What about Grammar? Using BERT Embeddings to Explore Functional-Semantic Shifts of Semi-Lexical and Grammatical Constructions

Lauren Fonteyn

Leiden University Centre for Linguistics, Department of English Language and Culture, Arsenaalstraat 1, 2311CT, Leiden, the Netherlands

Abstract

The aim of this short paper is to extend the application of embedding-based methodologies beyond the realm of lexical semantic change. It focuses on the use of unsupervised BERT-embeddings and uncertainty measures (Classification Entropy), and assesses whether (and how) they can be used to (semi-)automatically flag possible functional-semantic changes in the use of the construction [BE *about*] in the Corpus of Historical American English (COHA).

Keywords

Distributional Semantics, Corpus Linguistics, Grammatical change, embeddings, BERT

1. Introduction

Given its long tradition in computational and statistical research, it comes as no surprise that text-based humanities have embraced the use of distributional-semantic 'vectors' or 'embeddings' – i.e. (compressed) numeric vector representations of a word's contextual distribution that serve as a proxy of that word's meaning [e.g. 3]. In particular, in fields such as Corpus Linguistics – a subfield of Linguistics which grew around the computer-aided retrieval, annotation, and later also categorization of textual data [22] – there seems to be an unprecedented interest in vector-based distributional semantic models, which offer a quantifiable and data-driven means of studying meaning. This interest has been fueled further with the arrival of models equipped to create contextualized token vectors, which have eliminated the problems associated with polysemy/homonymy conflation [e.g 7].

In recent years, we have also witnessed a growth in the number of studies that have utilized either "count" or "predictive" vector models [2] to study historical and diachronic corpus data [also see 19]. Such studies, which often involve examination of nearest neighbours and cosine similarities between type- and/or token-vectors over time, have provided the key to a data-driven means of detecting and describing the diachronic trajectory of, predominantly, lexical change [e.g. 11, 9, 10, 13].

One consequence of this focus on lexical change is that, at present, the number of computational distributional semantic studies that consider the functional-semantic properties of more abstract, *grammatical* constructions seems disproportionate compared to the interest in

the phenomenon within the (Corpus) Linguistic community. Much like lexical semantics, the function(s) and underlying meaning(s) of grammatical structures – which are often notoriously polysemous and cover a broad range of nuanced, abstract meanings – are prone to change, and many linguists believe that the continued discovery, description and analysis of such changes plays an essential role in fleshing out our understanding of the mechanisms and motivations of language change. A logical and necessary continuation in the pursuit to automated semantic (shift) detection in large diachronic corpora would therefore involve further, more in-depth explorations of the extent to which embeddings can be employed to capture the (changing) functional-semantic properties of grammar. Given that that different components of language have differing diachronic dynamics [21, 12], they may pose different challenges in the development of unsupervised, embedding-based means of detecting diachronic change in large corpora – and it is only by exploring the functional-semantic properties of grammar (at whatever scale and level of detail is deemed reasonable) that these challenges and, consequently, their solutions, can be discovered.

The present study, then, sets out to do precisely that: explore some possible avenues of (semi-)automatically detecting diachronic functional-semantic changes of grammatical constructions. In that sense, the aim of this study is to extend the application of embeddings-based methodologies beyond the realm of lexical semantics, and further add to the budding research on whether and how token vectors [e.g. 14, 29] or contextualized embeddings [e.g. 4] can be used to study the functional-semantic properties of grammar. More specifically, the study, which focusses on embeddings created by means of BERT [8] makes the following contributions:

1. It demonstrates that unsupervised BERT-embeddings can successfully be employed to identify the different functions or 'usage types' of the grammatical construction [BE *about*] in English. **2.** It formulates three expected functional-semantic developments of [BE *about*] based on linguistic literature, and assesses whether (and how) BERT-embeddings can be used in combination with Entropy Difference measures and time-sensitive t-SNE plots to (semi-)automatically detect these changes in the Corpus of Historical American English [COHA, 5]. **3.** It discusses potential pitfalls associated with the 'present-day' bias of the explored methods.

2. Data and Methodology

2.1. Corpus and Data

As a case study, I focus on the recent diachronic development of [BE *about*]. The data for this study has been gathered from COHA, a corpus containing over 400 million words of American English text written between 1810 and 2009. The corpus is balanced for genre (Fiction, Non-Fiction, Magazine, Newspaper (after 1860)) and subgenre (e.g. prose, poetry, drama, etc. (Fiction)). Such (sub-)genre balance is said to ensure that any changes observed in the corpus will not simply be "artifacts of a changing genre balance"[5].

While [BE *about*] can be used with a wide range of more abstract, grammatical meanings (which vary substantially in frequency as well as distinctiveness), the present study will focus on its three most frequent usage types: the futurate (e.g. *I am about to leave*), approximative (e.g. *There were about ten cats in the room*), and descriptive use (e.g. *This song is about love*). These usage types align with the higher-order sense categories distinguished in the Oxford English Dictionary (OED) [1], which includes a group of tokens expressing approximation, another signalling connections/relations (in descriptions), and another containing tokens (predominantly) followed by a *to*-infinitive expressing intention or imminent future.

Figure 1 shows a two-dimensional t-SNE mapping of the embeddings of 1,000 examples of [be *about*], collected from COHA (decade 2000-2009). The examples were manually annotated at the level of granularity outlined above. The sample contained 448 futurate, 279 approximative, and 225 descriptive uses of [BE *about*]. The other 48 examples include irrelevant structures (e.g. due to mistakes in COHA's tagging of the possessive marker *'s* as a finite form of the verb *be*), as well as three much more infrequent usages of [BE *about*]. These infrequent types consist of spatial uses (e.g. *She must be somewhere about*), the fixed expression *that's about it*, as well as an actional use which can be paraphrased as 'occupied/dealing with' (chiefly found in more or less fixed expressions such as *be about your business* and *know what one is about*).
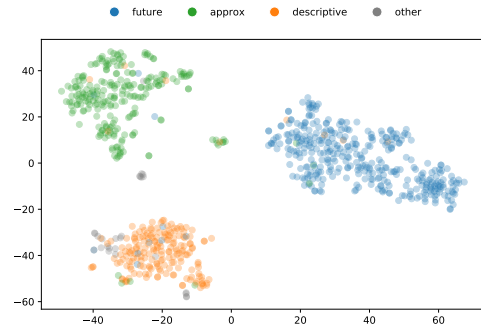

Figure 1: t-SNE of [BE *about*] token embeddings. There are three largely distinct usage types: futurate, approximative, and descriptive. Less frequent senses (e.g. spatial) are marked as 'other'.

To assess whether BERT-embeddings can be used to identify and distinguish the three usage types under scrutiny, we can conduct a simple 'sense distinction task'. For this task, the embeddings of the 952 'relevant' examples and their accompanying usage type labels were used as the training set. The procedure involved fitting a logistic regression classifier with L2 regularization (as implemented in Scikit-learn [25]) on the embeddings created for the labelled training set. Subsequently, the classifier was applied to unseen test set of 200 examples from each of the 20 decades covered in COHA (with the exception of the first decade (1810-1819), which contained only 73 tokens of [BE *about*]). In sum, the test set includes 3,873 unseen examples for which a usage type label was predicted. The predicted labels were then assessed against the true usage types of the tokens.

Based on the manual assessment of the predicted usage type labels, it appears that the model performs quite well at distinguishing the three main usage types of [BE *about*]. For the final decade of COHA, only 5 out of 200 tokens had been mislabelled, indicating a classification accuracy of 0.975. Notably, the classification accuracy of the model remains high with older data, ranging between 0.904 and 0.975 (as is shown in Figure 2). At the same time it can, perhaps unsurprisingly, be noticed that the accuracy (slightly) decreases as the linguistic data ages.
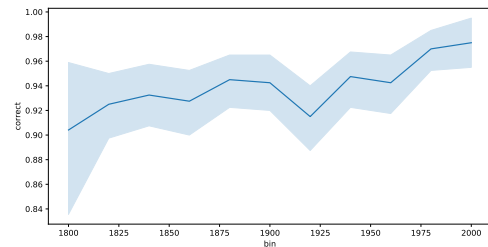

Figure 2: Accuracy based on labelled test set of 3,873 unseen examples from the 20 decades included in COHA (1810-2009). The classification accuracy ranges between 0.904 in the earliest decade, and 0.975 in the most recent decade.

Overall, these accuracy scores are encouraging, in that they highlight the efficiency of BERT-embeddings in identifying specific usage types of a grammatical construction, such as [BE *about*], in a large diachronic corpus. While that is an interesting observation in itself, the remainder of this paper will focus on assessing whether BERT-embeddings can be used to (semi-)automatically detect functional-semantic changes of grammatical constructions.

2.2. Method

This paper's methodological set-up starts from the assumption that changes in a word or construction's distribution, which can be captured in compressed numerical representations such as BERT-embeddings, may indicate changes in its functional or semantic range. Following earlier proposals using embeddings to detect lexical-semantic change in large diachronic corpora [e.g. 11], this study will investigate whether known changes that have affected the [BE *about*] construction can be detected by means of an embedding-based methodology combined with Entropy Difference measures. More specifically, the task is conceptualized as follows: in the case of lexical items, expansions or reductions in their distributional properties are often equated with expansions or reductions of their possible interpretations – or, in other words, with increases or decreases *uncertainty* regarding the exact interpretation of the lexical item. To measure whether the "uncertainty over possible interpretations varies across time intervals", then, one can "compute the difference in entropy between the two usage type distributions in these intervals" [11]. An increase in entropy over time could be used to signal that the number of interpretations of a word has increased (e.g. due to the emergence of a new usage type), whereas a decrease in entropy would signal that the opposite has occurred (e.g. loss of a usage type). In principle, it is possible to extend this approach to the study of grammatical items.

However, unlike the distributional changes that accompany lexical-semantic changes of well-known examples such as *broadcast* or *gay*, the distributional changes witnessed for [BE *about*] – and many grammatical constructions like it – seem to proceed in a protracted sequence of small steps, often spanning over several centuries [21]. The question, then, is whether we can manipulate the use of Entropy Difference measures to detect not only the emergence or loss of an entire usage type, but also to detect any small-scale shifts within a usage-type.

The approach tested here works assumes that the researcher is interested in determining whether any of the usage types they distinguished has changed in the time span covered by their corpus with respect to a single reference point (for instance: Present-day data). For [BE *about*], the procedure involved fitting a logistic regression classifier on the embeddings created for the 952 present-day tokens, and applying it to a test set of 200 examples from each decade included in COHA (cf. Section 2.1). Subsequently, for each test token x_i and each label $y \in Y$, the conditional probability $p(y|x_i)$ is computed to assess the uncertainty of the classifier in labelling the unseen examples. The resulting conditional probability over each label for each test token is then summarized in an entropy score, H:

$$H(x_i) = -\sum_{y \in Y} p(y|x_i) * \log p(y|x_i) \tag{1}$$

If the entropy score changes over the 20 decades included in COHA, one could take this as an indication that the distributional properties of the test tokens in a particular category have shifted over time. Such shifts could be indicative of proper (subtle) functional-semantic change, or of increased or decreased use of the construction under scrutiny in different genres or text types. Conversely, if the distributional properties of a linguistic item or construction have not changed over the time span covered by the corpus, we should not expect to witness any changes in the certainty by which the model classifies tokens to its true usage type category (i.e. entropy).

Of course, it is difficult to assess the effectiveness of this method if we do not know whether [BE *about*] has actually undergone any functional-semantic changes in the period covered by COHA. As a means of assessment, then, I will first formulate the expected results for each of

the construction's three main usage types based on prior literature.

2.2.1. futurate [BE *about*]

The first usage type under consideration is the futurate use of [BE *about*], illustrated in examples (2)-(4). In present-day English reference grammars, the phrasal expression *be about to* is commonly described as a so-called 'quasi-auxiliary' which can be used to make a temporal reference to the future [28]. As such, *be about to* can be considered near-synonymous to other English (quasi-)auxiliaries such as *will*, *shall*, and *be going to* (see example (2)). Notably, however the use *be about to* is commonly said to convey a strong sense of immediacy, and it has been suggested that the construction may in fact have more affinity with "aspectualizing expressions such as *begin to/start to V* than form expressing futurity" [16].

(2) Sheen, yes THAT Charlie Sheen, *is about* to become the best-paid actor in a comedy on TV. (2006, COHA)

(3) Just as I *am about* to step into the shower, the phone rings. It is George Stephanopoulos. (2002, COHA)

(4) But when he saw the sneer on St. Exeter's face, Logan knew things *were about* to get much worse. (2004, COHA)

From the relatively sparse number of accounts on the diachronic development of [BE *about*], it can be concluded that the construction had already grammaticalized into a marker of (immediate) future by the 19th century [23, 16] (the suggested time of the first, full-blown futurate uses ranging between the late 15th or 16th century [17] and the late 18th century [31], when the construction started occurring with, for example, inanimate subjects and non-intention verbs, as in (4)). As such, the semantic and distributional changes typical of a grammaticalizing construction (e.g. bleaching [30] or host-class expansion [15]) most likely pre-date COHA.

However, the [BE *about*] future did undergo a distributional change: in Present-day English, the phrase almost exclusively occurs with a to-infinitive complement clause, whereas 19th century and early 20th century texts also contain a variant with an *ing*-clause complement (as in (5)-(6)).

(5) I really thought all my bones were disjointed, and that my soul *was about* taking a last farewell of my poor body. (1812, COHA)

(6) I was trying to sleep, and just as I *was about* succeeding Henderson called out: '[...]'. (1902, COHA).

If accurate, the method should detect that the distributional properties of futurate [BE *about*] have narrowed slightly over the course of the 19-20th century.

2.2.2. approximative *about*

[BE *about*] often occurs as a marker of approximation. Note that the use of approximative *about* is not restricted to contexts with the verb *be*. The current sample therefore only represents a subset of the possible occurrences of approximative *about*, some examples of which are listed in (7)-(9):

(7)   When he *was about* 12, his parents left him and his siblings at an orphanage for five months. (2006, COHA)

(8)   Though you hate to say anyone is recession-proof, U2 *is about* as close to that as you can get. (2009, COHA)

(9)   It *was about* at that time that Takemore disappeared from the township too. (2001, COHA)

With respect to its diachronic development, it has been shown that the spatial preposition *about* approximative use of *about* – much like the near-synonymous use of *around* and various other adpositions in other languages – developed into "approximative qualifiers of numerical expressions and other amount expressions" [27] sometimes called "rounders" [24]. This process is shown to have started around the beginning of the Middle English period (1250-1500) [6], and the establishment of approximative *about* in the wide range of contexts in which it can presently occur pre-dates COHA with a very large margin. The inclusion of approximative [BE *about*] is therefore not so much motivated by the fact that it has been stable over the course of the 19th–20th century. As such, no changes in classification uncertainty should be attested.

2.2.3. descriptive [BE *about*]

In the third and final category, we find a group of what we could call 'descriptive' uses of [BE *about*], in which the phrase [BE *about*] can be roughly paraphrased as 'regards', 'is (primarily) concerned with'. Its occurrences commonly involve clarifications of why situations are occurring (e.g. (10)), as well as descriptions of the theme, topic (or plot) of conversations, books, and films (e.g. (11)-(12)):

(10)   ... he wonders what the fighting *is about* and and who is fighting whom. Is it North against South again? (2008, COHA)

(11)   The latest news *is about* Amanda! Haven't you heard? (2000, COHA)

(12)   *Mars and Venus Collide* is not just about men understanding women. It *is also about* women understanding themselves and learning how to ask effectively for the support they need. (2008, COHA)

Unlike with the previous two usage types, the number of historical and diachronic accounts that treat the descriptive use of [BE *about*] are not sparse, but virtually non-existent. Still, a quick scan of the dated examples listed in the Oxford English Dictionary suggests that the use of [BE *(all) about*] with an animate subject (e.g. a person, organization, or company) to describe what the subject is 'primarily concerned with' or 'fond of' constitutes a relatively recent (i.e. mid-20th century) phenomenon (e.g. (13)-(14)).

(13)   ... give him your authenticity spiel and how radio should be *all about* the music. (2009, COHA)

(14)   I'm *all about* the blindfold. There's something intensely sensual about not knowing where you're going (2005, COHA).

In other words, the types of elements that can occur in the subject slot in the descriptive subtype of [BE *about*] appear to have expanded during the time period captured by COHA, which may indicate that the descriptive use of [BE *about*] came to cover a broader semantic range. Given that the distributional properties of the descriptive use in the early 19th century were quite different from the present-day, we would again expect to find a shift classification uncertainty.

2.2.4. Summary: expected shifts

In sum, the (in)stability of the classification entropy should reflect the following: The distributional properties of futurate [BE *about*] have changed slightly during the 20th century. Having lost the ability to occur with *ing*-complements, it seems that the futurate use has narrowed. As with the futurate use, the distributional properties of the descriptive use have changed. More specifically, the descriptive use has expanded or broadened. By contrast, the distributional properties of the approximative use have remained stable.

3. Results: detecting changes in [BE *about*]

As a starting point, it is worth considering the distributional range of the [BE *about*] construction as a whole. It appears that entropy has indeed increased over time (from 0.76 to 1.09), and, as such, it could be flagged as a case of semantic broadening [11]. To understand what has been captured here precisely, it is worth considering the relative frequency of the different usage types across time: with the descriptive category growing more frequent, there are effectively there major usage types by the end of the 20th century (whereas there were only two at the start of the 19th century).

What such a general test does not reveal is whether there have been any changes within these major usage types. For instance, there is no immediate indication that anything has changed about the descriptive use besides its overall frequency.

Figure 3: Entropy based on logistic regression classifier, fitted on the embeddings created for a training set of 952 labelled present-day tokens, and applied to a test set of 200 examples from each decade in COHA. For each test token and each label, the conditional probability has been computed to assess the uncertainty of the classifier in labelling the unseen examples. The resulting conditional probability over each label for each test token is then summarized in an entropy score, which gradually decreases for the descriptive use, and, to a lesser extent, for the futurate use.

3.1. changing usage types

To assess whether embeddings can be used to detect changes within the three major usage types, one could explore treating the semantic change detection task as a classification experiment. As explained in Section 2.2, the goal of the classification entropy test is to establish whether any of the usage categories of [BE *about*] has changed compared to a present-day reference point. We expect to find Entropy Differences over time in two of the three usage types: the descriptive use, and, to a weaker extent, the futurate use. This expectation appears to be borne out (Figure 3).

A fair point that can be raised is that the attested decrease in uncertainty does not in fact signal that the usage of a construction has changed, but rather reflects a decrease in the extent to which embeddings created by unsupervised BERT (trained on predominantly Present-day English data) as the linguistic material more generally becomes older (and, consequently, less familiar). Still, the latter explanation loses at least some of its plausibility when considering the expected, relative stability of the approximative use. It is furthermore reassuring that the slight
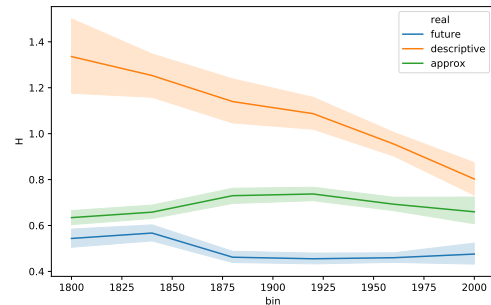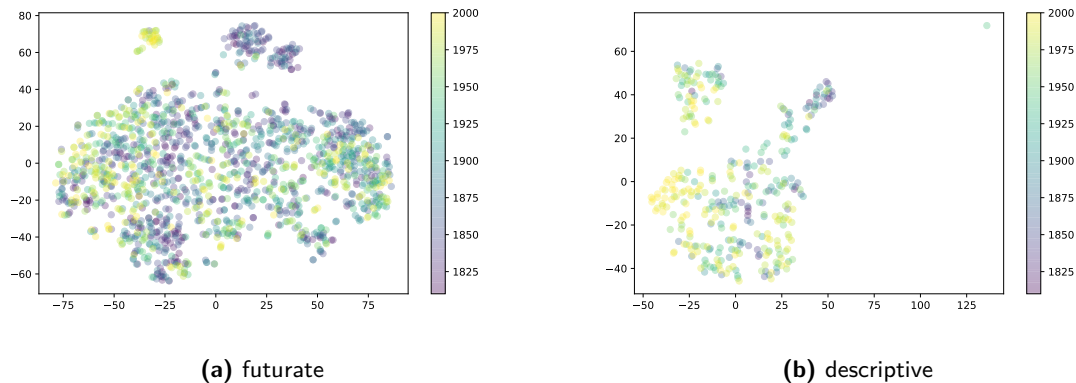
(a) futurate   **(b)** descriptive

Figure 4: Time-sensitive t-SNE of [BE *about*]. For the futurate use, two archaic/obsolete clusters and one recent cluster can be discerned. For the descriptive use, the pattern suggests expansion with more recent token groupings.

increase in classifier certainty appears to coincide with the decline of [BE *about* V*ing*], which renders futurate [BE *about*] more uniform and, consequently, more unambiguously recognizable.

The suggested shifts are also evident from the visualizations in Figure 4a and 4b, which present what one could call a 'time-sensitive t-SNE' representation of the futurate and descriptive use of [BE *about*]. If one wishes to avoid the use of an indirect, present-day reference point to query a corpus for potential constructional changes, it would of course also be possible to examine token groupings (as apparent from a time-sensitive version of t-SNE [20] or another type of dimension reduction technique) that appear to be specific for a particular time-period.

Using Figure 4, for example, one can examine the token groupings which are dated towards the beginning of the corpus (representing groupings of the [BE *about* V*ing*] pattern), or the smaller, markedly recent grouping top center left (containing solely negative uses, expressing absence of intent, e.g. *Wang's not about to forgive you* (2007, COHA)). Note that the relative size of the time-specific token groupings discussed here may affect the extent to which summary statistics (such as the average pairwise distance between tokens or the silhouette score of clusters over time) capture their emergence or disappearance.

4. Discussion and Conclusion

All in all, the present assessment of the use BERT embeddings and uncertainty measures to detect functional-semantic change in grammatical constructions seems largely positive. However, there are a number of potential pitfalls that must be addressed.

A first, smaller point that could be raised concerns the attested change in the [BE *about*] futurate. One may argue that this is, in fact, a formal rather than a functional-semantic change. What we witness is a reduction in the variability of (near-synonymous) complement clause types following the futurate, but this distributional change does not (straightforwardly) mark a shift in the futurate's meaning. Still, what has been detected is of value to linguists, as the model has picked up a reduction in semasiological variation between a current and a currently obsolete complementation pattern.

Second, one may, as pointed out earlier, not be fully convinced that uncertainty measures determined by an unsupervised, present-day neural language model are the most reliable mea-

sure to detect semantic shift. First, it is unclear to what extent the fact that BERT has been pre-trained on present-day English material affects its performance with respect to the gradually aging data. Second, if tokens from a decade d are flagged as yielding a high degree of classifier uncertainty with respect to the reference point r, one should not be too quick to assume semantic change proper has taken place: in fact, this may be due to a difference in how the tokens are distributed across (sub)genres in d and r. In this study, I minimized both problems somewhat, but the concerns are still legitimate. With respect to genre variation, it helps to work with a carefully balanced corpus such as COHA (if available), or to try and incorporate meta-information on (sub)genre in the model [e.g. 26]. With respect to the possible 'present-day bias' of the pre-trained model, it is reassuring to see that attest stability with usage types that are not known to have changed, and that similar conclusions on possible distributional shifts can be arrived at by examining time-sensitive t-SNE plots.

However, it is important to stress that the explored approach solely considers uncertainty with respect to the present-day reference point: while the classification entropy test successfully pointed out that the descriptive use of [BE *about*] had undergone some changes, the decrease in entropy cannot be equated to semantic narrowing. Instead, given that the descriptive use of [BE *about*] rather seems to have shifted and broadened, the classification entropy test merely indicates that the tokens have become more like the present-day examples and linguistic material the model has been trained with. Second, the explored approach is limited in the sense that the number (and nature) of usage types is imposed anachronistically to non-present-day data. Since the procedure relies on a single reference point, it will not straightforwardly flag any usage types that are absent in the training set, and it may erroneously impose the pre-defined category labels onto tokens representing obsolete usages. A further indication that the method may be problematically biased towards present-day language can be found when the model's actual classification errors are considered in more detail. In the case of the [BE *about*] futurate, the overall classification accuracy is remarkably high at 0.985, with only 22 of the 1517 examples not being recognized as futurates. On closer inspection of those 22 mistakes, it appears that 19 of them involve the now obsolete [BE *about* V*ing*] pattern. Given that there are 102 examples of [BE *about* V*ing*] in the test set, this amounts to an error rate of 18.6%. Furthermore, the mistakes are of the type illustrated in (15), where a Present-day descriptive interpretation (i.e. Napoleon was fond of going to war with England) is erroneously imposed:

(15)   Jefferson obtained the consent of Congress to make an effort to buy New Orleans and West Florida, and sent Monroe to aid our minister in France in making the purchase. When the offer was made, Napoleon *was about going to war* with England, and, wanting money very much, he in turn offered to sell the whole province to the United States. (1897, COHA)

Because the use of data-driven, automated methods of semantic annotation and analysis is appealing to researchers precisely because it could help avoid such anachronistic interpretations of historical language [e.g. 29], it is of course unfortunate that they still occur at a reasonably high rate. Yet, it should still be acknowledged that the very fact that word and phrase embeddings created by BERT did succeed in recognizing different grammatical usage types in Present-day language inspires hope that these problems can be tackled when models such as these are trained on contemporary linguistic material and (following proposals such as [18]) made dynamic.

Acknowledgments

References

[1] "about, adv., prep.1, adj., and int." In: *Oxford English Dictionary Online*. Oxford University Press, 1990. URL: oed.com/view/Entry/527.

[2] M. Baroni, G. Dinu, and G. Kruszewski. "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors". en. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 238–247. DOI: 10.3115/v1/P14-1023. URL: http://aclweb.org/anthology/P14-1023 (visited on 01/26/2020).

[3] G. Boleda. "Distributional Semantics and Linguistic Theory". en. In: *Annual Review of Linguistics* 6.1 (Jan. 2020). arXiv: 1905.01896, pp. 213–234. ISSN: 2333-9683, 2333-9691. DOI: 10.1146/annurev-linguistics-011619-030303. URL: http://arxiv.org/abs/1905.01896 (visited on 04/15/2020).

[4] S. Budts and P. Petré. "Putting connections centre stage in diachronic construction grammar". In: *Nodes and Networks in Diachronic Construction Grammar*. Ed. by L. Sommerer and E. Smirnova. Amsterdam: John Benjamins, 2020, pp. 317–352.

[5] M. Davies. *Corpus of Historical American English (COHA)*. Version V1. 2015. DOI: 10.7910/DVN/8SRSYK. URL: https://doi.org/10.7910/DVN/8SRSYK.

[6] H. De Smet. "The course of actualization". en. In: *Language* 88.3 (2012), pp. 601–633. ISSN: 1535-0665. DOI: 10.1353/lan.2012.0056. URL: http://muse.jhu.edu/content/crossref/journals/language/v088/88.3.de-smet.html (visited on 01/26/2020).

[7] G. Desagulier. "Can word vectors help corpus linguists?" en. In: *Studia Neophilologica* 91.2 (May 2019), pp. 219–240. ISSN: 0039-3274, 1651-2308. DOI: 10.1080/00393274.2019.1616220. URL: https://www.tandfonline.com/doi/full/10.1080/00393274.2019.1616220 (visited on 05/17/2020).

[8] J. Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". en. In: *Proceedings of NAACL-HLT 2019*. Minneapolis, Minnesota, June 2019, pp. 4171–4186.

[9] H. Dubossarsky et al. "Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 457–470. DOI: 10.18653/v1/P19-1044. URL: https://www.aclweb.org/anthology/P19-1044 (visited on 06/28/2020).

[10] S. Eger and A. Mehler. "On the Linearity of Semantic Change: Investigating Meaning Variation via Dynamic Graph Models". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 52–58. DOI: 10.18653/v1/P16-2009. URL: https://www.aclweb.org/anthology/P16-2009 (visited on 07/21/2020).

[11] M. Giulianelli, M. Del Tredici, and R. Fernández. "Analysing Lexical Semantic Change with Contextualised Word Representations". In: *arXiv:2004.14118 [cs]* (Apr. 2020). arXiv: 2004.14118. URL: http://arxiv.org/abs/2004.14118 (visited on 06/28/2020).

[12] S. J. Greenhill et al. "Evolutionary dynamics of language systems". en. In: *Proceedings of the National Academy of Sciences* 114.42 (Oct. 2017), E8822–E8829. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1700388114. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1700388114 (visited on 07/21/2020).

[13] W. L. Hamilton, J. Leskovec, and D. Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change". In: *arXiv:1605.09096 [cs]* (Oct. 2018). arXiv: 1605.09096. URL: http://arxiv.org/abs/1605.09096 (visited on 06/28/2020).

[14] M. Hilpert and D. Correia Saavedra. "Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims". en. In: *Corpus Linguistics and Linguistic Theory* 0.0 (Sept. 2017). ISSN: 1613-7027, 1613-7035. DOI: 10.1515/cllt-2017-0009. URL: http://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2017-0009/cllt-2017-0009.xml (visited on 05/16/2020).

[15] N. P. Himmelmann. "Lexicalization and grammaticization: opposite or orthogonal?"" en. In: *What Makes Grammaticalization: A Look from Its Components and Its Fringes.* Ed. by W. Bisang, N. P. Himmelmann, and B. Wiemer. Berlin: Mouton de Gruyter, 2004, pp. 21–44.

[16] S. Höche. "I am about to die vs. I am going to die: A usage-based comparison between two future-indicating constructions". In: *Converging Evidence: Methodological and Theoretical Issues for Linguistic Research.* Ed. by D. Schönefeld. Amsterdam: John Benjamins Publishing Company, 2011, pp. 115–142.

[17] B. Jirsa. "Synchronic Applications for Diachronic Syntax: The Grammaticalization of *to be about to* in English". en. In: *Colorado Research in Linguistics* 15 (1997). ISSN: 1937-7029. DOI: 10.25810/5h5t-xg32. URL: https://journals.colorado.edu/index.php/cril/article/view/231 (visited on 07/21/2020).

[18] Y. Kim et al. "Temporal Analysis of Language through Neural Language Models". en. In: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science.* Baltimore, MD, USA: Association for Computational Linguistics, 2014, pp. 61–65. DOI: 10.3115/v1/W14-2517. URL: http://aclweb.org/anthology/W14-2517 (visited on 06/28/2020).

[19] A. Kutuzov et al. "Diachronic word embeddings and semantic shifts: a survey". In: *Proceedings of the 27th International Conference on Computational Linguistics.* Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1384–1397. URL: https://www.aclweb.org/anthology/C18-1117 (visited on 07/20/2020).

[20] L. van der Maaten and G. Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[21] C. Mair and G. Leech. "Current Changes in English Syntax". en. In: *The Handbook of English Linguistics.* Ed. by B. Aarts and A. McMahon. Malden, MA, USA: Blackwell Publishing, Jan. 2006, pp. 318–342. DOI: 10.1002/9780470753002.ch14. URL: http://doi.wiley.com/10.1002/9780470753002.ch14 (visited on 09/19/2020).

[22] T. McEnery and A. Hardie. *The History of Corpus Linguistics*. en. Oxford University Press, Mar. 2013. DOI: 10.1093/oxfordhb/9780199585847.013.0034. URL: http://oxfordh andbooks.com/view/10.1093/oxfordhb/9780199585847.001.0001/oxfordhb-97801995858 47-e-34 (visited on 06/29/2020).

[23] J. Mee. "The evolution of constructions: The case of *be about to*". en. MA dissertation. University of New Mexico, 2013, p. 138.

[24] W. Mihatsch. "The Diachrony of Rounders and Adaptors: Approximation and Unidirectional Change". en. In: *New Approaches to Hedging*. Ed. by G. Kaltenböck, W. Mihatsch, and S. Schneider. BRILL, Jan. 2010, pp. 93–122. ISBN: 978-90-04-25324-7. DOI: 10.1163 /9789004253247_007. URL: https://brill.com/view/book/edcoll/9789004253247/B97890 04253247-s007.xml (visited on 07/20/2020).

[25] F. Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[26] V. Perrone et al. "GASC: Genre-Aware Semantic Change for Ancient Greek". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change* (2019). arXiv: 1903.05587, pp. 56–66. DOI: 10.18653/v1/W19-4707. URL: http://arxiv.org/abs/1903.05587 (visited on 09/19/2020).

[27] F. Plank. "Inevitable reanalysis: From local adpositions to approximative adnumerals, in German and wherever". en. In: *Studies in Language* 28.1 (2004), pp. 165–201. ISSN: 0378-4177, 1569-9978. DOI: 10.1075/sl.28.1.07pla. URL: http://www.jbe-platform.com/c ontent/journals/10.1075/sl.28.1.07pla (visited on 07/20/2020).

[28] R. Quirk et al. *A Comprehensive Grammar of the English Language*. London: Longman, 1985.

[29] E. Sagi, S. Kaufmann, and B. Clark. "Tracing semantic change with Latent Semantic Analysis". en. In: *Current Methods in Historical Semantics*. Ed. by K. Allan and J. A. Robinson. Berlin, Boston: DE GRUYTER, Jan. 2011, pp. 161–183. ISBN: 978-3-11-025290-3. DOI: 10.1515/9783110252903.161. URL: https://www.degruyter.com/view/boo ks/9783110252903/9783110252903.161/9783110252903.161.xml (visited on 01/26/2020).

[30] E. E. Sweetser. "Grammaticalization and Semantic Bleaching". In: *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society* (1988), pp. 389–405.

[31] T. Watanabe. "Development and grammaticalization of Be About To: An analysis of the OED quotations". In: *Aspects of the History of English Language and Literature*. Ed. by Y. Nakao and M. Ogura. Frankfurt am Main: Peter Lang, 2010, pp. 353–365.