Let's get into it: using contextualized embeddings as retrieval tools

Fonteyn, L.

Let's get into it

Using contextualized embeddings as retrieval tools

Lauren Fonteyn
Leiden University

This squib briefly explores how contextualized embeddings – which are a type of compressed token-based semantic vectors – can be used as semantic retrieval and annotation tools for corpus-based research into constructions. Focusing on embeddings created by the Bidirectional Encoder Representations from Transformer model, also known as 'BERT', this squib demonstrates how contextualized embeddings can help counter two types of retrieval inefficiency scenarios that may arise with purely form-based corpus queries. In the first scenario, the formal query yields a large number of hits, which contain a reasonable number of relevant examples that can be labeled and used as input for a sense disambiguation classifier. In the second scenario, the contextualized embeddings of exemplary tokens are used to retrieve more relevant examples in a large, unlabeled dataset. As a case study, this squib focuses on the INTO-INTEREST construction (e.g. *I'm so into you*).

Keywords: distributional semantics, BERT, corpus linguistics, data retrieval, prepositions

1. Introduction

Over the past few decades, the Constructionist Approach has become known as one that studies constructions – the basic buildings blocks of language (e.g. Croft 2001) – in a way that harmonizes the study of their *form, function*, and *frequency* (e.g. Hilpert 2013). The interest in frequency is particularly evident in constructionist studies that rely on a corpus-based methodology, which aim to empirically verify existing or propose new hypotheses based on the statistical exploration or analysis of formally and functionally/semantically annotated data sets (e.g. the contributions in Yoon and Gries 2016). In tackling the challenge of harmonizing such

https://doi.org/10.1075/bjl.00035.fon
Belgian Journal of Linguistics Volume 34 (2020), pp. 66–78. ISSN 0774-5141 | E-ISSN 1569-9676

a quantitative approach with the study of more elusive aspects of constructions, such as their functional-semantic properties, the corpus-constructionist approach has turned to, for instance, the nowadays established technique of collostructional analysis (e.g. Stefanowitsch and Gries 2003), and, most recently, the use of vector-based distributional semantic models (henceforth DSMs).

The application of vector-based DSMs has helped bring the wealth of functional-semantic information enclosed in corpus data to the fore (Louwerse and Zwaan 2009; Gupta et al. 2015; Sommerauer and Fokkens 2018; Bolukbasi et al. 2016), and DSMs have been used as tools to empirically test hypotheses on, for instance, syntactic productivity (Perek 2016) and asymmetric priming (Hilpert and Correia Saavedra 2017). Initially, the potential of DSMs seemed somewhat limited because the distributional properties of polysemous and homonymous items were conflated into a single vector, but recently these issues have been eliminated with the coming of so-called 'contextualized' models or 'token-based' vector models (e.g. Heylen et al. 2015; Peters et al. 2018; De Pascale 2019).

In this squib, I will briefly explore how contextualized embeddings – which are a type of compressed token-based semantic vectors – can be used as semantic retrieval and annotation tools. More specifically, I will focus on the use of the Bidirectional Encoder Representations from Transformer model, also known as 'BERT' (Devlin et al. 2019). Homing in on BERT's ability to distinguish different senses of constructions involving the preposition *into*, I aim to demonstrate how BERT could serve as a powerful tool to retrieve relevant examples from large corpora.[1]

2. Vector-based distributional semantic models

In general terms, vector-based DSMs can be divided into Count Models and Predict(ive) Models (Baroni, Dinu, and Kruszewski 2014). Most Constructionist applications of semantic vectors have thus far relied on count models, which offer a relatively transparent, straightforward means of operationalizing the distributional hypothesis (for an accessible explanation, see Heylen et al. 2015). Put simply, building count models typically involves retrieving all the co-occurrence counts of a target construction in a corpus, which can be represented in a vector format which is optimized in some way (e.g. by reweighting/omitting function words, applying dimensionality reduction, etc.). By contrast, context-predicting models (a cover-term for an extremely varied and rapidly growing group of models with different architectures) are designed to construct and optimize vectors as part of a learning

1. A jupyter notebook with all data and python code to replicate the analysis is available at https://github.com/LFonteyn/CxG_squib_into.

task, resulting in compressed numeric vector representations often referred to as 'embeddings'.

Predictive (neural) DSMs are particularly popular in the Machine Learning community, as they often outperform count models in a range of NLP tasks (e.g. Baroni, Dinu, and Kruszewski 2014). However, the improved performance brought by neural models has "come at the cost of our understanding of the system" (Linzen et al. 2019, iii), and this lack of transparency may partially explain why linguists have been more wary to use them. Still, it may be worth adopting predictive DSMs as retrieval tools, as they are likely to achieve accuracies in sense disambiguation tasks that are hard to achieve with count models, and, unlike count models, they may even produce usable representations for function words (Boleda 2020, 7), as demonstrated in this squib.

This squib focuses on the use of one such predictive model, BERT, which is a deep contextualized model based on a particular type of neural architecture called 'the Transformer' (Vaswani et al. 2017). While it is possible to train BERT from scratch, one can also use a version pre-trained on approximately 3.3 billion words (800 million words from the BooksCorpus, and 2.5 billion words from the English Wikipedia). During this pre-training task, BERT processes unlabeled data in a masked word prediction task, where the model predicts randomly masked input tokens based only on the context in which they occur. Like other Transformers, BERT consists of multiple layers (or 'transformer blocks'), all of which contain multiple self-attention heads (mechanisms that look at other 'words' in the input sequence to find clues that can lead to a better encoding for a target word), which behave similarly within their layer. In each of these layers the *n* tokens in the input sequence are captured as numerical vector representations (or 'embeddings'). While attention heads within these layers have been probed for the specific linguistic phenomena they focus on (e.g., valency patterns, dependency relations; Clark et al. 2019), the present squib will simply use the Spacy implementation of $\text{BERT}_{\text{base}}$, which currently only offers the model's 12th (and final) hidden layer.

3. The challenge: Finding INTO-INTEREST

To demonstrate how BERT can be used as a sense disambiguation tool and retrieval tool, I will focus on the construction exemplified in the sentences in (1), taken from the Oxford English Dictionary (OED):

(1) a. I tend to like the stuff the rock groups are doing because they're creative and original, and *that's something I'm very much into.* (1969, OED)
b. Matt is the dopest guy in school and *he's into you.* (1991, OED)

The examples listed here illustrate one of at least 23 senses that the OED distinguishes for the preposition *into*, which is used here to indicate that the subject (usually a person) is involved in, knowledgeable about or (romantically) interested in someone or something.

Drawing on insights from Cognitive Linguistics, CxG generally seems to treat prepositions as linking a word form with an abstract (spatial) relation schema. A detailed account of the constructional properties of *into* can be found in Bergen and Chang's (2005) outline of Embodied CxG. The function of *into*, much like that of other prepositions, is to capture an asymmetric (spatial) relation between a trajector (TR) with respect to a landmark (LM). In (2), for example, the position of the trajector (*Sali*) is defined relative to the landmark (*the pool/Amsterdam*):

(2) Sali jumped into the pool / drove into Amsterdam.

The INTO-CXN is said to evoke the Source-Path-Goal schema, where the TR moves, in some way, along a path towards a goal (i.e. the LM). Considering that constructions constitute categories with a radial structure (e.g. Goldberg 1995; Croft 2001), the proposed spatial image schema of the INTO-CXN can be thought of as its 'basic meaning', which can be metaphorically extended to non-spatial target domains, including time periods (3a) and states (3b):

(3) a. She partied into the night.
b. No one could help her as she slid into madness. (Goldberg 1995, 83)

In such cases, the image schema underlying the source domain is mapped onto the target, with implications of motion resulting in the TR crossing a boundary and being 'contained' in the LM (e.g. it was not night yet when she started partying). With respect to INTO-INTEREST, then, it seems relatively straightforward to conceive of a similar mapping from space to 'areas of interest'. However, INTO-INTEREST may be subject to some formal restrictions that do not apply to other senses of *into*, such as those in (2) and (3). Based on the examples in the OED, it seems fair to conclude that we are perhaps not dealing with a sense of *into,* but a sense of the complex phrasal expression consisting of a copula (predominantly *BE*, but also *GET*) combined with *into*.[2]

Supposing that one would like to conduct a corpus-based study of a construction like INTO-INTEREST (e.g. mapping its distribution across text types and/or across time), it will soon become clear that a simple formal query of BE/GET *into* (see Section 4) or *into* alone (see Section 5) will yield rather disappointing precision rates. In what follows, I will suggest how BERT can be employed as a disambiguation/retrieval tool to counter such issues.

4. A solution: BERT as a disambiguation tool

Let us first consider a scenario where a formal query yields a large number of hits, which contain a relatively small but reasonable number of relevant examples. Using the final two decades of the Corpus of Historical American English (approx. 58 million words) and querying them for the combination of any form of BE or GET and *into* (with max. one intermittent adverb) yields 3,494 results. It is difficult to determine recall (as we do not know how many verbs besides *BE* and *GET* occur in the INTO-INTEREST construction, and to what extent multiple intermittent elements are possible), but the precision of this formal query can easily be checked: a random sample of 1,000 tokens (366 *BE into*, 634 *GET into*) yields only 287 examples represent INTO-INTEREST (precision = 0.29). The remaining 713 examples constitute a relatively heterogenous group of concrete, spatial uses, as well as a number of abstract uses. Ultimately, the different uses of *BE/GET into* were assigned to 8 different categories. The categories, which were determined by the annotator in a bottom-up manner, are illustrated in Table 1, along with their token frequencies.

It is at this point that a model such as BERT could be employed to further narrow the set of examples that may be relevant, as it could help separate concrete, spatial uses of BE/GET *into* from more abstract, metaphorical uses. From previous studies employing DSMs, it seems that DSMs are apt at distinguishing lexical items with concrete meanings from those with more abstract meanings (e.g. Heylen et al. 2015; Perek 2016, 18; Schlechtweg et al. 2017; Giulianelli, Del Tredici, and Fernández 2020), and could help distinguish lexical and grammaticalized uses of the same word (e.g. Hilpert and Correia Saavedra 2017).

To assess BERT's performance in disambiguating different senses of BE/GET *into,* the above-mentioned set of examples was used as the input for a sense disambiguation task. First, we can use BERT to create embeddings for each of the 1,000

2. INTO-INTEREST also seems to allow adjective-like modification (e.g. *I'm very/so into you* vs. **Sali jumped very/so into the pool*).

Table 1. Usage categories of BE/GET *into*

Domain	Usage type	Examples	#
Concrete	SPACE *LM is a concrete space*	– She ***was** into* the passageway outside her quarters before she realized she 'd moved – I watched him ***get** into* an especially small and tinny brown car.	322
Concrete	CLOTHES *'put on', 'wear'*	– it was May and the police ***were** already into* short sleeves – I can't ***get** into* the dress Mother gave me anymore.	20
Abstract	ACCESS *TR gains access to selective group*	– you will never ***get** into* university with this attitude – Now our clique has about 20 girls in it. I ***got** into* it because the girl who's the head of the group is my play sister.	47
Abstract	ACT_SIT_STATE *TR partakes in activity, situation, or state* (LM)	– She***'s** deep into* her next novel-about a couple who fall back in love after getting divorced. (2005, COHA) – I worked with one actress who would randomly ***get** into* a rage and scream (2007, COHA)	269
Abstract	MIND *TR occupies a person's mind or feelings*	– It's ***got** into* your subconscious, hasn't it, son? – They are all just so peaceful and funny that they ***get** into* your heart.	12
Abstract	FIXED EXPRESSION	– What's ***gotten** into* you? – Eventually you begin to wonder what in the world has ***got** into* Bernardo Bertolucci.	17
Abstract	TIME *LM is a time period*	– As we *get* ***into*** next year there's going to be increasing risk in the Semiconductor names. – he ***was** into* his early eighties and perhaps had forgotten promising the space to a grandkid.	26
Abstract	INTEREST *TR is involved/interested in LM (business, hobby, person, etc.)*	– Okay, so not every guy ***is** into* football. Some love basketball, baseball, or even luge – if you want to know what boys ***are** really into*, here's a tip: try girls. – Maybe that's when Tommy ***got** into* all the mystic stuff. – How did you ***get** into* painting?	287

examples in the dataset. As a means of visualization, the embeddings have been mapped into a two-dimensional space (by means of t-SNE) in Figure 1. Note that the embeddings were created solely based on contextual information (a window of 20 words preceding and 20 words following *into*); it was only after the embeddings and the relative distances between them were mapped in the two-dimensional space that the manually assigned category labels (presented in Table 1) were added to the figure. The two-dimensional plot therefore visualizes the overall correspondence between the output of the model and the manually assigned labels.

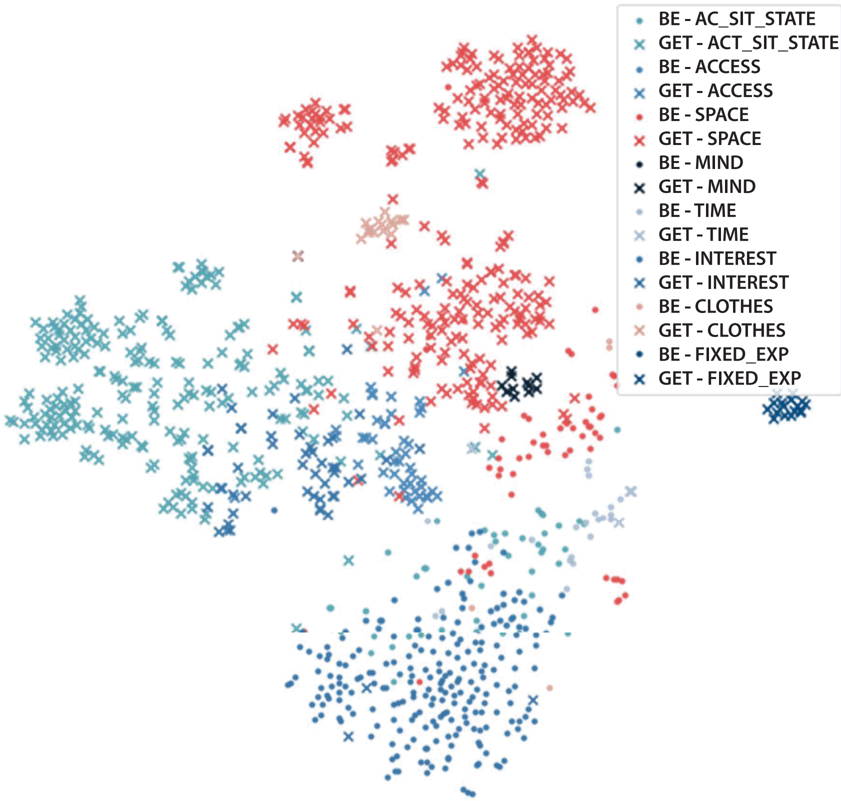

Figure 1. t-SNE of BE/GET *into* embeddings (KL divergence after 1,000 iterations: 0.83). The dots represent instances of *BE into*, while *GET into* is represented by means of crosses. Concrete uses are in red shades, while abstract uses are colored in blue shades

After creating the embeddings, the examples are divided into 10 sets or 'folds'. Then, a logistic regression classifier can be fitted on a labeled training set compiled from 9 of the 10 folds, which is then used to predict the labels of a test set (the remaining 10th fold). To avoid overfitting, a regularization function can be used that helps control the trade-off between fitting the training data well and keeping the number of parameters small. In the present case, L2 or Ridge Regularization was used. The procedure of training and fitting is repeated with each of the 10 folds as test set (10-fold cross-validation). Finally, the predicted labels are assessed against the manually assigned labels of the test sets to evaluate whether BERT can help assign the correct category label to unseen tokens. For each fold, a precision, recall, and F_1-score (the harmonic mean of precision and recall) is computed for each category. The average of those scores across the 10 folds is reported in Table 2.

Table 2. Classifier residuals

	Precision	Recall	F_1-score	# Tokens
ACCESS	0.80	0.85	0.82	47
ACT_SIT_STAT	0.91	0.89	0.90	269
CLOTHES	0.80	0.80	0.80	20
FIXED_EXP	1.00	1.00	1.00	17
INTEREST	*0.90*	*0.91*	*0.91*	*287*
MIND	0.73	0.92	0.81	12
SPACE	0.95	0.93	0.94	322
TIME	0.86	0.96	0.91	26
WEIGHTED AVERAGE	0.91	0.91	0.91	1,000

With an average F_1-score of 0.91, the performance of BERT disambiguating the various uses of *BE/GET into* is very good. In light of its potential as an annotation aide (e.g. building a classifier by means of a labeled training set to annotate further, truly unseen examples), the positive recall scores (which indicate that relatively few examples are 'missed') are also reassuring.

5. BERT as an exemplar-based retrieval tool

A second scenario worth considering is one where it proves difficult to obtain a sufficient number of relevant examples of a construction (e.g. to build a classifier) by means of a formal query. This would be the case, for instance, when no a priori stance is taken on whether INTO-INTEREST is associated with a specific (set of) verb(s), as such agnostic premises easily translate into poor precision rates. Querying the final decade of COHA for *into* yields 56,511 results, and a quick check of a randomly drawn sample of 1,000 examples yields only 5 relevant hits (Precision = 0.005).[3] Again, we could consider employing BERT to tackle such retrieval inefficiency issues.

The procedure described here assumes that the researcher is simply interested in retrieving a particular usage of *into* in a large, unlabeled dataset that would be difficult to find with formal queries alone. In this case, BERT can be used to create embeddings for all tokens of *into* in COHA (for instance using a context window of 20). The next step involves using an embedding of a representative (or 'exemplary') token (from the corpus, or for instance taken from the OED) as a query to search for similar vectors by means of a distance metric (e.g. Euclidean, Cosine) ideally combined with a clustering algorithm like k-means to create a ranking. A useful application for large datasets, which combines Euclidean distances and k-means, is FAISS (Johnson, Douze, and Jégou 2017).

Table 3 shows an example of the output of a similarity search. The example at $k=0$ is the input token, which, in this case, was taken from the OED and considered to be a representative, exemplary token. All subsequent examples represent its k-nearest neighbors, ranked according to the Euclidean distance between the input and the retrieved token. The ranking is relatively consistent up to $k=20$, when the occasional ambiguous example appears (e.g. $k=22$ can be read as an example of INTO-INTEREST or TIME). Examples are mixed between $k=30$ and $k=60$, after which they become more consistently irrelevant.

In the absence of a manually labeled list of all relevant tokens, we cannot evaluate the performance of BERT by means of any metrics that rely on both precision and recall. We can, however, check the quality of the ranked lists it produces: when we query a corpus by means of an exemplar embedding and rank its nearest neighbors from 'closest' to 'most distant', we hope to find all relevant cases in the top ranks of that list. We therefore measure the model's precision at a fixed level of retrieved results by means of an evaluation metric called 'Precision at k' (Manning, Raghavan, and Schütze 2009, 161), where k stands for the range of the ranking taken into consideration. If all examples in that list are relevant, precision at k will be 1. For each irrelevant example, the score lowers, and the metric punishes mistakes more harshly at high ranks: if the example at rank 1 is irrelevant, the score will immediately be lowered to 0.5. In the present case, a set of 10,000 randomly selected examples with *into* was queried by means of 10 exemplary tokens

3. Note that the number of hits yielded by the same query may vary slightly depending on whether the online interface of COHA (at English-corpora.org/coha) or the offline, downloadable version of COHA is used. The number of hits will be smaller in the offline version, as sections of the corpus are replaced by means of @-signs for copyright reasons. The number reported here is based on the offline version of COHA.

Table 3. Sample of similarity ranking (DISTANCE = Euclidean)

k	Distance	Example	Relevance Y/N/AMB
0	QUERY	This should have been the high-light of the evening, but the audience just wasn't **into** it.	QUERY
1	116.19	've given you hints on the best first moves to make when you're just not sure if he's **into** it.	Y
2	123.20	i knew it was trendy and all. the women at the country club are really **into** it.	Y
3	128.05	i thought you weren't **into** that stuff," i said to brian. "doesn't mean i wasn't paying attention	Y
4	131.50	generation y, bennet said, is into it, and the young kids are really, really **into** it. they're accepting of it	Y
5	137.71	they'll say: "azalia nelson: bright, very bright, but we hear she's a bit **into** this black thing..."	Y
…	…	…	
20	188.97	"well," she said, "it still doesn't mean he's **into** strippers. he could have gotten that	Y
21	193.60	i'm a person that if i love somebody or if i'm **into** something, i want to give it my all	Y
22	194.98	"i was **into** that moment," she said. "it was neat to feel my body reacting	AMB
23	195.50	we are coming back to appreciate motherhood. i see more and more men getting **into** the parenting thing	Y
24	196.11	it adds to the whole kiss mystique," scott says. "a lot of people know i'm **into** kiss, but not everyone	Y
…	…	…	

of INTO-INTEREST (8 taken from the examples listed in the OED, and the remaining 2 taken from the annotated dataset described in Section 4). For each of the 10 queries, the first 100 ranks ($k = 100$) were checked to calculate the 'Precision at 100'. With an average of 0.90, the model performed well across all 10 trials, and hence may serve as a helpful tool to quickly identify relevant examples of a construction in a large, unlabeled dataset.

6. Conclusion

A growing number of studies, including this squib, positively report on the performance of contextualized DSMs in meaning disambiguation tools, which makes them an appealing addition to our toolkit for studying form-meaning pairings. Focusing on INTO-INTEREST, this squib demonstrates how one particular model, BERT, can be a helpful disambiguation tool and an 'exemplar-based' retrieval tool.

Of course, when it comes to integrating these models into construction grammar research, there are still some hurdles to overcome. First, by means of exemplar-based querying methods, the use of embeddings can be helpful to make the retrieval of constructions based on broad, agnostic corpus queries more manageable. However, the procedures described here still require at least one element in the construction to be lexically specified (e.g. *into*), and determining how we can employ these models to effectively retrieve more schematic, lexically underspecified constructions is still challenging. Second, this squib focused on the use of BERT as a retrieval tool, and has steered clear of any reflection on whether it can also be employed as an analytic tool to, for instance, empirically test hypotheses regarding the semantic similarity between different uses of (*BE/GET*) *into* (or, more generally, to what extent different DSMs enable "researchers in Construction Grammar to create explicit corpus-based models of speakers' knowledge of constructions" (Hilpert and Correia Saavedra 2017, 29)). I still hope, however, that this squib has helped demonstrate the value of these models for corpus-based research, and that it will entice construction grammarians into not only using, but also improving and/or manipulating different types of DSMs to make them a more well-rounded tool for semantic analysis.

References

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. "Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. by Kristina Toutanova, and Hua Wu, 238–247. Baltimore, Maryland: Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1023

Bergen, Benjamin K., and Nancy Chang. 2005. "Embodied Construction Grammar in Simulation-Based Language Understanding." In *Construction Grammars: Cognitive Grounding and Theoretical Extensions*, ed. by Jan-Ola Östman, and Mirjam Fried, 147–190. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/cal.3.08ber

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

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." *ArXiv:1607.06520 [Cs, Stat]*, July. http://arxiv.org/abs/1607.06520

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. "What Does BERT Look At? An Analysis of BERT's Attention." *ArXiv:1906.04341 [Cs]*, June. http://arxiv.org/abs/1906.04341. https://doi.org/10.18653/v1/W19-4828

Croft, William. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198299554.001.0001

De Pascale, Stefano. 2019. "Token-Based Vector Space Models as Semantic Control in Lexical Sociolectometry". PhD dissertation, KU Leuven.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, ed. by Jill Burstein, Christy Doran, and Thamar Solorio, 4171–4186. Minneapolis: Association for Computational Linguistics.

Giulianelli, Mario, Marco Del Tredici, and Raquel Fernández. 2020. "Analysing Lexical Semantic Change with Contextualised Word Representations." *ArXiv:2004.14118 [Cs]*, April. http://arxiv.org/abs/2004.14118. https://doi.org/10.18653/v1/2020.acl-main.365

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

Gupta, Abhijeet, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. "Distributional Vectors Encode Referential Attributes." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, ed. by Lluís Màrquez, Chris Callison-Burch, and Jian Su, 12–21. Lisbon: Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1002

Heylen, Kris, Thomas Wielfaert, Dirk Speelman, and Dirk Geeraerts. 2015. "Monitoring Polysemy: Word Space Models as a Tool for Large-Scale Lexical Semantic Analysis." *Lingua* 157 (April): 153–172. https://doi.org/10.1016/j.lingua.2014.12.001

Hilpert, Martin. 2013. "Corpus-Based Approaches to Constructional Change." In: *The Oxford Handbook of Construction Grammar*, ed. by Thomas Hoffmann, and Graeme Trousdale, 458–475. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195396683.013.0025

Hilpert, Martin, and David Correia Saavedra. 2017. "Using Token-Based Semantic Vector Spaces for Corpus-Linguistic Analyses: From Practical Applications to Tests of Theoretical Claims." *Corpus Linguistics and Linguistic Theory* 16(2): 393–424. https://doi.org/10.1515/cllt-2017-0009

Johnson, Jeff, Matthijs Douze, and Hervé Jégou. 2017. "Billion-Scale Similarity Search with GPUs." *ArXiv:1702.08734 [Cs]*, February. http://arxiv.org/abs/1702.08734

Linzen, Tal, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes (eds). 2019. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-4800

Louwerse, Max M., and Rolf A. Zwaan. 2009. "Language Encodes Geographical Information." *Cognitive Science* 33 (1): 51–73. https://doi.org/10.1111/j.1551-6709.2008.01003.x

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Perek, Florent. 2016. "Using Distributional Semantics to Study Syntactic Productivity in Diachrony: A Case Study." *Linguistics* 54 (1): 149–188. https://doi.org/10.1515/ling-2015-0043

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

Schlechtweg, Dominik, Stefanie Eckmann, Enrico Santus, Sabine Schulte im Walde, and Daniel Hole. 2017. "German in Flux: Detecting Metaphoric Change via Word Entropy." In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, ed. by Roger Levy, and Lucia Specia, 354–367. Vancouver: Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-1036

Sommerauer, Pia, and Antske Fokkens. 2018. "Firearms and Tigers Are Dangerous, Kitchen Knives and Zebras Are Not: Testing Whether Word Embeddings Can Tell". In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, ed. by Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, 276–286. Brussels: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5430

Stefanowitsch, Anatol, and Stefan Gries. 2003. "Collostructions: Investigating the Interaction between Words and Constructions." *International Journal of Corpus Linguistics* 8 (2): 209–243. https://doi.org/10.1075/ijcl.8.2.03ste

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need". *ArXiv:1706.03762 [Cs]*, December. http://arxiv.org/abs/1706.03762

Yoon, Jiyoung, and Stefan Th Gries (eds). 2016. *Corpus-Based Approaches to Construction Grammar*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/cal.19

Author's address

Lauren Fonteyn
Leiden University Centre for Linguistics (LUCL)
Leiden University
Arsenaalplein 1
2312BM Leiden
The Netherlands
l.fonteyn@hum.leidenuniv.nl