Discursive Quads: New kinds of lexical co-occurrence data with linguistic concept modelling

Seth Mehl

Abstract

This paper introduces linguistic concept modelling, a new computational approach to humanities-driven analysis of meaning in large text collections, and presents illustrative examples of the approach applied to over one billion words of printed Early Modern English contained in Early English Books Online (Text Creation Partnership edition). Linguistic concept modelling methods and innovations are described in detail, and justified as a unique, new, powerful means for studying meaning in texts in relation to individual lexical items, sets of lexical items, and entire text collections. Linguistic concept modelling is compared to established approaches of distributional semantics and topic modelling, to show how linguistic concept modelling overcomes the limitations of those approaches to support humanities research. Examples demonstrate how our approach can be used to explore and analyse semantic, pragmatic, and discursive meaning in texts, highlighting the novel observations that our approach affords, and its benefits for humanities scholars.

Introduction

The *Linguistic DNA* (LDNA) project arose from ongoing discussions about key social and cultural words and concepts in Early Modern English, and the possibility of moving beyond traditional approaches to studying such key words and concepts, which rely on the historian or linguist to determine which words and concepts are and are not worthy of attention. The project aimed to circumvent that traditional approach by computationally analysing meaning relatively inductively, based on lexical co-occurrence, in Early English Books Online (Text

Creation Partnership edition, EEBO-TCP). The research team defined the aim early in the project of computationally identifying complex constellations of co-occurring lemmas. The team wanted to identify not just co-occurring pairs, but also co-occurring trios and quads, i.e. sets of three or four non-adjacent lemmas that actually occur together in specific spans of text. The team wanted to identify such co-occurrences at the level of discourse, rather than at the level of the phrase, clause, or sentence. It was deemed necessary to be able to sort and rank such trios and quads, and to link directly from the lists of trio and quad data to view the co-occurring lemmas in the specific span of text where they co-occur, in order to manually analyse the text examples, and to investigate semantic and pragmatic meaning in co-text. Finally, it was imperative that all such constellations be identified for large numbers of lemmas in the mass of textual data, rather than for a few pre-selected lemmas. The project succeeded in pioneering a method of identifying and ranking non-adjacent co-occurring lemma trios and quads within large discursive spans of running text, with statistical measures for each trio or quad, resulting in ranked lists of billions of significant co-occurring trios and quads in EEBO-TCP, linking to the examples themselves in co-text. We call this process linguistic concept modelling.

For example, the trio *reason-nature-law* occurs extremely frequently in the data and passes statistical thresholds (discussed further below), such that linguistic concept modelling forwards it as potentially worthy of scrutiny. Examples 1 and 2, from EEBO-TCP, contain the trio:

1. But then we are to learn our Duty, not from any express **Law** of God and **Nature**, but from the **Reason** and **Nature** of things. [Thomas Wagstaffe. 1691. Untitled. TCP ID B06596.]

2. Man should make much of Life, as **Nature's** table, wherein she writes the Cipher of her glory, Forsake not **Nature**, nor misunderstand her: Her mysteries are read without Faith's eyesight. She speaks in our Flesh, and from our Senses, delivers down her wisdoms to our **Reason**. If any man would break her **laws** to kill, **Nature** doth, for defence, allow offences. [Anonymous. 1680. Untitled. TCP ID B06293.]

The complexities of the discourses around *reason*, *nature*, and *law* are apparent even in just two examples; linguistic concept modelling has identified over eight thousand examples of *reason-nature-law* in EEBO-TCP, each of which can be closely inspected and analysed. Examples 1 and 2 both refer to *law* of *nature*, presented as *law of God and nature* in example 1, and as *her laws* in example 2. The polysemy of *nature* is apparent in example 1, which plays with the two meanings 'the physical world' and 'innate character' (*nature*, n. 11a, III, OED Online). In example 1, *nature* as 'innate character' is juxtaposed with *reason*, in *the reason and nature of things*; while in example 2, *our reason* is juxtaposed against *her* [*nature's*] *wisdoms*. Using linguistic concept modelling, it is possible to identify the complex discourses and concepts represented by this trio and then to closely analyse a broad range of instantiations. The ability to identify this trio as representative of a noteworthy discourse is an important offering of linguistic concept modelling; the ability to then immediately extract examples from EEBO-TCP, or from specific sub-sets of EEBO-TCP, renders linguistic concept modelling extremely valuable for careful linguistic, philological, or historical inquiry.

Linguistic concept modelling has yielded new lexical co-occurrence data, including data on billions of co-occurring trio and quad lemmas in EEBO-TCP, their frequencies and associated statistical information, across large proximity windows. No previous research project has generated this sort of data for text collections, nor posited that such observations are valuable for analysing discursive and conceptual meaning. Linguistic concept modelling is groundbreaking because it represents only the first steps in what we see as the fruitful field of interrogating actually co-occurring non-adjacent lexical constellations of three or more lemmas in large spans of text. Our data represents a novel way in to exploring texts, which indicates usage and meaning in a manner not previously possible.

In this paper, I first describe linguistic concept modelling and present its methodological innovations in contrast to established techniques in distributional semantics and topic modelling. Linguistic concept modelling is by design essentially different from those approaches; indeed, linguistic concept modelling was engineered to overcome limitations of distributional semantics and topic modelling, which would have obstructed the aims of the LDNA project. I then provide examples of the sorts of innovative data that are produced by linguistic concept modelling. I discuss examples broadly here; more deeply worked case studies can be found in Fitzmaurice (*this volume*), who investigates pragmatic routines for interpreting quads including *life-death-soul-spirit*; and Fitzmaurice and Mehl (*forthcoming*), who investigate how semantic underspecification and pragmatic enrichment facilitate change in discourses around the quads *life-death-spirit-body*, *world-heaven-earth-power*, and *sin-son-father-heart*. Examples in this paper evince the innovation in this work, and support the argument that linguistic concept modelling represents a step change in text analysis and in corpus linguistics for semantics, pragmatics, and discourse anlaysis, which will be of particular interest to linguists and historians.

Linguistic concept modelling

One of the key innovations of linguistic concept modelling is its identification of co-occurrences beyond the level of the pair. Traditional approaches to lexical co-occurrence analyse co-occurring pairs, in various proximity windows. Adjacent pairs are generally termed *collocations* and include typical pairings such as *strong tea* or more grammatical combinations such as *because of* (cf. Manning and Schuetze 1999: 151). Corpus linguists for decades have demonstrated the social and cultural import of some pairings, including the reinforcement of associations through such pairs (cf. Schneider, *this volume*; Baker 2006; Sinclair 2004, 1991; Stubbs 1996).

Popular corpus linguistic interfaces such as Mark Davies's BYU interface allow users to identify co-occurring pairs for any term in a corpus: that is, a search for lemma *a* yields a list of all co-occurrences of the type *ab* (cf. Davies 2008). The co-occurrences need not be adjacent: in the BYU interface, for example, co-occurrences can be identified up to nine tokens to the left and/or right of term *a*. Likewise, the popular corpus tool AntConc can identify co-occurrences up to twenty tokens to the left and/or right of term *a* (cf. Anthony 2018). Neither of these tools generates data on all co-occurrences for all terms simultaneously, but instead provides co-occurrence pair lists for a single search term at a time.[1]

According to Sahlgren (2006; cf. Heylen et al. 2015) and Turney and Pantel (2010), co-occurring pairs within a narrow proximity window (for example, up to ten tokens to the left and right of word *a*) can best indicate paradigmatic relations, or attributional similarity, like that between *hospital* and *clinic*. We might also see these relations as synonymy or co-hyponymy (cf. Murphy 2010, Geeraerts 2010: 82-87). On the other hand, identifying lexical co-occurrences across large proximity windows, such as paragraphs, is most useful for identifying syntagmatic and associative relations, or relational similarity, such as that

[1] There is also research into second-order collocates or co-occurrences, such that for a given node *a*, all co-occurrences *b* are identified and then used as new nodes, for which all co-occurrences *c* are in turn identified (cf. Baker 2016, summarising this work in corpus linguistics; and Schuetze 1998, for seminal applications in Natural Language Processing). Second-order pairs are not trios as defined in linguistic concept modelling.

between *doctor* and *hospital* or *car* and *drive* (Turney and Pantel 2010).[2] We might see these relations as representing a semantic field or semantic domain (cf. Kay and Allan 2015: 188).

LDNA aimed to identify sets of three or four lemmas representing syntagmatic or associative relations, relational similarity, or semantic fields, which encompass semantic, pragmatic, and topical relationships, and we therefore planned to identify co-occurrences across large proximity windows (cf. Fitzmaurice *et al.* 2017). The project experimented with proximity windows of 5, 10, 30, 50, and 100 tokens to each side of a node word (cf. Mehl 2019). As expected, the smaller windows of 5 or 10 tokens indicate lexical relationships mainly at the level of the phrase or clause; the larger window of 100 tokens demands too much computationally – in terms of processing time and storage. The project settled on a default of 50, a reasonably large discursive span that is not prohibitively demanding computationally; we continue to recognise the value of a 30-token window as well, which yields consistently different results from a 50-token window.

In the past, research on co-occurrences beyond the pair has been limited to adjacent collocations, also known as multi-word expressions, clusters, or *n*-grams (cf. Anthony 2018; Wahl and Gries 2018; Manning and Schuetze 1999: 154). These are often grammatical in nature, such as *because of the* or *in spite of*; they may also be compositional, with transparent meaning conveyed by content words, such as *sun ripened tomatoes* or *overseas development aid* (cf. Manning and Schuetze 1999: 161); or, indeed, they may be proper nouns, such as *Martin Luther King* (cf. Wahl and Gries 2018). AntConc calculates and ranks lists of all *n*-grams in a corpus, as a representation of usage in the corpus (Anthony 2018). Unlike the established techniques for studying such adjacent trigrams, linguistic concept modelling is the

[2] There is still a remarkable shortage of precise findings on the proximity windows, direction of co-occurrence, and other co-occurrence specifics that best indicate particular semantic relations for different word frequency ranges, parts of speech, semantic attributes, and languages; see Heylen *et al*. 2008 for an example of a valuable example of such findings.

first tool to identify and rank lists of all non-adjacent trios or quads in text collections, which can spread across proximity windows of 50 tokens to the left and right of lemma *a*.

Each trio or quad is thus a data point with associated characteristics, including frequency of occurrence (how many times the trio or quad actually co-occurs in the text collection; as well as how many individual texts the trio or quad appears in), MI score, and chi-square score. MI scores are interpreted as measuring the strength of the co-occurrence trio or quad; chi-square scores are used to remove examples for which data size or effect size are small enough that we cannot be confident generalising our observations; combining MI and chi-square is standard in corpus linguistics (cf. Mehl 2019). The details of the equations for MI and chi-square are explained in Mehl (2019), including linguistic concept modelling's unique application of a part-of-speech baseline. The LDNA team has been focusing on the highest frequency statistically significant trios and quads, as indications of dominant discourses, and the examples in this paper reflect that focus.

There is no precedent in the literature for calculating MI and chi-square scores for trios and quads in language data, so we base our calculations on Fano's (1960) seminal definition of MI in information theory. Both MI and chi-square calculations require a measure of observed (O_1) and expected (E_1) probabilities. For trios, (O_1) is the probability of lemma *c* occurring within the window around lemma *a*, given that lemma *b* also occurs in the window around lemma *a*; and (E_1) is the probability of lemma *c* occurring in all of EEBO-TCP. Likewise, for quads, (O_1) is the probability of lemma *d* occurring within the window around lemma *a*, given that lemmas *b* and *c* also occur in the window around lemma *a*; and (E_1) is the probability of lemma *d* occurring in all of EEBO-TCP. Chi-square calculations also require a measure of (O_2) and (E_2). For trios, (O_2) is the number of times that lemma *c* does not occur in the window around lemma *a*, given that lemma *b* is present; and (E_2) is the number of times that lemma *c* does not occur in EEBO-TCP. Likewise, for quads, (O_2) is the number of times that lemma *d* does not occur in the window around lemma *a*, given that lemma *b* and *c* are present; and (E_2) is the number of times that lemma *d* does not occur in EEBO-TCP.

Traditional pair data is identical across the two permutations *ab* and *ba*, and scholars are accustomed to navigating that data easily. Trio data includes six permutations of each unique trio of words *abc* (i.e. *abc, acb, bac, bca, cab, cba*), while quad data presents 24 permutations. Each permutation has a different chi-square score and MI score. These differences arise for two reasons. First, a trio may occur as follows:

- *lemma a* [49 tokens] *lemma b* [49 tokens] *lemma c*

In the above example, the trio *bac* and *bca* both occur once. But, the trios *cba* and *cab* do not occur, nor do the trios *abc* or *acb*. This is because lemma *a* and lemma *c* do not occur within a 50-token span of each other. The same issue extends to quads. Thus, raw counts of each trio or quad permutation will not necessarily be identical. In addition, chi-square and MI scores differ because they depend on a calculation of expected probabilities for the third lemma in the trio (or the fourth lemma in the quad). Thus, for example, *reason-nature-law* and *nature-law-reason*, even if they have identical raw counts, will have different chi-square and MI scores, because the expected probability for lemma *c* will differ. Again, the same issue extends to quads. These differences have not been an object of linguistic study because co-occurrences beyond the pair have not been an object of linguistic study, and they thus constitute new categories of data with new and unexplored – and indeed never before considered – characteristics.

We use linguistic concept modelling to analyse lemmas, rather than forms or tokens. Analysing forms can be useful in discerning how inflectional forms of a single root can be employed in different discourses (McEnery *et al.* 2015); while analysing tokens for discerning multiple meanings of a given root form (De Pascale 2019). We acknowledge that future implementations of linguistic concept modelling could experiment with analysing forms or tokens, though such experiments would increase computational demands considerably – perhaps prohibitively. Our version of EEBO-TCP is part-of-speech tagged and lemmatised using MorphAdorner (Burns 2013), with additional manual correction by the MorphAdorner team (p.c. Martin Mueller 2018), such that it represents the state of the art in the digital preparation of Early Modern English texts.

Two key difficulties in creating data for large numbers of non-adjacent trios or quads across large co-occurrence windows are the huge demands on computational processing, and the immense size of the output data. For processing, the team of developers at the University of Sheffield Digital Humanities Institute (DHI) has employed Apache's big data software utilities Hadoop, to process data across up to 40 virtual machines, and Hive, to search and retrieve information from the data outputs. Thresholds are applied to the data at various stages, in order to ease computational load and storage demands, and in accordance with research questions. In various analyses of EEBO-TCP, we have analysed trios or quads around node lemmas occurring from 2 to 5,000 times each in EEBO-TCP; a higher threshold means fewer lemmas for analysis, and therefore decreased computational load and storage demands. In various iterations, we have experimented with limiting our analysis to the 1,000 most frequent lemmas; restricting lemmas for analysis to specific parts of speech; limited stored outputs to trios or quads that occur a minimum of 2 to 50 times in EEBO-TCP; or identified only trios for pairs that have passed higher or lower chi-square score thresholds. Specific parameters for examples in this paper are described alongside data presented below.

Besides expanding from pairs to trios and quads, and analysing wide proximity windows, a key characteristic of linguistic concept modelling is the transparency of trio and quad data: each trio or quad is a data point with associated statistical measures, and is linked to the texts in which it occurs, i.e. the precise spans of 50 tokens to the left and right of node *a*, containing all three or four lexical items, so that the co-text can be manually analysed by researchers. The transparency of co-occurrence data is possible because linguistic concept modelling analyses actual co-occurring trios or quads in actual co-text; this is in contrast to distributional semantics and topic modelling, which I discuss in the next section.

Linguistic concept modelling: A step change from distributional semantics and topic modelling

The existing state of the art in computational semantics and in modelling textual meaning revolves around two methods: distributional semantics and topic modelling (cf. Schneider, *this volume*). Both generally rely on supervised or unsupervised machine learning. In this section, I summarise the basic methods of each; the summary is necessarily concise and based on principles of the methods rather than details of the many particular implementations. I contrast those methods with linguistic concept modelling, and explain why established approaches were not entirely satisfactory for the specific purposes of LDNA (but see Schneider, *this volume*, for valuable applications of both techniques in analysing historical meaning in text collections).

Distributional methods were pioneered in the 1970s (Salton *et al*. 1975), but have only been widely used since the statistical turn in linguistics, from the 1990s (cf. Geeraerts 2010: 157; for an introduction to distributional methods, see Turney and Pantel 2010). It might thus be said that these methods have been popular for nearly a generation. The principle behind distributional methods is that words occurring in similar co-texts will have similar meanings; that co-texts can be quantified by counting and listing for each word *a* in a corpus all individual second words (i.e. pairs) that co-occur with *a*; that lists of pair co-occurrences for each word *a* can be compared statistically; and, in turn, that similar lists (modelled as vectors of co-occurrences for word *a*, word *b*, etc.) represent words with similar meanings in the corpus.[3] Among the revolutionary benefits of these methods (cf. Schneider, *this volume*) is the move from identifying two words that co-occur together, to identifying words that co-occur in similar contexts, even if they never co-occur together. For example, *heatwave* and *blizzard* both co-occur in close proximity with *weather* and with *extreme*, respectively. However, in texts, *blizzard* tends not to co-occur with *heatwave* because in practical discussions, and in real-world contexts, people may not often discuss blizzards and heatwaves at the same time. Distributional methods build lists of co-occurring pairs for each word *a*, such as *blizzard* or *heatwave*, and use vector analyses to compare those lists and group the words with similar lists.[4] Generally, these lists contain tens of thousands of pair items for each word *a*, and often compare tens of thousands of words to each other, grouping them into categories based on similarity. This co-occurrence data has tended to be unmanageably big, so it is generally reduced algorithmically in opaque ways. For example, some lexical co-occurrences are particularly distinctive – they co-occur very strongly with only a small number of words – and these are retained, while less distinctive co-occurrences are eliminated from the data; alternatively, data may be reduced by randomly eliminating a set percentage of co-occurrences (cf. Turney and Pantel 2010). Finally, co-occurrence data may be lost incrementally once each word is placed into semantic groups. One of the limitations of existing distributional semantics techniques, therefore, is that, even while these methods build from measures of relationships between words in context, they

[3] Pado and Lapata (2007) introduced the possibility of incorporating syntactic dependency information into co-occurrence measures. This is now the basis for tools including Sketch Engine (cf. Kilgarriff *et al.* 2014). Because the *LDNA* project was, from the beginning, interested in discursive relations beyond the level of the phrase, clause, or sentence, such syntactic dependency information was not relevant. In any case, automatic syntactic parsing of Early Modern English is not currently effective.

[4] Researchers sometimes evaluate these groups of words' viability as 'concepts' (cf. McGregor *et al*. 2015, Heylen *et al*. 2008). Such studies define 'concepts' as meaning categories in ontologies or thesauri such as WordNet.

render the initial observations of the relationships between words, and their contexts of use, obscure – words in output lists may never occur together, and when they do, the relations in context are irretrievable from the process outputs. Distributional semantic methods are effective for many purposes, including computational tasks such as identifying synonyms or retrieving information effectively for keyword searches in online search engines, but they are not entirely satisfactory for LDNA's goal of manually analysing the nature of semantic, pragmatic, or discursive relations across individual co-occurrences in contexts of use.[5]

Topic modelling refers to a range of text categorisation procedures which have been used since at least the early 2000s (cf. Blei *et al*. 2003). Rather than identifying 'topics', which have not generally been defined or theorised prior to operationalising topic models (Brookes and McEnery 2018), this collection of methods categorises texts based on the lexical items contained in them, and thus *topic modelling* may be considered something of a 'misnomer' (Murakami *et al.* 2017: 244). The principles behind topic modelling are comparable to those behind distributional semantic methods, but in topic modelling, the primary object of study is texts and groups of texts rather than lexical items (cf. Steyvers and Griffiths 2007).[6] In topic modelling, texts with similar lexical content are seen as belonging to a shared category, and categories can be represented by their distinctive lexical content. Texts can thus be characterized by counting and listing all lexical items contained in each text; measuring similarities between each text's lexical items; grouping texts into categories based on similarity in their lexical content; and representing those categories with a word list

[5] Recent research is using distributional semantic methods to model meanings of phrases and clauses (cf. Weir *et al*. 2016); and to investigate semantics within grammatical constructions (cf. Perek 2016). In addition, other recent work combines distributional methods with topic-modelling (cf. Rönnqvist 2015). The issues in the opacity of the methods, which cause problems for LDNA, are still present in these developments.

[6] In fact, 'text' in this sense can refer to modelling entire documents, or smaller sections of documents. Schneider (*this volume*), for example, has performed topic modelling on 1000-token segments of EEBO-TCP. To our knowledge, topic modelling is not generally applied to spans of around 100 tokens, like the proximity windows used in linguistic concept modelling.

composed of lexical items that are distinctive to each category. A word list may contain any number of words, but for practical purposes, lists are generally presented with ten to twenty items, such as, for example, with the popular topic modelling tool Mallet (McCallum 2002). Topic modelling methods generally aim to group texts that are more similar to each other than to the rest of the text collection, and a family resemblance model will often hold, such that text *a* is very similar to text *b*, *b* to *c*, and *c* to *d*, but *a* and *d* may be only minimally similar; it suffices that *a* and *d* are more similar to each other than to the rest of the collection. In turn, out of a typical topic model word list's twenty representative words, some texts in the group may contain all twenty and some may contain only a few; again, it suffices that the texts are more similar to each other than to the rest of the collection. Topic modelling approaches do not fully satisfy LDNA's research aims, for two reasons. First, a topic model word list – as a whole – does not represent discursive co-occurrence: that is, the full list of words do not necessarily occur together in actual discursive spans of text. Identifying such discursive co-occurrences is a major research target for LDNA. In addition, as with distributional semantic methods, the relationships between words and their contexts tend to be irretrievable from topic modelling outputs: in the final outputs, the initial observations on which words occur in which texts, and in which quantities, are rendered obscure, as are the precise co-textual relationships between the words, and their contexts of use.[7] Thus, like distributional methods, topic modelling techniques are not fully satisfactory for meeting LDNA's aim of manually exploring semantic, pragmatic, and discursive meaning in co-text.

Trio and Quad Data: Describing lemmas

[7] Moreover, topic modelling processes are generally subject to some degree of randomness, and outputs may therefore be different each time a procedure is run, even given identical inputs (Brookes and McEnery 2018).

In order to understand the new possibilities raised by discursive trio data, it is important to compare this new type of data to more traditional pair outputs. For example, in EEBO-TCP, *diversity* occurs most frequently in the following ten statistically significant (*p*<0.05) *noun-noun* pairs. The second lemma appears within 50 tokens to the right or left of *diversity*.

- *diversity - faith*
- *diversity - life*
- *diversity - way*
- *diversity - person*
- *diversity - kind*
- *diversity - variety*
- *diversity – law*
- *diversity - world*
- *diversity - manner*
- *diversity - gift*

In the pairs listed here, relations span discursive and associative relations, and multiple semantic and pragmatic fields, topics, and real-world contexts, rather than just traditional semantic relations or attributional similarity (see above). In fact, it is only *diversity-difference* and *diversity-variety* that might indicate traditional relations such as synonymy or co-hyponymy.

Diversity occurs in three statistically significant (*p*<0.05) *noun-noun-noun* trios, with each noun occurring at least 5,000 times in EEBO-TCP:[8]

[8] Here and in subsequent examples, I have ordered lemmas within each trio or quad in order to neatly show similarities. For example, the permutation *diversity-spirit-gift* rather than

- *diversity – opinion - religion*
- *diversity – spirit - gift*
- *diversity – spirit – operation*

These trios constitute new observations not readily available from pair data: *faith* was a strong lemma in the top ten pair list, but *religion* was not present; *gift* formed a strong pair, but *spirit*, *operation*, and *opinion* did not. The crucial point here is not just that *diversity* co-occurs with *gift* at a high frequency, as was evident from the list of pairs, but that *diversity* co-occurs with *gift* and *spirit* together. If we accept that the co-occurrence pair of *diversity* with *gift* is meaningful, then we can see that the trio of *diversity*, *gift*, and *spirit* is a richer representation of discursive meaning. Here, again, we see not traditional semantic relations – there are no synonyms or antonyms, for example, in each trio – but indications of semantic fields, pragmatic processes, discourses, topics, and real world contexts. On one hand, in this example, *spirit* may be seen as specifying – to a considerable degree – the discourse that might be represented by *diversity* and *gift*; on the other hand, the polysemy and vagueness (or underspecification) in *spirit* may be seen to further destabilise the discourse that might have been inferred from *diversity* and *spirit* alone. This enrichment, consisting simultaneously of both potential specification and potential destabilisation via underspecification, is explored further by Fitzmaurice (*this volume*).

Most importantly, this new trio data is transparent, which allows us to meaningfully address the nature of the relationships within and between trios by analysing co-texts of each trio. Trios are actual co-occurrences in real spans of text, and the original texts can be

diversity-gift-spirit tidily foregrounds the similarity with *diversity-spirit-operation* insofar as both begin with the pair *diversity-spirit*.

accessed and studied – this possibility represents a dramatic difference from topic modelling and distributional semantics, as discussed above.

Example 3 contains the trio *diversity-opinion-religion*.

3. To know the causes of false **opinions** is the only means to break the strength and root out the force of false **opinion**. Profit, honour, loss, and dishonour are four causes of disjoined **opinions**. Shame breeds variation in **opinions**; yet not tumultuously, or without order. Great **opinions** alter not at one instant, but leave their strength by degrees, by little and little, except they be violent. Dissimilitude being a **diversity** of **opinions** in **religion**, is cause of civil war. The **diversity** of **opinions** in subjects is most dangerous to estates and sovereigns. [Nicholas Ling. 1598. *Politeuphuia: Wits Commonwealth*. TCP ID A05562.]

In example 3, within a larger discussion of opinions – and particularly *false opinions* – we see not only the phrase *diversity of opinions in religion*, but also *diversity of opinions in subjects*. The automatic extraction of examples of the trio *diversity-opinion-religion* facilitates specific investigation of the concepts around *diversity-opinion-religion*. The discourse around this trio, in this example, also encompasses terms representing falsehood and vice. If a trio can be seen to have a kind of semantic preference or discourse prosody, like individual words (cf. Stubbs 2001), then *diversity-opinion-religion* may be negative, associated with conflict, disorder, and danger (cf. Fitzmaurice and Mehl *forthcoming* on considering trios and quads as possessing semantic-like attributes).

The discourses represented by example 3 are very different from those represented by example 4, which shows the trio *diversity–gift–spirit.*

4. There are **diversity** of **gifts**, but the same **spirit**, so all receive the name of Oil. [Thomas Johnston. 1630. *Christ's Watch-Word*. TCP ID A04596.]

This trio is deemed a strong one using linguistic concept modelling largely because of the relatively fixed phrase *diversity of gifts, but the same spirit*, which is extremely frequent in EEBO-TCP. It is worth noting that with this trio, linguistic concept modelling has automatically extracted a discourse of *diversity* that does not present falsehood and vice, conflict, disorder, and danger, and therefore contrasts *diversity-opinion-religion* above.

Finally, example 5 shows the trio *diversity-spirit-operation*.

5. The conditions of men are exceeding various, and so are capable of several sorts of temptations. The temper of men's **spirits** we know is diverse, and so is capable of **diversity** of suggestions. Men of melancholy and jealous **spirits**, he plies with reasonings and suggestions that will most take with their **spirits**. And again the **operations** of graces, as of sin, are various in those several tempers. [Thomas Goodwin. 1636. *A child of light walking in darkness*. TCP ID A01898.]

Example 5 contrasts with example 4 insofar as it presents a different sense of *spirit*: example 5 presents *men's spirits*, whereas *spirit* in example 4 is divine. Again, we have automatically extracted a different discourse from examples 3 and 4, in a way that would have been impossible to extract automatically with a simple keyword search for *diversity*, or even a more advanced search for the pair *diversity-spirit*. Linguistic concept modelling automatically identifies and extracts over 500 examples of *diversity-opinion-religion*; over 800 examples of *diversity-spirit-gift*; and over 500 examples of *diversity-spirit-operation*, all of which can be manually examined by the researcher.

Building from trios to quads, the following are the five highest-frequency statistically significant (p<0.05) quads around *diversity*, composed of nouns, adjectives, or verbs occurring at least 5,000 times each in EEBO-TCP.

- *diversity-spirit-gift-word*
- *diversity-spirit-gift-operation*
- *diversity-spirit-gift-lord*
- *diversity-spirit-gift-work*
- *diversity-spirit-gift-administration*

We observe here a high-frequency constellation around the trio *diversity-spirit-gift*, which extends to a list of fourth lexical items. Each fourth lexical item can be seen as further enriching or narrowing the discourse indicated by the trio, but also as further destabilising the trio. That is, the fourth item may narrow our interpretation of the discourse around the trio, into more specific contexts of use; but may also destabilise our interpretation of the discourse trio, by introducing additional polysemy and vagueness in the fourth lexical item (cf. Fitzmaurice, *this volume*).

Diversity-spirit-gift was a high-frequency trio, but of the fourth items in this list, only *world* and *way* were present in the top 10 pairs; and only *operation* appeared in the significant trios. None of the fourth items are in a traditional semantic relationship with *diversity*, but instead indicate syntagmatic relationships and practical contexts of use. By closely reading examples, we can see that *diversity-spirit-gift* occurs most commonly in the relatively fixed phrase *diversity of gifts, but the same spirit*; and we can see each fourth item here as indicating common contexts, and pragmatic enrichment or specificity to the usages around that fixed phrase.

Given that the highest-frequency statistically significant (*p*<0.05) quads around *diversity* all include the trio *diversity-spirit-gift*, we might then identify all quads based around another significant trio, *diversity-opinion-religion*. Below are the highest-frequency statistically significant (*p*<0.05) quads built from the trio *diversity-opinion-religion*, composed of nouns, adjectives, or verbs occurring at least 5,000 times each in EEBO-TCP

- *diversity – opinion – religion – true*
- *diversity – opinion – religion – time*
- *diversity – opinion – religion – touch*
- *diversity – opinion – religion – world*
- *diversity – opinion – religion - way*

Among the fourth lexical items, *world* and *way* were among the highest frequency pairs; none of the fourth items was among the significant trios; and none of the fourth lexical items is in a traditional semantic relationship with any of the first three. Examples 6 and 7 illustrate the quad *diversity-opinion-religion-true*.

6. The kyngnes highnes by the aduise of his moste entierly beloued Vncle the Duke of Somersette Gouernor of his moste royall persone and Protector of all his Realm es Dominions and Subiectes and others of his Counsaill Consideryng nothyng so muche to tende to the disquietyng of his realme as **diuersitie** of **opinions** and varietie of Rites and Ceremonies concerning **Religion** and worshippyng of almightie God and therefore studiyng all the waies meanes which can be to directe this Churche and the Cure committed to his highnes in one and m

oste **true** doctrine Rite and Vsage… [England and Wales Sovereign Edward VI. 1548. *A proclamation against those that doeth innouate*. TCP ID A69318.]

7. Whereas Wee out of Our care to conserue and maintaine the Church committed to Our Charge in the vnity of **true Religion** and the bond of peace and not to suffer v nnecessary disputes which may trouble the quiet both of Church and State haue lat ely caused the Articles of **Religion** to bee reprinted as a rule for auoyding of **diuersities** of **opinion** and for the establishing of consent in **true Religion** [Charles I. 1628. *By the King, a proclamation for the suppressing of a booke intituled Appello Cæsarem*. TCP ID A22494.]

The fact of different perspectives, i.e. *diversity* of *opinion*, is presented as a problem by these writers because, they argue, there is one *religion* with its associated *doctrines*, *rites*, and *usage*, which is *true*. In Example 3, above, we interpreted the trio *diversity-opinion-religion* as indicating a dangerous problem, and we noted through close reading the importance of the additional co-occurring word *false*. In this quad, the fourth item *true* represents an additional key concept that defines the discourse as dangerous and problematic – for early modern writers, diversity threatens truth.

Example 8 illustrates the quad *diversity-opinion-religion-time*.

8. I could name many more **opinions** of
men who were all great and glorious lights in the Church and most illustrious instr uments for the advancement of Christian **Religion** and yet they have in some point s differed one from the other as Wickliffe Luther Beza Calvin Bucer Melancton O eclampadius yet for all other great **diversities**
they have alwayes agreed in the main Fundamentall points of Christian Doctrine s

> o that the outsides of Ceremonies of **Religion**
>
> did not shake the peace of the Church But in these **times** the Church and Church-Government is not only shaken but shattered… [John Taylor. 1642. *A cluster of coxcombes.* TCP ID A64161.]

Time in Example 8, as *in these times*, conveys that the dangerous problem of difference of opinion in religion is a pressing characteristic of the present moment. The fourth item here underlines the urgency of the threat, in the discourse indicated by the trio *diversity-opinion-religion*.

Within the quads presented here, each fourth item specifies a feature of the discourse for emphasis: the fact of one true religion in Examples 6 and 7; and the urgency of the immediate threat in Example 8. It is useful to compare these quads to the trios on which they are built. For some research purposes, trios will likely represent the perfect balance of generality and specificity in a pinpointed discourse; for other research purposes, the fourth item of the quad may usefully specify the discourse further. Linguistic concept modelling has identified 606 examples of *diversity-opinion-religion-true* in EEBO-TCP, and 360 examples of *diversity-opinion-religion-time*, each of which can be closely read and analysed. Moreover, the full list of quads around a trio, such as all quads built on the trio *diversity-opinion-religion*, can be seen as illustrating the specific issues that early modern authors tend to emphasise around the discourses of that trio.

A researcher can use this trio and quad data around a given word, like *diversity*, in two ways. First, trio data can be used as an exploratory tool, to represent and retrieve discourses around *diversity* from the very large text archive of EEBO-TCP. In this way, discourses that might not otherwise have been accessed can be explored in systematic ways. Ultimately, this can be used for knowledge discovery – either as the discovery of otherwise unrecognised discourses, via otherwise unconsidered trios; or as the discovery of otherwise unrecognised texts or text fragments containing a given discourse. Second, trio and quad data can be used to retrieve discursive content that the researcher has already defined. For example, a historian might begin with an interest not just in the keyword *diversity,* which is extremely frequent and extremely complex, but in the specific discourses around *diversity of opinion in religion*. Searching for the trio *diversity-opinion-religion* can retrieve instances of those discourses. As with other corpus linguistic tools, a user can revise and refine subsequent searches to retrieve through iterative or cyclical processes a more complete representation of the discourses that are sought.

Trio data: Describing texts

The examples so far began with a single lemma and presented co-occurring trios and quads to explore discursive relationships. Those examples demonstrate how linguistic concept modelling circumvents some limitations of the standard distributional semantic methods described above. Linguistic concept modelling can also be used to create meaningful representations of texts and text collections, by circumventing some limitations of topic modelling. Given a text collection (for example, EEBO-TCP, or a twenty-year span within EEBO-TCP, or an authorial oeuvre, or a specific text type or genre), linguistic concept modelling can identify trios or quads in the collection, ranked according to various measures.

For example, LDNA has curated a collection of early modern sermons and related texts from EEBO-TCP data. The following are the five statistically significant (p<0.05) *noun-noun-noun* trios that occur in the largest number of sermons and related texts, given that each lemma occurs at least 5,000 times in EEBO-TCP. Each trio occurs in over 600 texts, and each example can be investigated closely with its co-text.

- *soul-body-life*
- *heaven-earth-world*
- *earth-heaven-power*
- *soul-body-spirit*
- *soul-body-sin*

This might appear unsurprising for content in sermons, but it is important to note what could have been prominent, but is not. To do this, we need only view some of the trios that occur in far fewer documents drawn from sermons and related texts. Far less frequent statistically significant (p<0.05) *noun-noun-noun* trios, containing lemmas that occur at least 5,000 times each in EEBO-TCP, are listed below; each occurs in only around 100 texts:

- *heart-joy-hope*
- *place-world-darkness*
- *heaven-favour-earth*
- *will-life-son*
- *sin-heaven-end*

It is clear that each of these trios represents rich conceptual and discursive information, and each might be readily associated with sermons, but they occur in only a small fraction of the documents compared to the higher frequency list above. The rankings here reveal useful information about the most frequent, and the less frequent, discourses in this text collection.

The above lists can be compared to trios drawn from the science texts in the Visualizing English Print Super Science Collection, a collection of scientific texts carefully curated from EEBO-TCP.[9] The following are the five statistically significant (p<0.05) *noun-noun-noun* trios that occur in the largest number of science texts, given that each lemma occurs at least 5,000 times in EEBO-TCP:

- *body-part-nature*
- *body-part-water*
- *body-part-place*
- *body-part-spirit*
- *body-part-reason*

The differences between the trios drawn from sermons and those drawn from science texts are stark, and they indicate the differences in the prominent discourses and topics in the two text collections. Again, these lists are dramatically different from distributional semantic lists or topic models of these text collections, because these lists represent actually co-occurring trios, within specific spans of text, allowing researchers to move from the trio itself to the text containing it, for a thorough manual analysis of linguistic meaning in the co-text.

This trio data offers another newly enriched category of information, immediately apparent in the lists above, which requires further investigation. Every trio in the science data contains the pair *body-part*. The top five trios in sermons include the pair *soul-body* in three trios and the pair *heaven-earth* in two trios. Logically, there must be a category of lemmas *a* that co-occurs significantly with a relatively large number of second lemmas (to form pairs), but only a relatively small number of third lemmas (to form trios); and vice versa. These pair/trio/quad ratios constitute an entirely new category of quantitative linguistic data, which could not have been observed without linguistic concept modelling’s statistical analysis of

[9] https://graphics.cs.wisc.edu/WP/vep/vep-early-modern-science-collection/

co-occurrences beyond the level of the pair. What is the range of these pair/trio/quad ratios? What are the semantic and pragmatic phenomena that are represented by different ratios? These questions require further investigation.

Ongoing research with linguistic concept modelling

This paper has aimed to present breadth of examples to demonstrate the innovations of linguistic concept modelling, to explain and justify the linguistic concept modelling methodology, and to illustrate the need for these new kinds of analyses. I have argued that linguistic concept modelling is a valuable tool for exploring discourse in large text collections. Studies of specific discourses, represented by sets of trios and quads, are one focus of ongoing research (cf. Fitzmaurice *this volume*; Fitzmaurice and Mehl *forthcoming*).

An important unfolding area of exploration for the LDNA team is semantic and pragmatic 'volatility' in trios and quads (cf. Fitzmaurice and Mehl *forthcoming*). Generally, a pair of co-occurring pairs such as *strong tea* is idealised as stable across space and time. However, it is of course possible for the meanings of both items to vary or change, such that a pair – or trio or quad – does not indicate a constant meaning, or an unchanging discourse. With high-frequency statistically significant trios and quads in EEBO-TCP, such as *reason-nature-law* or *diversity-opinion-religion,* the component parts are themselves polysemous and underspecified or vague in multiple ways. Whereas it might be intuitive that *strong tea* remains strong tea through the centuries, and across regional varieties, it is apparent that *reason-nature-law* can vary and change not only across centuries and regions, but from one writer to the next within the same time period, and within the oeuvre of a single author. We propose, first, that this variation and change in co-occurrences is an important object of linguistic analysis; and second, that the semantic underspecification of co-occurring terms, which facilitates such variation and change, also renders these co-occurrences indispensable in shifting, conflicting, competing, and contested discourses.

Given that trio and quad data presents pragmatically and discursively richer information than pair data, it will be possible to build distributional semantic models around trio data rather than pair data. For that research, new vectors would be created around a given lemma's co-occurrence trios or quads, rather than its co-occurrence pairs, and lemmas with similar vectors of trios would be seen to be semantically, pragmatically, or discursively similar. Of course, such processes will have the same limitations described above for distributional semantic approaches, but initial experiments, currently being planned, will test whether trio and quad data can improve existing distributional semantic methods – and will also test the feasibility of using trio data in this way, given the immense computational demands. Similarly, it will be possible to categorize sections of text in a new sort of topic modelling technique, based not on raw lexical content as in typical topic modelling systems, but based instead on the trios that texts contain. Again, such processes will face the same limitations described for topic modelling approaches above, but experiments are underway with EEBO-TCP, which may be expected to categorise texts with more semantically, pragmatically, and discursively rich similarities than existing topic modelling methods (cf. Brookes and McEnery 2018).

Conclusion

I have argued that linguistic concept modelling is innovative and groundbreaking, and that it offers new ways in to linguistic investigation, creates new categories of data, and facilitates new insights into language. The LDNA project continues to develop linguistic concept modelling for use not only by the LDNA team, but also by linguists, historians, other scholars, and non-academic partners. The LDNA team, and the DHI, have applied various instantiations of linguistic concept modelling to multiple datasets, including an analysis of the BBC News Scripts with partners at the BBC; an analysis of specific dictionary headwords in EEBO-TCP with partners at the OED; an analysis of millions of YouTube comments related to militarisation, with the Militarisation 2.0 project; and an analysis of selected social science journal publications with the Ways of Being in a Digital Age project. Details of each implementation of linguistic concept modelling and interfaces for selected datasets, as well as other information on ongoing LDNA research, can be viewed via www.linguisticdna.org

References

ANTHONY, LAURENCE, 2018. AntConc (Version 3.5.7), Tokyo, Japan: Waseda University. http://www.laurenceanthony.net/software. (27 August, 2020.)

BAKER, PAUL, 2006. *Using corpora in discourse analysis*, London: Continuum.

BAKER, PAUL, 2016. 'The shapes of collocation', *International Journal of Corpus Linguistics* 21 (2): 139-164.

BLEI, DAVID M., ANDREW Y. NG & MICHAEL I. JORDAN, 2003. 'Latent Dirichlet Allocation', *Journal of Machine Learning Research* 3: 993-1022.

BURNS, PHILIP R., 2013. 'MorphAdorner v2: A Java library for the morphological adornment of English language texts', Evanston, Illinois: Northwestern University. https://morphadorner.northwestern.edu/morphadorner/download/morphadorner.pdf. (27 August, 2020.)

DAVIES, MARK, 2008. *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. https://www.english-corpora.org/coca/. (27 August, 2020.)

DE PASCALE, STEFANO, 2019. *Token-based vector space models as semantic control in lexical lectometry*, Leuven: KU Leuven Dissertation.

FITZMAURICE, SUSAN, 2021. ‘From Constellations to Discursive Concepts: The historical pragmatic construction of meaning in Early Modern English’, *Transactions of the philological society* 119 (S1).

FITZMAURICE, SUSAN & SETH MEHL, forthcoming. ‘Volatile discursive concepts: Co-occurrence quads as indicators of semantic and pragmatic change in Early Modern English’.

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: 21-38.

GEERAERTS, DIRK, 2010. *Theories of lexical semantics*, Oxford: Oxford University Press.

HEYLEN, KRIS, YVES PEIRSMAN, DIRK GEERAERTS & DIRK SPEELMAN, 2008. ‘ModellingWord Similarity: An Evaluation of Automatic Synonymy Extraction Algorithms’, *Sixth International Conference on Language Resources and Evaluation, LREC*: 3243-3249.

HEYLEN, KRIS, THOMAS WILFAERT, DIRK SPEELMAN & DIRK GEERAERTS, 2015. ‘Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis’, *Lingua* 157: 153-172.

KILGARRIFF, ADAM, VÍT BAISA, JAN BUŠTA, MILOŠ JAKUBÍČEK, VOJTĚCH KOVÁŘ, JAN MICHELFEIT, PAVEL RYCHLÝ, & VÍT SUCHOMEL, 2014. ‘The Sketch Engine: Ten years on’, *Lexicography* 1: 7-36.

MANNING, CHRISTOPHER & HINRICH SCHUETZE, 1999. *Foundations of statistical natural language processing*, Boston: MIT Press.

MCCALLUM, ANDREW KACHITES, 2002. 'MALLET: A Machine Learning for Language Toolkit', Amherst: UMass Amherst, http://mallet.cs.umass.edu. (27 August, 2020.)

MCENERY TONY, MARK MCGLASHAN & ROBBIE LOVE, 2015. 'Press and social media reaction to ideologically inspired murder: The case of Lee Rigby', *Discourse and Communication* 9 (2), 237–259.

MCGREGOR, STEPHEN, KAT AGRES, MATTHEW PURVER & GERAINT A. WIGGINS, 2015. 'From Distributional Semantics to Conceptual Spaces: A Novel Computational Method for Concept Creation', *Journal of Artificial General Intelligence* 6 (1): 55-86.

MEHL, SETH, 2019. 'Measuring lexical co-occurrence statistics against a part-of-speech baseline', in Hanna Parviainen, Mark Kaunisto & Päivi Pahta (eds), *Corpus approaches into World Englishes and language contrasts*, Helsinki: VARIENG e-series, http://www.helsinki.fi/varieng/series/volumes/20/mehl/. (27 August, 2020.)

MURAKAMI, AKIRA, PAUL THOMPSON, SUSAN HUNSTON & DOMINIK VAJN, 2017. 'What is this corpus about?: using topic modelling to explore a specialised corpus', *Corpora* 12 (2): 243-277.

PADÓ, SEBASTIAN & MIRELLA LAPATA, 2007. 'Dependency-based Construction of Semantic Space Models', *Computational Linguistics* 33(2): 161-199.

PEREK, FORENT, 2016. 'Using distributional semantics to study syntactic productivity in diachrony: A case study', *Linguistics* 54 (1): 149-188.

RÖNNQVIST, SAMUEL, 2015. 'Exploratory Topic Modeling with Distributional Semantics', *The Fourteenth International Symposium on Intelligent Data Analysis*, https://arxiv.org/abs/1507.04798. (27 August, 2020.)

SAHLGREN, MAGNUS, 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-dimensional Vector Spaces*, Stockholm: Stockholm University dissertation.

SALTON GERARD, ANDREW WONG, YANG CHUNGSHU, 1975. ‘A vector space model for automatic indexing’, *Communications of the ACM* 18 (11): 613-620.

SCHNEIDER, GEROLD, 2021. ‘Systematically detecting patterns of social, historical and linguistic change: the framing of poverty in times of poverty’, *Transactions of the philological society* 119 (S1).

SINCLAIR, JOHN, 1991. *Corpus, concordance, collocation*, Oxford: OUP.

STEYVERS, MARK & TOM GRIFFITHS, 2007. ‘Probabilistic topic models’, in T. LANDAUER, D. MCNAMARA, S. DENNIS & W. KINTSCH (eds), *Latent Semantic Analysis: A Road to Meaning*, Mahwah, New Jersey: Laurence Erlbaum, 427-448.

SCHUETZE, HINRICH, 1998. ‘Automatic word sense discrimination’, *Computational Linguistics* 24 (1): 97-123.

STUBBS, MICHAEL, 1996. *Text and corpus analysis*, Oxford: Blackwell

TURNEY, PETER D. & PATRICK PANTEL, 2010. ‘From Frequency to Meaning: Vector Space Models of Semantics’, *Journal of Artificial Intelligence Research* 37: 141-188.

WAHL, ALEXANDER & STEFAN TH. GRIES, 2018. ‘Multi-word expressions: A novel computational approach to their bottom-up statistical extraction’, in Pascual Cantos-Gomez and Moisés Almela-Sanchez (eds), *Lexical collocation analysis*, New York City: Springer International, 85-109.

WEIR, DAVID, JULIE WEEDS, JEREMY REFFIN & THOMAS KOBER, 2016. ‘Aligning packed dependency trees: A theory of composition for distributional semantics’, *Computational Linguistics* 42 (4): 727-761.