Transactions of the Philological Society
PhilSoc
The Philological Society

Transactions of the Philological Society Volume 120:3 (2022) 399–426 doi: 10.1111/1467-968X.12250

COGNITIVE SOCIOLINGUISTIC VARIATION IN THE OLD BAILEY VOICES CORPUS: THE CASE FOR A NEW CONCEPT-LED FRAMEWORK

By Justyna A. Robinson and Julie Weeds
University of Sussex

(Submitted: 30 August, 2020; Accepted: 2 October, 2022)

Abstract

The current paper contributes to a greater understanding of concepts in the context of cognitive sociolinguistic variation. We show how concepts, which we root in thesauri-based units, co-occur with other concepts and how they vary between and within groups of speakers as represented in historical transcripts from the Old Bailey court. In order to carry out distributional sociosemantic analysis at a conceptual level we have developed a tool which combines word sense disambiguation with dependency analysis, which we refer to as *Conceptual Dependency Corpus*. A case study of gendered concepts from the Old Bailey Voices Corpus demonstrates that there are differences in the usage of concepts both within and between sociodemographic groupings of speakers. In particular, the co-occurrence distributions of the concepts MAN, WOMAN and CHILD vary according to a speaker's gender and role in the court. By introducing methodological innovations, we empirically elicit socioconceptual polysemy, that is different speakers using the same concept in different ways and propose a framework to analyse and interpret this phenomenon. The results also provide support for engaging computational semantics in researching cognitive and sociolinguistic phenomena.

1. Introduction

The most intuitive unit of a language is a word. To a large extent this is reflected in the way in which linguistics has developed its analytical tools. Most of these tools are based on different ways of seeing a word, that is, either a word on its own or as a component in a sentence. However, the word-centred view of linguistic analysis fails to fully accommodate the main purpose of language, which is to communicate ideas, not words. While many ideas can be encapsulated by individual words, for example *apple*, *library*, other ideas exist in a sphere that goes beyond individual lexical items. Ideas, concepts and attitudes can be expressed using a whole range of linguistic resources available to speakers. These may involve using particular groups of words, for example *the House of Commons*, *in spite of*, or *swimming pool*; words used in a given grammatical relationship, for example the *keep* V-*ing* construction; particular pronunciations, for example BATH vowel variation across English dialects; intonation, for example high rising intonation, just to name a few. In the context of complexity of ways through which concepts can be expressed, in the current paper we assume a thesaurus-based view of a concept in that a particular concept or idea may be referred to by a variety of different lexical items. For example, the conceptual category of WOMAN, in the Historical Thesaurus of English (hereafter, HTE), can be expressed as for example, *woman*, *wife*, *skirt* (1578), *bit of muslin* (1823), *she-woman* (1703), *riding-hood* (1718), *piece of goods* (1727).

In this paper, we show that moving away from a word-level analysis towards a more aggregate category of a concept provides new insights into the ideas that are in use by a speech community. At a basic level, we operationalise *concept* as a semantic category from a thesaurus together with its usage patterns. These usage patterns involve a range of context-governed conditions for the usage of the thesaurus semantic category, including dependency relations, co-occurrences with other concepts, stylistic and sociodemographic contexts. We also acknowledge concepts to be represented by more complex structures of language and cognition, but this broad view of concepts is beyond the remit of the current paper.

In the current paper we, focus on concept usage in a corpus of transcripts from the Old Bailey court from 1800–1820. Starting with Historical Thesaurus of English, we root concepts in thesaurus-based categories assigned to each word sense. Next, we identify models of the concepts usage by exploring distributional semantic properties of the thesaurus-based categories and sociodemographic characteristics of the speakers using given concepts, such as the context of speech, gender, and social position of a speaker. In result, we develop a usage-based view of concepts characterising in nineteenth-century courtroom. In the current paper, we report on a case study of concepts MAN, WOMAN and CHILD and how they are used in male and female speech in the nineteenth century court.

The usage-based definition of concept we advocate in the current study allows us to capture cognitive sociolinguistic variation. Thus, we explore not only whether speech communities use different words for the same concepts but also whether they use different concepts to reflect different lived-in experiences. Further, we explore whether certain concepts tend to take on different usage meanings, as reflected by their different co-occurrences, for different speech communities. We refer to this phenomenon as *socioconceptual polysemy*.

2. Scope of the research

The current research focuses on conceptual variation and explores it through examining grammatical relations and sociocontextual information. Studies focused on lexical semantics have shown that a meaningful correlation exists between word usage and key sociolinguistic categories, such as geographical region (Geeraerts et al. 1994), demographics (Robinson 2010, 2011), ideology (Fitzmaurice 2017), stance (Sandow & Robinson 2018). In this research, we discover underlying macro-patterns of thinking operating across different communities of speakers, which are manifested through entire concepts, not just individual words.

Two main hypotheses guide the current research. First, drawing on ideas from cognitive sociolinguistics *which aims to find cognitive explanations of usage patterns observable in a speech community* (Pütz et al. 2012, 2014), we hypothesise that different speakers use the same concepts with different probabilities. We ask whether a given concept, regardless of its written representation, occurs with a higher probability in a group of speakers (a subcorpus) than one would expect given its probability across the whole community (the entire corpus). For example, considering traditional gender roles in the family, one might expect the concept CHILD to have a higher probability of occurrence in female speech than in male speech. By investigating extralinguistic parameters conditioning concept use, we aim to better understand the relationship between sociocognitive factors and linguistic representation.

Second, we hypothesise that the same concept may be comprised of different meaning components for different speakers, even if the probability of the concept usage for those speakers is the same/similar. In other words, two groups of people may use the same concept with the same probability but develop different meaning components for that concept. We refer to the cluster of uses making up the concept as *socioconceptual polysemy*. These different meaning components of a concept emerge from the varied ways in which each individual experiences and perceives the world. For example, in modelling gendered language use, men and women might refer to a concept CHILD with similar probability, but the concept of CHILD may mean different things to men and women depending on how these two groups perceive and use the concept. The usage of a concept can be modelled through a variety of factors. In this paper, we model variation in the meaning components of a concept from dependency relationships that the concept forms with other concepts. By assuming that one knows the concept by the company it keeps, we extend the distributional hypothesis (Harris 1954; Firth 1957) to concepts. We hypothesise that patterns in conceptual polysemy can be captured through usage patterns, which in this paper are investigated through concept co-occurrence and dependency relations between concepts. This variation is further explored in the context of the sociodemographic characteristics of speakers.

In order to address these hypotheses and to carry out distributional semantic analysis at the conceptual level, we have developed a framework which combines word sense disambiguation with dependency analysis, which we refer to as Conceptual Dependency Corpus (CDC). We apply this tool to the analysis of Old Bailey transcripts. More specifically, we put the original transcripts of the nineteenth-century Old Bailey Voices (OBV) data into the CDC frame and we name the resulting resource the Old Bailey Voices Conceptual Dependency Corpus (OBVCDC). The process of building the OBVCDC starts with disambiguating historical language taken from the OBV with the Samuels Tagger (Piao et al. 2017), which is sensitive to historical word usage. While providing part-of-speech information, it does not provide grammatical dependency relations. Therefore, in order to relate concepts grammatically within a sentence, we use the SpaCy dependency parser (Honnibal & Johnson 2015). The outcome of this process is an enhanced representation of the original OBV corpus which, instead of words related within sentences, contains concepts which are grammatically related within sentences. We query this new OBVCDC using techniques from corpus linguistics and distributional semantics. Additionally, we prepare the corpus to be queried from the perspective of historically relevant metadata which is included with the OBV corpus. Detailed biographical data about each speaker in the court has been recorded in the proceedings, including gender, age and occupation (where available). We analyse gendered perceptual differences of the concepts MAN, WOMAN AND CHILD. More specifically, we identify three gendered groups, that is, males with a legal position in the court (lawyers), males without a legal position in the court (witnesses and defendants) and females without a legal position in the court (witnesses and defendants).[1] In the current paper, we refer to these groups as *legal males*, *non-legal males*, and *non-legal females*, respectively. We consider how each group understands the concepts of MAN, WOMAN and CHILD via their distributional semantic profiles.

The rest of this paper is organised as follows. In section 3, we discuss related work, while in section 4, we introduce the Old Bailey Voices Corpus (OBVC). In section 5, we describe the details of the method used to build a conceptual dependency corpus from the OBV dataset. In section 6, we illustrate the value of the conceptual dependency corpus by presenting an analytical case study focused on gendered perceptual differences of the concepts MAN, WOMAN and CHILD. In section 7, we conclude key findings and present our directions for further work.

3. Related work

The idea to model language use primarily through concepts has been present in science for at least the past 300 years. In the early eighteenth century, Gottfried Wilhelm Leibnitz proposed

[1] There were insufficient numbers of females holding legal positions in the transcripts analysed for us to consider this group.

the notion of the ‘alphabet of human thoughts’, which assumes that larger ideas are composed of smaller ideas which can be decomposed further (Bunnin & Yu 2004). One attempt to revive this idea was made by Anna Wierzbicka and colleagues (Wierzbicka 1992; Goddard & Wierzbicka 1994), who proposed the theory of Natural Semantic Metalanguage (NSM). This approach involves analysing concepts through decomposition until the most basic semantic component, called a *semantic primitive* or *prime*, is reached. Because the semantic primitive cannot be decomposed any further in any language, it forms the basic operational unit for the analysis of concepts within and between languages. However ambitious the NSM research programme is, it heavily relies on introspection, which leaves its empirical validity in some doubt. In contrast, we provide an empirical framework based on usage to investigate the meaning of a concept.

Introspection as to the relative importance of given concepts has also guided historians in identifying concepts through *cultural keywords*. An example of this approach is Raymond Williams’ *Keywords: A Vocabulary of Culture and Society* (Williams 1983), which has recently been updated and further developed by the Keywords Project (Maccabe & Yanacek 2018). Researching keywords involves first identifying concepts that are salient to a researcher, for example EDUCATION, QUEER, DEMOCRACY and then documenting their histories through semantic change of main lexicalisations of those concepts, that is, words such as *education*, *queer*, *democracy*, as well as through engagement with cultural and political debate records.

Another way to look at the history of a concept is through exploring words that have been used to express it. This onomasiological perspective is typically incorporated in studies of lexical fields or another thematically organised domain, such as field of emotions (Geeraerts et al. 2012; Diller 2014), conceptualisations of intelligence (Allan 2009), concept of intensification (Lenker 2007). This perspective is often complemented by a semasiological approach which traces meanings mapped by a given word over a period of time (Tucker 1972; Robinson 2011, 2012). Studies that focus on the history of a word, and through it concepts have increasing been relying of usage data.

Studies of words which grew out of historical semantics tradition now typically use corpus evidence to investigate semantic variation and change (e.g. Allan & Robinson 2011). These studies also typically interpret the findings in the context of cognitive linguistics (e.g. Glynn & Robinson 2014). An emerging trend in this context is incorporation of sociolinguistic evidence alongside corpus evidence in understanding changes in semantic categories (e.g. Pütz et al. 2012, 2014). In this context, cognitive-sociolinguistics explores word usage as result of work of both cognitive and social factors. For example, Robinson (2010) shows that the prototypical usage of the adjective *awesome* across a speech community differs from the prototypical sense distinct generations of speakers attach to that adjective. This apparent discrepancy at a cognitive porotype level turns out to be an expected stage in language change in progress when a sociolinguistic construct of apparent time is incorporated.

More recently a more bottom-up view of concept analysis emerged, a view which looks for concepts in groups of words co-occurring across larger passages of text. This approach relies on the assumption that concepts are not always coterminous with a word, but may be represented by a combination of words, phrases, or constructions (Fitzmaurice et al. 2017a: 56–8). Therefore, in order to research a discursive concept, one needs to look for syntagmatic relations that characterise meaning in text. This approach generates sets of associated words, phrases, or constructions referred to as *co-occurrence clusters*, which may constitute discursive concepts or fragments thereof. The application of that methodology brings to the surface findings that go beyond traditional semasiology and onomasiology (Fitzmaurice et al. 2017a, 2017b; Fitzmaurice 2022; Mehl 2022). The potential of the LDNA methodology is indisputable, but it currently does not incorporate other meaning-making linguistic or extralinguistic information associated with data, such as those involving grammatical parameters or text creation contexts.

In the empirical analysis and comparison of corpora, a lot can be learnt from the field of corpus linguistics. In this field, the standard method used to compare corpora is to look at word frequency distributions in those corpora (see e.g. Glynn & Robinson 2014). Statistical measures, such as relative entropy (Kullback & Leibler 1951), the log-likelihood ratio (LLR) (Dunning 1993) and keyness (Bondi & Scott 2010) are frequently used to identify words which are particularly characteristic of a particular subcorpus. For example, Degaetano-Ortlieb (2018) uses relative entropy to examine gender and social class variation in the Old Bailey Corpus. In the current research, we use methods from corpus linguistics to compare corpora in terms of their conceptual content rather than their word content.

Analysing frequency distributions of concepts rather than words is also not entirely new. For example, USAS tagger draws on an ontology of concepts originally developed for lexicography to identify key concepts in the text (see Rayson 2008). This work elicits a list of key concepts in a given text and has been shown useful in comparing texts, such as political manifestos (Rayson 2008) or genres of courtroom speech (Klingenstein et al. 2014). Klingenstein et al. (2014) use Roget's Thesaurus categories to 'coarsegrain' the words in 112,485 Old Bailey trial records from 1760–1930 (Hitchcock et al. 2012; Huber et al. 2012) before carrying out a frequency analysis of the semantic categories. The main purpose of this task is to reduce the number of lexical entries in order to find which concepts were indicative of a type of crime in Old Bailey records, for example violent versus non-violent. For example, *purse* is found in two categories, that is SET 800 MONEY and SET 802 TREASURY. An occurrence of this noun in the text is split between these two semantic categories. Other similar words, such as *coin*, are also found in SET 800 MONEY. The distributional analysis of SET 800 vs other sets shows that this set is more distinctive of non-violent crime than of violent crime.

However, while splitting counts between the different possible semantic categories serves the purpose of reducing the number of lexical entries, it overlooks the fact that, absent a literary pun or device, each word or phrase occurrence is typically a reference to a single concept. Each occurrence of the word *purse* is referring **either** to SET 800 MONEY **or** SET 802 TREASURY, and the most likely sense in each case depends on the context. Word sense disambiguation (WSD) has been a key challenge in computational linguistics in recent years and there has been considerable development of automatic WSD algorithms (Raganato et al. 2015). Methods range from knowledge-based (Banerjee & Pedersen 2003) to fully supervised and semi-supervised machine-learning approaches (Pilehvar & Navigli 2014). State-of-the-art systems often report accuracies in the region of 90%. However, most practical research has focused on present-day language due to the need for and availability of sense inventories and corpora. Algorithms also generally expect clean spelling and grammatical usage. Deviation from these expectations results in a rapid decrease in the performance of most off-the-shelf solutions. Here, we use a semantic tagger, the Samuels Tagger (Piao et al. 2017), which is sensitive to historical spellings and usages.

Further, in the current research, we go beyond identifying lists of characteristic words or concepts (as in Rayson 2008 or Klingenstein et al. 2014) by examining the dependency relationships between concepts and the concept co-occurrence. We demonstrate the value that can be added by investigating relations between concepts. We show that just as words are polysemous, having different meanings in different contexts, concepts are also polysemous, that is meaning different things to different people living in the same time and space.

The area of word-level distributional semantic analysis has also received a lot of attention in computational linguistics recently. It would be straightforward to extend state-of-the-art low-dimensional word embedding methods (Mikolov et al. 2013; Pennington et al. 2014; Tian et al. 2018) to operate at the conceptual level rather than the word level. Of particular note, Tian et al. (2018) propose a new tensor decomposition method which incorporates extralinguistic information into the distributional semantic model, and thus they create and analyse covariate-specific word embeddings. However, in general, word embedding methods only consider the proximity of two words and do not consider the grammatical relationship between two co-occurring words. On the other hand, we consider grammar to be an essential component of meaning making (Langacker 1987, 1991). In other words, we consider the co-occurrence of two concepts in a given dependency relation to be more meaningful than the co-occurrence of two concepts within a given window of proximity. Further, we also wish to be able to interpret the conceptual representations, which is not possible with state-of-the-art low-dimensional word embeddings. Thus, we combine dependency analysis with conventional distributional semantic analysis over high-dimensional representations (Weeds 2003; Sahlgren 2006; Weir et al. 2016).

4. Data: The Old Bailey Voices Corpus

We address the above-mentioned challenges by analysing functional, that is grammatical and sociolinguistic, distributions of concepts in the Old Bailey Proceedings as represented in the Old Bailey Voices Corpus (OBVC). The choice of the corpus is primarily motivated by our broader interests in the sociolinguistic reality of nineteenth-century speech. The OBVC is a particularly good fit to test our hypotheses related to conceptual content since it contains historical speech that is rigidly controlled for context. The OBVC represents a historically real, yet linguistically controlled, dataset restricted to one genre. Since the OBVC takes a consistent generic form (the trial), it is well suited for testing methods of automatic concept identification.

This corpus was also chosen because it is rich with sociolinguistic metadata. The OBVC contains tags for a variety of linguistic and sociocontextual characteristics such as the offence type, the verdict, the sentence, the gender of the defendant, and the gender and role of speakers at the trial. It also contains information about the length of utterances and the number of utterances in a trial and whether a defendant speaks or is silent. A recent illustrative example of sociolinguistic study of language in the Old Bailey court has been carried out by Degaetano-Ortlieb et al. (2021). Degaetano-Ortlieb et al. (2021) show the roles men and women assume in language change, which is a result which further strengthens the focus of the current paper on gendered language.

The OBVC is derived from two sources: the Old Bailey Corpus (version 2) (OBC) (http://fedora.clarin-d.uni-saarland.de/oldbailey/) (Huber et al. 2012) and the Old Bailey Online (OBO) (http://www.oldbaileyonline.org) (Hitchcock et al. 2012). It contains data from all single-defendant trials (21,023 defendants) in 227 sessions of the Old Bailey Proceedings between 1780 and 1880 which have had linguistic markup added by the Old Bailey Corpus project. The dataset has been created in order to explore the Voices of Authority research theme of the Digital Panopticon (http://www.digitalpanopticon.org) project. Single-defendant trials were chosen in the Digital Panopticon project since the records are linked to the outcomes of the defendants (via transportation and imprisonment records). For our purposes, it further controls the context of the utterances, which is particularly important when testing a method. From a methodological point of view, the methods described in the current paper could be applied to other corpora.

5. Methodology: From corpora to conceptual dependency corpora

We have developed a computational tool that uses a pipelined approach to annotate corpora with conceptual and dependency information. We refer to this tool as Conceptual Dependency Corpus (CDC).

There are four major steps in the process of building and analysing the CDC. Assuming that we have the speech data and metadata from the OBV as input, then the procedure, shown in Figure 1, is as follows:

(1) Subcorpora creation
(2) Automatic semantic annotation
(3) Automatic dependency annotation
(4) Co-occurrence extraction

The CDC also makes it possible to return from the co-occurrence statistics to the actual text of the corpus, enabling further linguistic and historical analysis. Each step is now described in detail below.

5.1. *Subcorpora creation*

The OBVC is annotated with metadata at both the trial and the utterance level. In the first step of our pipeline, this metadata is queried to build subcorpora with given characteristics. The subcorpora contain solely the raw text and no further metadata, making it possible for them to be processed by off-the-shelf text processing tools.

In the case study presented in Section 6, we have further controlled the context of speech by restricting the time period and offence type of the trial. We have chosen to study the period

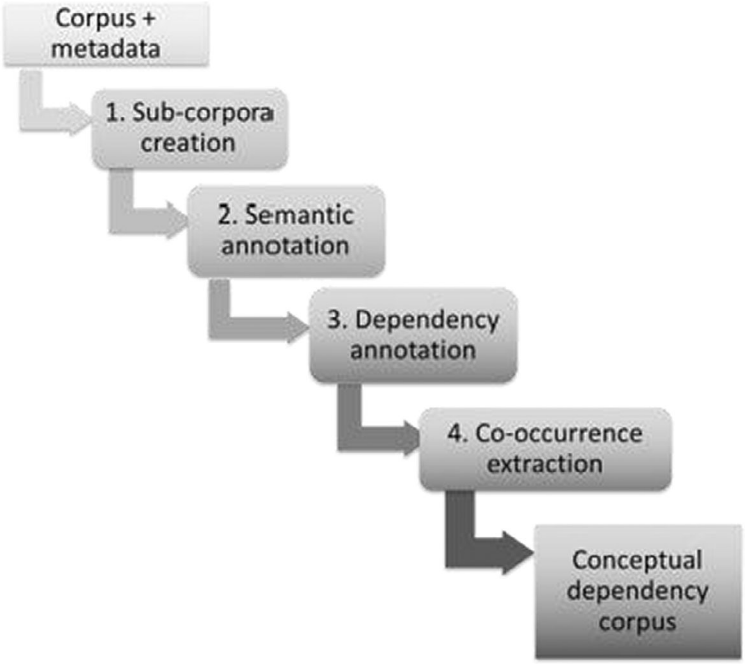


Figure 1. Pipeline for building the conceptual dependency corpus

Table 1. Number of tokens in each subcorpus

Key	Description	# Tokens
Leg	male AND lawyer	165,824
Mnonl	male NOT lawyer (i.e. witnesses and defendants)	1,083,003
fnonl	female NOT lawyer (i.e. witnesses and defendants)	219,407

1800–1820 and only trials of theft offences.[2] This is one of the larger categories of crime and by restricting to one crime we are limiting the domain of intra-genre variation. We wish to study the differences between female and male speech. However, we note that of the roles played by the speakers (lawyer, defendant, witness), there are no female lawyers and therefore it is necessary to further separate the corpora by speaker role.

We then created separate subcorpora matching each of the following conditions on the speaker of an utterance: male AND lawyer, male NOT lawyer, female NOT lawyer. Table 1 shows the amount of data in each subcorpus in terms of the number of tokens. We note that the witness to defendant ratio is approximately 4:1 for both males and females.[3]

5.2. Semantic annotation

The Samuels Tagger (Piao et al. 2017) is a historically sensitive semantic tagger. It takes raw text as input and carries out tokenisation and variant detection (VARD), lemmatisation, part-of-speech (POS) tagging and word sense disambiguation. In this process, each token in the input is mapped to its most likely conceptual category in the Historical Thesaurus of English (Kay et al. 2016), given both the context of the sentence and the historical nature of the input. For example, the word *purse* might be tagged as any of BB.07.f [Symbol of office/authority], ZA01 [Personal Name], BJ.01.m.01 [Funds/pecuniary resources] and BJ.01.m.09.c [Money-bag/−purse/−belt, etc] depending on the sentential context.

The Samuels Tagger uses automatic word disambiguation methods and is generally reported (Piao et al. 2017) to have an accuracy of between 77% and 91% in individual test texts. We found that the Samuels Tagger had a token coverage of 98.3% across all three subcorpora, that is the Samuels Tagger was able to assign a conceptual category in 98.3% of cases but this may not have been the correct tag. Regarding accuracy, we have not performed an extensive quantitative evaluation of the tagger on our corpus. However, we have manually checked the tags for errors and have restricted our case study (section 6) to conceptual categories which are largely error-free. However, we note there may be lexical choices for conceptual categories which have been missed by the tagger.

Table 2 illustrates how the input sentence, *I am the wife of Thomas Hopwood* would be annotated by the Samuels Tagger.

Each semantic tag refers to a category in the Historical Thesaurus of English. Note that these categories are hierarchical in nature. Accordingly, the category AY.01.g.01 [Married woman] is subsumed by the category AY.01.g [Marriage/wedlock], which is in turn subsumed by the category AY.01 [Kinship/relationship], which in turn is subsumed by the category AY [Society].

[2] In the 1800–20 period close to 60% of crimes at courts like the Old Bailey were larceny, so the reduction to only larceny cases does not limit the corpus in an extreme way. The category "theft with violence" is a subcategory of "theft" within the OBCV. Authors thank Reviewer 2 for this observation.

[3] Although approximately 25% of defendants are females the % of speech from females is less than that, about 15% (to add reference and comments from Reviewer 2, p. 7, Table 1)

Table 2. Example of semantic annotation

Token	Lemma	POS	SEMTAG3
I	*i*	PPIS1	ZF [Pronoun]
am	*be*	VBM	AK.01.g [State/condition]
the	*the*	AT	ZC [Grammatical item]
wife	*wife*	NN1	AY.01.g.01 [Married woman]
of	*of*	IO	ZC [Grammatical item]
Thomas	*thomas*	NP1	ZA01 [Personal Name]
Hopwood	*hopwood*	NP1	[]

5.3. Dependency annotation

In order to be able to carry out distributional analysis of concepts, it is first necessary to extract grammatically related concepts. SpaCy (Honnibal & Johnson 2015) is a Python library providing reliable and robust Natural Language Processing capabilities. Using this library, it is possible to automatically carry out sentence segmentation, tokenisation, lemmatisation, part-of-speech tagging, dependency analysis and named entity recognition.

In order to preserve the tokenisation carried out by the Samuels Tagger, we passed pre-tokenised sentences into SpaCy and inspected the dependency analyses carried out. Note that SpaCy provides exactly one dependency label per token, which describes its relationship to its head constituent. These dependency labels are strictly exclusive. For example, _DET does not overlap with _POSS despite a possessive being considered a determiner in traditional linguistic analysis. Since we use pre-tokenised sentences (produced by the VARD component of Samuels), it is straightforward to align the outputs of Samuels and SpaCy. Table 3 shows the SpaCy dependency analysis for the sentence *I am the wife of Thomas Hopwood*.

From Table 3, we can reconstruct dependency relationships between words. For example, we can see that token 0 (*I*) is the dependent in a non-clausal subject (NSUBJ) relationship with token 1 (*am*). Similarly, we can see that token 3 (*wife*) is the dependent in an attributive (ATTR) relationship with token 1 (*am*).

Table 4 shows the subsequent combination of semantic and dependency annotations in the CDC.

Through the combination of annotations, we can infer that the concept ZF [Pronoun] is the dependent in a non-clausal subject relation (NSUBJ) with the concept AK.01.g [State/Condition]. Similarly, we can also infer that the concept AY.01.g.01 [Married woman] is the dependent in an attributive relation (ATTR) with the concept AK.01.g [State/Condition].

Table 3. Example of dependency annotation

	Token	Head	rel
0	*I*	1	NSUBJ
1	*am*	9	CCOMP
2	*the*	3	DET
3	*wife*	1	ATTR
4	*of*	3	PREP
5	*Thomas*	6	COMPOUND
6	*Hopwood*	4	POBJ

Table 4. Combination of semantic and dependency annotation

	Token	Lemma	POS	SEMTAG3	Head	rel
0	*I*	*i*	PPIS1	ZF [Pronoun]	1	NSUBJ
1	*am*	*be*	VBM	AK.01.g [State/condition]	9	CCOMP
2	*the*	*the*	AT	ZC [Grammatical item]	3	DET
3	*wife*	*wife*	NN1	AY.01.g.01 [Married woman]	1	ATTR
4	*of*	*of*	IO	ZC [Grammatical item]	3	PREP
5	*Thomas*	*thomas*	NP1	ZA01 [Personal Name]	6	COMPOUND
6	*Hopwood*	*hopwood*	NP1	[]	4	POBJ

5.4. *Co-occurrence extraction*

In order to test our hypotheses, we inevitably need to use computational or statistical techniques that are capable of analysing data in a bottom-up fashion.

In this work, we apply the very popular log-likelihood ratio (LLR) (Dunning 1993; Rayson & Garside 2000) to discover concepts which are observed to occur more often than one would expect given some hypothesised probability distribution. Intuitively, if a concept such as WOMAN occurs more frequently in one subcorpus than one would expect, given its probability in the whole corpus, then WOMAN would be considered a characteristic concept for that subcorpus as compared to the reference corpus.

More concretely, the hypothesised distribution is the probability distribution in the combined corpus and the observed distribution is the probability distribution in a given subcorpus. If we have a 2 × 2 contingency table showing the observed frequency of concept w in corpus c ($O_{1,1}$), the observed frequency of concept w NOT in corpus c ($O_{1,2}$), the observed frequency of NOT concept w in corpus c ($O_{2,1}$) and the observed frequency of NOT concept w in NOT corpus c ($O_{2,2}$), then the log-likelihood ratio λ is given by Equation 1.

Equation 1: Log-likelihood ratio

$$-2\log\lambda = 2\sum_{i,j} O_{i,j}\log\frac{O_{i,j}}{E_{i,j}},$$

where $E_{i,j}$ is the expected frequency of concept i in corpus j given by Equation 2:

Equation 2: Expected frequency in contingency table

$$E_{i,j} = \frac{\sum_i O_i \times \sum_j O_j}{\sum_{i,j} O_{i,j}}.$$

One advantage of using the log-likelihood ratio over some other tests is that there is a straightforward way to determine the number of significant concepts as well as the most significant concepts. In our setting and using a Bonferroni Correction (Shaffer 1995), any log-likelihood ratio scores greater than 32 are statistically significant at the 0.1% level, that is, there is less than a 0.1% chance of this score or higher occurring given the hypothesised probability distribution. We use the term *highly characteristic* to refer to words with scores, which are statistically significant at the 0.1% level.

We also want to extract and analyse patterns in the co-occurrences of grammatically related concepts. Traditional distributional semantic analysis (Weeds 2003; Sahlgren 2006; Weir et al. 2016) relies on representing words in a very high-dimensional space. Each dimension of this space is co-occurrence (in some relationship) with a given word in the vocabulary. Geometric and statistical measures can then be applied to determine words which are similar (close) to each other in this space.

Table 5. Examples of concepts being represented in terms of co-occurring dependency relationships and concepts

Concept	Feature representation based on co-occurrences with other concepts in named relationships	Feature representation based on co-occurrences in named relationships
ZF	{NSUBj+AK.01.g: 1}	{NSUBJ: 1}
AK.01.g [State/condition]	{_NSUBJ+ZF: 1, _ATTR:AY.01.g.01: 1}	{_NSUBJ: 1, _ATTR: 1}
ZC [Grammatical item]	{DET + AY.01.g.01: 1, PREP+AY.01.g.01: 1}	{DET: 1, PREP: 1}
AY.01.g.01 [Married woman]	{ATTR+AK.01.g: 1, _DET + ZC: 1, _PREP+ZC: 1}	{ATTR: 1, _DET: 1, _PREP: 1}
ZA01 [Personal Name]	{COMPOUND+[]: 1}	{COMPOUND: 1}
[]	{_COMPOUND+ZA01: 1, POBJ+ZC: 1}	{_COMPOUND: 1, POBJ: 1}

Here, we apply distributional semantic analysis to concepts rather than words. Therefore, we represent concepts in terms of their co-occurrences with other concepts in certain grammatical dependency relations. First, for each concept, we build two co-occurrence feature vectors (1) by counting how many times it occurs with another concept in a named dependency relation and (2) by counting how many times it occurs in each named dependency relation. Table 5 shows the co-occurrence features extracted for each concept from the sentence *I am the wife of Thomas Hopwood.*

Note that, in the first case, the name of each feature or dimension contains both the dependency relation and the concept (therefore the features _DET+ZC and _PREP+ZC are distinct). Note also that the direction of the dependency relation (from head to dependent or from dependent to head) is also marked by the presence or absence of the prefix symbol '_'. For example, the determiner relation from a determiner to a noun is labelled as DET whereas the inverse relation from the noun to the determiner is labelled as _DET.

Once co-occurrence features have been extracted and counted over the entire subcorpus, the weight of each feature is transformed from raw frequency into a measure of association. Here, following Evert (2008), we use a normalised localised version of positive pointwise mutual information. In Equation 3, the pointwise mutual information between a concept, c, and a feature, f, is defined as the log of the ratio between the conditional probability $P(f|c)$ and the marginal probability $P(f)$.

Equation 3: Pointwise mutual information

$$PMI(c,f) = \log\left(\frac{P(f|c)}{P(f)}\right).$$

Intuitively, this captures how much more likely a feature is to occur with a given concept than it is to occur in general. For example, a very frequent feature such as DET:ZC has a high probability of occurrence with many concepts, which results in it having a lower weight of association with concepts than less frequent (i.e. more informative) features such as COMP: ZA01.

In Equation 4, positive pointwise mutual information (PPMI) ignores features which have occurred less frequently than would be expected. From a practical point of view, this allows us to side-step the mathematical issues in taking the log of 0 and, from a theoretical point of view, focuses subsequent comparisons on features which have occurred significantly more than would be expected.

Equation 4: Positive pointwise mutual information

$$PPMI(c,f) = \max(PMI(c,f), 0).$$

While it is very common to use PPMI in distributional semantic analysis, it has been noted by a number of researchers that it gives too much weight to low-frequency features. This is particularly problematic in our case when we have relatively small corpora and we wish not only to determine similarity between concepts but also to discover significant differences between subcorpora. We do not wish our evidence for linguistic variation to be based on just a handful of co-occurrences in the respective subcorpora. Evert (2008) recommends reweighting PMI with the conditional probability of the feature given the word (or concept in our case), as shown in Equation 5. The result is a weight of association between each concept and feature which is more balanced between frequently occurring features and informative features.

Equation 5: Localised pointwise mutual information

$$lPMI(c,f) = P(f|c)PMI(c,f).$$

Finally, we renormalise the weights of association in each feature vector so that each feature vector has unit length. This does not affect similarity calculations such as the cosine measure which effectively renormalise when they are applied. However, by pre-computing and storing the normalised vectors, we make subsequent similarity calculations faster (since cosine can be computed with a single dot product) and it means that our visualisations of feature representations have a consistent scale. Cosine similarity between vectors for concepts c_1 and c_2 can then be calculated as shown in Equation 6.

Equation 6: Cosine similarity

$$\cos(v_{c_1}, v_{c_2}) = \frac{v_{c_1}.v_{c_2}}{\|v_{c_1}\|\|v_{c_2}\|}.$$

Assuming that vectors are unit length, this simplifies to being equivalent to the dot product of the two vectors.

We are now in a position to compare feature representations for different concepts both within and across subcorpora. Our results are presented in section 6.

6. Case study

In this section, we explore conceptual variation that exists in the speech of legal males, non-legal males and non-legal females with a particular attention to the concepts MAN, WOMAN and CHILD. The case study presented here showcases the analysis which is possible using the conceptual dependency corpus-based methodology described in section 5. First, we demonstrate the gains which can be made by moving from a word-based analysis to a concept-based analysis (section 6.1) and second, we demonstrate the gains which can be made by moving from individual concepts to concept co-occurrences (section 6.2).

6.1. *From word-based analysis to concept-based analysis*

Typically, word-based analysis focuses on identifying words which are characteristic of a particular subcorpus using various statistical measures. For example, we can ask, which words occur most frequently in each subcorpus? Or we can ask, which words occur more often than one would expect in each subcorpus given their general frequency distribution?

Table 6. Frequencies of most frequently occurring words in each subcorpus

	Female	Male	Legal
the	13,951	70,388	10,567
i	11,375	56,776	605
and	5916	29,728	1,542
of	4,353	24,410	4,222
in	4,265	19,999	2,724
to	4,176	22,352	3,811
a	4,152	21,373	1,929
was	3,784	18,553	3,192
he	3,318	21,244	1,882
my	2,901	9,813	65
it	2,761	14,077	2,161
she	2,523	4,693	465
prisoner	1,920	10,930	1,709
him	1,636	13,540	1,174
that	1,332	7,426	3,442
did	749	3,116	3,704
any	311	1,330	1,972
you	250	1,477	11,069
your	62	245	2,106

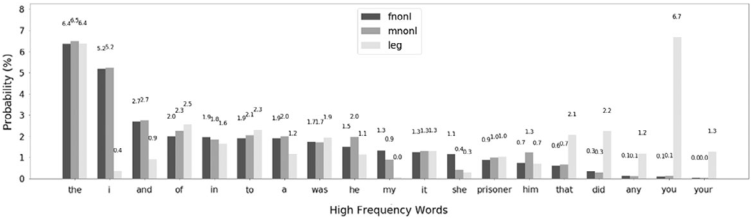


Figure 2. Comparison of probabilities of high-frequency words across the three subcorpora

Table 6 shows the frequencies of the most frequently occurring words in both the female non-legal corpus, the male non-legal corpus and the (male) legal corpus. These words are ordered by frequency in the female corpus but include the ten most frequently occurring word tokens in each of the three subcorpora. For example, *you* is not one of the ten most frequently occurring words in the female subcorpus but since it is in the top ten most frequent words in the legal corpus, it is included in the analysis presented here. These frequencies are shown as probabilities, that is proportions of the entire subcorpus, in the graph in Figure 2. For example, the word *the* makes up approximately 6% of the entire female non-legal sub-corpus, that is six in every 100 words is the word *the*.

This type of analysis highlights the different types of pronouns used in each subcorpus. We can see that legal speakers refer far more frequently to *you* and *your* and the non-legal speakers refer far more frequently to *I* and *my*. This is hardly surprising due to the nature of the discourse. Regarding genders, we can see that both male and female non-legal speakers use *I* with equal probability but male non-legal speakers are more likely to use *he* than female non-legal speakers while, conversely, female non-legal speakers are more likely to use *she*. Therefore, it appears that the courtroom narratives told by men are more likely to have male characters and the courtroom narratives told by women are more likely to have female characters. This is consistent with men and women having well-defined gendered roles and mixing predominantly with their own gender.[4]

[4] Note that the OBVC trials are only single-defendant trials and therefore do not contain any mixed gender crimes.

Comparing the female non-legal subcorpus to the other subcorpora, using the LLR, we find 50 highly characteristic words, that is words with scores which are statistically significant at the 0.1% level. Comparing the male non-legal subcorpus against the other subcorpora, we find 38 highly characteristic words. Comparing the male legal subcorpus against the other subcorpora, we find 121 highly characteristic words. The much higher number of highly characteristic words perhaps suggests that this subcorpus is more distinctive in its use of vocabulary.

The top ten most characteristic words for each subcorpus together with their LLR scores and frequencies are listed in Table 7 and illustrated in Figure 3. This analysis highlights which words occur more frequently than one would expect but not how these words relate to each other conceptually.

Table 7. 10 Highest LLR scores for words for each of the three subcorpora

Female non-legal			Male non-legal			Male legal		
Word	LLR	Freq.	Word	LLR	Freq.	Word	LLR	Freq.
she	773	2,523	i	1,572	56,776	you	19,606	11,069
husband	600	519	am	439	6,975	your	3,721	2,106
her	450	2,045	and	398	29,728	did	3,324	3,704
my	271	2,901	him	382	13,540	any	2,016	1,972
room	184	617	he	319	21,244	what	1,589	1,748
stairs	150	425	his	199	6,281	do	1,454	1,608
things	131	513	took	168	4,536	that	1,312	3,442
wife	101	340	a	125	21,373	how	1,262	970
mother	98.5	181	officer	118	1,599	are	684	1,062
me	97.5	2,155	said	117	6,522	know	589	1,112

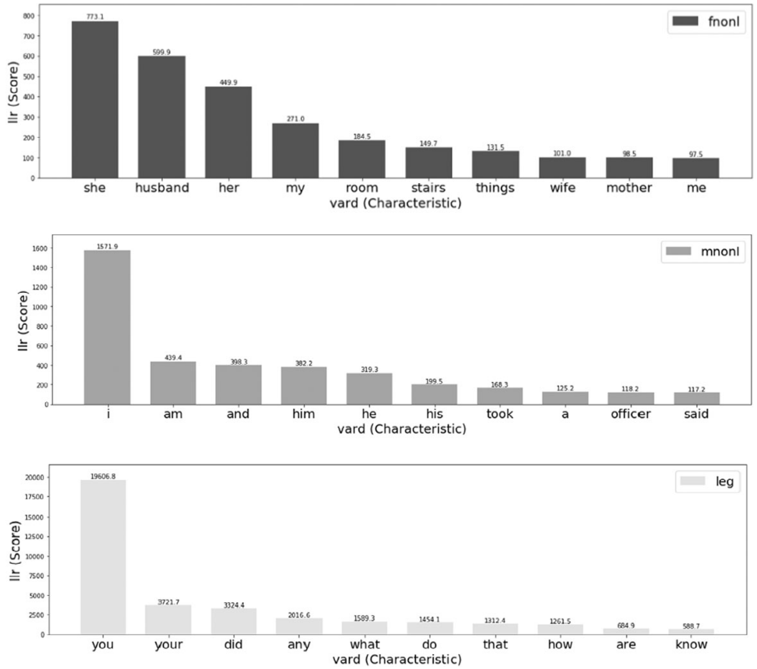


Figure 3. 10 Highest LLR scores for words in each of the three subcorpora

We now ask, what do we get by moving to concepts? Accordingly, we carry out the same analysis as above but at the concept level using the CDC created in section 4.

Table 8 shows the frequencies of the most frequently occurring concepts, as tagged by the Samuels Tagger, in the female non-legal corpus, the male non-legal corpus and the (male) legal corpus. In keeping with our word-based analysis, these concepts are ordered by frequency in the female corpus but include the ten most frequently occurring semantic concepts in each of the three subcorpora. For example, AW.15 [Taking] is not one of the ten most frequently occurring concepts in the female non-legal subcorpus but since it is one of the ten most frequently occurring concepts in the male non-legal subcorpus, it is included in the analysis presented in Table 8.

The frequencies from Table 8 are shown as probabilities, that is proportions of the entire subcorpus, in Figure 4. For example, grammatical items (ZC) make up 23.9% of the female non-legal sub-corpus, that is 23.9 in every 100 words is a grammatical item (ZC). Figure 4 shows that grammatical items (ZC) and pronouns (ZF) are more frequent in the legal speech than in the non-legal speech. Non-legal speakers refer more to the concepts of NUMBER (AP.04) and OWNING (AW.01). Looking at just these most frequent concepts, the differences between male and female non-legal speech appear less marked.

We use the log-likelihood ratio (LLR) to discover which concepts occur more frequently in one subcorpus than one would expect, given their frequency in the combined corpus. We find that there are 37 highly characteristic concepts in the female non-legal subcorpus, 23 highly characteristic concepts in the male non-legal subcorpus and 54 highly characteristic concepts in the male legal subcorpus. We have a smaller number of highly characteristic concepts than characteristic words in each case, since we are probing the underlying conceptual ideas rather

Table 8. Frequencies of the most frequently occurring concepts in each of the three subcorpora

	Female	Male	Legal
ZC [Grammatical Item]	52,409	269,044	44,079
ZF [Pronoun]	28,680	140,523	24,808
AK.01.g [State/Condition]	7,719	40,193	6,741
AP.04 [Number]	5,194	25,141	2,231
AW.01 [Owning]	4,737	19,641	1,068
ZA01 [Personal Name]	3,276	16,314	2,712
AXE.03 [Speech]	2,809	13,148	1,362
ZD [Negative]	2,320	9,913	2,064
BB.12.e.02 [Prisoner]	1,810	10,385	1,598
Al.14.b [Seeing/looking]	1,680	7,954	1,744
AN.05.k [Going/coming out]	1,192	4,766	383
AW.15 [Taking]	1,044	5,856	554
AR.36 [Knowledge]	753	3,178	1,368
AP.04.a [One]	391	1,759	1,947

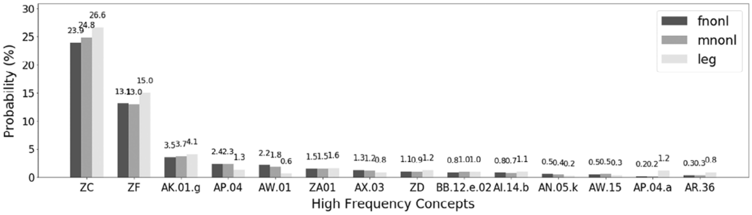


Figure 4. Probabilities of the most frequently occurring concepts in each of the three subcorpora

than individual word forms chosen by speakers. The top ten concepts for each subcorpus, together with the LLR scores and their respective frequencies, are given in Table 9 and illustrated in Figure 5.

By grouping related words into concepts, a clearer picture emerges of the discourse of the speakers. For example, one of the highly characteristic concepts for female non-legal speakers

Table 9. Top ten highest LLR scores for concepts in the three subcorpora

Female non-legal			Male non-legal			Male legal		
Concept	LLR	Freq.	Concept	LLR	Freq.	Concept	LLR	Freq.
AY.01.g.02 [Married Man]	602	523	BC.07.b [Law enforcement officer]	157	1,770	AP.04.a [One]	1,677	1,947
BF.25.j [Vestments]	155	335	AP.04 [Number]	105	13,607	AM.08.c [The Past]	926	647
AZ.06.f.02.f [Room]	142	680	BA.14.k.08.a [Officer/soldier]	95	1,613	AS.04 [Enquiry]	781	875
AZ.06.f.02.h [Stairs]	131	383	ZA02 [Geographical name]	95	1,822	AR.36 [Knowledge]	430	1,356
AW.01 [Owning]	129	4,737	AW.01 [Owning]	81	19,641	BA.06 [Defence]	415	207
AY.01.g.01 [Married woman]	119	339	AO.16.a.04 [Watching/keeping guard]	68	530	AM.03 [Particular time]	386	1,321
AA.04.g [Flowing water]	107	380	AO.11.b [Searching/seeking]	58	1,423	AM [Time]	354	796
AY.01.e.01.b [Mother]	105	188	BB.03 [Control]	57	343	AM.02 [Duration]	298	322
AD.03.d [Child]	104	206	AL.06 [Direction]	53	3,576	AK.01 [Existence]	276	1,196
AY.01.e.05.b [Sister]	95	127	AW.15 [Taking]	46	5,849	ZF [Pronoun]	241	24,808

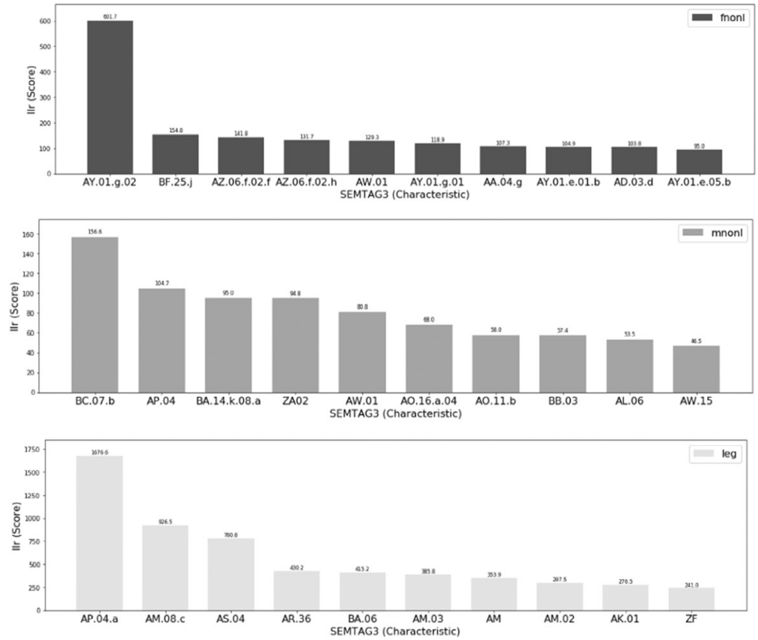
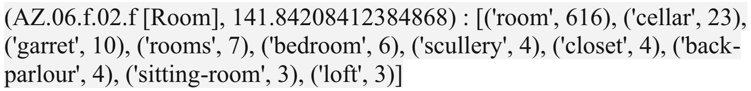


Figure 5. 10 Highest LLR scores for concepts in each of the three subcorpora



```
(AZ.06.f.02.f [Room], 141.84208412384868) : [('room', 616), ('cellar', 23),
('garret', 10), ('rooms', 7), ('bedroom', 6), ('scullery', 4), ('closet', 4), ('back-
parlour', 4), ('sitting-room', 3), ('loft', 3)]
```

Figure 6. Words (with frequencies) tagged as the concept AZ.06.f.02.f [Room] in the fnonl subcorpus

is BF.25.j [Vestments], which is used to tag (predominantly) words such as *gown*, *gowns*, *apron* and *aprons*. In the word-based analysis, *gown* was placed 13th in the list of highly characteristic words with a LLR of 96, *gowns* 33rd with a LLR of 46 and *apron* 49th with a LLR of 33. However, with the word-based analysis, the conceptual link between these words is ignored and the fact that it is a concept, rather than individual words, which is significant is obscured.

Another of the highly characteristic concepts for female non-legal speakers is AZ.06.f.02.f [Room]. This conceptual tag has been applied to words in this subcorpus (see below) which would not have been deemed frequent or characteristic enough on their own, given a purely word-based analysis. Of these words, only the word *room* had a significant LLR (>32) in the word-based analysis. However, grouping together infrequent words such as *back-parlour*, *closet* and *garret* into the single concept AZ.06.f.02.f [Room] further highlights the significance of this concept (rather than the individual word forms) in the language of our female non-legal speakers, as illustrated in Figure 6.

Our analysis also shows that other related concepts (e.g. AZ.06.f.02.h [Stairs]) are also highly characteristic of the female non-legal speech. The co-hyponymy relationship between this concept and AZ.06.f.02.f [Room] can be inferred from the structure of the tags since they share the prefix AZ.06.f.02.

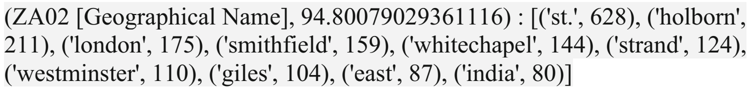
In general, we see that many of the highly characteristic concepts of the female non-legal speech are related to the home or family members. In the 37 highly characteristic concepts for female non-legal speech there are seven which refer to people and of these seven (AY.01.g.02 [Married Man], AY.01.g.01 [Married Woman], AY.01.e.01.b [Mother], AD.03.d [Child], AY.01.e.05.b [Sister], AY.01.e.02.b [Daughter], AD.03.b [Woman]), five refer specifically to females.

Conversely, we see that many of the highly characteristic concepts of the male non-legal speech are suggestive of a narrative being told that involves law enforcement officers (e.g. BC.07.b [Law enforcement officer], BA.14.k.08.a [Officer] and AO.16.a.04 [Keeping guard]) and which is taking place outside the home (e.g. ZA02 [Geographical Name] and AL.06 [Direction]). Looking more closely (Figure 7) at the words tagged with the concept of ZA02 [Geographical Name] also sheds further light on the locations mentioned in these narratives.[5]

It is worth noting that none of these individual words were deemed highly characteristic (LLR > 32) in the word-based analysis of the male non-legal subcorpus. It is only through grouping them together as a concept that their significance becomes apparent.

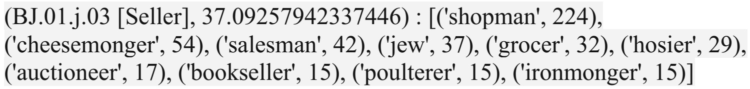
In the 23 characteristic concepts of male non-legal speech there are only three concepts (BC.07.b [Law enforcement officer], BA.14.k.08.a [Officer] and BJ.01.j.03 [Seller]) which refer specifically to people. All of these are references to professions which would have been held by males at the time, as illustrated by looking more closely (Figure 8) at the words tagged with the concept BJ.01.03 [Seller]:

[5] Also notice that the non-legal males use ZA02 [Geographical Name] words while females use of AA.04.g [Flowing water] words, which are primarily references to the Thames river. The conceptualisation of space across genders differ since women use more relative positional place identifiers. Authors thank Reviewer 2 for this suggestion.



```
(ZA02 [Geographical Name], 94.80079029361116) : [('st.', 628), ('holborn',
211), ('london', 175), ('smithfield', 159), ('whitechapel', 144), ('strand', 124),
('westminster', 110), ('giles', 104), ('east', 87), ('india', 80)]
```

Figure 7. Words (with frequencies) tagged as the concept ZA02 [Geographical Name] in the mnonl subcorpus



```
(BJ.01.j.03 [Seller], 37.09257942337446) : [('shopman', 224),
('cheesemonger', 54), ('salesman', 42), ('jew', 37), ('grocer', 32), ('hosier', 29),
('auctioneer', 17), ('bookseller', 15), ('poulterer', 15), ('ironmonger', 15)]
```

Figure 8. Words (with frequencies) tagged as the concept BJ.01.j.03 [Seller] in the mnonl subcorpus

Finally, we note that the highly characteristic concepts in the legal subcorpus include a lot of concepts relating to time (AM [Time], AM.02 [Duration] and AM.03 [Particular Time]) as well as the concept of AS.04 [Enquiry], which is used mainly to tag question words such as *how* and *whether* as well as the words *question*, *questioned*, *interview* and *ask*. The only one of the 54 characteristic concepts for legal speech which refer to people is the generic concept AD.03 [Person].

6.2. *Conceptual-dependency-based analysis*

The conceptual analysis presented so far only allows us to answer questions about whether different groups of people use different concepts. It does not allow us to answer questions about how different communities of people use the same concepts. By moving from a conceptual corpus to a CDC, we probe the idea of socioconceptual polysemy. We investigate whether the same concept has the same meaning for different communities of people by applying the distributional hypothesis (Harris 1954; Firth 1957) and analysing the co-occurrence distributions of related concepts.

For example, we have noted that AD.03.B [Woman] and AD.03.d [Child] occur as characteristic concepts for the female non-legal speakers in this study, whereas AD.03.a [Man] is not characteristic of any of the three communities of speakers. By moving to the CDC, we can now examine whether these concepts are used in different ways by different groups of speakers, which would support the hypothesis that these communities conceptualise these concepts in different ways.

First, we calculate localised PPMI, as described in section 5.4, between concepts and dependency relations in order to determine which dependency relations are used with each concept more often than one would expect if concepts and dependency relations occurred independently. Here we focus on the dependency relations which occur in the top five most salient dependency relations for any of our target concepts. These dependency relations together with examples from the female non-legal corpus are listed in Table 10 for reference.

The results of this analysis for the concepts AD.03.b [Woman], AD.03.a [Man], and AD.03.d [Child] are summarised in Figures 9, 10 and 11, respectively.

Figure 9 shows the dependency profiles for the concept AD.03.b [Woman] in each of the three subcorpora. We note the relative salience of the _DET, AMOD and ATTR relations in all three subcorpora, since these relations are only applied to noun concepts. For example, statements from female non-legal speech include '*She is a married woman, therefore it is not likely*', '*and when the lady came to lay the cloth there were no spoons*'. The comparison of the subcorpora shows that the use of adjectival modifiers (_AMOD) and the attributive relation

Table 10. Most frequently occurring dependency relations with target concepts across the three subcorpora

Relation	Relation description	Example
_AMOD	concept is modified by an adjective	***young*** *woman*
_CASE	concept has a case modification	*woman* **‘s**
_DET	concept is modified by a determiner	***that*** *woman*
_NUMMOD	concept is modified by a number	***two*** *women*
_POSS	concept is modified by a possessive	***his*** *child*
_RELCL	concept is modified by a relative clause	the *woman* ***that*** I live with
ATTR	concept is direct object of the attributive relation	she **is** a false swearing *woman*
CONJ	concept is conjoined	the prisoner **and** an old *woman*
DOBJ	concept is direct object of a verb	**saw** the young *woman*
NSUBJ	concept is non-clausal subject of verb	a *woman* **told** me
POSS	concept is a possessive modifier for some noun	the *woman*’s **house**

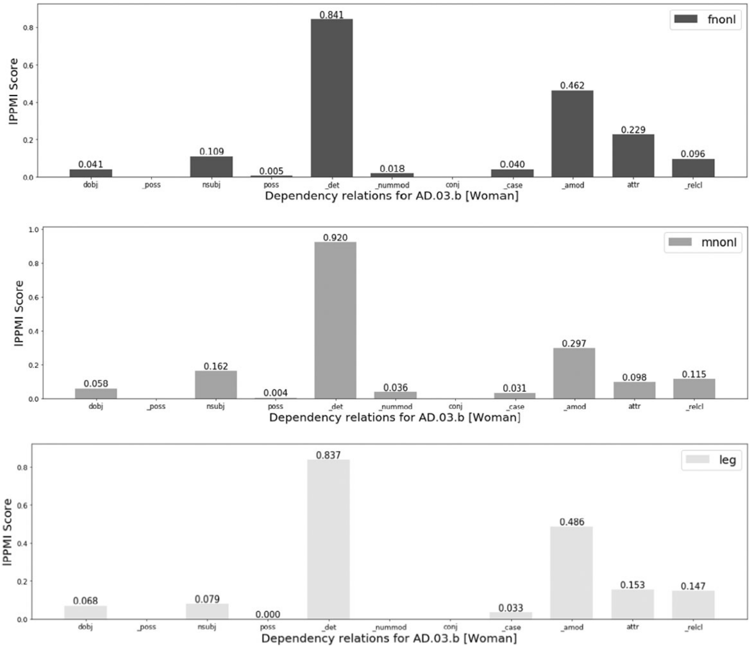


Figure 9. Salience of dependency relations with AD.03.b [Woman]

(ATTR) with the AD.03.b [Woman] concept are both more salient in the female non-legal corpus and the legal corpus than in the male non-legal corpus.

Figure 10 shows the dependency profiles for the concept AD.03.a [Man] in each of the three subcorpora. Again, we note the relative salience of the _DET, ATTR and _AMOD relations. For example, statements from female non-legal speech include ‘*I was out of work at the time and the young man too*’, ‘*and then the other man drove the cart away*’, ‘*the constable and a young man stopped me*.’ More interestingly, we note that these profiles appear to be more similar across the three different subcorpora. The female speakers do use the attributive relation (ATTR) with the concept AD.03.a [Man] more than either group of male speakers but adjectival modification (_AMOD) of the concept AD.03.a [Man] is used with the same probability.

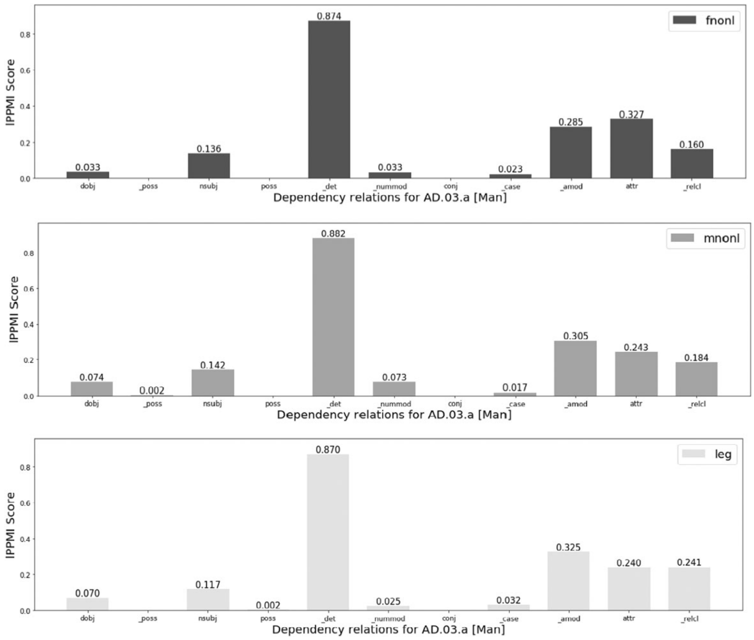


Figure 10. Salience of dependency relations with AD.03.a [Man]

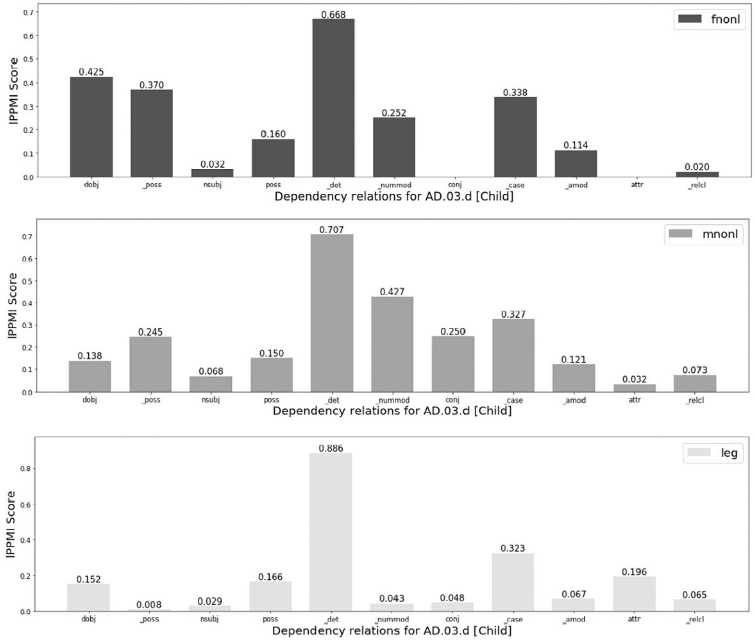


Figure 11. Salience of dependency relations with AD.03.d [Child]

Figure 11 shows the dependency profiles for the concept AD.03.d [Child] in each of the three subcorpora. These profiles are markedly different to the profiles for AD.03.a [Man] and AD.03.b [Woman]. Consistent with being a noun concept, we do see a high use of _DET. However, we also observe significant use of the POSS, _POSS and _CASE relations indicating that a child is both possessed and a possessor, that is when a child is talked about, their relationship to other people is also generally stated. For example, statements from male non-legal speech include '*she said it was for a baby's cap she wanted it*', '*I went one way to look for my child, and my wife the other*'. We also see highly significant use of _NUMMOD since the number of children will often also be stated. Finally, particularly in the female non-legal corpus, we see that the AD.03.d [Child] concept is often the direct object of a verb (DOBJ) as in examples '*I have three small children, and am pregnant now*', '*I had come three hundred miles, with three fatherless children, and one buried on the road*'.

As well as considering differences across the different subcorpora, it is also helpful to look more closely at differences between concepts in a single corpus, which are presented in Figure 12.

Figure 12 shows the differences in lPPMI for each of the dependency relations between the concept AD.03.b [Woman] and AD.03.a [Man] (where *AD.03.b [Woman]* is shown as positive) in each of the three subcorpora. This highlights the fact that the male non-legal speakers use adjectival modification with almost equal probability for the AD.03.b [Woman] concept as for the AD.03.a [Man] concept. However, both female non-legal speakers and male legal speakers are much more likely to use adjectival modification with the AD.03.b [Woman] concept than with the AD.03.a [Man] concept.

In conclusion, this analysis suggests that in this court and at this period individual men and women were differentiated through adjectival modifiers whereas children were made reference to through possessive relationships (*my child*, *the child's father*). However, both legal speakers and female non-legal speakers use more adjectival modifiers when speaking about women than about men. Therefore, we now query the CDC to discover the most salient adjectives used by each group of speakers when referring to men and women.

For a given subcorpus, target concept and dependency relation, we can build a profile of the concepts which co-occur with the target concept in the given dependency relation in the

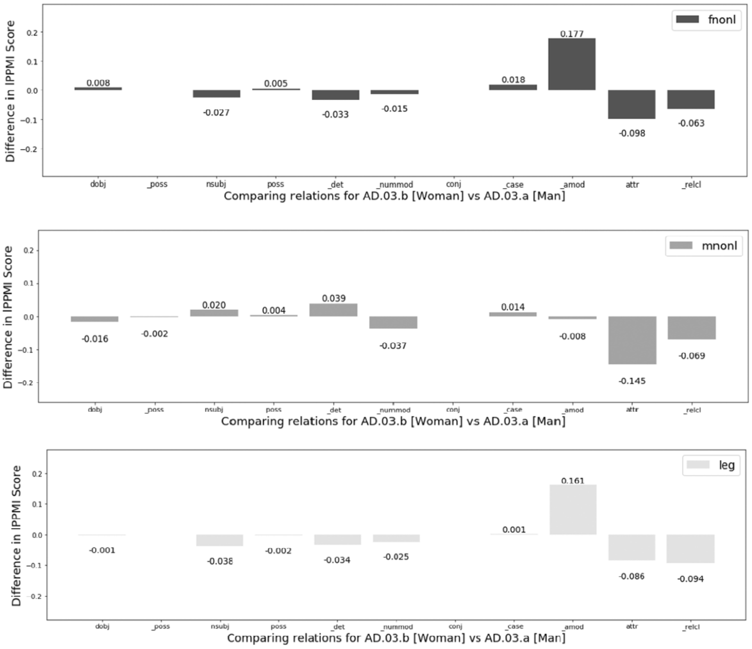


Figure 12. Difference in salience of dependency relations with AD.03.b [Woman] and AD.03.a [Man]

given subcorpus. As before, we use lPPMI to weight the salience of each co-occurring concept or feature. Standard vector methods, such as cosine similarity, can then be used to determine the similarity of different profiles.

We are now in a position to ask questions relating directly to *socioconceptual polysemy*. To answer such questions, we compute inter-subcorpora and intra-subcorpora similarities, as defined below.

(1) *Inter-subcorpora similarity*: How similar are the adjectival modifier profiles for a given concept across the different corpora?
(2) *Intra-subcorpora similarity*: How similar is the adjectival modifier profile for *AD.03.a [Man]* and *AD.03.b [Woman]* in each subcorpus?

Inter-subcorporal cosine similarities between adjectival modifier profiles for given individual concepts are shown in Table 11.

Table 11 shows that the adjectival modifier profiles for the concept AD.03.a [Man] is almost identical across the three subcorpora, with cosine similarity scores close to 1. We see that the adjectival modifier profiles for the concept AD.03.b [Woman] are less similar across the three subcorpora. In particular, there are clearly differences in the adjectives used by male and female non-legal speakers when talking about women. The profile of the male legal speakers appears to be a blend of the female and male non-legal speakers, presumably using adjectives used by both groups of non-legal speakers. Finally, we see that the adjectival modifier profile for AD.03.d [Child] in the legal subcorpus is very different to that of both groups of non-legal speaker.

Intra-subcorporal cosine similarities between adjectival modifier profiles for given concept pairs are shown in Table 12.

From the results presented in Table 12, it is apparent that the male non-legal speakers and male legal speakers use largely the same adjectives when talking about AD.03.a [Man] and AD.03.b [Woman] (as signified by cosine similarity scores >0.9). The female non-legal speakers, on the other hand, have moderately similar adjectival modifier profiles for AD.03.a [Man] and AD.03.b [Woman], which means these profiles overlap but not to the same extent as for male non-legal speakers and male legal speakers.

As well as calculating the similarity of the profiles, we can establish the key features which contribute to both their similarity and their dissimilarity. Figure 13 shows the most salient features in the intersection and difference for the adjectival modifier profiles for AD.03.b [Woman] and AD.03.a [Man] in each of the three subcorpora in turn: female non-legal, male non-legal and legal.

Table 11. Inter-subcorpora similarities for individual concepts

	fnonl	mnonl	leg
AD.03.a [Man]			
fnonl	1	0.98	0.94
mnonl	0.98	1	0.95
leg	0.94	0.95	1
AD.03.b [Woman]			
fnonl	1	0.67	0.78
mnonl	0.67	1	0.87
leg	0.78	0.87	1
AD.03.d [Child]			
fnonl	1	0.80	0.06
mnonl	0.80	1	0.12
leg	0.06	0.12	1

Table 12. Intra-subcorpora similarities for pairs of concepts

	AD.03.a [MAN]	AD.03.b [WOMAN]	AD.03.d [CHILD]
fnonl			
AD.03.a [MAN]	1	0.71	0.09
AD.03.b [WOMAN]	0.71	1	0.08
AD.03.d [CHILD]	0.09	0.08	1
mnonl			
AD.03.a [MAN]	1	0.95	0.05
AD.03.b [WOMAN]	0.95	1	0.06
AD.03.d [CHILD]	0.05	0.06	1
leg			
AD.03.a [MAN]	1	0.90	0.51
AD.03.b [WOMAN]	0.90	1	0.55
AD.03.d [CHILD]	0.51	0.55	1

Figure 13 shows that both male and female non-legal speakers and legal speakers use the concept AF.05 [Age/cycle] with a very high saliency when referring to both men and women. This concept includes adjectives such as *young* and *old* and so it is apparent that all groups of speakers use age to differentiate men and women. For example, statements from female and male witnesses and defendants include, '*I was out of work at the time and the young man too.*' and '*I bought these taps of a young man*'; legal questioning includes '*This young man at the bar drove for his father?*'. Looking at the differences between the profiles, we see that female non-legal speakers are more likely than male non-legal speakers to use AP.02.b [Individual character] and AW.04.a [Poverty] when talking about women. These concepts include adjectives such as *single* and *poor*. For example, we see statements from female defendants and witnesses such as '*I am a single woman.*' '*I live in Church Lane.*', '*She is a married woman, therefore it is not likely*' and '*I am a poor unfortunate woman*'. Male non-legal speakers are more likely than female speakers to use AP.01.f [Difference] when talking about both men and women. This concept includes adjectives such as *different* and *other*. For example, statements from male defendants and witnesses include '*The boy told me to follow the other man, he wanted to crimp me as did the other two men* and *the other woman was with her then*'. The legal speakers use similar adjectives to the female speakers when speaking about women, particularly AO.19.a.1 [Calamity/misfortune] and AY.01.g [Marriage]. For example, we see many statements referring to *this poor woman* and *that unfortunate woman* as well as questions asking '*Are you a married woman?*'. When speaking about men, the legal speakers use the same adjectives as the male non-legal speaker, particularly AP.01.d [Identity] and AP.01.f [Difference]. For example, there are questions such as '*Did you see any other man about the premises?*' and '*Look at the prisoner, is that the same man?*'

7. Conclusions

The current research offers a better understanding of concepts and conceptual variation. We have explored the ways in which different speakers use different concepts; and the ways in which different speakers use the same concepts differently. Motivated by previous work in the area (section 3), we experimented with methodological innovations in cognitive and computational semantics and developed a framework to address claims regarding conceptual variation empirically.

One methodological advantage promoted by the current work is that we facilitate the application of semantic distributional analysis to small corpora. Moving from a traditional, word-based corpus to a CDC allows for infrequent words which share a meaning to cluster

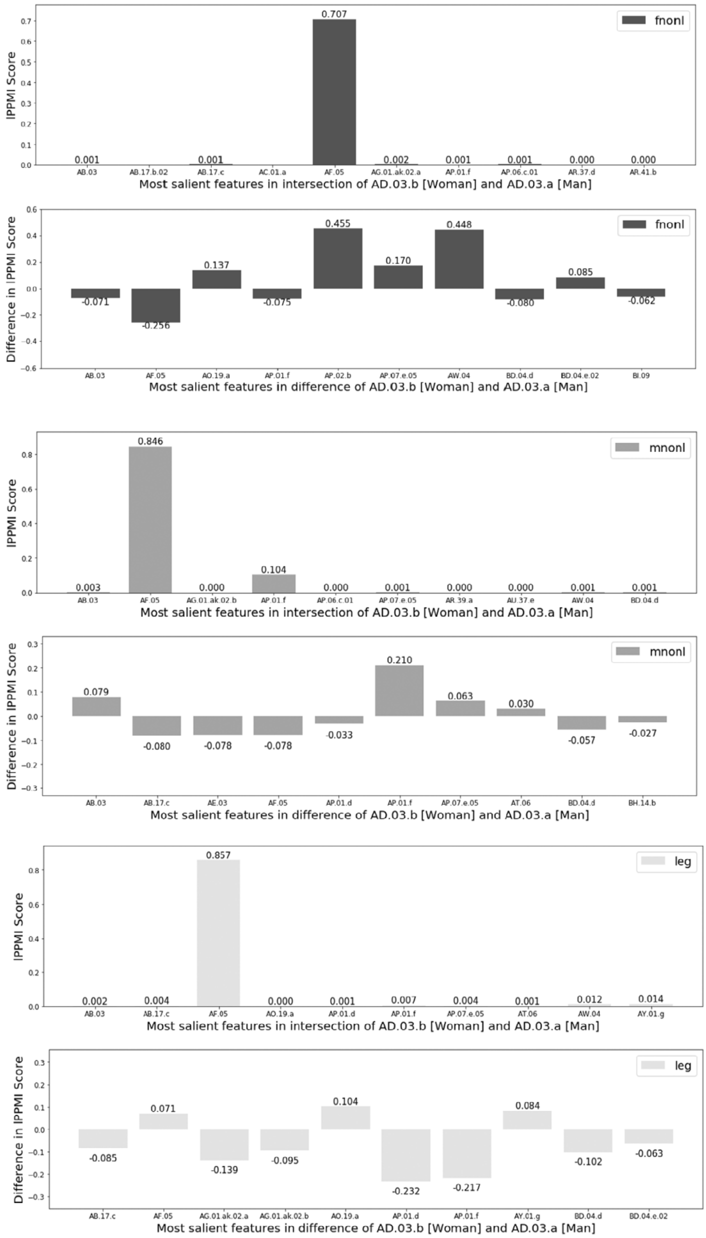


Figure 13. Comparing adjectival modifier profiles for AD.03.b [Woman] and AD.03.a [Man] in each subcorpus

under the umbrella of a concept. This concept is more frequent, which enables the use of more powerful computational techniques.

The move from word-based analysis to concept-based analysis has theoretical as well as computational implications. By querying the nineteenth-century OBV dataset presented as a conceptual dependency corpus we discover underlying macro-patterns of thinking operating across different communities of speakers, which are seen not just through individual words, but whole concepts. The most characteristic concepts in female speech in the data are related to the household, clothes and people's roles in the house/family. The most characteristic concepts in male speech are related to out-of-house roles and activities, referring to watchmen, officers, places in London and using verbs of control, creation and keeping guard. These findings are unsurprising since courtroom narratives are grounded in and reflect real-world situations in which speakers are engaged. Speakers' different, shared experiences of the world have given these communities distinctive ways of viewing and categorising the world, which are then reflected in their speech. With this in mind, the current research provides a contribution to cognitive sociolinguistics research (Pütz et al. 2012, 2014).

Another key contribution of the current study is that, we have extended the distributional hypothesis (Harris 1954; Firth 1957) to concepts and have shown that concepts are characterised by other co-occurring concepts and that same concepts can form different dependency relations with other concepts. The usage data indicates the differentiation of MAN and WOMAN by qualifying them with concepts of AGE/CYCLE, for example *young, old*; CALAMITY/MISFORTUNE, for example *unfortunate*, *unlucky*; MARRIAGE, for example *married*; IDENTITY AND DIFFERENCE, for example *same/different*; or DRUNKENNESS for example *drunken*. The concept of CHILD is differentiated in terms of their relationship to other people, which is seen by the co-occurrence with the concept of OWNING as in *his child*, *child's mother*.

Our analysis (section 6.2) also provides strong evidence for the existence of socioconceptual polysemy, that is different speakers using the same concept in different ways. Concepts are made of a cluster of uses which are not only situationally and contextually flexible (cf. discursive concept, Fitzmaurice et al. 2017c) but also differ from person to person. In modelling gendered language use we have shown that the concept of WOMAN means different things to legal men, non-legal men and women. These cross-sectional differences emerge both through the dependency profile as well as the adjectival modifier profile of WOMAN (section 6.2). For each community of speakers the concept WOMAN is composed of various characteristic uses. Some of these uses are overlapping across the speech of legal men, non-legal men, and women, and others are characteristic to a single community of speakers. For example, in section 6.2, we observe that adjectival modification of the concept WOMAN by concepts related to 'being married' or not and 'to being poor' or not, is particularly characteristic of female non-legal speech and, to a lesser extent, male legal speech. We observe that the same concept develops different meaning components for different speakers, which provides new evidence for established claims as to the perspectival and experiential nature of meaning (Geeraerts 2010). The differences in meaning components can be explained by reference to the given sociohistorical reality in which individual conceptualisations develop. Further investigation should determine the internal structure of the conceptual polysemy, for example, from the perspective of cognitive assumptions of category structure (cf. Geeraerts 1997). Mapping concept co-occurrence patterns in semantic networks could offer relevant investigative opportunities if we assume a semantic network can be a given representation of a concept.

From a methodological point of view, including a grammatical component in the semantic analysis has been instrumental in identifying socioconceptual polysemy. For example, female witnesses and defendants are far more likely to use adjectival modifiers when talking about a woman than a man. In particular, they are far more likely to comment on the marital or wealth status of a woman, thus demonstrating that they conceptualise the concept of WOMAN differently (Figure 13, Tables 11 and 12).

The analysis of concepts in Old Bailey recorded speech gives us a reasonably realistic window into the life and events of nineteenth-century London life. The preoccupation with a woman's marital status in the OBV is not surprising given the laws at the time which assigned different rights to women according to their marital status (Shanley 1993). However, being able to clearly see this historical fact from the computational analysis of the language serves to indicate a large potential of our methodology in approaching history in a data-driven or bottom-up fashion. Our analysis also demonstrates that this preoccupation with a woman's status existed primarily in the minds of female and legal speakers rather than in the minds of non-legal male speakers. Therefore, from the standpoint of socioconceptual polysemy, it appears that whilst females and legal speakers have multiple senses of the concept WOMAN, non-legal male speakers do not. Female and legal speakers clearly differentiate married and unmarried women, suggesting that for these communities of speakers there are (at least) two meaning components (MARRIED WOMAN and UNMARRIED WOMAN), which are part of the more generic concept of WOMAN. Further research in the area would involve identifying parameters of the inter-concept relationships.

The computational linguistic analysis presented in this paper also contributes to the debate over whether the transcripts of the Old Bailey are true transcripts of the courtroom speech or summaries made by the courtroom reporter (Huber 2007). In particular, the analysis of adjectives used by women evidences the existence of different voices in the courtroom and these differences are systematic and statistically significant (Section 6). It would be possible to conjecture that the court reporter recorded speech according to his own preconceptions about men and women. However, modifying conceptual content by the reporter would be far more cumbersome than modifying other features of language. A more verbatim recording is a much simpler explanation when faced with linguistic evidence for female characters and male characters using much the same set of modifiers but with different probability distributions.

Meaningful findings from the current analysis of the Old Bailey support the empirical approach to concept modelling taken in this paper. The decision to model concept usage from semantic-grammatical distributions of thesaurus-based concepts and from sociodemographic information reveal many possible ways of how ideas and information permeate linguistic tools available to speakers.

The current research provides a proof of concept for further investigative opportunities which are addressed in our forthcoming work. One such avenue is tracking the evolution of the concepts discussed in the current paper over time and testing further sociolinguistic hypotheses regarding language variation and change. In our future work we also wish to use semantic network approach to explore a cognitive representation of a concept. We also aim to use the hierarchical nature of the tagging to establish which conceptual categories are most subject to polysemy.

Acknowledgements

We thank Fraser Dallachy (Glasgow University) for providing the semantically annotated versions of the subcorpora used in this paper and Tim Hitchcock (University of Sussex) for many helpful discussions. We also thank the audiences of the workshops (*Making Effective Use of Metadata of Historical Texts and Corpora*, Saarbrucken 2017; *Corpora and Discourse*, Lancaster 2018; and *20th International Conference on English Historical Linguistics*, Edinburgh 2018) at which earlier versions of this paper were presented for their enthusiasm and comments. Last, but not least, we are grateful to two anonymous reviewers for all the insightful comments that greatly improved the current paper.

Correspondence
Justyna A. Robinson
University of Sussex
Email: Justyna.Robinson@sussex.ac.uk

References

Allan, Kathryn, 2009. *Metaphor and Metonymy: A Diachronic Approach*. Oxford: Blackwell.

Allan, Kathryn & Justyna A. Robinson (eds.), 2011. *Current Methods in Historical Semantics*. Berlin/Boston, MA: de Gruyter Mouton.

Banerjee, Satanjeev & Ted Pedersen, 2003. 'Extended gloss overlap as a measure of semantic relatedness', in *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. San Francisco: Morgan Kaufmann Publishers Inc. 805–810. https://dl.acm.org/citation.cfm?id = 1630775

Bondi, Marina & Mike Scott, 2010. *Keyness in Texts* (Studies in Corpus Linguistics 41). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.41

Bunnin, Nicholas & Jiyuan Yu, 2004. *The Blackwell Dictionary of Western Philosophy*. Oxford: Blackwell.

Degaetano-Ortlieb, Stefania, Säily Tanja & Bizzoni Yuri, 2021. 'Registerial adaptation vs. innovation across situational contexts: 18th century women in transition', *Frontiers in Artificial Intelligence* 4. https://doi.org/10.3389/frai.2021.609970

Degaetano-Ortlieb, Stefanie, 2018. 'Stylistic variation over 200 years of court proceedings according to gender and social class', in Julian Brooke, Lucie Flekova, Moshe Koppel, & Thamar Solorio (eds.), *Proceedings of the Second Workshop on Stylistic Variation (NAACL)*. New Orleans: Association for Computational Linguistics. 1–10. http://aclweb.org/anthology/W18-1600

Diller, Hans-Juergen, 2014. *Words for Feelings: Essays towards a History of the English Emotion Lexicon*. Heidelberg: Winter.

Dunning, Ted, 1993. 'Accurate methods for the statistics of surprise and coincidence', *Computational Linguistics* 19 (1). 61–74.

Evert, Stefan, 2008. 'Corpora and collocations', in Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, Article 58. Berlin: Mouton de Gruyter. 1212–1248.

Firth, John R., 1957. 'A synopsis of linguistic theory 1930–1955', in *Studies in Linguistic Analysis*. Oxford: Philological Society. 1–32.

Fitzmaurice, Susan, 2017. 'When natives became Africans: A historical sociolinguistic study of semantic change in colonial discourse', *Journal of Historical Sociolinguistics* 3(1). 1–36.

Fitzmaurice, Susan, 2022. 'The historical pragmatic construction of co-occurrence clusters as discursive concepts; Evidence from EEBO-TCP', *Transactions of the Philological Society*, 120.

Fitzmaurice, Susan, Justyna A. Robinson, Marc Alexander, Iona C. Hine, Seth Mehl & Fraser Dallachy, 2017a. 'Reading into the past: Materials and methods in historical semantics research', in Tanja Säily, Arja Nurmi, Minna Palander-Collin, & Anita Auer (eds.), *Exploring Future Paths for Historical Sociolinguistics* (Advances in Historical Sociolinguistics 7). Amsterdam: John Benjamins Publishing. 53–82. https://doi.org/10.1075/ahs.7.03fit

Fitzmaurice, Susan, Justyna A. Robinson, Marc Alexander, Iona C. Hine, Seth Mehl & Fraser Dallachy, 2017b. 'Linguistic DNA: Investigating conceptual change in early modern English discourse', *Studia Neophilologica* 89(sup1). 21–38.

Fitzmaurice, Susan, Justyna A. Robinson, Iona C. Hine, Fraser Dallachy, Kathryn Rogers, Marc Alexander, Martin Pidd, Seth Mehl, Matthew Groves & Brian Aitken, 2017c. *The Seven Words of the Virgin: Identifying change in the discourse context of the concept of virginity in Early Modern English*, 225–227. Available at https://dh2017.adho.org/abstracts/016/016.pdf

Geeraerts, Dirk, 1997. *Diachronic Prototype Semantics: A Contribution to Historical Lexicology*. Oxford: Clarendon Press.

Geeraerts, Dirk, 2010. *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Geeraerts, Dirk, Caroline Gevaert & Dirk Speelman, 2012. 'How *anger* rose: Hypothesis testing in diachronic semantics', in Kathryn Allan & Justyna A. Robinson (eds.), *Current Methods in Historical Semantics*. Berlin: Walter de Gruyter. 109–132.

Geeraerts, Dirk, Stefan Grondelaers & Peter Bakema, 1994. *The Structure of Lexical Variation: Meaning, Naming, and Context*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110873061

Glynn, Dylan & Justyna A. Robinson (eds.), 2014. *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy* (Human Cognitive Processes). Amsterdam: Benjamins. https://doi.org/10.1075/hcp.43

Goddard, Cliff & Anna Wierzbicka (eds.), 1994. *Semantic and Lexical Universals: Theory and Empirical Findings*. Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/slcs.25

Harris, Zelig, 1954. 'Distributional structure', *Word* 10(23). 146–162.

Hitchcock, Tim, Robert Shoemaker, Clive Emsley, Sharon Howard, John McLaughlin, et al., 2012. *The Old Bailey Proceedings Online, 1674–1913* (www.oldbaileyonline.org, version 7.0, 24 March 2012).

Honnibal, Matthew & Mark Johnson, 2015. 'An improved non-monotonic transition system for dependency parsing', in Eric Brill & Kenneth Church (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1373–1378. http://anthology.aclweb.org/W/W96/W96-0200.pdf

Huber, Magnus, 2007. *The Old Bailey Proceedings, 1674–1834: Evaluating and annotating a corpus of 18th and 19th century spoken English*. http://www.helsinki.fi/varieng/series/volumes/01/huber/

Huber, Magnus, Magnus Nissel, Patrick Maiwald & Bianca Widlitzki, 2012. *The Old Bailey Corpus*. Spoken English in the *18th and 19th centuries*. www.uni-giessen.de/oldbaileycorpus

Kay, Christian, Jane Roberts, Michael Samuels, Irene Wotherspoon & Marc Alexander (eds.), 2016. *The Historical Thesaurus of English*, Version 4.2. Glasgow: University of Glasgow. http://historicalthesaurus.arts.gla.ac.uk (last accessed on 10 December 2018).

Klingenstein, Sara, Tim Hitchcock & Simon DeDeo, 2014. 'The civilizing process in London's old bailey', *Proceedings of the National Academy of Sciences of the United States of America* 111(26). 9419–9424.

Kullback, Solomon & Richard A. Leibler, 1951. 'On information and sufficiency', *Annals of Mathematical Statistics* 22(1). 79–86.

Langacker, Robert W., 1987. *Foundations of Cognitive Grammar Vol. 1: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Langacker, Robert W., 1991. *Foundations of Cognitive Grammar Vol. 2: Descriptive Application*. Stanford, CA: Stanford University Press.

Lenker, Ursula, 2007. *Soþlice, forsooth, truly* – Communicative principles and invited inferences in the history of truth-intensifying adverbs in English. Susan Fitzmaurice and Irma Taavitsainen (eds.), 81–106.

Maccabe, Colin, Holly Yanacek & The Keywords Project (eds.), 2018. *Keywords for Today: A 21st Century Vocabulary*. Oxford: Oxford University Press.

Mehl, Seth, 2022. 'Discursive Quads: New kinds of lexical co-occurrence data with linguistic concept modelling', *Transactions of the Philological Society*, 120(3).

Mikolov, Tomas, Irma Sutskever, Kai Chen, Greg S. Corrado & Jeffrey Dean, 2013. 'Distributed representations of words and phrases and their compositionality', *Proceedings of Neural Information Processing Systems* 26. 3111–3119.

Pennington, Jeffrey, Richard Socher & Christopher Manning, 2014. 'Glove: Global vectors for word representation', in *Proceedings of EMNLP*. Doha, Qatar: Association for Computational Linguistics. 1532–1543. https://www.aclweb.org/anthology/D14-1162

Piao, Scott, Fraser Dallachy, Alistair Baron, Jane Demmen, Steve Wattam, Philip Durkin, James McCracken, Paul Rayson & Marc Alexander, 2017. 'A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation', *Computer Speech & Language* 46. 113–135.

Pilehvar, Mohammad T. & Roberto Navigli, 2014. 'A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation', *Computational Linguistics* 40(4). 837–881.

Pütz, Martin, Justyna A. Robinson & Monika Reif (eds.), 2012. Cognitive sociolinguistics: Social and cultural variation in cognition and language use. *Special Issue of Annual Review of Cognitive Linguistics* 10.

Pütz, Martin, Justyna A. Robinson & Monika Reif (eds.), 2014. *Cognitive Sociolinguistics. Social and Cultural Variation in Cognition and Language Use* (Benjamins Current Topics 59). Amsterdam/Philadelphia, PA: John Benjamins. https://doi.org/10.1075/bct.59

Raganato, Alessandro, Jose Camacho-Collados & Roberto Navigli, 2015. 'Word sense disambiguation: A unified evaluation framework and empirical comparison', in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia: The Association for Computational Linguistics. 99–110. http://aclweb.org/anthology/E17-1000

Rayson, Paul, 2008. 'From key words to key semantic domains', *International Journal of Corpus Linguistics* 13(4). 519–549.

Rayson, Paul & Roger Garside, 2000. 'Comparing corpora using frequency profiling', in *WCC '00: Proceedings of the Workshop on Comparing Corpora*, Volume 9. Hong Kong: Annual Meeting of the Association for Computational Linguistics (ACL 2000). 1–6.

Robinson, Justyna A., 2010. '*Awesome* insights into semantic variation', in Dirk Geeraerts, Gitte Kristiansen, & Yves Piersman (eds.), *Advances in Cognitive Sociolinguistics*. Berlin/New York: de Gruyter Mouton. 85–109. https://doi.org/10.1515/9783110226461.85

Robinson, Justyna A., 2012. 'A *gay* paper: Why should sociolinguistics bother with semantics?', *English Today* 28(4). 38–54.

Robinson, Justyna A., 2011. 'A sociolinguistic approach to semantic change', in Kathryn Allan & Justyna A. Robinson (eds.), *Current Methods in Historical Semantics*. de Gruyter Mouton: Berlin/Boston, MA. 199–230.

Robinson, Justyna A., 2014. 'Quantifying polysemy in cognitive sociolinguistics', in Dylan Glynn & Justyna A. Robinson (eds.), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*. Amsterdam: Benjamins. 87–115. https://doi.org/10.1075/hcp.43.04rob

Sahlgren, Marcus, 2006. *The Word-Space Model*. Ph.D. dissertation, Stockholm University.

Sandow, Rhys & Justyna A. Robinson, 2018. ''Doing Cornishness' in the English periphery: Embodying ideology through Anglo-Cornish dialect lexis', in Natalie Braber & Sandra Jensen (eds.), *Sociolinguistics in the UK*. Basingstoke: Palgrave. 333–361. https://doi.org/10.1057/978-1-137-56288-3_13

Shaffer, Juliet P., 1995. 'Multiple hypothesis testing', *Annual Review of Psychology* 46(1). 561–584.

Shanley, Mary L., 1993. *Feminism, Marriage and the Law in Victorian England*. Princeton, NJ: Princeton University Press.

Tian, Kevin, Teng Zhang & James Zou, 2018. 'CoVeR: Learning covariate-specific vector representations with tensor decompositions', in *International Conference on Machine Learning*, Jul 3, Stockholm, Sweden. 4926–4935. https://proceedings.mlr.press/v80/

Tucker, Susie, 1972. *Enthusiasm: A Study in Semantic Change*. Cambridge: Cambridge University Press.

Weeds, Julie, 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph.D. dissertation, University of Sussex.

Weir, David J., Julie Weeds, Jeremy Reffin & Thomas Kober, 2016. 'Aligning packed dependency trees: A theory of composition for distributional semantics', *Computational Linguistics* 42(4). 727–761.

Wierzbicka, Anna, 1992. *Semantics, Culture, and Cognition: Universal Human Concepts in Culture-Specific Configurations*. Oxford: Oxford University Press.

Williams, Raymond, 1983. *Keywords: A Vocabulary of Culture and Society* (2nd ed.). Oxford: Oxford University Press.