JS Journal of Historical Syntax
Volume 6, Article 17: 1–41, 2022

EXPLORING MORPHOSYNTACTIC VARIATION & CHANGE WITH DISTRIBUTIONAL SEMANTIC MODELS*

Lauren Fonteyn
Leiden University

Enrique Manjavacas
Leiden University

Sara Budts
University of Antwerp

Abstract This paper surveys how computational distributional semantic models (DSMs) have thus far been employed to study morphosyntactic variation and change in (Early) Modern English. Using a case study on the development of the Early Modern English auxiliary *do*, we illustrate how computational DSMs can be used to flag areas of functional-semantic overlap between the class of modal verbs and auxiliary *do* in a data-driven manner. The paper will be concluded with a summary and critical assessment of how computational DSMs can complement (and be complemented by) other approaches to morphosyntactic variation and change in the Early Modern period.

1 Introduction

In studies that address the functional-semantic side of linguistic variation and change, there is an increased interest in finding "objective means to make observations on linguistic change and test hypotheses in a way that does not depend on the researcher's intuitive judgment" (Sagi, Kaufmann & Clark 2011: 161). As a result, a growing number of studies have started to explore how distributed meaning representations in the form of 'semantic vectors' or 'embeddings' – which can be understood as (compressed) numeric

* We would like to thank the editors and contributors of this volume for their suggestions. We furthermore thank the anonymous reviewers for their comments and feedback on earlier versions of this article.

representations of a word's contextual distribution that serve as a proxy of that word's meaning – can be employed as methodological tools in historical linguistics (see, among many others: Sagi et al. 2011, Hamilton, Leskovec & Jurafsky 2016b, Mitra, Mitra, Riedl, Biemann, Mukherjee & Goyal 2014, Hamilton, Leskovec & Jurafsky 2016a, Bamler & Mandt 2017, Rosenfeld & Erk 2018, Rudolph & Blei 2018, Kutuzov, Øvrelid, Szymanski & Velldal 2018, Hu, Li & Liang 2019, Dubossarsky, Hengchen, Tahmasebi & Schlechtweg 2019, Tahmasebi, Borin & Jatowt 2019, Del Tredici, Fernández & Boleda 2019, Schlechtweg, McGillivray, Hengchen, Dubossarsky & Tahmasebi).

While the specific aims of individual studies are varied, this recent computational work with distributional semantic models (henceforth: DSMs) generally aims to present a fully automated, 'data-driven' means of detecting and describing the diachronic trajectory of linguistic change. Previous work has proposed various ways of using computational DSMs to determine laws of semantic change (e.g. Hamilton et al. 2016b, Dubossarsky, Weinshall & Grossman 2017a), developing statistical measures that help detect different types of semantic change (e.g. specification vs. broadening; cultural change vs. linguistic change) in a data-driven manner (e.g. Sagi et al. 2011, Mitra et al. 2014, Hamilton et al. 2016a, Del Tredici et al. 2019, Dubossarsky et al. 2019, Schlechtweg et al., Giulianelli, Del Tredici & Fernández 2020), mapping changes in more specific (groups of) concepts in particular lexical domains (e.g. 'racism', 'knowledge'; Sommerauer & Fokkens 2019, Betti, Reynaert, Ossenkoppele, Oortwijn, Salway & Bloem 2020), and improving the models that generate (diachronic) word embeddings to attain these goals (e.g. Rudolph & Blei 2018, Rosenfeld & Erk 2018).[1] Specifically for (Early) Modern English, projects such as Linguistic DNA (Fitzmaurice, Robinson, Alexander, Hine, Mehl & Dallachy 2017) set out to employ distributed meaning representations to investigate the development of concepts that shape thought, culture and society in the Modern English period in a bottom-up manner; and studies such as Bizzoni, Degaetano-Ortlieb, Fankhauser & Teich (2020) and Sun, Liu & Xiong (2021) focus on the development of scientific terminology as new scientific fields arose during the Early Modern period.

Notably, the majority of these studies and projects focuses exclusively on capturing and quantifying some aspect of *lexical* or conceptual change. One consequence of this focus on lexis is that, at present, the number of computational studies that consider the functional-semantic properties of morphosyntactic constructions from the Early Modern period seems disproportionate to the interest in morphosyntactic change within historical linguistics more

1 For a survey of recent advances and remaining challenges of computational semantic change research, see Hengchen, Tahmasebi, Schlechtweg & Dubossarsky (2021).

generally (also see Fonteyn 2020). As such, it remains somewhat of an open question to what extent (changes in) the functional-semantic properties of morphosyntactic constructions can be captured and analysed by means of DSMs.

The aim of this paper is to shed some light on how DSMs have thus far been employed to study the functional-semantic properties of morphosyntactic constructions in diachronic corpora. We first provide some background on the different types of computational DSMs in Section 2. In Section 3, we discuss earlier work to survey the different types of questions on morphosyntactic variation and change that have been tackled by means of DSMs. Subsequently, to exemplify how the output of computational DSMs can be employed in more detail, we home in on the development of the Early Modern English auxiliary *do* in Section 4. More specifically, we will illustrate how DSMs can be used to flag areas of functional-semantic overlap between the class of Early Modern English modal verbs and auxiliary *do* in a bottom-up manner. The surveyed case study is based on work by Budts (2020a), who employs a DSM based on Neural Networks to generate probability distributions of modal verbs and auxiliary *do* across different input sentences. We will extend this analysis by also offering insight into the contextual cues that the model relies on to predict whether *do* or a modal verb should be used in a given input sentence.

We wish to note here, at the outset, that the background, survey and case study are discussed in a conceptual rather than a technical manner. For readers who are interested in learning more about the technical details of computational DSMs, we will refer to sources where such details can be found. The paper will be concluded in Section 5, where we summarize and critically assess how research involving computational DSMs can complement (and be complemented by) manual analyses of morphosyntactic variation and change.

2 Background: distributed meaning representations

Computational models that generate distributed meaning representations for linguistic forms rely on the idea that the meaning of words (or, more generally, of constructions) can be conceptualized as a function of their lexical and/or grammatical context. Over the last three decades, a plethora of ways to operationalize contextual distributions as a proxy to lexical meaning have been developed, resulting in a wide variety of computational DSMs (for extensive overviews, see, e.g., Lenci 2018, Boleda 2020, Young, Hazarika, Poria & Cambria 2018). Within these DSMs, a distinction can be made between 'count' and 'predict(ive)' models (Baroni, Dinu & Kruszewski (2014); also

see 'explicit' and 'implicit' models in Dubossarsky, Weinshall & Grossman (2017b: 1136)).

Count models most straightforwardly operationalize the idea that the meaning of a given word or construction can be approximated by its collocates. These models are essentially derived from large, comprehensive co-occurrence count matrices (for an accessible explanation of how count models are built, see Heylen, Wielfaert, Speelman & Geeraerts (2015)). Given a number of examples involving, for instance, the target words *legs* and *arms* – as in examples (1)-(4), the first step of constructing a count model involves compiling a co-occurrence matrix in which target words as represented as rows, and the words they co-occur with (within a certain context window) are represented as columns – as in Table 1.

(1) Ulcers on the **legs**, or any other part of the body, require pretty much the same treatment (Lind, 1753; PPCMBE)

(2) his Arms and **Legs** are dwindled away with Whoring; and his Body 's decay 'd with Intemperance. (Stevens, 1745; PPCMBE)

(3) her pretty little white hands and **arms** were almost covered with rings and bracelets. (Montefiore, 1836; PPCMBE)

(4) their Confederates, did assemble in a warlike Manner, and procured **Arms**, Ammunition, and other Instruments of War; (Townley, 1746; PPCMBE)

The co-occurrence matrix maps which words the targets *legs* and *arms* co-occur with (e.g. *ulcers, on, the, his, pretty*, etc.), and how often they do so. Subsequently, to extract numeric vector representations for the target words, the raw counts in the co-occurrence matrix are pruned or transformed in various ways: the numeric information can, for instance, be optimized by reweighting or dropping columns with uninformative function words such as *the* or *and* (see Section 2.1). Which specific transformations are applied in optimizing the resulting vectors is determined by the analyst.

With context-predicting models, a cover term for a varied set of artificial neural networks, the construction of distributed meaning representations is similar, but approached as a training task (Baroni et al. 2014: 238). While optimizing count models involves tuning and reweighting the quantitative information contained in the co-occurrence matrix after it has been compiled, predictive models construct vectors as part of a machine learning task, in which these vectors are set and adjusted to optimally predict target words. Because

	ulcers	*body*	*arms*	*hands*	*bracelets*	*ammunition*	*war*	...
legs	1	2	1	0	0	0	0	...
arms	0	1	0	1	1	1	1	...

Table 1 Simple co-occurrence matrix for the word types *legs* and *arms*

	ulcers	*body*	*arms*	*hands*	*bracelets*	*ammunition*	*war*	...
legs in (1)	1	1	0	0	0	0	0	...
legs in (2)	0	1	1	0	0	0	0	...
arms in (3)	0	0	0	1	1	0	0	...
arms in (4)	0	0	0	0	0	1	1	...
...	...	...	...	...	...	...	...	...

Table 2 Simple co-occurrence matrix for individual tokens of *legs* and *arms*

predictive models reduce the analyst's involvement in the vector transformation and optimization process, and because they generally yield better performance than count models in a range of Natural Language Processing (NLP) tasks (though see Lenci, Sahlgren, Jeuniaux, Gyllensten & Miliani 2021), the relatively recent emergence of predictive models is often portrayed as an attractive advancement (Baroni et al. 2014).[2] Furthermore, predictive models may be particularly appealing for studies beyond lexical semantics, as predictive models seem more successful than count models in providing useful representations of function words (e.g. Bullinaria & Levy 2012, Boleda 2020: 7).[3]

A second distinction that can be made is between models that generate word type representations and models that generate token representations (or 'static' and 'contextualized' models Lenci et al. 2021). Given the examples

2 This is not to say that predictive models involve absolutely no parameter tuning, and it has been suggested that, given comparable settings and tuning, count models may be as effective as predictive models (Levy, Goldberg & Dagan 2015, Lenci et al. 2021).

3 There is, furthermore, some evidence that human language processing is shaped by a drive to predict future inputs, a behaviour mimicked by predictive language models. The finding that there are strong correspondences between at least some types of predictive language models and human representations also renders these models appealing tools to test hypotheses about the human mind (Schrimpf, Blank, Tuckute, Kauf, Hosseini, Kanwisher, Tenenbaum & Fedorenko 2020).

involving the words *legs* and *arms*, a type-based model will provide a single numeric representation for all of the context in which these words occur (see Table 1). Generally speaking, type-based models work from the assumption that a word has a single, constant, 'core' meaning (which can be understood as a prototype, see Erk & Padó 2010: 92). For a word such as *arms*, for instance, the contextual information that suggests that the word refers to (a set of) body parts (3) would therefore be conflated with contextual information that suggests that it refers to a type of object (i.e. weapons, (4)). A potential problem with such vector representations is that, for certain research questions, they may render unsatisfactory or problematic vector representations in cases of polysemy or homonymy (see, e.g. Desagulier 2019, De Pascale 2019). In response to this issue, models that generate token-specific representations were developed. These token-based distributional models – in which individual vectors are assigned to, for instance, the examples of *arms* in (3) and (4) (see Table 2) – are better equipped to handle the complex internal semantic structure of words, and, hence, are better suited for specific tasks such as word sense disambiguation Heylen et al. (see, e.g. 2015).[4] As they constitute means to different ends, the modeling requirements of token-based and type-based embedding approaches differ. For example, word order can be relevant for distinguishing between senses of the same construction (e.g. *He used to eat soup with a fork* vs. *He used a fork to eat soup*), but not as much for identifying synonymy and antonymy relations between different construction types. For this reason, approaches that compute type embeddings, can effectively rely on so-called 'bag-of-words' representations, in which word order is ignored. By contrast, (predictive) token-based models more commonly encode word order, and some have been shown to infer syntactic structure from the textual material they have been trained on to a surprising extent (Conneau, Kruszewski, Lample, Barrault & Baroni 2018, Goldberg 2019, Jawahar, Sagot & Seddah 2019, Manning, Clark, Hewitt, Khandelwal & Levy 2020, Linzen & Baroni 2021).[5]

4 The way in which different models obtain distributed meaning representations of individual tokens varies between architectures. For a relatively transparent, accessible explanation of how distributed token representations can be obtained by means of count models, see Heylen et al. (2015) or Hilpert & Correia Saavedra (2017). The construction of token representations in those studies is similar to the procedure adopted in Section 4, albeit that the latter relies on predictive models.

5 See Lenci et al. (2021) for an extensive comparison of different types of count and predictive models, and type-based versus token-based models in a range of semantic tasks (semantic similarity and synonymy detection, analogy completion, sentiment classification, etc.).

2.1 Modelling historical data: corpora and pre-processing

An important aspect of any type of DSM, regardless of whether it is applied to present-day or historical data, relates to how the data should be pre-processed before the model can be applied. In order to collect co-occurrence counts or predict target words, a series of pre-processing steps first determine what counts as a word. Common pre-processing steps include tokenization (which helps to isolate punctuation while keeping abbreviations in place), as well as segmentation of hyphened words, lowercasing, removing non-alphanumeric tokens that carry little lexical information, and regularizing spelling.

(5) **original**
let him make an Incision, eyther right or straight, or somewhat crooked, on the necke vnder the Jaw-bones (William Clowes, PPCEME, 1602)
processed
let him make an incision , either right or straight , or somewhat crooked , on the neck under the jaw bones

Furthermore, researchers can opt to ignore words in the corpus based on their part-of-speech (e.g. any word that is not a noun, a verb or an adjective), or ignore any word below a particular absolute frequency threshold.

When it comes to Early Modern linguistic data, many of the mentioned pre-processing steps are less trivial then they might seem. For English, regional spelling and writing systems only just started to give way to a more general standard (for more on standarization in English, see Oudesluijs, Gordon & Auer, this volume), and the period was characterized by considerable spelling variation between (and even within) individual texts and authors (Scragg 1974). The extent to which spelling should be regularized is certainly not trivial: in some cases, spelling variation in fact expresses a meaningful distinction that is no longer made in Present-day language variants [6] and some spelling variation may also reflect systematic inter- and intra-individual variation (Scragg 1974, Evans & Tagg 2020). Pre-processing historical data is, in short, an exercise in striking a balance between retaining what may be meaningful information and making sure the model one intends to use can provide useable representations. [7]

6 The nominative second-person plural pronoun *ye*, for instance, contrasted with *you* before the pronominal paradigm was simplified in Modern English. In such cases, it makes sense to regularize the spelling of *ye*/*yee* to *ye*, but not to conflate *ye* with *you* (see Baron 2011: 161).

7 More practical information on how to perform such pre-processing can be found in, for instance, Piotrowski (2012: Ch. 6) and Bollmann (2019).

In that respect, it is interesting to know that the number of steps applied in pre-processing in current research largely depends on the type of model at hand. For example, for predictive models that generate distributed type representations, it is common to apply many of the previously mentioned word filtering steps. The goal is generally to increase the signal-to-noise ratio in the training corpus by increasing the number of occurrences of each relevant word while reducing the number of non-informative distributed representations that must be computed (e.g. so that common spelling variants such as *iudge* and *judge* or *neck* and *necke* are not assigned separate representations; to this end lemmatization is often applied to the training corpus).

Token-based approaches resort to subword tokenization algorithms like Byte Pair Encoding (BPE) (Gage 1994, Sennrich, Haddow & Birch 2016) or WordPiece (Schuster & Nakajima 2012). These methods learn an optimal basis subword vocabulary of a user-specified size and a set of merge rules over this vocabulary. With these two components, any given sentence can be tokenized into words and subwords (e.g. `playing` is split into `play` and `##ing`), provided no characters are used that were not present in the original training corpus. If a word is not present in the inferred basis vocabulary, it will be (sub-)tokenized into subword tokens. For example, provided the model can generate representations for subwords `walk` and `##er`, a representation can still be computed for `walker` even if that word was not seen in the training data. By doing this, token-based embedding approaches can compute contextualized meaning representations for tokens that were not present in the training corpus (even though the meaning representation for unknown tokens still needs to be computed by aggregating the vectors of each subword token).[8] This automatic (sub-)tokenization procedure reduces the need for extensive spelling regularization and other fine-grained filtering pre-processing steps (e.g. for certain languages, the induced subword tokenization rules resemble segmentations into stems, prefixes and suffixes, thus inducing some sort of morphological segmentation; but see Bostrom & Durrett (2020)). However, since token-based approaches take sentential context into account, it is often necessary to perform sentence segmentation (a step that is not as important in type-based approaches). An example of a pre-processing (and model evaluation) procedure for a predictive, token-based embeddings model of Early and Late Modern English can be found in the description of MacBERTh (Man-

8 The subword tokenization and basis vocabulary approach is ultimately motivated by the fact that token-based models require very large training corpora (in the order of several billion words) in order to produce high quality representations, and due to the Zipfian nature of language, increasingly larger training corpora would require increasingly larger vocabulary sizes.

javacas Arévalo & Fonteyn 2021, 2022).[9]

2.2 *Motivation*

While the specific model choice (count vs. predict; type vs. token) may differ from study to study depending on the research aims, the motivation for adopting a computational approach to constructional meaning over a 'manual' or 'introspective' one generally tends to be framed in the same way (for some examples of studies that engage in an elaborate motivation of the approach, see Sagi et al. 2011, Perek 2016, Hengchen et al. 2021).

First and foremost, introspective corpus data annotation is not only extremely labour-intensive (and hence costly), but also potentially subjective if conducted by a single annotator. Yet, a single-expert-annotator set-up tends to be the default modus operandi in historical linguistics, given that native speakers of historical language varieties are no longer available to provide semantic judgments, and that the number of historical language experts capable of reliably annotating historical data is also relatively small. Computational models that enable a bottom-up extraction of functional-semantic information from corpus data are thus considered valuable tools, which can help minimize not only the subjectivity, but also the laboriousness and cost of functional-semantic annotation.

Second, computational models that generate distributed meaning representations can be considered particularly appealing for function-oriented approaches to morphosyntactic change (e.g. Grammaticalization Theory, Diachronic Construction Grammar), where functional-semantic processes such as metaphor, metonymy, subjectivity, and invited inferencing play a central role (Hopper 1991, Traugott 1989, Traugott & König 1991, Traugott & Dasher 2002, Traugott 2003, 2010, Traugott & Trousdale 2013). Because distributional models produce numeric and hence measurable information on a construction's meaning – which is notoriously difficult to achieve with introspective analysis – the models also allow researchers to operationalize functional-semantic concepts in quantifiable terms, which could enable them to statistically verify or falsify hypotheses on the nature and causes of semantic change.

3 Applications to morphosyntactic change

At present, there is no abundance of work in historical linguistics involving distributed meaning representations to study morphosyntactic variation and

9 The MacBERTh model, as well as the historical Dutch model GysBERT, are freely available through the Huggingface repository. More information on these (Early) Modern language models can be found at https://macberth.netlify.app/.

change,[10] but the work that does exist has already led to some interesting findings. Existing work ranges from more exploratory work, which focuses on the extent to which we can rely on distributed meaning representations in retrieval and automated change detection, to more explanatory work, in which word embeddings are utilized to help test hypotheses regarding the nature of and mechanisms underlying constructional variation and change.

Exploratory work using embeddings often investigates whether functional-semantic changes of linguistic items can be detected in a data-driven, fully bottom-up fashion, and whether it is possible to automatically trace different types of functional-semantic change. Hamilton et al. (2016a), for instance, look into whether and how functional-semantic change can be automatically flagged in large-scale corpora. In doing so, they use type embeddings to consider lexical/conceptual change, but also illustratively tackle the subjectification of English *actually* (e.g. *... dinners which you have actually eaten* > *I actually agree*), the development of *must* from a deontic (e.g. *You must listen!*) to an epistemic modal (e.g. *The bread must have been stale*), and the grammaticalization of *promise* (e.g. *I promise to pay you* > *The weather promises to be good*). The aim of Hamilton et al. (2016a) is not to test any hypotheses regarding the development of *actually*, *must* or *promise*, but to demonstrate that the automated detection of 'linguistic change' and 'cultural' or conceptual change requires different metrics. Linguistic change (such as the functional-semantic developments underlying grammaticalization), which tends to be systematic and regular, is best captured by combining word embeddings with a global measure (i.e. the cosine distance between a word's embeddings in two consecutive time stages). Conceptual change, which tends to be irregular, is best captured by a local neighbourhood measure (using cosine distances between second-order vectors based on the target word and its immediate nearest-neighbours in two consecutive time stages). Depending on the type of change the researcher wishes to detect or flag in a corpus, then, they may wish to resort to different measures. To automatically flag which type of change has affected particular linguistic constructions, researchers may also resort to distributed token representations, as demonstrated by, for instance, Sagi et al. (2011), who briefly touch on the development of English *do* over the course of the Middle and Early Modern period (see Section 4).

Further exploratory work has also focused on how distributed token representations can be employed to automate certain aspects of linguistic annotation. Hilpert & Correia Saavedra (2017) set out to employ distributed to-

10 Within diachronic construction grammar, however, collocational and collexeme analyses are very popular (e.g. Hilpert 2011, Coussé 2014). In some ways, collexeme analyses can be considered the predecessor of the distributional approaches surveyed here (Heylen et al. 2015).

ken representations to automatically disambiguate between lexical uses of the verb form *used* (e.g. *Is that the hammer she used to kill him?*) and the grammaticalized habitual construction *used to* (e.g. *She used to like me*). [11] In a similar vein, Fonteyn (2020) employs a predictive language model (BERT; Devlin, Chang, Lee & Toutanova 2019) to automatically disambiguate and trace the development of the various senses of 'BE about' (e.g. descriptive *The lyrics were about her childhood*, *Life is about money and success* vs. grammaticalized uses such as approximative *She was about twenty years old* and futurate *The train was about to leave*) in Late Modern English (1800-2000). Again, this work predominantly focuses on showing *that* token-based distributional models can be used to classify different uses of a construction across time into sense categories (and analysing any classification errors to flag areas of improvement). As such, these studies lay the groundwork for more explanatory work, which uses the models to ultimate address specific questions about constructional change.

An interesting example of work that can be considered more explanatory is Perek (2016, 2018), which employs count models to home in on the development of *hell*-construction (e.g. [*beat*/*scare*/*hug*] *the hell out of someone*) and the *way*-construction (e.g. [*swim*/*beat*/*smile*] *one's way to something*) in Late Modern English. In doing so, Perek demonstrates that computational methods can be used to disentangle changes in productivity (measured by the number of unique lexical items that occur in the open slots of a morphosyntactic construction) from changes in a construction's schematicity (measured by the semantic diversity of those lexical items), and help provide support for previously proposed hypotheses on the determinants of syntactic productivity.

With respect to predictive models, it appears that historical linguists have been more wary to use them in explanatory work. This may be due to the fact that their increased performance has come at the cost of model transparancy (Linzen, Chrupała, Belinkov & Hupkes 2019: iii). Still, there is some notable work that focuses more prominently on explaining linguistic change. Using type embeddings (word2vec; Mikolov, Chen, Corrado & Dean 2013), Luo, Jurafsky & Levin (2019) formalize and measure the 'semantic bleaching' of grammaticalizing adverbs by comparing 250 adverbial intensifiers (e.g. *awfully*) to those of their source adjectives (e.g. *awful*) over time. Besides suggesting that 'semantic bleaching' can be operationalized as a decrease in similarity to the source adjective (and an increase in similarity to the 'fully

11 Unfortunately, the analysis "does not yield a satisfactory classification accuracy" (Hilpert & Correia Saavedra 2017: 25). Part of the issue could be, as Hilpert & Correia Saavedra (2017) suggest, the imbalance in the data set, as the lexical uses of *used* vastly outnumber the habitual ones. It could also possible that a model that integrates or infers more structural information from the context will achieve a higher accuracy.

bleached' intensifier *very*), their findings also lead them to suggest that 'bleaching' is triggered when the adverb collocates with adjectives that have a highly similar meaning at a given point in time (e.g. *awfully disgusting*).[12]

A more narrow-scoped study on intensifying expressions is Fonteyn & Manjavacas (2021), which homes in on the grammaticalization of *to death* from a phrase that expresses the result of an action (e.g. *He was beaten/shot to death*) to an 'amplifying' expression (e.g. *We were shocked/pleased to death to see you*). Over the course of the 16th and 17th centuries, *to death* sporadically started occurring in contexts where a literal, death-resulting reading is ruled out (e.g. *That book bored me to death.*). These non-literal, intensifying cases started to appear more frequently the 18th century onwards (Margerie 2011, Hoeksema & Napoli 2008). To probe into this expansion of *to death*, Fonteyn & Manjavacas (2021) draw on the research design of Perek (2016, 2018) and operationalize the process as a change in the structure of the semantic space that the collocate verbs of *to death* occupy, relying on hierarchical cluster analysis over the type embeddings (obtained with `word2vec`) of *to death*'s collocate verbs. Additionally, as the expansion of *to death* also involved increased co-occurrence with verbs with progressively more positive connotations, Fonteyn & Manjavacas (2021) devised a way to quantify the average polarity of verbs over time using word embeddings.

Focusing on Early Modern Scientific writing, Bizzoni et al. (2020) and Bizzoni, Degaetano-Ortlieb, Menzel, Krielke & Teich (2019) employ a predictive type-embedding model that is sensitive to the word order of the context (Ling, Dyer, Black & Trancoso 2015) to trace patterns of lexical as well as grammatical change – that is, changes in the use of function words (i.e. determiners, conjunctions and adpositions) and deverbal *ing*-forms (extracted by means of the 'gerund' and 'participle' part-of-speech tags VVG [e.g. *making, smiling*, etc.], VBG [e.g. *being*], and VHG [e.g. *having*] – in the Royal Society Corpus (Kermes, Degaetano-Ortlieb, Khamis, Knappen & Teich 2016). They find that lexical and function words behave differently: while the average lexical word in the underlying corpus undergoes specialization, function words do not. With respect to *ing*-forms, their word embeddings are used to compute nearest-neighbour clusters for each *ing*-form per decade in the corpus to examine not only the average size of their clusters, but also the average similarity between (and extent to which *ing*-forms cluster together) over time in scientific writing. The effect they observe, which can be summarized as a diminishing of cluster sizes across time, can essentially be linked back to the generally observed specialization of their lexical base verbs: as the base

12 While Luo et al. (2019) suggest causation by calling such examples 'bridging contexts', an alternative explanation may be that their co-occurrence is the result of the adverb's bleaching.

verbs becomes more specific over time, "less overlap between their contexts is observed" (Bizzoni et al. 2019: 180). At the same time, the analysis also reveals that the base verbs of gerundive and participial *ing*-forms generally fall into three semantic clusters (i.e. 'academic verbs' [e.g. *ascertaining, determining, examining*, 'change-of-state verbs' [e.g. *purifying, warming, cooling*] and 'motion verbs' [e.g. *passing, extending, running*] that occur as either gerunds or participles to different extents over time, revealing some more subtle tendencies of how scientific writing developed stylistically. In a similar vein, Krielke, Fischer, Degaetano-Ortlieb & Teich (2019) employ diachronic word embedding to explore change in the *wh*-relativizer paradigm in Early and Late modern English scientific writing.

Finally, token-embeddings are increasingly often employed to investigate the degree to which different constructions can in fact be distinguished based on their distributional properties. Examining the distributional overlap between the grammaticalized English core modals, Hilpert & Flach (2020) use distributed token representations to show that *may* and *might* in Present-day English can only be distributionally distinguished to a moderate degree, linking their findings to Correia Saavedra (2019: 98)'s observation that a high degree of grammaticalization correlates positively with a high collocate diversity. Embeddings generated by predictive models have also been utilized to chart the degree of mutual similarity between the English core modal auxiliaries and periphrastic *do* in Early Modern English, revealing, among other things, a divergence in the behaviour of paradigmatic competitors *doth* and *does*: while the distribution of *doth* has always been relatively similar to that of modal auxiliaries like *will*, *does* appears to have resisted such auxiliary uses until at least the mid-seventeenth century (Budts & Petré 2020).

In what follows, we will continue to focus on the development of periphrastic *do* in Early Modern English. In doing so, we demonstrate how predictive models can facilitate linguistic annotation and analysis, as well as examine and assess the way in which the predictive model captures functional-semantic overlap between morphosyntactic constructions.

4 Case Study: Auxiliary DO

One of the few cases of morphosyntactic change that has received some more elaborate attention in prior work with computational DSMs is the development of English *do*. In Present-day English, When no auxiliary is present, the English verb *do* functions as a semantically empty but syntactically obligatory operator in contexts of Negation, Inversion (in questions), Coding and Emphasis (the "NICE" contexts Huddleston 1976: 333-334):

(6) I do not eat meat. (*I eat not meat.)

(7) Do you eat meat? (*Eat you meat?)

(8) You really don't eat meat, do you? (*You really don't eat meat, eat you?)

(9) I DO eat meat! (*I EAT meat)

The syntactically regulated use of *do*, which goes by the names 'periphrastic *do*', 'do-support', 'dummy *do*' or 'auxiliary *do*', is a salient property of English syntax, not in the least because it is so uniquely English: in no other Germanic language has the periphrastic structure reached the level of categoriality it has obtained in Present-day English. With respect to its historical development, it has been suggested that *do* did occur as an auxiliary in early records, but it appeared to be restricted to a limited set of constructions where the verb straightforwardly contributed to the meaning of the clause (e.g. evoking causative meaning, as in *did him gyuen up* 'made him give up', the Peterborough Chronicle, ca. 1154). The rise of auxiliary *do* as a semantically empty, periphrastic structure, then, started in the Middle English period, and its usage was regulated over the Early and Late Modern English period (also see Oudesluijs, Gordon & Auer, this volume).

Perhaps the most canonical study on the history of periphrastic *do* dates back to the 1950s. On the basis of a dataset of a then unprecedented scale, Ellegård (1953) tallied the attestations of *do* and simple verb forms in random samples drawn from texts written between 1400 and 1700. Subsequently, he worked out their relative frequencies (as compared to simple verb forms) in a broad range of clause types corresponding with the above-mentioned NICE-contexts (affirmative declaratives, negative decalaratives, affirmative questions, negative questions, negative imperatives). As such, he found that around 1550 *do* was present in about 80% of all negative questions, but only in 50% of affirmative questions and a mere 35% of negative declaratives. By 1700, these rates went up substantially for all syntactic environments, with the exception of affirmative declaratives.

Because Ellegård's work has been so influential, the counts he obtained and vantage point from which he approached the regulation of *do* has tacitly been reused in subsequent studies.[13] Sagi et al. (2011), for instance, use Ellegård's account to evaluate the performance of a count DSM to automatically

13 Some noteworthy case studies on the regulation of periphrastic *do*, which often involve novel types of statistical methods, include Kroch (1989), Ogura (1993), Vulanovic (2005), Kauhanen & Walkden (2017)

flag semantic broadening. They start from the observation that, as *do* developed into an auxiliary, its contextual distribution becomes more varied over the course of the Middle and Early Modern English period (1150-1700). Approaching this development by means of Latent Semantic Analysis (a variation of count models), Sagi et al. (2011) constructed distributed meaning representations of each individual token of *do* in the Penn-Helsinki suite of Historical corpora (Kroch 2020). Subsequently, they take the cosine distance between individual tokens to measure the semantic density of *do* in different time bins, and show that tokens in earlier time periods in fact constitute a denser group than those in later time periods. These changes in semantic density, Sagi et al. (2011) show, appear to correspond with Ellegård (1953)'s manually annotated data (see Figure 1).

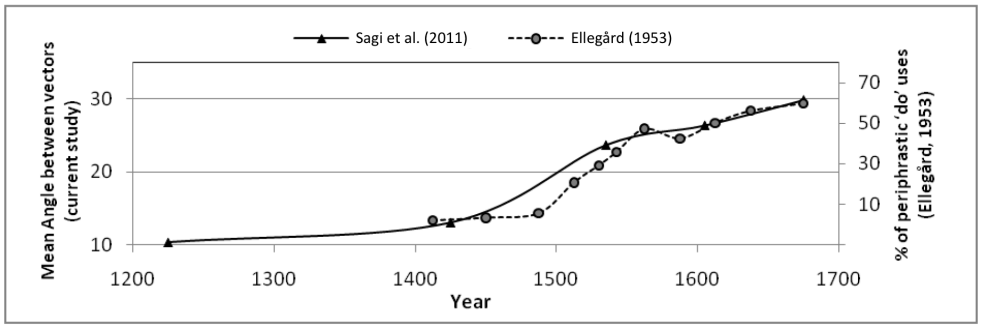

Figure 1 Figure from Sagi et al. (2011: 20): A comparison of the rise of periphrastic *do* as measured by semantic density and the proportion of periphrastic uses of do by Ellegard (1953).

The authors take this correspondence to indicate that the method will be valuable to statistically support previously attested developments in established cases of semantic change (such as the semantic development of auxiliary *do*), but also to identify new cases of functional-semantic change in a bottom-up manner.

Beyond applications where the main aim is to see whether distributed meaning representations can be used to automatically trace the semantic broadening of *do* as it developed from a lexical verb to auxiliary, it may also be possible to manipulate computational DSMs to investigate the causes underlying the observed distributional changes (Budts 2020b,a, Budts & Petré 2020). In the case of periphrastic *do*, one hypothesis as to why the construction acquired its present day distribution revolves around analogy with the English core modals –*can, may, must, shall* and *will*. As the core modals had become signif-

icantly distinct from main verbs around 1550 – shortly before periphrastic *do* settled in its eventual distribution – the closeness in timing has raised questions about whether the developments are connected (Warner 1993: 221): perhaps the newly established modals served as an analogical model for *do*, and collectively helped it to acquire full auxiliary status in the subsequent century.[14] Yet, the hypothesis that the modal auxiliaries steered the regulation of periphrastic *do* as proposed by Warner (1993) is highly speculative, and it has proven difficult to provide empirical evidence to support it. The lack of evidence is most likely due to pragmatic considerations: not only is analogy a notoriously-difficult-to-measure concept, the constructions involved in this particular case study are also among the most frequent items in the English language. Thus, even if a measure of analogical influence were to be found, investigating the analogical influence of the modal auxiliaries on periphrastic *do* by means of manual annotation is bound to become an extremely laborious endeavour.

4.1 Method

To establish whether it is reasonable to postulate analogy between *do* and *can, may, must, shall* and *will*, one essentially needs to establish whether it is reasonable to postulate functional-semantic equivalence between them, and find a means of measuring the extent of functional equivalence. We take functional-semantic equivalence between *do* and any of the English core modals to mean not only that they are used to convey similar (modal) meanings, but also that they occur in similar text types, as the latter has been shown to be an important factor in earlier work on the development of periphrastic *do* (see Nurmi (1999); Oudesluijs, Gorden & Auer, this volume). To address this question in a data-driven and quantifiable manner, the distributional properties of the constructions under scrutiny are taken to serve as a proxy of their functional features, and no a priori assumptions will be made of where *do* and the core modals may differ or overlap.

The procedure is as follows: first, the `word2vec` algorithm was used to compute word type embeddings based on Antigoon, a large training corpus of 16th and 17th century English (Budts 2020b). The raw material included in Antigoon has been drawn from the EEBO-TCP database. EEBO, short for Early English Books Online, is a comprehensive database that comprises nearly all texts printed in England between 1473 and 1700 that survived the passage of time. The texts have been manually transcribed by the Text Cre-

14 The hypothesis can either serve as an alternative or a supplement to account relying solely on systemic pressures, such as the V-to-I raising account proposed by for instance Kroch (1989).

ation Partnership (TCP) and are available in TEI-compliant xml-format. Because of their scope, availability and uniformity, EEBO data are well suited for a large-scale, partly automated corpus study into the diachronic evolution of high-frequency items such as auxiliary verbs. To assess the eligibility of individual EEBO texts for their inclusion in Antigoon, four criteria were considered: (i) the main language of the text had to be English; (ii) the text had to be published between 1580 and 1700; (iii) it could not be published posthumously; and (iv) the text could not be identical to another text in EEBO that was published earlier. Unfortunately, the EEBO source material lacks consistent genre tags, which means genre cannot be considered as a factor in the actual analysis without extensive manual enriching of the EEBO corpus data.[15]

Then, all attestations of periphrastic *do* and the core modals are collected from the same corpus using a fixed context window of 50 words.[16] The Antigoon corpus is divided into six 20-year periods. In the present case study, we will focus on the attestations in three periods: 1580-1600 (henceforth: Period 1), 1620-1640 (henceforth: Period 2) and 1680-1700 (henceforth: Period 3). Subsequently, each word in the collected utterance sequences is replaced by their word type embedding. In other words: the input utterances are conceptualized as a sequence of words, which can be represented in embedding form. After this transformation, all occurrences of *do*, *can*, *may*, *must*, *shall* and *will* are replaced by a generic <target> vector in order to mask them.

These transformed masked input sequences are then fed to a Convolutional Neural Network (CNN), which is tasked with reconstructing which of the candidate forms (i.e. *do* or one of the core modals) is placed in the masked target slot. This method can be framed as just one example of the type-based predictive approaches surveyed in Section 2. Similarly to `word2vec` – the paradigmatic algorithm for extracting type-based predictive meaning representations – the current approach relies on co-occurring words in order to compute a vector for the target word. However, this approach differs from `word2vec` in that (i) it uses a CNN layer (instead of a linear layer) in order to produce vector representations that are sensitive to word order, and (ii) in that it focuses on the model's predictions on the identity of the masked word to make (instead of relying on arithmetic on the generated vector representations in order to infer semantic properties of the represented words).

15 Budts (2020b) does provide an estimate of genre balance across the periods of Antigoon. For more detailed information on Antigoon, see Budts (2020b: Ch. 3).

16 A context window of 50 words around the masked "is short enough to be computationally tractable, but long enough for a human annotator to make sense of the context and make a reasonable guess as to the identity of the missing target"(Budts 2020b: 135), which helps in the manual evaluation of the model's output.

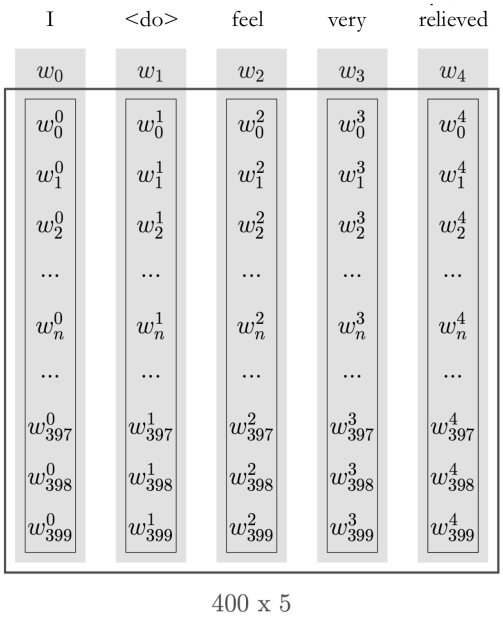

Figure 2 Horizontal concatenation of word embeddings (400 dimensions) turns 5-word sentence into 2D grid

To do this, the distributed (type) representations of the ordered words in the input sequences are concatenated into a two-dimensional matrix (for similar applications with CNNs, albeit not with historical language, see Collobert & Weston 2008, Dauphin, Fan, Auli & Grangier 2017, Vanni, Ducoffe, Aguilar, Precioso & Mayaffre 2018). In this matrix, one axis represents the input sequences, whereas the other axis contains the compressed numerical information of the n-dimensional embeddings of each word in the utterance (or: the numerical representation of the context in which each word in the utterance occurs in the corpus on which the `word2vec` model has been trained; see Figure 2). Once the utterances have been arranged as a two-dimensional matrix, the matrix can be 'convolved' with 'filters' of varying lengths. These filters can be understood as N-gram windows sliding down over the axis of the input words (see Figure 3).

The output of the procedure is exemplified in Table 3, where the context of two masked input sequences, one originally containing *do* and one originally containing *may*, are shown. The masked input sequences were fed to the CNN, which was forced to learn which (groups of) words in the con-

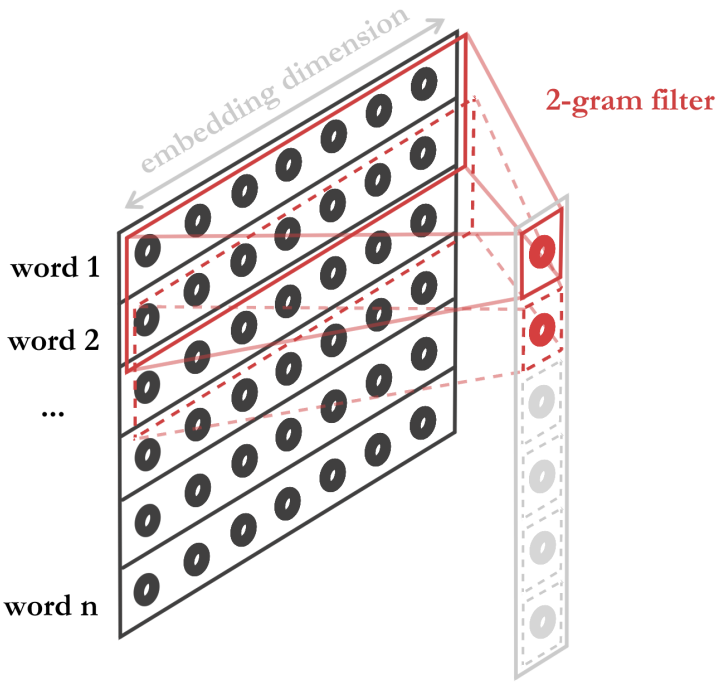

Figure 3 Visual representation of 2-gram filter

text are most predictive of, respectively, *do* and *may*. If we repeat this procedure for all collected sequences containing either a modal or a form of *do*, the algorithm is able to generalize over the individual attestations and pick up on more abstract contextual features that maximally discriminate *do* from each of the modals, as well as the modals among themselves. To illustrate, the first example is a context that originally contained *may*. Here, the model correctly assigned a significantly higher probability to *may* (0.98) than to all other forms (0.02). In the second example, which originally contained *do*, the model recognized *do*' s suitability (0.46), but it puts forward *can* as a likely candidate too (0.45). Other forms, such as *may*, are considered to be much less likely options (0.02).

In order to make this prediction, the CNN has essentially learned to find contextual features that help distinguish 16th and 17th century auxiliary *do*

input sequence	output label	predicted (prob. dist.)
unto his will in all things. Amen. O Lord increase my faith. O Lord open thou my lips, that my mouth **<target>** extol thee with praise, and be thankful unto thee for my benefits, & grant that I speak nothing but that which may	***may***	*may* (0.98)
amongst which the first is our sins, not only those that be mortal, but also venial sins, because these, albeit they **<target>** not extinguish charity in vs, yet do they slack and make cold the fervor of charity, which is as it were devotion	***do***	*do* (0.46) *can* (0.45) *may* (0.02)

Table 3 Probability distribution over competitors (do vs. core modals)

and modal verbs from one another, and this information can serve as the basis for further analysis:

- During training, the CNN learns to compute for each context a probability distribution over the competing variants. For every context, the algorithm assigns a score to each competitor, as an indicator of the degree to which the competitor fits in that particular context. Probability distributions as quantitative means of flagging of 'prototypical' cases, where only one form is deemed suitable, as well as ambiguous cases where multiple forms seem suitable (i.e. potential cases of overlap). By inspecting a large sample of cases that were highly likely to host one form, one can use the output of the CNN to reconstruct the prototypical use(s) of that form in opposition to competitive forms in the alternation (as done by Budts 2020b,a). In Section 4.2, we focus on mapping the extent to which *do* and the core modals are considered to overlap in the 16th and 17th century.
- To gain more insight into what the CNN considers to be distinctive

contextual features of *do* versus the English core modals, it is possible to examine (groups of) filters. This could be a valuable source of information from a theoretical point of view, because it provides insight into the factors that determine the choice between all forms involved in a completely bottom-up fashion. However, as will be pointed out in Section 4.3, the patterns the CNN considers discriminatory may not always be interpretable.

More details on the procedure (as well as a reference to the code to perform the analysis) can be found in the Appendix.

4.2 Mapping overlap

To attain a rough picture of how the functional overlap between *do* and the modal auxiliaries changed from the 16th to the 17th century, the proportion of attestations where both a modal and a form of *do* had been assigned a score between 0.25 and 0.75 were computed for each modal. A manual evaluation of these 'overlap cases' revealed that the vast majority of flagged cases turned out to be contexts that could actually host both forms selected, suggesting that the models output is reliable, and that the 0.25-0.75 margin is likely to be a conservative range.

From the manual inspection of the overlap cases reveals in the late 16th century, it also becomes evident that the functional similarities between *do* and the core modals were rooted in *do*'s use in affirmative declaratives (Budts 2020b,a). Affirmative declarative *do* served as an emphatic marker of truth (also see Nurmi 1999, Stein 1990), either strengthening the inherent truth value of universally valid propositions, or imposing a sense of truthfulness onto propositions whose truth value is not beyond doubt. In these essentially modal uses, the construction showed clear functional parallels with certain uses of the other modal auxiliaries: the universal/habitual sense of affirmative declarative *do* aligned well with, for instance, *must* as a marker of logical entailment, with *shall* for predictions in legal texts, and above all with *will* in universal truths and complaints about persistent habits. A more specialised use of affirmative declarative *do* in scientific writing was reminiscent of the use of *will* in the same genre, as well as the use of *may* in tentative lists of causes of scientific phenomena.

(10) The same corruption **must** of necessity happen unto the flesh of Christ as well as into ours [**0.42 *must*; 0.46 *doth***]

(11) And if the said T. N. his heirs, executors or administrators, **do** fail or

make default, and do not well and truly acquit, discharge, or save harmless the said T. S. G. F [⋯] [**0.52 *do*; 0.45 *shall***]

(12) ... the juice whereof **will** cause the skin to blister: some call it the travellers joy. [**0.33 *will*; 0.63 *doth***]

(13) This impediment **doth** come of corrupt gross flume, certain times it **doth** come of caterva, some times of a pleurisy, it **may** come of superabundance of other gross humours. [**0.61 *doth*; 0.36 *may***]

Another specialised use of *do* that was commonly found in argumentative prose linked the construction to evidential *may*, where *do* and *may* explicitly invite the reader to draw a conclusion based on the evidence provided, while its occurrence with first person subjects and verbs of communication echoed the use of first person *will* and *must* with performatives.

(14) From which definition we **may** clearly gather, that the cause and fountain of contingency is the free will of man [**0.71 *may*; 0.27 *do***]

(15) Which here I **do** omit for brevity sake. [**0.34 *do*; 0.61 *will***]

(16) We **do** confess that we do believe in Iesus our lord. [**0.31 *do*; 0.63 *must***]

Finally, the universal/habitual sense of *do* also aligned well with *can* when it expressed generic negation. Both *doth* and *do* overlap with *can* in clauses with third person subjects expressing universal impossibility. Indeed, in sentences like (17) and (18), the two forms are naturally close in meaning: as *can* expresses the impossibility of a situation to occur or the inability of a person to perform an action. In terms of truth value, it is equivalent to epistemic *do* expressing emphatically that a situation does not hold or that someone did not perform an action. Even though *can* and *do* themselves have different semantics, they by and large occur in similar environments.

(17) Our bodies and souls **do** not make vs members of Christ, but our faith and obedience. [0.41 *do*; 0.49 *can*]

(18) They which **can** not valiantly expose themselves to dangers, become slaves to those which assail them. [0.40 *can*; 0.52 *do*]

Having evaluated the reliability and conservativeness of the model's output and the estimate range, it is then possible to look into the estimated overlap proportions between *do* and the modals over time. Table 4 provides a summary of the overlap estimates for each pair of forms between 1580 and 1600 (Period 1), 1620 and 1640 (Period 2) and 1680 and 1700 (Period 3).

Modal	overlap with	Period 1	Period 2	Period 3	trend
can	do	2.44%	1.99%	3.33%	none
	doth/does	2.12%	1.69%	1.81%	decrease
may	do	1.27%	0.71%	0.62%	decrease
	doth/does	1.4%	0.74%	0.48%	decrease
must	do	1.22%	0.5%	0.62%	decrease
	doth/does	0.53%	0.37%	0.27%	decrease
shall	do	1.05%	0.54%	0.43%	decrease
	doth/does	1.13%	0.55%	0.25%	decrease
will	do	2.81%	1.6%	1.47%	decrease
	doth/does	2.02%	1.59%	1.22%	decrease

Table 4 Proportion of modal attestations ambiguous with forms of do

The figures in Table 4 suggest that the similarity between *do* and the modals systematically decreased over the three periods examined. Their pairwise overlap in the late 17th century tends to be just half the size of their late sixteenth century counterparts for nearly all pairs involved. The only exception is *can*: while the overlap between *can* and *does*/*doth* decreases as well, the change is only moderate in comparison to corresponding shifts with the other modals, and while the overlap between *can* and *do* slightly drops between the first and second period, this loss is made up for by a sharp increase during the transition to the final period under scrutiny.

Returning to the question whether the English core modals played a role in the regulation process of periphrastic *do*, the data suggest that the estimated distributional similarity between *do* and the modals – which can be taken to serve as a proxy for their functional similarity – had peaked by the late 16th century, and dwindled after. As such, Budts (2020b) suggests, the figures do not support the hypothesis that the 17th century regulation of periphrastic *do* was governed by analogical forces resulting from similarity with the modal auxiliaries – as it appears their functional similarity had become quite limited at the time.

4.3 Filters and contextual patterns

When the CNN generates probability distributions of competing variants in specific contexts, it provides output that can subsequently be examined manually, as illustrated in Section 4.2. This manual analysis can focus on determining the quality of the model's predictions. If the predictions are sensible, the analyst may continue to examine what sort of contextual cues appear to characterize prototypical contexts for a given form, or which sort of contexts are likely to host multiple forms. It is interesting to note, however, that the algorithm employed to generate the probability distributions also stores information regarding contextual cues that it finds most informative to predict which form can be placed in the masked <target> slot.

In classifying the contexts according to the form they are most likely to host, the CNN gradually grows sensitive to contextual features that are predictive of one of the competitors, but not the others. These contextual features come in the shape of N-grams. Importantly, though, the N-grams do not need to be the same literal string every time. Instead, the model grows sensitive to (combinations of) sub-word features, such as verb semantics or syntactic category. This flexibility in terms of feature extraction stems directly from the input representation: because the model operates on the embeddings of the input words rather than the input words themselves, it has access to information about the behaviour of the words in a large corpus. This implies that the model grows sensitive to abstract N-gram templates that are tailored for the alternation at hand, in that they encode exactly the patterns that allow the model to discriminate best between the competitors in the alternation. An interesting question that arises, then, is which patterns that the CNN considers discriminatory in the case study at hand, and to what extent these patterns correspond to the patterns that humans would find informative. For the procedure adopted in this case study, one could, for instance, try and gain insight into what the filters of the CNN in fact attend to when they (successfully) discriminate between *do* and each of the modals.[17]

When assessing the patterns the CNN has grown sensitive to, it seems reasonable to state that the model's judgements have been informed by recurrent usage templates. From a computational angle, these templates are essentially just non-linear combinations of many low-level cues. From a linguistic perspective, the templates could be seen as bundles of formal or distributional characteristics at different levels of specificity that jointly evoke certain semantic and pragmatic features. Each target is typically associated with several of these templates, and the same template may be involved in the

17 For a detailed explanation on how to extract this information, see Budts (2020b: Ch. 5).

prediction of various targets. In some cases, the model appears to have grown sensitive to lexically specific patterns. They include all patterns where the filter has grown maximally sensitive to a specific word in one position in a given N-gram and is heavily underspecified for all other positions. One example is a 5-gram filter that has grown sensitive to the generic <target> vector at position 2 and *needs* (or a near-synonym) at position 3 (e.g. *vengeance <target> needs be when*, *they <target> needs find out*).

A second group of filters has grown sensitive to N-grams that are lexically varied but semantically coherent. The first filter has grown sensitive to markers of evidentiality. The optimal stimuli for this filter all encode cases where the writer reports the source of a claim made elsewhere. The second filter has grown sensitive to N-grams where the subject or speaker expresses the desirability of some underspecified action. This group of filters is, to a certain extent, more flexible than the fully lexically specified group of filters, as it allows for more variation as to how its target meaning is conveyed.

Evidentiality	Desirability
by this testimony you <target>perceive what	indeed it is fit we <target>have
as <target>plainly appear by the examples	lest any man <target>boast himself:
as hath been plainly declared. what	that it was fit men <target>get
as hath been already showed, in	: it was necessary wee <target>have
the third canon you <target>find mention	lest with the world they <target>be
and by it we <target>discern this	broken, it is requisite it <target>be

Table 5 Semantically specified 7-gram filters

A third group of filters has come to attend to syntactic patterns. The filter on the left has grown sensitive to a combination of an auxiliary and a main verb, followed by a second person object. There is variation in the nature of the objects, as the filter seems to select both prepositional objects and indirect objects, with and without prepositions. This clearly reveals that the model has not made a full syntactic parse of the sentence, but at the same time it shows that flat parses and low-level generalisations can serve as useful approximations of high-level syntactic structures that would be more difficult to parse in their entirety. The second syntactic pattern contains one lexically specific slot – it requires a form of *do* at position 5 – but the other slots have remarkably abstract selection preferences. Here, it appears the filter has grown sensitive to unusual word order, where an NP subject is separated from the finite *do* form by means of an intervening constituent. This word order is common in poetry, where *do* is used for metric reasons. Note that the form of *do*

Second person objects	Unusual word order
oil as wee <target>show you hereafter	indeed it is fit we <target>have
as he <target>slay from you	silver swans by me did ride
you god <target>fight for you	then thy stone invisibly doth fall
i <target>not open unto thee:	true faith in christ doth breed
peace <target>they add to thee	the muses nine, do take
my general history <target>show you	god by adversities doth make

Table 6 Syntactially specified 7-gran and 6-gram filters

selected by slot 5 is not masked. This means that the forms attested here are not the targets of the context they occur in, but that they have been retrieved from the context of another target. This unmasked occurrence of *do* allowed the model to associate the construction directly with uncommon word order, an association which is likely to help it in identifying masked forms of *do* as well, when they occur with intervening constituents in target position.

Another category of filters appears to have grown sensitive to lexical items that are not straightforwardly related to the alternation between *do* and the modals, but whose discriminatory relevance is indirect. The lexical items the filters singled out serve as markers of a textual genre that correlates more with some competitors than others. A first example of this is a filter that learned to attend to unusual word order patterns. This filter probably served as an indicator of poetry, a genre that appeared to correlate with the use of *do*. Another filter learned to attend to *the said N*, which probably served as a marker of the legal genre. As legal texts correlate with *shall* and *do*, these competitors are assigned a higher probability whenever *the said N* is attested in the context. This indicates that not all features identified by the filters should be regarded as immediate regulators of the alternations: while they are all genuine tendencies in the data set, some of them are proxies for another, non-explicitly fed variable underlying variable that governs the alternation.

Finally, a minority of filters does not display clear tendencies at all. From these stimuli, it is hard to deduce what exactly the filter has grown sensitive to. One such uninterpretable pattern is illustrated in Table 7.

While we do not take the inspection of the filters to be a proper means of evaluating the reliability of the method employed, it does offer a interesting picture of what sort of features the model attends to. The wide array of features reflects the semantic and pragmatic pluriformity of *do* and the modals

Lexical register marker	Uninterpretable
patents unto the said subjects	other course turn away his just judgements
examined by the said commissioners	i know you would have pardoned him
, until the said fifth	he <target>without impeachment of his justice
officers of the said garrison	i <target>send her safe unto my
themselves within the said shires	duke of lorraine <target>be restored to

Table 7 5-gram lexically specific register marker (i.e. 'the said'; left) and 7-gram uninterpretable pattern (right)

in Early Modern English, and most features extracted by the model can be related to genuine high-level semantic and/or syntactic patterns that either help set apart *do* from the modals or indicate equivalence between them.

Yet, it should be kept in mind that whatever the model considers informative does not (necessarily) map onto meaningful cues for humans. This is evidently clear by the fact that some patterns are uninterpretable – but it should also be stressed that the interpretable features extracted by the model are at most low-level approximations of the high-level structures commonly posited by linguists when they engage in functional-semantic analyses. A telling example is the model's perceived equivalence between *do* and *may* – both of which easily combine with *well* (a lexically specific pattern) and are thus deemed functionally equivalent. However, the alleged functional equivalence only exists superficially. With *may*, *well* functions as an epistemic adverb that qualifies the proposition at hand as likely, whereas *do well* is a construction in its own right that implements *do* as a lexical verb.

A similar phenomenon occurs with superficial similarities in the wider context. If a given input sequence constitutes a snippet from a legal document (including formulaic sequences typical of the genre), it will always receive a relatively high score for *shall*, even if the immediate context of the target is more compatible with and typical of another potential target. In other words, unlike the expert linguist annotator, the computational model is more easily fooled by local lexical overlap. Yet, at the same time, the examination of the filters also indicates that many of the high-level concepts linguists employ when analysing data can be approximated by computationally light-weight patterns that naturally emerge from raw input corpus data.

5 Conclusion and Remaining Challenges

Over the past few years, computational DSMs seem to have solidified their position as important retrieval, annotation and analytic tools in lexical semantic change research. In this paper, we point out that computational DSMs may prove just as valuable in the realm of morphosyntactic change. For a number of tasks, they can be seen as means to alleviate manual labour: computational DSMs can be used to automatically trace morphosyntactic constructions that have undergone functional-semantic change in a given diachronic corpus, or they can be used to automate certain tasks of linguistic annotation such as sense disambiguation. In both cases, computational models offer a data-driven alternative to tasks that would otherwise require a great deal of time-consuming manual labour from (multiple) annotators – and even if the output they generate still requires manual post-correction, they could prove a welcome addition to the historical linguist's toolkit. For other tasks, computational DSMs can provide a means of tackling questions that would be difficult to address by means of introspective annotation alone: the models can provide quantifiable, measurable information that can be manipulated, for instance, to measure (and distinguish) changes in a construction's productivity, schematicity and polarity (Perek 2016, 2018, Fonteyn & Manjavacas 2021), or, as demonstrated by the case study on *do* and the core modals, to quantify the extent of functional-semantic equivalence between different linguistic forms over time.

Of course, while the adoption of these models to research morphosyntactic variation and change has many merits – the survey and case study also pointed to some potential drawbacks, which warrant further discussion. The first and perhaps most obvious drawback concerns a practical issue. At present, essentially all computational DSMs require lots of data to produce usable output.[18] Fortunately, when it comes to Early Modern Germanic Languages, there is a reasonable amount of digitized textual data, including a few very large library dumps (e.g. EEBO; Google Books; Delpher; Deutsches Text Archiv) – and while the preparation and pre-processing of these large-scale data sets (or aligning and balancing a collection of smaller, cleaner corpora) may still prove costly and labour-intensive, it should, in principle, only be done once if the processed data can be made freely accessible.

A related point to be made here is that the linguistic data the models process is exclusively representative of written language. Given that the liter-

18 There are some differences in this respect depending on the chosen approach. If the approach relies on predictive type-based models, one can already generate sufficiently high quality with a few hundred million words, whereas token-based embeddings from predictive language models are rarely trained on less than a few billion words.

acy rates were much lower in the Early Modern period than they are today (Stephens 1990: 555), any patterns the model detects therefore cannot be considered representative of the language at large. Of course, this rings true for the vast majority of historical corpus linguistics. Yet, because computational researchers in historical linguistics may have to prioritize quantity over balance to meet the data demands of the models they work with, their findings are more likely to be (heavily) biased towards certain (high-register) genres (e.g., non-fiction prose, religious treatises, sermons and legislation). A desideratum for computational research into Early Modern Linguistic Variation is, in short, the development of models and suitable procedures that can produce good quality results with smaller data sets.

Second, even if such optimized data and/or models become available, it is important to be aware of which sort of information can and cannot be extracted from (raw) corpus data by a computational DSM. Inspecting the CNN's filters, it became clear that many features extracted to compute the probability distribution of *do* and the core modals were relatable to genuine high-level syntactic patterns, but they remain low-level approximations of those syntactic patterns nonetheless. A possible way to integrate syntactic structure to construct distributed meaning representations that do not (solely) consist of lexically specific collocates, but (additionally) include morphosyntactic tags, as was done in Jenset (2013)'s study on Old to Early Modern English locative and existential adverbs (e.g. *there, here*). Be that as it may, the models and procedures adopted in our case study as well as the vast majority of the studies surveyed in Section 3 still require that the morphosyntactic constructions the researcher targets are at least partially lexically specified, because most models do not straightforwardly compute vector representations for schematic structures (e.g. word order patterns, passives, cleft-clauses, N-N compounds) or bound morphemes. This is not to say that it is entirely impossible to employ computational DSMs to study fully schematic morphosyntactic patterns, but, at present, it is difficult to say how effective the models will be in this respect, and what a suitable procedure for such endeavours would be. Relatedly, the case study also indicated that the models appear to pick up on text type variation, as it starts to associate certain modals with genre-specific N-grams. Again, such patterns can be seen as low-level approximations of a higher-level (in this case extra-linguistic) variable. Yet, neither the higher-level variable 'genre' nor the situational characteristics that define it were integrated into the procedure (as they are in Demske this volume; Oudesluijs, Gordon & Auer this volume) – but they could be if the model were explicitly exposed to corpus data enriched with such information.

Furthermore, there may also be some dimensions of meaning that are

simply not encoded in corpus data at all, and hence may not be detectable by fully corpus-driven computational methods (e.g. Bender & Koller 2020, Fonteyn 2021). While current work with DSMs has already highlighted that a considerable amount of lexico-semantic information can be extracted from pure word co-occurrence in corpora, the models still lack embedding in a real-world communicative setting where speakers/writers and hearers/readers may draw on extra-linguistic context, invited inferencing, or perceived similarity between concepts to convey or process a meaning. As such, it remains doubtful that the models and procedures we currently adopt can also encode the extra-linguistic information that lies at the basis of cognitive and communicative processes that drive language understanding and change.

In light of the preceding discussion, and given that we cannot simply assume that a given computational model will arrive at certain conclusions in the same way as a language user (or language expert) would, perhaps the most important question has become how the output of computational distributional models should be evaluated. NLP models for Present-day language are commonly evaluated by comparing their output to human 'gold standards' – that is, large datasets containing the judgements of native speakers of a given language variety (on semantic similarity, sense disambiguation, etc.). For linguistic data from the early modern period, such human gold standards do not exist. As such, the evaluation of the output with respect to historical data commonly takes the form of a confirmatory study, where a computational model has value inasmuch as it replicates results that corpus-based work that historical linguists have previously unearthed – or, in absence of extant work, the output requires thorough manual assessment by experts. Thus, we currently find ourselves at a stage where the very work these models are meant to evaluate (at unprecedented scales) is the same work that we use to evaluate them, and we are still exploring what types of questions can appropriately addressed with the models available to us. The key to moving beyond this stage lies in a collaborative effort between computational and historical linguists from different backgrounds and approaches. Such collaboration will be essential to provide (continued) support for the reliability of computational distributional modeling for a given type of research question, but also to devise robust external means of evaluation for historical language data.

References

Bamler, Robert & Stephan Mandt. 2017. Dynamic word embeddings. In Doina Precup & Yee Whye Teh (eds.), *Proceedings of the 34th international conference on machine learning*, vol. 70 Pro-

ceedings of machine learning research, 380–389. PMLR. Tex.pdf: http://proceedings.mlr.press/v70/bamler17a/bamler17a.pdf.

Baron, Alistair. 2011. *Dealing with Spelling Variation in Early Modern English Texts*. Lancaster: Lancaster University dissertation.

Baroni, Marco, Georgiana Dinu & Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (*volume 1: Long papers*), 238–247. Association for Computational Linguistics. doi:10.3115/v1/P14-1023. https://www.aclweb.org/anthology/P14-1023.

Bender, Emily M. & Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5185–5198. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.463. https://aclanthology.org/2020.acl-main.463.

Betti, Arianna, Martin Reynaert, Thijs Ossenkoppele, Yvette Oortwijn, Andrew Salway & Jelke Bloem. 2020. Expert Concept-Modeling Ground Truth Construction for Word Embeddings Evaluation in Concept-Focused Domains. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6690–6702. Barcelona, Spain (Online): International Committee on Computational Linguistics. doi:10.18653/v1/2020.coling-main.586.

Bizzoni, Yuri, Stefania Degaetano-Ortlieb, Peter Fankhauser & Elke Teich. 2020. Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach. *Frontiers in Artificial Intelligence* 3. 73. doi:10.3389/frai.2020.00073.

Bizzoni, Yuri, Stefania Degaetano-Ortlieb, Katrin Menzel, Pauline Krielke & Elke Teich. 2019. Grammar and Meaning: Analysing the Topology of Diachronic Word Embeddings. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 175–185. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/W19-4722.

Boleda, Gemma. 2020. Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics* 6(1). 213–234. doi:10.1146/annurev-linguistics-011619-030303. ArXiv: 1905.01896.

Bollmann, Marcel. 2019. A large-scale comparison of historical text normalization systems. 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*), 3885–3898. Minneapolis, Min-

nesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1389. https://aclanthology.org/N19-1389.

Bostrom, Kaj & Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the association for computational linguistics: Emnlp 2020*, 4617–4624. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.414. https://aclanthology.org/2020.findings-emnlp.414.

Budts, Sara. 2020a. A connectionist approach to analogy. on the modal meaning of periphrastic do in early modern english. *Corpus Linguistics and Linguistic Theory* doi:10.1515/cllt-2019-0080.

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

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

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

Collobert, Ronan & Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on machine learning* ICML '08, 160–167. New York, NY, USA: Association for Computing Machinery. doi:10.1145/1390156.1390177.

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault & Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2126–2136. Melbourne, Australia: Association for Computational Linguistics. doi:10.18653/v1/P18-1198. https://www.aclweb.org/anthology/P18-1198.

Correia Saavedra, David. 2019. *Measurements of Grammaticalization: Developing a quantitative index for the study of grammatical change*. Neuchâtel & Antwerpen: l' Université de Neuchâtel & Universiteit Antwerpen PhD dissertation.

Coussé, Evie. 2014. Lexical expansion in the HAVE and BE perfect in Dutch: A constructionist prototype account. *Diachronica* 31(2). 159–191. doi:10.1075/dia.31.2.01cou.

Dauphin, Yann N., Angela Fan, Michael Auli & David Grangier. 2017. Lan-

guage modeling with gated convolutional networks. In *Proceedings of the 34th international conference on machine learning - volume 70* ICML'17, 933–941. JMLR.org.

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

Del Tredici, Marco, Raquel Fernández & Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. 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)*, 2069–2075. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1210.

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

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

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi & Dominik Schlechtweg. 2019. Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 457–470. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1044. https://www.aclweb.org/anthology/P19-1044.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2017a. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, 1136–1145. Copenhagen, Denmark: Association for Computational Linguistics. doi:10.18653/v1/D17-1118. https://www.aclweb.org/anthology/D17-1118.

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

Ellegård, Alvar. 1953. *The auxiliary do. the establishment and regulation of its use in english*. Stockholm: Almqvist & Wiksell.

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

//www.aclweb.org/anthology/P10-2017.

Evans, Mel & Caroline Tagg. 2020. *Women' s spelling in early modern english: Perspectives from new media* 191–218. Cambridge University Press. doi:10.1017/9781108674171.010.

Fitzmaurice, Susan, Justyna A. Robinson, Marc Alexander, Iona C. Hine, Seth Mehl & Fraser Dallachy. 2017. Linguistic DNA: Investigating Conceptual Change in Early Modern English Discourse. *Studia Neophilologica* 89(sup1). 21–38. doi:10.1080/00393274.2017.1333891.

Fonteyn, Lauren. 2020. What about grammar? Using BERT embeddings to explore functional-semantic shifts of semi-lexical and grammatical constructions 12.

Fonteyn, Lauren. 2021. Varying abstractions: a conceptual vs. distributional view on prepositional polysemy. *Glossa: a journal of general linguistics* 6(1). doi:https://doi.org/10.5334/gjgl.1323.

Fonteyn, Lauren & Enrique Manjavacas. 2021. Adjusting scope: a computational approach to case-driven research on semantic change. In *Computational humanities research*, .

Gage, Philip. 1994. A new algorithm for data compression. *C Users Journal* 12(2). 23–38.

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

Goldberg, Yoav. 2019. Assessing BERT's Syntactic Abilities. *arXiv:1901.05287 [cs]* http://arxiv.org/abs/1901.05287. ArXiv: 1901.05287.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016a. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2116–2121. Austin, Texas: Association for Computational Linguistics. doi:10.18653/v1/D16-1229. https://www.aclweb.org/anthology/D16-1229.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 1489–1501. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-1141. https://www.aclweb.org/anthology/P16-1141.

Hengchen, Simon, Nina Tahmasebi, Dominik Schlechtweg & Haim Dubossarsky. 2021. Challenges for computational lexical semantic change,

doi:10.5281/ZENODO.5040322.

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

Hilpert, Martin. 2011. Diachronic collostructional analysis: How to use it and how to deal with confounding factors. In Kathryn Allan & Justyna A. Robinson (eds.), *Current Methods in Historical Semantics*, Berlin, Boston: DE GRUYTER. doi:10.1515/9783110252903.133.

Hilpert, Martin & David Correia Saavedra. 2017. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory* 0(0). doi:10.1515/cllt-2017-0009.

Hilpert, Martin & Susanne Flach. 2020. Disentangling modal meanings with distributional semantics. *Digital Scholarship in the Humanities* fqaa014. doi:10.1093/llc/fqaa014.

Hoeksema, Jack & Donna Jo Napoli. 2008. Just for the hell of it: A comparison of two taboo-term constructions. *Journal of Linguistics* 44(2). 347–378. doi:10.1017/S002222670800515X.

Hopper, Paul. 1991. On some principles of grammaticalisation. In Elizabeth Closs Traugott & Bernd Heine (eds.), *Approaches to grammaticalization*, vol. 1, 17–35. Amsterdam: John Benjamins.

Hu, Renfen, Shen Li & Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 3899–3908. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1379.

Huddleston, Rodney. 1976. Some theoretical issues in the description of the english verb. *Lingua* 40. 331—–383.

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

Jenset, Gard B. 2013. Mapping meaning with distributional methods: A diachronic corpus-based study of existential *there*. *Journal of Historical Linguistics* 3(2). 272–306. doi:10.1075/jhl.3.2.04jen.

Kauhanen, Henri & George Walkden. 2017. Deriving the constant rate effect. *Natural language and linguistic theory* 36(2). 483—–521.

Kermes, Hannah, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen & Elke Teich. 2016. The royal society corpus: From uncharted data to

corpus. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 1928–1931. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1305.

Krielke, Marie-Pauline, Stefan Fischer, Stefania Degaetano-Ortlieb & Elke Teich. 2019. System and use of wh-relativizers in 200 years of english scientific writing. In *Proceedings of the 10th international corpus linguistics conference*, Cardiff, Wales, UK. https://stefaniadegaetano.files.wordpress.com/2019/05/cl2019_paper_266.pdf.

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1. 199—244.

Kroch, Anthony. 2020. Penn Parsed Corpora of Historical English. https://www.ling.upenn.edu/hist-corpora/.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski & Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th international conference on computational linguistics*, 1384–1397. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

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

Lenci, Alessandro, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten & Martina Miliani. 2021. A comprehensive comparative evaluation and analysis of distributional semantic models.

Levy, Omer, Yoav Goldberg & Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics* 3. 211–225. doi:10.1162/tacl_a_00134. https://www.aclweb.org/anthology/Q15-1016.

Ling, Wang, Chris Dyer, Alan W Black & Isabel Trancoso. 2015. Two/Too Simple Adaptations of Word2Vec for Syntax Problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1299–1304. Denver, Colorado: Association for Computational Linguistics. doi:10.3115/v1/N15-1142. http://aclweb.org/anthology/N15-1142.

Linzen, Tal & Marco Baroni. 2021. Syntactic Structure from Deep Learning. *Annual Review of Linguistics* 7(1). 195–212. doi:10.1146/annurev-linguistics-032020-051035. ArXiv: 2004.10827.

Linzen, Tal, Grzegorz Chrupała, Yonatan Belinkov & Dieuwke Hupkes (eds.). 2019. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and In-*

terpreting Neural Networks for NLP. Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-4800.

Luo, Yiwei, Dan Jurafsky & Beth Levin. 2019. From Insanely Jealous to Insanely Delicious: Computational Models for the Semantic Bleaching of English Intensifiers. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 1–13. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/W19-4701. https://www.aclweb.org/anthology/W19-4701.

Manjavacas Arévalo, Enrique & Lauren Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the workshop on natural language processing for digital humanities (nlp4dh)*, 23–36. Association for Computational Linguistics. http://icon2021.nits.ac.in/resources/nlp4dh.pdf#page=35.

Manjavacas Arévalo, Enrique & Lauren Fonteyn. 2022. Adapting vs pre-training language models for historical languages. *Journal of Data Mining and Digital Humanities* .

Manning, Christopher D., Kevin Clark, John Hewitt, Urvashi Khandelwal & Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*. doi:https://doi.org/10.1073/pnas.1907367117.

Margerie, Hélène. 2011. Grammaticalising constructions: to death as a peripheral degree modifier. *Folia Linguistica Historica* 45(Historica vol. 32). doi:10.1515/flih.2011.005.

Mikolov, Tomás, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In Yoshua Bengio & Yann LeCun (eds.), *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, http://arxiv.org/abs/1301.3781.

Mitra, Sunny, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee & Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, 1020–1029. Baltimore, Maryland: Association for Computational Linguistics. doi:10.3115/v1/P14-1096.

Nurmi, Arja. 1999. *A social history of periphrastic DO*. Helsinki: Société Néophilologique.

Ogura, Mieko. 1993. The development of periphrastic do in english. a case of lexical diffusion in syntax. *Diachronica* 10(1). 51—85. doi:10.1075/dia.l0.1.04ogu.

Perek, Florent. 2016. Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics* 54(1). doi:10.1515/ling-2015-0043.

Perek, Florent. 2018. Recent change in the productivity and schematicity of the *way* -construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory* 14(1). 65–97. doi:10.1515/cllt-2016-0014.

Petré, Peter, Lynn Anthonissen, Sara Budts, Enrique Manjavacas Arévalo, Emma-Louise Silva, William Standing & Odile Aurora Oscar Strik. 2019. Early Modern Multiloquent Authors (EMMA): designing a large-scale corpus of individuals' languages. *ICAME Journal: computers in English linguistics* 43. 83–122.

Piotrowski, Michael. 2012. Natural Language Processing for historical texts. *Synthesis Lectures on Human Language Technologies* 5(2). 1–157.

Rosenfeld, Alex & Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)*, 474–484. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-1044.

Rudolph, Maja & David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 world wide web conference* WWW '18, 1003–1011. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. doi:10.1145/3178876.3185999. Number of pages: 9 Place: Lyon, France.

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

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky & Nina Tahmasebi. ???? SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection 23.

Schrimpf, Martin, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum & Evelina Fedorenko. 2020. Artificial neural networks accurately predict language processing in the brain. *bioRxiv* doi:10.1101/2020.06.26.174482.

Schuster, Mike & Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 ieee international conference on acoustics, speech and signal processing (icassp)*, 5149–5152. IEEE.

Scragg, D. G. 1974. *A History of English Spelling*. Manchester: Manchester University Press.

Sennrich, Rico, Barry Haddow & Alexandra Birch. 2016. Neural machine

translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 1715–1725. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-1162. https://aclanthology.org/P16-1162.

Sommerauer, Pia & Antske Fokkens. 2019. Conceptual Change and Distributional Semantic Models: An Exploratory Study on Pitfalls and Possibilities. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 223–233. Association for Computational Linguistics. doi:10.18653/v1/W19-4728. https://www.aclweb.org/anthology/W19-4728.

Stein, Dieter. 1990. *The semantics of syntactic change*. Berlin: Mouton de Gruyter. doi:10.1515/9783110846829.

Stephens, William B. 1990. Literacy in England, Scotland, and Wales, 1500-1900. *History of Education Quarterly* 30. 545—–571.

Sun, Kun, Haitao Liu & Wenxin Xiong. 2021. The evolutionary pattern of language in scientific writings: A case study of Philosophical Transactions of Royal Society (1665–1869). *Scientometrics* 126(2). 1695–1724. doi:10.1007/s11192-020-03816-8. http://link.springer.com/10.1007/s11192-020-03816-8.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2019. Survey of Computational Approaches to Lexical Semantic Change. *arXiv:1811.06278* [*cs*] http://arxiv.org/abs/1811.06278. ArXiv: 1811.06278.

Traugott, Elisabeth C. & Graeme Trousdale. 2013. *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.

Traugott, Elizabeth Closs. 1989. On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language* 57. 33–65.

Traugott, Elizabeth Closs. 2003. From subjectification to intersubjectification. In Raymond Hickey (ed.), *Motives for language change*, 124–139. Cambridge: Cambridge University Press.

Traugott, Elizabeth Closs. 2010. (Inter)subjectivity and (inter)subjectification: A reassessment. In Kristin Davidse, Lieven Vandelanotte & Hubert Cuyckens (eds.), *Subjectification, intersubjectification and grammaticalization*, 29–71. Berlin: Mouton De Gruyter.

Traugott, Elizabeth Closs & Richard Dasher. 2002. *Regularity in semantic change*. Cambridge: Cambridge University Press.

Traugott, Elizabeth Closs & Ekkehard König. 1991. The semantics-pragmatics of grammaticalization revisited. In Elizabeth Closs Traugott & Bernd Heine (eds.), *Approaches to grammaticalization*, vol. 1, 189–218. Amsterdam: John Benjamins.

Vanni, Laurent, Melanie Ducoffe, Carlos Aguilar, Frederic Precioso & Da-

mon Mayaffre. 2018. Textual deconvolution saliency (TDS) : a deep tool box for linguistic analysis. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 548–557. Melbourne, Australia: Association for Computational Linguistics. doi:10.18653/v1/P18-1051. https://aclanthology.org/P18-1051.

Vulanovic, Relja. 2005. The rise and fall of periphrastic do in affirmative declaratives: A grammar efficiency model. *Journal of quantitative linguistics* 12(1). 1—31.

Warner, Anthony R. 1993. *English auxiliaries: Structure and history*. Cambridge: Cambridge University Press.

Young, Tom, Devamanyu Hazarika, Soujanya Poria & Erik Cambria. 2018. Recent Trends in Deep Learning Based Natural Language Processing. *arXiv:1708.02709 [cs]* http://arxiv.org/abs/1708.02709. ArXiv: 1708.02709.

Lauren Fonteyn
Leiden University Centre for Linguistics
.....
l.fonteyn@hum.leidenuniv.nl

Enrique Manjavacas
Leiden University Centre for Linguistics
.....
enrique.manjavacas@gmail.com

Sara Budts
Universiteit Antwerpen
.....
sara.budts@uantwerpen.be

Appendix

The Convolutional Neural Networks (CNNs) described in this paper are based on the implementation of Vanni et al. (2018) and have been trained on data from the Antigoon corpus. Antigoon is an 800 million word corpus, largely based on EEBO, that covers the period from 1580 to 1700. It shares most of its preprocessing (i.e. tokenisation, language identification) with the EMMA corpus (Petré, Anthonissen, Budts, Manjavacas Arévalo, Silva, Standing & Strik 2019) but it is further enriched with a missing character completion algorithm and a spelling normalisation stage. These augmentation steps are discussed in detail in Budts (2020b: 80-89). The choices made in terms of model design and hyperparameter selection have been summarised in Budts (2020a: supplementary online material) and are discussed at length in Budts (2020b: 135-138). The pre-trained word embeddings used in the encoder part of the network result from earlier research on the same dataset. The technical details of their creation are provided in Budts (2020b: 99-103).

To reconstruct what each filter had grown sensitive to, we first extracted a random sample of 100,000 attestations from the model's training data. We then fed each of these attestations into the encoder part of our trained model but cut off the prediction process right before the max pooling step. As such, we turned each input sequence into an activation map that indicates for each filter in the model to what extent that filter matches all possible N-grams of appropriate size in the given input sequence. For each filter, we then kept track of the 10 N-grams that yielded the highest convolution score across the entire sample of 100,000 attestations. A manual inspection led to the categories described in Section 4.3. All code needed to train new models, to apply pretrained ones to unseen data and to reconstruct what they have learned is publicly available at https://github.com/srbdts/hyperdeep.