexicalised meanings and the story at hand can be clearly seen in the lexical substitution data we collected in Kremer et al. (2014), where annotators provided lexical substitutes for words in context. Many substitutes are not WordNet-synonyms, -hypernyms or -hyponyms of the original target but context-specific, story-specific. Table 9.1 shows two example sentences, both with the target noun *charge* in the sense of someone that you take care of or that you oversee. The substitutes reflect this commonality: we get *dependent* and *companion* in both sentences. But the first sentence seems to be from some kind of ballroom story, and the *charge* in question is likely a young girl. The second sentence seems to be part of a war story, where the *charges* are more likely to be prisoners or subordinates. With respect to word embeddings, this raises the question of whether contextualised word embeddings are fine-grained enough to capture not only frequent usage contexts but also these subtle effects that the context at hand exerts on word meaning.

[2] Trott et al. (2020) argue for the importance of *construal* in language, and in natural language processing. Construal is about how "linguistic choices subtly colour meaning." The notion of construal overlaps with what we call story influences, for example Trott et al. include metaphor in their list of dimensions of construed meaning, but they explicitly exclude background knowledge evoked by words and other effects that Fillmore includes in U-semantics. So construal has some overlap with story-like influences, but it is not the same.

9.3 STORIES, FROM COUNT-BASED WORD VECTORS TO CONTEXTUALISED EMBEDDINGS

Word embeddings computed from text corpora record statistical regularities about how the words are used. Patterns of language use encode many different kinds of information, about word senses, human concept representation and the mental lexicon, sizeable amounts of world knowledge, and social and cultural contexts of use. This is all well known. But we think it is instructive to look at story-like influences in word embeddings comprehensively, across all the tasks in which they have shown up, to see that they come together to paint a single picture: Word embeddings capture how humans engage with words and with the world, the typical stories they tell and kinds of value judgements they make.

To start, there is the distinction of *similarity* and *relatedness* in lexical items. *Similar* words have properties in common and are close to one another in a taxonomy. *Related* words pertain to the same topics. (We italicise the terms *similar* and *related* to distinguish this specific meaning of the terms from their general-language use.) A jug is *similar* to a bucket. A hammer is *related* to a nail. Distributional models reflect both kinds of relations. In fact, they can be tuned to focus on one or the other by changing the size of the context window. Count-based co-occurrence models which use smaller context windows are better at predicting similarity relations, while those with larger windows are better at predicting semantic relatedness judgements. Sahlgren (2006) noted this phenomenon anecdotally, and subsequent empirical testing has confirmed the context window effect both in Dutch (Peirsman, 2008) and English (Baroni and Lenci, 2011). Felix Hill and colleagues exploit the fact that distributional models can be tuned to focus on either similarity or relatedness to study lexical concreteness (Hill et al., 2013) and to tune distributional spaces towards particular tasks (Kiela et al., 2015). This difference between *similarity* prediction and *relatedness* prediction has been shown for contextualised language models as well. Chronis and Erk (2020) show that the final layers of BERT (Devlin et al., 2019), which contain more propagation of contextual information, are better at relatedness, but earlier layers best predict similarity. We will make use of this fact in our qualitative analysis in Section 9.4. Linking the notions of *relatedness* and *similarity* to the theories from the previous section, Fillmore's (1982) notion of frames is very general, so that both groups of *related* words and groups of *similar* words should be frames, though the frames in the FrameNet resource (Fillmore et al., 2003) are groups of *similar* words. The words that fill roles in a frame could be characterised as being *related* to the lexical units in the frame. *Relatedness* links words that tend to appear in the same stories, or in the same generalised events in the sense of McRae and colleagues.

As a second piece of the puzzle, when distributional models are used to extract context items that are interpretable in themselves, as in Baroni, Murphy, et al. (2010) and Baroni and Lenci (2010), they do not tend to extract typical attributes. Baroni, Murphy, et al. (2010) write about the extracted context items for *motorcycle*: "we clearly see here the tendency of [the model] [...] to prefer actional and situational properties (riding, parking, colliding, being on the road) over parts (such as wheels and engines)" (p. 233). These actional properties are similar to affordances and to

the event verbs that McRae, Hare, et al. (2005) found to be primed from the mention of nouns that are typical agents, patients, or instruments.

Again and again, story effects show up as noise in natural language processing tasks. Reisinger and Mooney (2010), doing word sense induction, notice that their clustering algorithm sometimes picked up on what they call "thematic polysemy" (p. 1174). For instance, it discovered two very distinct "senses" for the word *wizard*, one of them the "King Arthur wizard", the other "Hogwarts wizard". Is there an analytic difference between this thematic polysemy and the distinct related word senses like *running a computer program* and *running a marathon*? Intriguingly, humans feel that these usages clearly constitute the same sense of *wizard* but in different contexts. We will see similar story effects in clusters made from contextualised embeddings below in Section 9.4.

Story effects are also visible in diachronic studies. A similar issue appears as with word sense induction, in that if you try to detect changes in sense, you also observe changes in the social context of a word while the sense stays the same. This leads to spurious detected sense changes, as reported in Rosenfeld (2019) and Del Tredici (2020). Strikingly, Kutuzov et al. (2017) employ the same methods that are being used to find diachronic changes in word sense but use them to detect changes in cultural context, in particular outbreaks of armed conflict. The analysis of word embeddings for the cultural context of words is still in its beginnings, but is growing. In the social sciences, Kozlowski et al. (2019) analyse word embeddings over time for dimensions representative of social class.

There is much recent research on mitigating and detecting social bias in word embeddings, starting with Bolukbasi et al. (2016). Webson et al. (2020) characterised the de-biasing task as a task of distinguishing denotation from connotation. When we view the task in this way, bias can be viewed as just another type of story effect– where the stories that people tell often involve judgements, or are dependent on particular cultural associations. One problem with existing work on de-biasing is that it needs to define a bias in order to remove it, so it always focuses on some particular, well-defined type of bias, often gender bias, or left-versus-right political bias in the case of Webson et al. (2020). Maybe seeing bias as a story effect can help us find more general techniques to (at least partly) separate the categories of items, and their properties, from the stories that people tell about them.

9.4 ET TU, BERT?

As a language model, BERT (Devlin et al., 2019) is emblematic of the current transformer architecture paradigm. In contrast to type-level distributional models, BERT leverages representations of individual word occurrences, or tokens, which take into account the surrounding sentential context. For word token embeddings, as for word type embeddings, we can ask to what extent they reflect stories. We first again check the existing literature, before moving on to a qualitative analysis of BERT embeddings for a small sample of lemmas.

In the NLP literature, the natural first place to look is the task of sense disambiguation and sense induction. Contextualised language models achieve state of the

art accuracy on several word sense disambiguation benchmarks, achieving F1 scores 10 or 20 percent higher than the baseline of selecting the most frequent sense (Wiedemann et al., 2019). Contextualised language models also lead to massive leaps in the harder task of word sense induction, though the best models still perform well below human levels (Amrami and Goldberg, 2019). This gap in performance between language models and humans is intriguing; the error analysis in Amrami and Goldberg (2019) hints that the model sometimes creates clusters specific to particular topics.

In the realm of lexical semantics, contextual word embeddings have been employed most commonly to study semantic change. Giulianelli et al. (2020) coin the phrase "usage types" to describe the kinds of clusters formed by BERT tokens. In some cases, they found them to make distinctions based on metaphoricity, syntactic roles and argument structure, as well as phrasal collocations (conventional multi-word expressions) and named entities. Finally, looking at de-biasing, most research on this topic has been done on type embeddings, but biases have also been found, and de-biasing efforts undertaken, for contextualised language models (Bommasani and Cardie, 2020; Kaneko and Bollegala, 2021). This research indicates that many words "list" to one side of a cultural axis of meaning.

9.4.1 A Qualitative Analysis of BERT Tokens

To get a clearer sense of the kinds of story effects we find in contextualised embeddings, we perform a qualitative analysis of word token embeddings in BERT for a small number of noun and verb lemmas. As our main basis of the analysis, we use clusters of BERT token embeddings.[3] Following our previous work (Chronis and Erk, 2020), where we found that multi-prototype embeddings derived from the middle layers of BERT best approximate word *similarity* (as opposed to *relatedness*), we look at clusters for the 8th layer of the 12 layer BERT. If any layer were to distinguish senses while not being overly sensitive to story effects, we would expect it to be layer 8, based on its ability to predict word similarity. As in 2020, we sample a number of tokens for each lemma, drawn from the British National Corpus (BNC; here, we use a maximum of 100 tokens per lemma), and use a fixed number of five clusters per lemma, obtained using k-means clustering. In order to counteract distortions introduced by the fixed number of clusters, we additionally visualise token vectors in the Context Atlas visualisation tool, which shows T-SNE plots of BERT tokens from Wikipedia (Coenen et al., 2019).[4]

[3]For this study, we use the 12 layer `bert-base-uncased` model available from the HuggingFace Python API (Wolf et al., 2020). The model is pre-trained with standard training objectives on English Wikipedia and the BookCorpus (Zhu et al., 2015). We interpret the hidden representations above a token as the contextual embeddings for that token. Since there are twelve layers, this yields 12 vector representations of each token. For words consisting of more than one sub-word token, like 'in #depend ##ence', we follow the precedent of averaging the embeddings of each of its tokens.

[4]Interactive demo available at https://storage.googleapis.com/bert-wsd-vis/demo/index.html. The clusters apparent from observing the Context Atlas are neither more comprehensive, nor identical to the clusters we generate. The kinds of organisation we see according to genre/topic/framing are not as prevalent in Context Atlas, which is limited to usages from Wikipedia. This goes to show that when conducting statistical research into lexical semantics, one must be very mindful of the distribution one draws from and whose language it represents.

Our main analysis comprises 45 nouns, sampled from nouns that appear at least 25 times in the BNC. We sampled 5 nouns each from three polysemy bands (1 sense, 2-5 senses, 6 and more senses) in WordNet 3.0 crossed with three bands of concreteness according to the concreteness norms of Brysbaert et al. (2014) (average concreteness ratings 0-2.3, 2.3-4.5, 4.5-5.0). In addition, we inspected some verbs from the stimulus list of Bicknell et al. (2010). This study, which we discussed in Section 9.2, studies effects of generalised event knowledge on expectations in sentence processing, with stimuli pairing, for example *the mechanic checked the brakes* with *the journalist checked the spelling*. Here we want to know whether BERT will separate the Bicknell agent/verb/patient combinations. We again focus on BERT layer 8.

In the following, we discuss particularly interesting lemmas as well as general trends we observe across all the lemmas. This study has all the advantages and disadvantages of a qualitative study: Because we look at the data 'by hand', we may notice subtleties that a quantitative analysis may miss, and patterns we had not anticipated. The downside is that the analysis is subjective. We even observed differences in our interpretation of the clusters between the two authors. It is also small in scale, and we may simply not be able to perceive some higher-order statistical regularities underlying the data because they are too different from surface co-occurrences. However, in most cases, the rough groupings have natural interpretations and/or particular features, both syntactic and semantic, which distinguish them from one another. In our analysis, we see some story effects, especially in verbs and high-polysemy nouns. In addition to distinguishing senses, BERT notices subtle distinctions which look a lot like affordances, narrative and topic effects, and is also sensitive to syntactic and collocation effects.

9.4.2 Sense Separation

Among mildly polysemous nouns (2-5 senses), there is great variation on the extent to which the clusters reflect WordNet senses. For some concrete words, senses are clearly separated into different clusters: a smoking *pipe* is distinguished from a plumbing *pipe*, programming *libraries* are distinguished from lending *libraries*. For *bottle*, one cluster focuses on the physical object while others focus on the contents. *Shoe* behaves similarly. Most of its WordNet senses are very rare. Here the clusters organise according to topic/genre as well as to grammatical cues. There is one cluster dedicated to actions done by or to a shoe, such as kicking in a door or getting-some-sticky-substance-stuck-to-the-bottom-of. This cluster could be described as "shoe in a narrative." Another cluster focuses topically on the shoe industry. This cluster could alternatively be described as involving compound nouns with shoe, such as *shoe store, shoe repairer, shoe factory*. Interestingly, a separate compound noun cluster isolates the topic of shoe-related items like *shoe laces* and *shoe polish*. Syntactic patterns and topic jointly yield clusters with definite perspectives on shoes: as a commodity, as an

Some kinds of story effects are more prevalent in scientific literature, others in fiction. The topics we can discern are not necessarily definitive or most significant to the meaning of the English word, but they are the most significant in our corpus.

TABLE 9.2 WordNet senses for *load*.

Sense	Example
(a) fill or place a load on	*load the truck with hay*
(b) put (something) on a structure or conveyance	*load hay onto the truck*
(c) supply (a device) with something necessary to function	*load the gun*
(d) transfer from storage to a computer's memory	*load the software*
(e) corrupt, debase, or make impure	*load the servers* (as in overload)

object functionally consisting of several components, or as a participant in a narrative event.

Abstract nouns with a mild degree of polysemy tend to have closely related senses. WordNet lists four senses for *admission*, where the first is a confession, and the other three are to do with entry to an event or location. Admission can refer to the cost of entry, the permission to enter, or the act of entering. BERT distinguishes the confession sense quite clearly. However, it mixes the other three senses between different clusters.

Verbs tend to display a greater degree of polysemy than nouns, and on average are more abstract. We find that for verbs, grammatical cues are a stronger organising principle than topics. In Layers 7 and 8, clusters can typically be recognised as distinguishing different grammatical constructions in which the verb is used. In many cases this corresponds to sense separation—for instance, in a given construction that is only used for one or two of a verb's many senses. We observe that more topical separation is sometimes reached in the final layer.

Take for example the word *load*, whose five WordNet senses are shown in Table 9.2. At layer 8, there are two distinct clusters which capture the computer loading sense (d) and one cluster which captures the "put" sense (b). The clusters seem to group both by syntactic and topical similarity. Examples of *load* in the "put" cluster (Table 9.3, cluster 3) all have an overt direct object corresponding to the Patient argument, as well as a locative phrase denoting the Location or Goal of movement. Sentences in the computer cluster of *load* all have topical words related to computers, and the direct object position is always overt. Cluster 1 contains most of the "fill" senses (in which the Location or Goal occupy direct object position) as well as sentences which do not contain enough context to disambiguate between uses. In the final layer of the network, we see separations that more closely correspond with the topic and less with grammatical cues. Sometimes this results in the separation of a sense, and sometimes it results in the collapse of several senses. The "fill" and "put" senses are merged in a cluster to do with loading vehicles. The "corrupt/overload" sense is separate.

In attempting to relate the degree to which BERT distinguishes dictionary senses, we were forced to countenance the many ways in which BERT organises meaning; it is difficult to talk about BERT's separation of senses without talking about other kinds of stories. Story effects crop up at all levels of polysemy and concreteness. In *library*, there are clusters corresponding to different metonymic senses of library, as

TABLE 9.3 Tokens of *load* from the BNC clustered according to their representations at layer 7 of BERT base.

Cluster	Description	Examples
0	adjectival	... that 's not even a half **load**.
	nominal	straps at top of shoulder slings for fine adjustment to **load** carrying.
1	ambiguous	Others will **load** at Mareham in Norfolk.
	"fill"	to **load** our jeep up from the magazine
	"charge"	Although slow to **load** and fire, hand guns can penetrate armour more easily than arrows.
		The alternative, when plutonium is seen as a fuel, is to **load** it into existing Russian reactors ...
2	computer	It only took six seconds to **load** each room but some puzzles require you to operate buttons in several rooms, so you would obviously be running between rooms for a while.
	infinitival/modal	If it does, they will be overwritten when you **load** the line editor.
3	"put"	Standing up, she began to **load** the tea-things on to a tray.
		The defendant helped to **load** the goods into Ballay's van.
		Some way away a couple of humans were using some sort of machines to **load** boxes into a hole in the side of the plane.
4	computer	**Load** it from starting Window.
	imperative	Bulk **load** a wrong formulation or a not a wrong formulation.
	first position	**Load** it up, arrange your features into a smirk and invite the nearest Mac user to come and have a butcher's.

an organisation or as a building. We also see a cluster that focuses on the books and people in the library. This latter distinction corresponds to different attitudes or perspectives surrounding a single sense of the word. There are two separate clusters for the "hollow tube" sense of *pipe*: one in which a pipe is seen as a medium through which a substance moves (such as water, waste, in one case sea otters!), and the other in which pipes are seen as raw material for constructing or engineering a system. It is hard to tease apart different kinds of story effects, because they tend to interact. Story effects also often coincide with syntactic cues. In some cases, when BERT manages to isolate a sense (as with admission to a hospital), it is a consequence of isolating a

topic which only makes use of one of those senses. To see how the different types of story effects interact, we take a detailed look at one word: *independence*.

9.4.3 Interaction between Topic, Sense, and Narrative Type: *Independence*

The noun *independence* has one main WordNet sense: freedom from control over others. This encompasses both individual freedoms and state sovereignty, though the line between encyclopedic and lexical meanings is drawn differently in different lexical resources.[5] Of the five clusters generated for *independence* (Table 9.4), three pertain exclusively to political independence. Of the other two, one small cluster pertains to collective independence of non-political bodies and abstract bodies ("the press," "phonemes," "interpreters.") The final, large cluster, pertains to independence of individuals ("one's manhood and independence," "her independence and organising ability.")

With the separation between state and individual independence, the clusters for *independence* demonstrate something like sense distinctions (even if WordNet does not list state sovereignty and the personality trait as separate senses—perhaps it should!) The three clusters relating to political independence exhibit differences in the type of narrative. We mentioned that *independence* has two clusters of state independence. Of these, one seems to be anticipatory ("432 votes in favour of party independence," "vague promises about independence," "The vote for independence") and the other seems to reflect on an accomplished event ("The DEMOS...had led Slovenia to independence," "smaller than anticipated flows of aid after independence," "Qatar gained full independence"). In both clusters is the same sense of independence, but the perspective on the event of independence is different. This does not correspond with any single grammatical cue. This close look at *independence* demonstrates how senses, selectional preferences, and topical information all interact to organise meanings.

9.4.4 Affordances, Narratives, Events

Affordances, in the sense of Gibson (1966), are the properties of an object which allow it to function in different capacities. Embeddings organised according to affordances encode manners in which the word is experienced by a subject. Embeddings are subjective not just in the sense that they encode emotional and affective meaning (which they do), but also in the sense that they encode different phenomenological perspectives on a word. This phenomenon is clearly observed for the word *desk*. Only one sense for desk is recorded in WordNet, but The New Oxford American Dictionary distinguishes a subsense for the counter where one checks in or receives information (the front desk) (in addition to two more uncommon senses). The BERT clusters for *desk* (Table 9.5) mirror these sub-senses, and make further distinctions seemingly related to topic. Two of the five clusters (1, 3) contain the information desk subsense. Two clusters (0, 2) seem to refer to desks as a surface holding objects, where cluster

[5] WordNet also lists a second sense with a highly limited meaning, for references specific to the successful conclusion of the American war for independence from Britain.

TABLE 9.4 Tokens of *independence* from the BNC clustered according to their representations at layer 8 of BERT base.

Cluster	Description	Examples
0	sovereignty	At the Estonian CP congress there were 432 votes in favour of party **independence**, three votes against and six abstentions. They were then on the verge of **independence**. The politicians made vague promises about **independence**. By 1814 [...] the prospects for **independence** were gloomy; Bolívar had been driven out of Venezuela and New Granada was about to be recaptured.
1	political entity (other than state)	The **independence** of the cities was effectively strangled. The test by this time was no longer '**independence** of Moscow' but 'Human Rights'. The older though continuing tensions between cultural authority and cultural **independence** have been transformed by the increasingly dominant social relations of the new means of production and reproduction. In his resignation speech, Gorbachev asserted that although he had been in favour of "the **independence** and self-determination of peoples and the sovereignty of republics," he had also believed in "preserving the union state and the country's integrity"
2	organisation	The existence of these characteristics required the **independence** of the profession from interference by government. Other systems have used different strategies [...], for example, the HMM 's erroneous assumption about **independence** between adjacent phonemes means that the acoustic evidence is underestimated. That is that practices in theoretical, moral-practical, and aesthetic spheres become 'contingent' in their **independence** from externally imposed order. The concept of the 'freedom of the press' must, therefore, be appraised with reference to the structural and organisational **independence** of the press from the state.
3	personality	His films reflect his background, a masculine world where one's manhood and **independence** can only survive through violence. Joshua's physical **independence** was achieved because he was young and motivated. Departments [...] have received instructions that–whatever the precise degree of **independence**–the minister is answerable to Parliament for whether the body is working efficiently and economically. Her **independence** and organising ability had been displayed early when she joined the Girl Scouts in 1909 and formed the first Bournemouth troop of guides.
4	colonialism	The Islamic Republic of Mauritania, formerly part of French West Africa, gained full **independence** in 1960. Launch of the Burma **Independence** Army (1942) **Independence** of Lithuania, Estonia and Latvia Qatar gained full **independence** from the United Kingdom in 1971.

TABLE 9.5 Tokens of *desk* from the BNC clustered according to their representations at layer 8 of BERT base.

Cluster	Description	Examples
0	surface for papers	Small machines like the Olivetti PCS 33 and the Packard Bell Elite 1000 don't take up much **desk** space, but they don't leave you with much room for future expansion either. In New Scotland Yard John McLeish was trying, increasingly irritably, to clear his **desk** so he could go home. Would you really, and think carefully about this, trust all your personal information; diary, telephone list and so on to the memory of that recalcitrant computer on your office **desk**? Nigel himself vacillated between belief in and total rejection of death, and busied himself with sorting out his **desk** and jettisoning surplus papers.
1	reception; location; direct object/ prepositional phrase	In the hotel lobby, Bodie came away from the **desk** and spoke briefly to Cowley . When she turned again she saw Mahoney approaching the **desk**. Madame Gauthier was perched on a stool at the reception **desk**, making up her accounts. She ushered them into a small cluttered room where a middle-aged woman in a navy dress sat behind a littered **desk**.
2	movement across surface; definite article	Edward 's fingers drummed on the edge of the **desk**. There was the thin voice beating at him across the **desk**. He showed her into another bare cubicle where two hatchet-faced men were scribbling, and pointed to a phone that lay scratched and bruised on the **desk**. The draft position paper goes sliding over the edge of the **desk** into the waste-paper basket as I snatch up the receiver.
3	reception; function	Rather, the offer is made by the customer when he takes the goods to the cash **desk** [...] The information **desk** was manned throughout the weekend, the timetable was strictly adhered to and everyone benefited from the efficiency. We do need a **desk** in both places–a desk to welcome people is important.
4	somebody's desk; workstation source/destination of movement	If you have to leave your **desk** [...] When he got to Kafka's **desk**, he spoke. Dyson took off his overcoat and went to his **desk**, frowning heavily. My mother was working at her **desk**. He arrived at his **desk**, squinted at the offering tucked into his blotter and sat down to read it properly: [...]

0 focuses more on clutter ("clear his desk so he could go home", "busied himself with sorting out his desk" but also "desk lamp") and cluster 2 more on movement across the surface ("the draft position paper goes sliding over the edge of the desk", "flipped the five ten-pound notes across the desk .", "fingers drummed on the edge of the desk", "reached over the desk for a bottle of pills", "the receiver was dangling under the desk"). While the clusters are only partly distinguishable topically, they differ grammatically, where cluster 2 but not 0 uses past tense and the definite determiner 'the' on *desk*. Comparing clusters 2 and 3, we see that tokens exemplifying the same sense of desk (a furniture with a flat surface at which one can work) cluster together according to their affordances: as a surface, or as a location at which a person is working. The same participants might be present in the situations described by these sentences—a person, the furniture, papers strewn about—but the aspect of the desk afforded to the reader shifts. And BERT picks up on these shifts.

Box, a concrete word with high polysemy, also pays attention to affordances. One cluster specifies boxes full of stuff—that is, boxes seen as containers ("box of matches", "box of dolls", "box of bullets". Another cluster specifies boxes that *do* things—boxes which contain or comprise some kind of mechanism ("phone box", "connection box", "junction box").

Context clues to affordances blend into context clues for associated events, which seem the organising principle for the word *letter*. The alphabet sense of the word is not attested enough in the BNC.[6] We almost exclusively see the epistle sense of the word. Letters in this sense participate in different kinds of events—sending, reading, producing. These events afford or emphasise different prototypical participants. The clusters for *letter* are given in Table 9.6. Cluster 0 contains uses that overtly refer to the act of sending or receiving. This event information interacts with topic information: cluster 0 is also largely limited to letters sent in an official capacity. Cluster 1 contains usages in which personal letters are exchanged between individuals. Here we find letters as texts that are being read, or else as physical objects with which people interact. Cluster 2 usages focus on the speech act of communicating information by way of a letter. Grammatically, they contain overt reference to both author and recipient/audience, as well as a subordinate clause detailing the thematic content of the letter.

We see a similar sort of topic/event interaction in the word *admission*. Recall that BERT recognises very distinct senses (admission of guilt vs. admission to an event/location), but does not distinguish between the metonymic senses listed in WordNet (cost of access, right to access, or the act of giving access). Rather, BERT distinguishes topic clusters (admission to an event, admission to a hospital, admission to an institution or school). What's interesting about these clusters is that in some of them, you see all three metonymic senses, but in some of them, only one or two senses apply: the hospital cluster contains only the act of entering/giving access, because in the hospital frame, admission does not prototypically incur a fee or require an application.

[6]The alphabet sense of *letter* however forms a distinct cluster at layer 8 in Context Atlas.

TABLE 9.6 Tokens of *letter* from the BNC clustered according to their representations at layer 8 of BERT base.

Cluster	Description	Examples
0	official letter;	They had received yet another **letter** from the school ...
	recipient named;	It is probably a surprise to you to receive a **letter** from me. The next morning a letter came for Matthew.
	sending event	**Letter** to the Editor: Tie is staying in place well
1	personal letter;	Please, old friend, come to my house at once with this **letter** in your hand.
	movement events;	It was a good **letter**, if a little pompous.
	reading events	He pulled a crumpled **letter** out of his shirt pocket, opened it and handed it to her to read.
2	sender named;	But, as Olga said in her **letter** to me, one weeps more often because one is not free.
	recipient named;	One **letter** from 12 senators, including Edward Kennedy, spoke of the 'continuing human rights abuses of British security forces'.
	speech act	Thank you for your **letter** of 2 November 1992 regarding the problems associated with the above property.
		In that connection, I should like to quote from a **letter** that the Secretary of State wrote to the right hon. Member for Selby (Mr. Alison)...
3	noun complement	The contract required the yard to open a **letter** of credit . . .
		POST OF CONFIRMATION OF **LETTER** OF OFFER
		A proper **letter** of instruction is vital.
		This month's **letter** of the month prize is a luxurious Cotman Watercolour Box from Winsor & Newton.
4	alphabet letter;	These consist of a capital **letter** followed by numbers.
	unclear	Points referred to in accompanying **letter**.
		This is a standard record update **letter** to be sent to a client
		In seperate [sic] incidents, two farmers were injured by **letter** bombs thought to be from animal rights extremists.
		Activity in the target **letter**'s detector will therefore be inhibited (switched off).

9.4.5 Framing

The lemmas in this study are too broadly chosen to show clear instances of the kind of bias targeted in the de-biasing literature. We do, however, see cases where BERT distinguishes separate framings of a noun, as when different contexts invite different value judgements. In the clusters for fire (Table 9.7), there is a distinction between dangerous, destructive fire (cluster 1), small domesticated fires (cluster 3), and fire which is characterised in terms of how it is physically experienced, both physically and metaphorically (cluster 0). The latter cluster contains both literal and figurative references to the experience of fire, but in each case, the fire is framed as an intense or transformative experience rather than a destructive one. Note that different fire-related experiences are allocated to different clusters: "He was tired, and the heat of the fire was making him sleepy," is assigned to the hearth cluster, and "I was as much arsonist as alchemist now, swinging the axe gleefully, impervious to everything but the fire's appetite," is assigned to the destructive event cluster. What we are seeing here is an organisation of the human emotions towards fire into several attitudes. Different kinds of fire are categorised into stories that come pre-packaged with judgements.

9.4.6 Syntactic cues and their interaction with senses, topics, and affordances

As noted in several above cases, BERT consistently groups common syntactic constructions together. Sometimes these syntactic differences seem to be the only distinguishing trait of clusters, but in many cases the differences in syntax coincide with differences in sense, or other kinds of story outlined above. In short, *subtly different uses of a word tend to have their own syntactic fingerprint.*

The noun *train* offers a good illustration of this phenomenon. One cluster groups together tokens of *train* which serve as the argument to a preposition ("on the train", "in the train", "straight off the London train"), as well as those which serve as an argument to the verb *board* ("She boarded the train in advance of him.") Semantically, these usages all convey the meaning of a train as a container, with people inside it. Movement of the train is not typically entailed in these usages. There is another cluster which groups together tokens of *train* which are syntactic subjects. This group clusters together sentences in which the movement of the train is foregrounded ("This train would take her there," "And the train is getting nearer," "The train stops at Greenwich"). This cluster also contains tokens which contain words related to movement, without the syntactic cues ("The sheep gazed through the bars at the departing train," "He was left behind by the rest of the wagon train"). So in the usages where the train is more likely to be the subject, it is also more likely to be in motion. In contrast with this story, which foregrounds the train, another cluster of *train* tokens entails motion from a different perspective: as a connection between two locations or a journey as a whole. These instances of *train* are often syntactically objects, with animate subjects ("get a train from Glasgow down to Girran," "Needs to catch uncrowded train," "they took the train to London.")

The close ties between syntactic behaviour and variations in meaning has been well studied. In her now standard book on verb classes and alternations, Levin (1993) hypothesised that differences in meaning influence syntactic behaviour. BERT, in

TABLE 9.7 Tokens of *fire* from the BNC clustered according to their representations at layer 8 of BERT base.

Cluster	Description	Examples
0	transformative fire;	An occasional shaft of sunlight penetrated the foliage and lit up the bronze trunks of the pines, touching them with **fire**.
	intense feeling;	Adam swiftly read the titles, most of which contained romantic words like 'love', 'heart', 'arrow', 'passionate', '**fire**', 'dream', 'kiss', and 'enchanted'.
	emotion	Changez said nothing, but shuffled backwards, away from the **fire** of Anwar's blazing contempt, which was fuelled by bottomless disappointment.
		Never again, except in the nostalgic hopefulness of a few—would the ceremonies be performed; gone were the offerings, the blood-shedding, the **fire** and incense, the gorgeous (and the plain) robes . . .
1	destructive fire; arson	There was a **fire** at Mr's store and they called it arson.
		An electrical short circuit started the **fire**, they think.
		Mr Green hopes the **fire** will provoke further pledges of aid to People In Need at 1113 Maryhill Road, Glasgow.
		A woman and seven children left a house in nearby Gudmunsen Avenue after a **fire** was discovered in a bedroom at around 7.30am on Saturday.
2	artillery;	Small-arms **fire** scorched a web of gaps through the foam.
	metaphorical artillery	Almost immediately, there was a brief burst of machine-gun **fire**, which destroyed the three remaining wheels.
		'The hierarchy are now under **fire** because of the team's performance and are seeking to deflect criticism by blaming me.
		Following an exchange of **fire** the Ju88 flew back to Sicily in a damaged condition and with one crewman wounded, but one Beaufighter—T3239 'B' crewed by Flt.Lt.
3	hearth	or reading in the shadow of a **fire**;
		They all went over to the **fire** for plates of meat and bread.
		The light from the **fire** bathed her in a warm flickering glow as he lay down beside her.
		The bar is warm and cosy, with an open **fire** and oak beams.
4	compound nouns;	Now add the top of the fireback, bedding it on top of the lower half with a layer of **fire** cement
	control over fire	That's when the **fire** brigade arrived.
		Mr Small said **fire** alarms were installed and special voice tapes would tell people to leave the premises.
		A **FIRE** station is to be put up for sale, a council report has revealed.

picking up on rather complex syntactic patterns, may be finding semantic groups of usages to which these patterns correspond.

We do not want to be overly optimistic. Sometimes, *all* you can see are syntactic patterns, without any specific meanings (story-related or otherwise) separated out. Sometimes not even that, for example in the clusters for the verb *fix* and the noun *analysis*.

9.4.7 Fine-grained Story Effects on Individual Occurrences?

In Section 9.2 we discussed how some theories focus on memorised, lexicalised story contexts, while others primarily look at dynamic, ad-hoc context effects on individual tokens. In our qualitative analysis in this section, we have seen plenty of evidence for something like memorisation of frequent senses: Frequent usage contexts form coherent clusters. But it remains unclear to what extent contextualised embeddings are sensitive to the fine-grained context effects like those that contextualism discusses and those that we observe in lexical substitutes (Table 9.1). Manual inspection of token clusters is a tool that is much too blunt to answer this question, but it remains an important question to be probed with quantitative measures in the future.

Looking specifically at the verbs in our study, which as mentioned above were pulled from the stimuli of Bicknell et al. (2010), we can ask whether we see the fine-grained event knowledge effects that Bicknell and colleagues found in their participants: does BERT make a difference between *journalist checks spelling* and *mechanic checks brakes*? Here we find that when the different usages each are sufficiently frequent (as with *worker operates machinery* versus *druglord operates cartel*), we see the stimuli falling in different usage groups,[7] but when a usage is not very frequent, or when many uses are blurred without clear clusters (which is the case with *check*), we cannot see a distinction. Again, this could be because some patterns influencing embedding locations may not be discernible in manual inspection.

9.4.8 Discussion

In this section we examined a number of nouns, both concrete and abstract, both polysemous and 'monosemous' (at least according to WordNet), along with some verbs, and explored how BERT organises tokens of these word in relation to each other. Our aim has been to get a clearer sense of the semantic generalisations (and their syntactic reflections) that contextualised language models are able to see. We found that tokens are often grouped by shared syntactic constructions or prototypical topics, where sometimes differences in syntax can be a cue to differences in topics. What is interesting for the study of word meaning is that division along these lines often leads to organisation according to particular stories, which are sometimes at odds with traditional sense distinctions. There are distinctions along the lines of prototypical affordances, events, and evaluative framing—the meaning of *train* (Section 9.4.6)

[7] Again, sometimes we see this in our BNC-based clusters, sometimes only in the Wikipedia tokens of the Context Atlas. These are genre effects, which we get even with such a general corpus as the BNC.

emerges from the situated perspectives from which trains are encountered—as containers, as conveyances, as moving objects.

This understanding invites new strategies for building applied machine learning systems in which it is important to control the types of meaning used. Work on the discovery and mitigation of biases in word embeddings is often framed in terms of untangling the 'truth' of the matter from spurious correlations in the data or separating out denotation from connotation. We hope that our analysis complicates the idea that there *is* some truth of the matter which can be abstracted from the perspectives according to which it is perceived.

Work on bias focuses on instances where the model exhibits discriminatory behaviour towards oppressed and minority groups. Ideally, debiasing language models constitutes a separation of fact from stereotype, denotation from connotation, events from their emotional framings. It is important to recognise that, given the close-knit relation between sense and story in word embeddings, denotation from connotation, debiasing efforts are likely to yield tools which allow for greater *control* over the narrative or story before they enable the elimination of story effects entirely. Further analyses should examine the kinds of stories by which contextualised language models organise culturally loaded terms, and experiment with tuning the model in or out to different kinds of story.

Through qualitative evaluation of what information BERT collapses and what distinctions are represented, we can approximate whether the representations accord with any extant theory of the lexicon. This is a question not of which theory is 'correct', but which one(s) the model has found suitable to its task of language modelling.

Some of the usage groups found by BERT match word senses that we find in dictionaries; others match distinct stories told around words in what might be called the same dictionary sense. But to what extent can we distinguish between word senses and stories? Story meanings are already inherent in the way many words are analysed into senses, for example in the senses of *admission*, which reflect admission as a price, as a permission, and as an action, and which are arguably distinguished by prototypical events. On the other hand, word embeddings miss generalisations like in the case of *wizard* discussed in Section 9.3, maybe because they miss information on real-world attributes that Hogwarts wizards and King Arthur wizards have in common.

In Section 9.2 we mentioned Elman's proposal of "lexical knowledge without a lexicon" (Elman, 2009). Is that what modern language models such as BERT are? The architecture of transformer encoders makes it possible to fulfill Elman's vision: tokens are never represented independently of their left and right neighbours, and representations can vary continuously across tokens of the same word.[8] It is certainly the case that BERT contextual embeddings encode lexical knowledge at the level

[8]This might also have been said of the RNN developed by Elman himself, but in an RNN the lexical knowledge is all mixed into a single vector which represents "the text so far". Transformer models such as BERT build representations which can be interpreted as corresponding to individual words in the text. Thus, the model produces on the fly representations of each token of a word on the basis of the surrounding context, without selecting a specific sense of the word.

of the lemma. BERT token vectors of a lemma vary, but they do tend to cluster together (apart from tokens of other lemmas). Mickus et al. (2020) measure how well BERT-space is divided by lemmas by calculating the silhouette score of token vectors with respect to other vectors of the same lemma.[9] In the last layer of `bert-large-uncased`, the authors determine, one quarter of BERT tokens would be better assigned to a word-type other than their own. Put another way, three quarters of tokens cluster together well with their own type. In the large majority of cases, the lemma would appear to be a significant level of organisation, much as with a lexicon. In our qualitative analysis, we observe some token groups that correspond to clear senses, another representation level important to the construction of a lexicon. But is it a lexicon, a collection of identifiable representations of senses that we can point to? This requires more analysis, with better tools than we currently have, but given how strongly some usage patterns stand out as token groups, it would be reasonable to assume that the language model has memorised frequent co-occurrence patterns as 'senses'.

9.5 CONCLUSION

Word embeddings that are computed from text capture statistical regularities about how words are used. It has often been noted, in the context of many different tasks, that this usage context includes what may be called stories connected to the words. When these pieces of evidence are put together, as we did in Section 9.3, they form a strong picture of a "story bias" that includes typical events, typical stories that people tell when they use a word, affordances (ways in which people interact with objects), as well as judgements and biases.

We think that it is useful to see all the different instances of "story bias" as part of a whole. With an eye on natural language processing applications, we see individual instances of bias and prejudice as types of stories that people tell around words, which may give us new ways of addressing bias. For linguistics and social science, word embeddings are also a resource, a distilled corpus, which we may be able to use to get additional insights about people's relation to words and objects from the way they talk about them; we have seen some glimpses of this in the qualitative analysis in Section 9.4.

There are some approaches in formal semantics that propose integrating word embeddings as fine-grained representations of word and phrase meaning, including Asher et al. (2016), Baroni, Bernardi, et al. (2014), Bernardy et al. (2018), Clark et al. (2008), Emerson (2018), McNally (2017), Muskens and Sadrzadeh (2018), and Zeevat et al. (2017). In these approaches, there is a tendency to interpret embeddings as denotational, or to create explicitly denotational embeddings (in Bernardy et al. (2018),

[9]The silhouette score is used to determine whether an observation $\mathbf{v}$ assigned to cluster $C_i \in C$ is well-assigned to that cluster, or whether it would be better off assigned to another cluster in C. It is calculated as a function of the 'cohesion' of the vector to other vectors in the assigned cluster and the 'separation' between the vector and other observations outside the cluster; a vector with a high silhouette score is maximally close to other observations in the cluster and maximally distant from observations in other clusters.

Herbelot (2020), and Herbelot and Copestake (2021) and Lappin (2021, ch. 6)). We think it will be interesting to integrate such frameworks with fine-grained representations of stories, and of story-modulated word meanings. As noted in Baroni, Murphy, et al. (2010), text-based word vectors seem not to notice object properties as strongly as human participants do, and denotational embeddings tend to foreground particularly those object properties. So it will be interesting to explore mechanisms for how object properties and "story context" interact in determining meaning in context – mechanisms that might underlie Fillmore's U-semantics: How is a "sentence story" constructed from the individual components of the sentence, and how can we both construct the whole from its parts and have the parts be modulated by the whole? Two formal/distributional approaches, Chersoni et al. (2019) and our own Erk and Herbelot (2020), explicitly model story effects at the sentence level, and could serve as a basis for studying these questions. In such a framework, word embeddings could serve either as representations of lexicalised, stored meanings, or as representations of meaning in a particular sentence context.

BIBLIOGRAPHY

Amrami, Asaf and Yoav Goldberg (May 2019). "Towards Better Substitution-Based Word Sense Induction". In: *arXiv:1905.12598 [cs]*. arXiv: 1905.12598 [cs].

Asher, Nicholas, Tim Van de Cruys, and Márta Abrusán (2016). "Integrating type theory and distributional semantics: a case study on adjective-noun compositions. Computational Linguistics". In: *Computational Linguistics* 42.4, pp. 703–725.

Baroni, Marco (2022). "On the proper role of linguistically-oriented deep net analysis in linguistic theorizing". In: *Algebraic Structures in Natural Language*. Ed. by Shalom Lappin and Jean-Philippe Bernardy. Taylor and Francis.

Baroni, Marco, Raffaella Bernardi, and Roberto Zamparelli (2014). "Frege in Space: A program for compositional distributional semantics". In: *Linguistic Issues in Language Technology* 9.6, pp. 5–110. URL: http://elanguage.net/journals/lilt/article/view/3746.

Baroni, Marco and Alessandro Lenci (2010). "Distributional Memory: A general framework for corpus-based semantics". In: *Computational Linguistics* 36.4, pp. 673–721.

— (2011). "How we BLESSed distributional semantic evaluation". In: *Workshop on GEometrical Models of Natural Language Semantics (GEMS)*. Edinburgh, Great Britain. URL: www.aclweb.org/anthology/W11-2501.

Baroni, Marco, Brian Murphy, Eduard Barbu, and Massimo Poesio (2010). "Strudel: A corpus-based semantic model based on properties and types". In: *Cognitive Science* 34.2, pp. 222–254. ISSN: 03640213. DOI: 10.1111/j.1551-6709.2009.01068.x.

Bernardy, Jean-Philippe, Rasmus Blanck, Stergios Chatzikyriakidis, and Shalom Lappin (2018). "A compositional Bayesian semantics for natural language". In: *First International Workshop on Language Cognition and Computational Models*. Santa Fe, NM, United States, pp. 1–10.

Bicknell, Klinton, Jeffrey L. Elman, Mary Hare, Ken McRae, and Marta Kutas (2010). "Effects of event knowledge in processing verbal arguments". In: *Journal of Memory and Language* 63.4, pp. 489–505. ISSN: 0749596X. DOI: 10.1016/j.jml.2010.08.004.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai (2016). "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29.

Bommasani, Rishi and Claire Cardie (Nov. 2020). "Intrinsic evaluation of summarization datasets". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8075–8096. DOI: 10.18653/v1/2020.emnlp-main.649.

Brysbaert, Marc, AB Warriner, and V Kuperman (2014). "Concreteness ratings for 40 thousand generally known English word lemmas". In: *BEHAVIOR RESEARCH METHODS* 46.3, pp. 904–911. ISSN: 1554-351X.

Burgess, Curt and Kevin Lund (Mar. 1997). "Modelling parsing constraints with high-dimensional context space". en. In: *Language and Cognitive Processes* 12.2-3, pp. 177–210. ISSN: 0169-0965, 1464-0732. DOI: 10.1080/016909697386844. URL: https://www.tandfonline.com/doi/full/10.1080/016909697386844 (visited on 04/23/2022).

Chersoni, E., E. Santus, L. Pannitto, A. Lenci, P. Blache, and C.-R. Huang (2019). "A structured distributional model of sentence meaning and processing". In: *Natural Language Engineering* 25.4, pp. 483–502. DOI: 10.1017/S1351324919000214.

Chronis, Gabriella and Katrin Erk (2020). "When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships". In: *Proceedings of CoNLL*.

Clark, S., B. Coecke, and M. Sadrzadeh (2008). "A compositional distributional model of meaning". In: *Proceedings of QI*. Oxford, UK, pp. 133–140.

Coenen, Andy, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, F. Viégas, and M. Wattenberg (2019). "Visualizing and measuring the geometry of BERT". In: *NeurIPS*.

Del Pinal, Guillermo (Apr. 2018). "Meaning, modulation, and context: A multidimensional semantics for truth-conditional pragmatics". en. In: *Linguistics and Philosophy* 41.2, pp. 165–207. ISSN: 0165-0157, 1573-0549. DOI: 10.1007/s10988-017-9221-z.

Del Tredici, Marco (2020). "Linguistic Variation in Online Communities: A Computational Perspective". PhD thesis. Universiteit van Amsterdam.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of NAACL*.

Elman, Jeffrey L. (2009). "On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon". In: *Cognitive Science* 33, pp. 547–582. ISSN: 03640213. DOI: 10.1111/j.1551-6709.2009.01023.x.

Emerson, Guy (2018). "Functional Distributional Semantics: Learning Linguistically Informed Representations from a Precisely Annotated Corpus". PhD thesis. University of Cambridge.

Erk, Katrin and Aurélie Herbelot (2020). How to marry a star: probabilistic constraints for meaning in context. arXiv 2009.07936. arXiv: 2009.07936 [cs.CL].

Erk, Katrin and Sebastian Padó (Oct. 2008). "A structured vector space model for word meaning in context". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pp. 897–906. URL: https://aclanthology.org/D08-1094.

Fillmore, Charles J. (1982). "Frame semantics". In: *Linguistics in the morning calm*. Ed. by The linguistic society of Korea. Seoul: Hanshin Publishing Co., pp. 111–137.

— (1985). "Frames and the semantics of understanding". In: *Quaderni di Semantica* 6, pp. 222–254.

Fillmore, Charles J., Christopher R. Johnson, and Miriam R.L. Petruck (2003). "Background to FrameNet". In: *International Journal of Lexicography* 16.3, pp. 235–250. ISSN: 09503846. DOI: 10.1093/ijl/16.3.235.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. London: Allen and Unwin.

Giulianelli, Mario, Marco Del Tredici, and Raquel Fernández (July 2020). "Analysing lexical semantic change with contextualised word representations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 3960–3973.

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni (2018). "Colorless green recurrent networks dream hierarchically". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1195–1205. URL: http://aclweb.org/anthology/N18-1108.

Herbelot, Aurélie (2020). "Re-solve it: simulating the acquisition of core semantic competences from small data". In: *Proceedings of CoNLL*.

Herbelot, Aurélie and Ann Copestake (2021). "Ideal words: A vector-based formalisation of semantic competence". In: *Künstliche Intelligenz*.

Hill, Felix, Douwe Kiela, and Anna Korhonen (Aug. 2013). "Concreteness and Corpora: A Theoretical and Practical Study". In: *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 75–83.

Kaneko, Masahiro and Danushka Bollegala (2021). "Debiasing pre-trained contextualised embeddings". In: *Proceedings of EACL*.

Kiela, Douwe, Felix Hill, and Stephen Clark (Sept. 2015). "Specializing word embeddings for similarity or relatedness". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 2044–2048. DOI: 10.18653/v1/D15-1242. URL: https://aclanthology.org/D15-1242.

Kozlowski, Austin C., Matt Taddy, and James A. Evans (Oct. 2019). "The geometry of culture: Analyzing the meanings of class through word embeddings". en. In: *American Sociological Review* 84.5, pp. 905–949. ISSN: 0003-1224. DOI: 10.1177/0003122419877135.

Kremer, Gerhard, Katrin Erk, Sebastian Padó, and Stefan Thater (2014). "What substitutes tell us - Analysis of an "All-Words" Lexical Substitution Corpus". In: *Proceedings of EACL*.

Kruszewski, Germán, Denis Paperno, and Marco Baroni (July 2015). "Deriving Boolean structures from distributional vectors". en. In: *Transactions of the Association for Computational Linguistics* 3, pp. 375–388. ISSN: 2307-387X. URL: https://transacl.org/ojs/index.php/tacl/article/view/616 (visited on 04/24/2022).

Kutuzov, Andrey, Erik Velldal, and Lilja Øvrelid (2017). "Tracing armed conflicts with diachronic word embedding models". In: *Proceedings of the Events and Stories in the News Workshop*. Vancouver, Canada.

Kuypers, Jim A. (2009). "Framing analysis". In: *Rhetorical Criticism: Perspectives in Action*. Ed. by Jim A. Kuypers. Lanham: Lexiongton Press, pp. 181–204.

Lake, Brenden and Marco Baroni (2018). "Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks". In: *35th International Conference on Machine Learning*. Vol. 7, pp. 4487–4499. ISBN: 978-1-5108-6796-3.

Lakoff, George (2004). *Don't Think of an Elephant!: Know Your Values and Frame the Debate*. Chelsea Green Publishing.

Lakoff, George and Mark Johnson (1980). *Metaphors we Live By*. Chicago: University of Chicago Press.

Lappin, Shalom (2021). *Deep Learning and Linguistic Representation*. London: Chapman and Hall.

Levin, Beth (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press. ISBN: 978-0-226-47532-5 978-0-226-47533-2.

Linzen, T., E. Dupoux, and Y. Goldberg (2016). "Assessing the ability of LSTMs to learn syntax-sensitive dependencies". In: *Transactions of the Association for Computational Linguistics* 4, pp. 521–535.

McNally, Louise (2017). "Kinds, descriptions of kinds, concepts, and distributions". In: *Bridging Formal and Conceptual Semantics. Selected Papers of BRIDGE-14*. Ed. by K Balogh and W Petersen, pp. 39–61.

McRae, Ken, Mary Hare, Jeffrey L. Elman, and Todd Ferretti (2005). "A basis for generating expectancies for verbs from nouns". In: *Memory and Cognition* 33.7, pp. 1174–1184. ISSN: 0090502X. DOI: 10.3758/BF03193221.

McRae, Ken and Kazunaga Matsuki (2009). "People use their knowledge of common events to understand language, and do so as quickly as possible". In: *Linguistics and Language Compass* 3.6, pp. 1417–1429. ISSN: 1749818X. DOI: 10.1111/j.1749-818X.2009.00174.x.

Mickus, Timothee, Denis Paperno, Mathieu Constant, and Kees van Deemter (Jan. 2020). "What do you mean, BERT?" In: *Proceedings of the Society for Compu-*

tation in Linguistics 2020. New York, New York: Association for Computational Linguistics, pp. 279–290. URL: https://aclanthology.org/2020.scil-1.35.

Mrkšić, Nikola, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young (June 2017). "Semantic Specialisation of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints". In: *arXiv:1706.00374 [cs]*. arXiv: 1706.00374. URL: http://arxiv.org/abs/1706.00374 (visited on 04/24/2022).

Muskens, Reinhard and Mehrnoosh Sadrzadeh (2018). "Static and dynamic vector semantics for lambda calculus models of natural language". In: *Journal of Language Modelling* 6.2, pp. 319–351.

Nickel, Maximilian and Douwe Kiela (May 2017). "Poincaré Embeddings for Learning Hierarchical Representations". In: *arXiv:1705.08039 [cs, stat]*. arXiv: 1705.08039. URL: http://arxiv.org/abs/1705.08039 (visited on 04/21/2022).

Padó, Sebastian and Mirella Lapata (July 2003). "Constructing semantic space models from parsed Corpora". In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan: Association for Computational Linguistics, pp. 128–135. DOI: 10.3115/1075096.1075113. URL: https://aclanthology.org/P03-1017 (visited on 04/24/2022).

Peirsman, Yves (2008). "Word Space Models of Semantic Similarity and Relatedness". In: *European Summer School in Logic, Language and Information (ESSLLI) Student Session*. Hamburg, Germany.

Potts, Christopher (2019). "A case for deep learning in semantics: Response to Pater". In: *Language*. DOI: 10.1353/lan.2019.0003.

Pustejovsky, James (Dec. 1991). "The Generative Lexicon". In: *Comput. Linguist.* 17.4, pp. 409–441. ISSN: 0891-2017.

Recanati, François (2003). *Literal Meaning*. Cambridge University Press.

— (2017). "Contextualism and Polysemy". In: *dialectica* 71.3, pp. 379–397. DOI: 10.1111/1746-8361.12179.

Reisinger, Joseph and Raymond Mooney (Oct. 2010). "A mixture model with sharing for lexical semantics". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, pp. 1173–1182.

Roller, Stephen, Katrin Erk, and Gemma Boleda (Aug. 2014). "Inclusive yet selective: Supervised distributional Hypernymy detection". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, pp. 1025–1036. URL: https://aclanthology.org/C14-1097 (visited on 04/24/2022).

Rosch, Eleanor (1973). "On the internal structure of perceptual and semantic categories". In: *Cognitive Development and the Acquisition of Language*. Ed. by T. E. Moore. New York: Academic Press, pp. 111–144.

Rosenfeld, Alex (2019). "Computational Models of Lexical Change". PhD thesis. University of Texas at Austin.

Sahlgren, Magnus (2006). "The Word-Space Model. Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-

dimensional Vector Spaces". PhD thesis. Stockholm University. URL: https://www.sics.se/~mange/TheWordSpaceModel.pdf.

— (2008). "The distributional hypothesis". en. In: *Rivista di Linguistica* 2.1, pp. 33–53.

Schütze, Hinrich (1998). "Automatic word sense discrimination". In: *Computational Linguistics* 24.1, pp. 97–123.

Thater, Stefan, Hagen Fürstenau, and Manfred Pinkal (Nov. 2011). "Word meaning in context: A simple and effective vector model". In: *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, pp. 1134–1143. URL: https://aclanthology.org/I11-1127.

Travis, Charles (1997). "Pragmatics". In: *A Companion to the Philosophy of Language*. Ed. by Bob Hale and Crispin Wright. Blackwell, pp. 87–107.

Trott, Sean, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider (2020). "(Re)construing Meaning in NLP". en. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5170–5184. DOI: 10.18653/v1/2020.acl-main.462. URL: https://www.aclweb.org/anthology/2020.acl-main.462 (visited on 10/21/2020).

Webson, Albert, Zhizhong Chen, Carsten Eickhoff, and Ellie Pavlick (2020). "Are "Undocumented Workers" the Same as "Illegal Aliens"? Disentangling denotation and connotation in vector spaces". In: *Proceedings of EMNLP*.

Wiedemann, Gregor, Steffen Remus, Avi Chawla, and Chris Biemann (2019). "Does BERT Make Any Sense? Interpretable word sense disambiguation with contextualized embeddings". In: *Konferenz zur Verarbeitung natürlicher Sprache / Conference on Natural Language Processing (KONVENS)*. Erlangen, Germany, p. 10.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Zeevat, Henk, Scott Grimm, Lotte Hogeweg, Sander Lestrade, and E. Smith (2017). "Representing the Lexicon". In: *Bridging Formal and Conceptual Semantics. Selected Papers of BRIDGE-14*. Ed. by K Balogh and W Petersen. Düsseldorf: düsseldorf university press, pp. 153–186.

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (Dec. 2015). "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books". In: *The IEEE International Conference on Computer Vision (ICCV)*.