Meaning in Poetry

Semantic Annotation of Verse with the *Historical Thesaurus of English*

Elina Koristashevskaya

BA (Hons), MRes

Submitted in fulfilment of the requirements of the degree of Doctor of Philosophy

School of Critical Studies

College of Arts

University of Glasgow

April 2021

Abstract

This thesis addresses the current gap in semantic annotation of poetry by presenting the first semantic tagging system specifically designed to disambiguate senses in a diachronic corpus of poetry. The 'ambiguity tagger' developed for this purpose utilises the hierarchical taxonomy of the *Historical Thesaurus of English* (HTE; Kay 2011: 42) to assign conceptual 'tags' to lexical items that denote the meaning of the word in context, with multiple meanings assigned to ambiguous words. The ambiguity tagger encompasses a configurable pipeline for semantic annotation, thus presenting a more flexible alternative to existing applications (Piao et al. 2005a; Rayson 2009a; Piao et al. 2017). To train the tagger, a corpus was curated from the Oxford Book of English Verse, containing poetry from the early 16th to the late 19th century (OBEV; Quiller-Couch 1919/1999).

As the ambiguity tagger allows multiple meanings to be assigned to individual words in the corpus, without restricting the maximum number of senses, the semantic metadata produced by the tagger is unique in its breadth. Correspondingly, the analysis sections of the thesis look at different techniques for interpreting the data, using case studies from the OBEV corpus. Both macro- and micro-level approaches to analysing the data are explored, highlighting the benefits of the ambiguity tagger at different levels of critical analysis. To further explore the capabilities of semantic annotation with HTE data, this research extends the interpretative analysis of the semantic metadata gained through the ambiguity tagger by presenting a systematic approach for analysing the significant co-occurrence of concepts in the text. This process borrows the framework for identifying significantly co-occurring words (collocates) and extends this into a measure of 'semantic collocation', thus significantly expanding on existing research in this field (Alexander et al. 2015a; Archer & Malory 2015; 2017). By shifting the focus from lexical collocation to the significant co-occurrence of 'meaning' in texts, this approach reveals a pattern of previously inaccessible textual data for analysis and marks a further methodological contribution of this research.

Contents

Tables

Figures

Poems

Acknowledgements

I would like to extend my sincere gratitude to my supervisors, Marc Alexander, Wendy Anderson, and Fraser Dallachy, who have tirelessly supported this research for so many years. I will be forever grateful for their guidance and valuable insight during the prolonged development of this project; this work would not be possible without them.

I owe a further debt of gratitude to my examiners, Dan McIntyre and Catherine Emmott, for their invaluable advice, feedback, and all the time that they have graciously devoted to this research. This thesis would not be complete without their constructive and extensive commentary, to which I can only hope to have done justice.

I will also be forever grateful to Christian Kay, for her work as Director of the *Historical Thesaurus of English*, and for granting me the scholarship that made my research possible. I was incredibly fortunate to have known Christian, and though her work has had a truly significant and tangible impact on my life, the impact of the informal chats we had at conference dinners is immeasurable.

In recognition of the varied help and guidance that I have received during my time at Glasgow, I would like to thank the staff in the School of Critical Studies and the College of Arts. In particular, I would like to acknowledge Alison Bennett, Brian Aitken, Kiran Faisal, Louise Pollock, and Claire Smith, all of whom have in some form or another helped me reach the finish line.

For kindling my passion for language and inspiring me to pursue further research, I would like to thank Michael Rodgers, Elspeth Jajdelska, Heather Froehlich, Nigel Fabb, Richard Niland, and Patrick Heart at the University of Strathclyde. My further thanks to all who inspired this research, whether directly or indirectly, over the many years I have been at Glasgow. Particularly, I learned so much from conversations with Dawn Archer, Jean Anderson, Andrew Prescott, John Corbett, Irené Wotherspoon, and Jennifer Smith. To Jeremy Smith I owe an even greater debt, for supervising my master's thesis, for encouraging me to pursue my PhD, for

being a supportive Chair during my viva, and for being enthusiastic about my research when I needed it the most.

I wish to thank my colleagues in the Learning Enhancement & Academic Development Service (LEADS), for their encouragement, stimulating discussions, friendship, and patience over the many years we have worked together. In particular, I am grateful to Andrew Struan, for being a friend, mentor, and more understanding as a line manager than I ever deserved; this thesis would not be complete without your help, and I expect that no one in my life will be more pleased to see it submitted. My further thanks to Jennifer Boyle, Scott Ramsay, and Micky Ross, who gave their time to review my draft and helped me prepare for my viva. A further thanks to Amanda Sykes, for giving me so many opportunities as a Research Assistant and for being a friend. And of course, my thanks to Jessica Bownes, for being the best office mate I could have imagined.

Finally, my personal thanks to my friends, who have suffered years of deadlines along with me. I would not have survived without them. In particular, my thanks to Stuart, for understanding everything, and to Levi, Ross, Duncan, and Quinny, for somehow managing to always take an interest in what I am doing. A special thank you to David, for always believing that I would finish. And to my family, who above all made this possible. To Grandma, who I know would be proud of me, to my Baba, for giving me the ambition to get here, to Aunt Fiona, Dad, Mosh, and to my Mother, who made me this way, I give my eternal love and gratitude.

Abbreviations and Conventions

Abbreviations

Frequently used abbreviations are listed below:

CL – Corpus linguistics

CLAWS – Constituent Likelihood Automatic Word-tagging System

HTE – *Historical Thesaurus of English*

HTST – Historical Thesaurus Semantic Tagger

LL – Log-likelihood

MWE – Multi-word-expression

OBEV – *Oxford Book of English Verse*

OED – *Oxford English Dictionary*

POS – Part-of-speech

SAMUELS – Semantic Annotation and Mark-Up for Enhancing Lexical Searches

SOI – *Songs of Innocence*

SOE – *Songs of Experience*

UCREL – University Centre for Computer Corpus Research on Language (Lancaster University)

USAS – UCREL Semantic Analysis System

VARD – Variant Detector

Research-specific abbreviations:

MV – Match value

RMV – Relative match value

SR – Semantic relevance

AV - Assigned value

EAV - Expected assigned value

Style conventions

Typographical symbols are used to distinguish between codes used in this thesis:

The Historical Thesaurus hierarchy is indicated by the category number and the description in brackets: e.g. 01 (The world), 02.04.22.12 (Encouragement). Where referencing words from a specific category, the category number is provided in brackets: e.g. *movement* (01.02.03.03.08), *biological processes* (01.02.03.03).

Numeric identifiers for category headings (CATID) are prefaced with a right-pointing guillemet (») and where necessary list the description in brackets, e.g. »1 (The world), »133417 (Encouragement).

The thematic headings hierarchy is presented in italics, with the description in brackets: *AA* (The world), *Au.47.d* (Encouragement)

Numeric identifiers for thematic headings (TID) are prefaced with an obelisk (†) and where necessary list the description in brackets, e.g. †1 (The world), †2139 (Encouragement).

The identifiers used to preserve the word order of the texts in the corpus are distinguished with a hash prefix, e.g. #1115, #1820.

Poems are numbered using lower case roman numerals: e.g. Poem i.

Where relevant, the OBEV poem number is given in square brackets before the title of the poem: e.g. [598] ‘For Music’

References to database fields are distinguished from the text with a change to a monospaced typeface, e.g. `corp_id`. The same font is used for inline code and any queries recorded in the appendix. References to database tables follow the same convention, e.g. `category`, `lexeme`.

Words and lemma forms are put in italics.

Cross-references to different sections in the thesis are preceded by the section symbol, e.g. §2.3.1.

Chapter 1 Introduction

'The language looks rather different when you look at a lot of it at once'.
(Sinclair 1991: 100)

1.1 Introduction

In the decades following Sinclair's statement, linguists continued expanding the discipline of investigating language at a macro or corpus level. Technological advances made computer-assisted research more feasible, and the ubiquity of digitised (and digital) text made the analysis of written language possible on a new scale. We now have access to corpora of hundreds of millions of words, a volume that far surpasses the expectations of early corpus researchers (Leech 1991: 10), and tools that allow users with no prior knowledge of corpus linguistics to obtain basic metrics (frequency lists, concordances) and carry out their own analysis (keywords, collocation) (Evison 2010; Taylor 2013). The value of these resources, however, is determined by their capacity to produce meaningful results, and this requires a robust system for interpreting this data (Leech 1991: 13).

One area where this research activity has been particularly visible is in the development of tools for semantic annotation of corpora (Kübler & Zinsmeister 2015: 83). The ability to systematically identify key themes in text is of obvious advantage to anyone working with large corpora, particularly when information retrieval is a priority (Kübler & Zinsmeister 2015: 83). The combination of disciplinary differences, varied research aims and motivations, and subject-specific training corpora guided the design and behaviour of these tools, leading to a proliferation of resources that behave differently but work towards a similar goal (Palmer 1990: 18). The specialised nature of many semantic annotation projects means that methodological success often varies between text types, with period and literary texts often presenting the biggest challenge to automated semantic analysis (Alexander et al. 2015b: 125–126). While recent developments have begun to bridge the gap in semantic annotation of non-standard text types, no approach has been designed to cope specifically with the automated semantic analysis of period literature (Archer et al. 2003; 2006; Alexander et al. 2015b; Piao et al.

2017). Indeed, this gap becomes more pronounced when looking at period poetry, which remains an underexplored genre within corpus linguistics and its subdisciplines. To address this research gap, the first part of this thesis presents a new method for semantic annotation of corpora, which allows for a flexible approach to identifying meaning in text.

1.2 Extending the scope of semantic annotation

In discussing the development of a new approach to semantic annotation, it is useful to acknowledge alternative methods of investigating meaning in corpora. This includes targeted application of general corpus methods, such as the use of collocation analysis to show the impact of co-occurrence on the meaning of collocating words (Evert 2009; McEnery & Hardie 2012),[1] or keyness analysis to classify texts through the significant appearance of certain words (Culpeper 2009).[2] Outside of corpus linguistics, computational approaches for investigating meaning in large textual data include automatic classification of lexical items based on a template of syntactic parameters (Palmer 1990),[3] and algorithmic approaches for grouping words into clusters (Jurafsky & Martin 2000) or topics (Blei 2012). While these methods allow users to explore corpora in different ways, they are either too general or too specialist to address the gap in corpus semantic analysis of non-standard corpora.

Semantic annotation tools and methods allow for enhanced corpus analysis research and give corpus linguists better access to the semantic properties of their textual data. Extending these benefits to a broader range of corpus types and research designs is a necessary step in expanding the scope of corpus semantic enquiry. The semantic annotation method proposed in this thesis represents a significant move towards achieving this goal. Of the five research questions (RQs) guiding this thesis, two initial questions were used to justify the need for a new semantic annotation method and direct the development of this new approach:

> RQ1. What are the barriers to using existing semantic annotation tools and methods in analysing meaning in non-standard corpora?

[1] See §2.3.2.
[2] See §2.3.3.
[3] See §2.4.1.

RQ2. What are the practical design parameters for overcoming these barriers?

Corpus linguistics research currently relies on two related semantic annotation tools: the UCREL Semantic Analysis System (USAS; Piao et al. 2005a), and the Historical Thesaurus Semantic Tagger (HTST; Alexander et al. 2015b: i16–i17). Thus, to answer the first research question (RQ1), the capabilities of both taggers were reviewed in relation to existing research and the requirements of non-standard, historical, and literary corpora (§2.4).

While most computational approaches for extracting meaning from texts require specialist knowledge,[4] the USAS (§2.4.3) and HTST (§2.4.5) systems made semantic annotation accessible to a broader audience. USAS excels in attributing general semantic labels in text, but it is primarily used to annotate contemporary texts as its semantic lexicon does not reflect historical meanings (Archer et al. 2003; Piao et al. 2004). Furthermore, as the USAS semantic lexicon represents 'a conception of the world that is as general as possible' (Piao et al. 2005a: 2), the tagger is less suited to annotating specialised texts, highlighting subtler semantic distinctions, and identifying precise meanings for further analysis. In contrast, the HTST attributes senses from its fine-grained historical semantic taxonomy, and allows users to filter available meanings by date, thus restricting the semantic output to period-appropriate senses (Alexander et al. 2015b: i21). This filter corresponds to the data contained in the *Historical Thesaurus of English* (HTE; Kay 2011: 42), where meanings are organised based on their first and last recorded usage, helping to constrain the search parameters of the expansive HTE taxonomy. Consequently, the HTST is better suited to annotating historical corpora, making it a useful complementary tool to the USAS tagger. In testing, however, the HTST tagger performed less effectively on non-standard texts (Alexander et al. 2015b: i20), while further technical limitations of the current version of the tagger restrict its use in corpus analysis research (Piao et al. 2017).

Two further key limitations are discussed in §2.4.6: the lack of configuration options in both disambiguation methods restrict the analysis of figurative language (Koller et al. 2008), and the decontextualised annotation process prevents the analysis of meaning based on different 'contextual scopes' in the corpus (Scott &

[4] See §2.4 & §2.4.1.

Tribble 2006: 9). The first limitation relates to the semantic metadata produced by both taggers; while both taggers demonstrate the ability to identify metaphorical expressions by assigning both the source and target domains to words in corpora (e.g. Koller et al. (2008) for USAS; Piao et al. (2017) for HTST), their performance in this regard is inconsistent, as the target domain is not always identified in the tagged output. Part of this issue relates to the USAS lexicon, which was not designed to annotate figurative language (Alexander & Bramwell 2014), and includes only popular metaphors as potential senses of lexical items. While the HTE taxonomy used by the HTST includes all senses recorded in the *Oxford English Dictionary* (OED), it still partially relies on a pre-determined tagset for disambiguating highly polysemous words. Furthermore, the HTST was designed to assign the 'correct' HTE category to each word (Alexander et al. 2015b: i19), and so the tagger attempts to reduce possible meanings to the most likely candidate meaning. In annotating poetic language, the tagger preferred literal senses, often failing to identify metaphorical usage in context (See §2.5.2, cf. §6.2). The tagger was not tested on poetry by its programmers, so its design does not account for the idiosyncrasies of the genre (Piao et al. 2017: 123).

The challenge of identifying meaning in context speaks to the second limitation of both taggers, as neither method allows users to define contextual elements in the corpora they are annotating. While both taggers rely on a series of disambiguation parameters, the distribution of senses across the corpus are used to prioritise candidate meanings as a proxy of contextual relevance (Piao et al. 2005a; 2017). For diachronic or semantically varied corpora, users must choose between decontextualising their corpus and annotating it as a whole, or annotating individual texts or elements based on the contexts they want to examine. Consequently, users of USAS and the HTST have mostly annotated smaller corpora or individual texts (Archer et al. 2006; McIntyre & Walker 2010; Culpeper 2014a; Alexander et al. 2015b).[5]

To demonstrate the combined impact of these limitations in practice, §2.5 explores the use of semantic annotation in the corpus stylistics research. The emerging field helpfully demonstrates both the demand for semantic annotation, with several studies employing both taggers in stylistic analysis of literary corpora (Archer et al.

[5] See §2.4.6.

2006; McIntyre & Walker 2010; Culpeper 2014a; Alexander et al. 2015b),[6] and the unmet potential of semantic annotation in exploring particular stylistic features across diachronic and varied corpora, as well as the analysis of literary texts. In particular, the difficulty of annotating poetic language is highlighted in §2.5.3, as well as the challenge of annotating a corpus containing multiple poems by different authors and from different periods and genres. Consequently, a diachronic corpus of poetry was used as a reference to for the design and analysis of the new semantic annotation method, as it encapsulates all of the elements that cannot be addressed by USAS and the HTST alone.

To address this gap, this thesis presents the first semantic tagging system specifically designed to identify and return appropriate candidate senses for a complex genre such as poetry. This design reflects the parameters identified in Chapter 3 as the necessary design features for overcoming the barriers to semantic annotation (RQ2). The tagger utilises the hierarchical taxonomy of the *Historical Thesaurus of English* (Kay et al. 2009) to assign conceptual 'tags' to lexical items that denote the meaning of the word in context, with multiple meanings assigned to ambiguous words. To reflect this, the term 'ambiguity tagger' is used as a representative title for the tagger. Conceptual tags are selected from a list of all possible, or 'candidate', meanings recorded in the HTE on the basis of 'semantic relevance', which is calculated in relation to the distribution of meaning across each poem. The full semantic annotation method is described in Chapter 4, while the code for calculating semantic relevance and subsequent disambiguation phases is included in Appendix II below. The annotation process is fully contained within a relational database management system (RDBMS), using MySQL queries to manipulate the data from the initial cross-referencing stage that combines the lemmatised version of the corpus with the HTE taxonomy, to the final 'summary view' that outputs a semantically annotated version of the corpus. In this respect, the tagger further differs from existing systems by presenting a customisable annotation pipeline, which can be configured to meet the needs of individual research designs. Furthermore, as the tagger relies on the richness of the HTE taxonomy to overcome the challenges of investigating meaning in verse, this research also showcases the HTE's capacity for contextual disambiguation without relying on external data and with minimal manual input.

[6] See §2.4.6.

1.3 Extending the scope of semantic analysis of corpora

The tagger designed for this thesis was built to overcome specific limitations of existing semantic annotation tools. In this respect, the primary contribution of this research is methodological, which is reflected in the structure by the expanded methodology chapter (Chapter 4) and the evaluative stance of the analytical case studies in Chapter 5 (Macro analysis) and Chapter 6 (Micro analysis). The semantic tagger presented in this research is unique, but it was developed to extend the scope of semantic analysis of corpora beyond the limits of existing tools. In identifying the extent to which the tagger addresses these gaps, the second part of the thesis examines the application of the tagger in different use-case scenarios. These analyses begin with the well-established point of looking at a lot of language at once, but the work is primarily directed towards understanding how this vantage point can reveal new insights. Accordingly, the evaluation chapters sought to address the following questions:

RQ3. To what extent does the ambiguity tagger address the barriers to semantic annotation of non-standard corpora?
RQ4. How can the tagger facilitate analysis of non-standard corpora using existing corpus analysis methods?
RQ4a. What kind of insights do these analyses provide?

To answer these questions, the analysis was carried out at two different levels of enquiry: a macro-level analysis in Chapter 5, and a micro-level approach in Chapter 6. Chapter 5 reviews the aggregated semantic metadata across the full corpus, while Chapter 6 looks at a smaller sample of tagged data for individual poems in the corpus. By exploring the tagged corpus at different levels, the analysis chapters demonstrated that a macro approach can be used to identify salient semantic features across a corpus and highlight potential research questions, while the micro-level results can aid the stylistic analysis of works by individual authors and have the potential to support comparative analysis. Thus, within the scope of this pilot study, the analysis of the results confirmed that the ambiguity tagger could overcome the key barriers to semantic annotation of non-standard corpora.

While the primary goal of this research was to overcome the restrictions to semantic annotation of corpora, the value of corpus analysis methods should be measured by how they can extend our knowledge beyond existing boundaries, not simply expedite the discovery of what we can already learn by other means (Sinclair 2004: 12). As such, the potential of this new semantic annotation method presented was examined through a final research question:

RQ5. What new research opportunities are opened by the ambiguity tagger?

To further explore the capabilities of semantic annotation with HTE, Chapter 7 extends the interpretative analysis of the semantic metadata gained through the ambiguity tagger by presenting a systematic approach for analysing the significant co-occurrence of concepts in the text. This process borrows the framework for identifying significantly co-occurring words (collocates) and extends this into a measure of 'semantic collocation'. By shifting the focus from lexical collocation to the significant co-occurrence of 'meaning' in texts, this approach reveals a pattern of previously inaccessible textual data for analysis and marks a further methodological contribution of this research.

1.4 Thesis structure

Chapter 2 of this thesis, the Literature Review, begins with a survey of significant developments in corpus research, establishing the foundation for the current state of the art in semantic annotation of corpora. Following this, Chapter 3 establishes the groundwork for the design of the ambiguity tagger, which is discussed in detail in Chapter 4. The semantic metadata is then analysed from a macro perspective in Chapter 5, and a micro perspective in Chapter 6, with both chapters designed to showcase the application of the tagger in a critical reading of poetry. The final analysis undertaken in this thesis is the investigation of semantic collocation in Chapter 7, which expands substantially on the current approaches to investigating meaning in verse. This chapter tests the hypothesis that extending the measure of co-occurrence to the semantic properties of a text can reveal thematic relationships beyond the scope of a lexical analysis. The hypothesis was tested by submitting two sample corpora to a comparative analysis, and then investigating key semantic collocates to explore the significant co-occurrence of themes in the texts in relation to existing criticism. To enable this, Chapter 7 begins by outlining

the methodological framework that made the analysis of semantic collocation possible. All three analysis chapters examine different elements of the ambiguity tagger and explore the strengths and limitations of the tagger in the analysis of verse. The concluding section in Chapter 8 provides a brief summary of these findings, highlights the key contributions of this work, alongside the known limitations and future development, while the supporting Appendices in Appendix I and Appendix II provide the necessary technical background and supporting materials that drive the investigative analysis.

1.5 Conclusion

The semantic annotation method presented in this thesis was developed to address the gaps in the capabilities of the USAS and HTST taggers. Notably, this method was not designed to replace these taggers, but as an addition to the semantic annotation toolkit already available to researchers. While elements of the tagging method are technical in nature, the process is described in detail to allow non-specialist users to replicate the approach and configure the tagger to suit their research parameters. Thus, while the current iteration of the ambiguity tagger is experimental, the release of this tool has the potential to significantly expand the scope of semantic annotation and corpus semantic analysis.

Chapter 2 Literature Review

2.1 Introduction

The primary contribution of this thesis is the design of a flexible approach to semantic tagging, which facilitates annotation of non-standard corpora, and increases the scope of research that looks at the development of meaning in texts. This literature review establishes the context for this research, outlines existing approaches to semantic annotation and the advances made in this research area, and highlights the remaining obstacles for corpus analysis of meaning in texts. While this research has applications in several sub-disciplines of corpus linguistics, its methodological contribution can be most accurately attributed to the area of corpus semantics as it aligns with the research aims of this emerging field.

This chapter surveys the development of corpus semantics as a field of research and addresses key developments in relevant disciplines that inspired contemporary approaches to semantic annotation of corpora. As corpus semantics is an epistemological branch of corpus linguistics (CL), this chapter begins by identifying key developments that advanced corpus-related research into a well-represented discipline (§2.2). The following section (§2.3) extends this discussion by covering significant developments in corpus analysis, with a particular focus on approaches for extracting meaning from corpora.

Section §2.4 introduces existing approaches for semantic annotation of corpora, and the strengths and limitations of different approaches to disambiguating the meanings of words in corpora. This section then discusses the UCREL Semantic Analysis System (USAS; Piao et al. 2005a), and the Historical Thesaurus Semantic Tagger (HTST; Alexander et al. 2015b: i16–i17), as primary examples of current semantic taggers that annotate corpora with disambiguated sense tags. While both systems looked towards identifying single meanings, resolving ambiguity altogether as the ideal, they were instrumental in providing the groundwork for the flexible ambiguity tagger developed for this thesis. This section concludes by discussing the limitations of current approaches to semantic annotation of corpora

and introducing the parameters of a new flexible tagging approach, which could extend the applications of semantic annotation in corpus analysis.

Section §2.5 presents corpus stylistic analyses of literature as a case study for flexible semantic annotation. This section briefly discusses the development of corpus stylistics, and then establishes the restrictions of decontextualised corpus analysis tools as a barrier to investigating stylistic features that belong to different 'contextual scopes' (Scott & Tribble 2006: 9). The chapter then proposes a diachronic corpus of collected poetry as a useful reference corpus for testing the design of the ambiguity tagger, noting the need for preserving contextual parameters and flexible sense disambiguation as beneficial to stylistic and semantic analysis of this type of corpus. Concurrently, the wider implications of a flexible approach to semantic annotation are then discussed, with reference to related disciplines that would benefit from a semantic tagger that can be modified to disambiguate senses based on different contextual scopes.

The literature review concludes with §2.6, which summarises the rationale and considerations of a flexible approach to semantic annotation and underlines the contribution of the ambiguity tagger in research involving corpora. In doing so, this section underlines the need for the semantic annotation approach developed as part of this thesis.

2.2 Corpus linguistics

Corpus linguistics (CL) is a well-established field, though the scope and purpose of CL research has changed in the decades that followed the creation of the first electronic corpus, 'later to be known as the Brown Corpus' (Leech 1997: 1).[7] Early research in corpus linguistics viewed the corpus as a tool for systematically

[7] While the Survey of English Usage, which began in 1959, marked the first modern major corpus linguistics project, there were 'no plans to computerise it until many years later' (Tognini-Bonelli 2010: 15). Development of the Brown Corpus began in 1961, and consisted 'of just over one million words, comprising 500 text samples of about 2,000 words each' when it became available in 1964 (Leech 1997: 1). However, while this milestone marked the beginning of 'computer corpus linguistics' (Leech 1997: 1), analogue corpora, or 'pre-computer corpora' (Biber & Reppen 2015: 2), have been in use for centuries. Early examples of researchers using 'collections of natural texts' to record and analyse language include 'Samuel Johnson's *Dictionary of the English Language*, published in 1755' (Biber & Reppen 2015: 2), and the 'corpus on slips of paper' that served as the 'meaningful body of text' from which the *Oxford English Dictionary* (OED) was compiled in the 1880s (McCarthy & O'Keeffe 2010: 4).

investigating language: the corpus was ‘a sufficiently large body of naturally occurring data of the language’, and it presented a methodological alternative to using ‘intuitive evidence’ in research (Leech 1991: 9). The conclusions drawn from corpus analysis could therefore rely on replicable evidence and methods, adding scientific validity to the researcher’s claims. This focus led to CL briefly falling out of favour in some parts of mainstream linguistics, coinciding with the shift away from using empirical data for linguistic analysis and towards introspective enquiry (Tognini-Bonelli 2001: 49–52). Instead of disappearing altogether, CL research diversified, expanding towards new subdisciplines by synthesising of ‘human processing, computer processing and corpus data’ (Leech 1991: 15–16) to ‘achieve the interaction of data coverage and insight’ that surpassed the expectations of both early corpus linguists and the generative linguists of the 1960s. This diversification persisted even after academic discourse shifted back towards ‘empiricism’ (Kübler & Zinsmeister 2015: 3), resulting in a rich theoretical and methodological discipline that extends well beyond using corpora as a source of empirical evidence.

For the purpose of this research, three observable changes within CL following the release of the Brown Corpus in 1961 (Leech 1997: 1) are of particular importance: the impact of technological advances on CL research, the theoretical developments within CL that expanded the scope of the discipline, and the diversification of CL methodologies to accommodate the growing needs of CL research pursuits. All three developments led to the current state of corpus semantics and the related work in semantic annotation of corpora.

As a result of its divisive history and fragmented developments, the methods employed by CL researchers have similarly undergone phasal changes. While the foundational corpus linguistics techniques of concordancing, collocation analysis, and even basic frequency lists are still employed in corpus analysis, the scope of CL research has broadened with increased access to online tools and digital texts (McEnery & Hardie 2012: 2; Taylor 2018: 22). Scholars took advantage of the ‘growth of the internet and fast download speeds’ (McCarthy & O’Keeffe 2010: 5) to share their research and data, while advances in digitisation replaced ‘clumsy text scanners of the early 1980s’ with ‘access to vast quantities of text already in electronic form’. Where previously corpus size was restricted by the time taken to manually catalogue physical records, and later the prohibitive cost and limitations

of hardware, technological advances have made these challenges effectively obsolete (McCarthy & O'Keeffe 2010: 4–5).

These advances have expanded the scope of CL research, particularly benefitting disciplines that rely on access to large volumes of textual data. McCarthy & O'Keeffe (2010: 6) note that lexicographers took the most advantage of the growing size capacities of corpora, seeking to record as much language use as possible. Extracting information from larger corpora required the development of new tools to allow researchers to analyse and query the data. However, the advances in software for corpus analysis have not been able to match the technological advances in the hardware capable of processing increasingly complex calculations at speeds that far outpace the capabilities of the software (Leech 1991: 11–12). As such, while technological developments and growing volume of data have expanded the scope of CL, the concerns of the field have shifted towards different approaches and methods for analysing this data and are still very much in development.

Of even greater significance to the current research were the epistemological changes within CL over the last few decades, which saw the field growing substantially from being 'limited to simple tasks, such as the discovery of English words classes by clustering words on the basis of their distribution' (Leech 1991: 15), and into a vibrant research field, capable of inspiring 'new theories of language [...] which draw their inspiration from attested language use and the findings drawn from it' (McEnery & Hardie 2012: 1). While widely acknowledged that developing new tools and approaches for analysing data is essential for the betterment of the field, the demand for these resources continues to outpace their development.[8] One particular challenge has been in extending the use of computational methods beyond word-searches, where the researcher would be responsible for identifying search parameters.

As CL methods became more widespread in linguistics, the impact that a research objective has on the analysis became of greater concern. Within discourse studies,

[8] The state of affairs is summarised quite aptly by McEnery & Hardie (2012: 42) with the proviso that 'if the toolset does not expand, then neither will the range of research questions that may be reasonably addressed using a corpus.' It is worth noting that these concerns are not new to the field, with Sinclair (1994: 13) cautioning that 'the change in the availability of information which we now enjoy makes it prudent for us to be less confident re-using accepted techniques'. The change, it would seem, is in how forcefully we are articulating the concern, rather than the issue itself.

for example, researchers turned to corpora to explore existing notions of how 'realities are constructed, represented and transmitted linguistically' (Marchi & Taylor 2018: 1). The blending of CL and discourse studies methodologies allowed researchers to broaden the scope of their analysis, but the enquiry was still guided by the concerns of the source discipline. In the same volume, Marchi & Taylor (2018: 2) encourage discourse linguists to consider corpora as more than a conduit to quantitative research methodologies, explicitly because of this relationship between the corpus and the researcher. To move beyond this dynamic, an alternative approach for setting the computational parameters becomes necessary, where a greater degree of flexibility enables observation and discovery of new patterns of information (McIntyre & Walker 2019). Leech (1991: 16) saw this as a human 'expert' encoding their knowledge into the programs that parse and annotate corpora. A natural extension of this was the desire to develop a way of extracting meaning from corpora without relying on pre-existing knowledge of the texts. These goals inspired much of what is currently practised within corpus semantics, including influencing the advanced semantic annotation of corpora, and the research carried out in this thesis.

There is, as yet, no 'unified body of social theory' for understanding the scope of these methods, and existing research often relies on statistical measures for supporting the significance of their findings (Stubbs 2010: 21). Without a transparent and accessible framework, only specialist users can extend their use of corpus techniques beyond the validation of existing findings. Reliance on existing corpus analysis tools further hinders innovation, as published applications often leave little room for customisation. These restrictions become even more prominent when working with non-standard texts, including period and literary corpora.

2.3 Extracting meaning from corpora

In response to the limitations of existing tools and approaches, much of the recent research activity within corpus linguistics and related fields has been directed towards developing new methods for the computational analysis of texts, and new approaches for interpreting the results of such analyses. The increasing popularity of natural language processing (NLP) techniques in research (Leech 1991: 10;

Clark et al. 2010: 1), the growth and expansion of corpus linguistics (Tognini-Bonelli 2010: 17), and the related developments in Digital Humanities (Arthur & Bode 2014: 4) all contribute towards narrowing the gap between information and interpretation. Research across these disciplines shares in the desire and curiosity expressed in Sinclair's (1991: 100) work: now that we have these new tools, we can see the world in a different way. However, any inquiry into the semantic properties of a corpus must first determine the process for attributing these properties. Multiple factors affect this decision, including availability of resources, the desired breadth of information and the related accuracy of the results, and the purpose for collecting this data. No one method encompasses all research goals, and each approach has different advantages and disadvantages.

A further problem in using computation to derive semantic properties is that the computer cannot *understand* concepts, but instead stores a collection of constructs and attributes those to words based on rules. However, these constructs are determined by the researcher, who will then determine the rules and parameters for how they are assigned (typically resolved by presenting example words within the constructs), thus presenting a circular problem (Palmer 1990: 5). A potential solution to this issue is to 'sidestep' the circular logic of using the computer to identify 'meaning', and instead ask the question of what we need to understand about the meaning of a word to complete the task at hand (Palmer 1990: 5). Indeed, much of the research undertaken within the loosely defined field of corpus semantics has assumed this inquisitive stance.

2.3.1 Corpus semantics

Corpus semantics, as an extension of corpus linguistics, is still an emerging field, and was described by Stubbs (2001: 19–20) as 'an approach to studying language in which observational data from large text collections are used as the main evidence for the uses and meanings of words and phrases.' The implications of this reach beyond academic research, such as allowing 'companies to finesse their business strategies' by better understanding the competition, and helping 'general Internet users to find the information they require more rapidly' by removing the need to search for exact phrases (Alexander et al. 2015b: i17). Of course, these benefits are also present in academic research, where better understanding of 'textual data' helps 'researchers to identify patterns in data sets too large to be

"read" by a human researcher.' (Alexander et al. 2015b: i17). Examples range from the use of semantic annotation in enhancing the quality of systematic literature reviews (Kreimeyer et al. 2017) to employing semantic metadata in sentiment analysis of users' posts on Twitter (Martínez-Cámara et al. 2014). However, while a lot of research has been moving towards establishing a coherent approach to identifying the 'meaning' in corpora, the most established approaches in corpus semantics include collocation and keyness analysis.

2.3.2 Collocation

The term 'collocation' was first introduced by Firth (1957) to denote the frequent co-occurrence of lexical items in natural language, whereby 'the meaning and usage of a word (the *node*) can to some extent be characterised by its most typical *collocates'* (Evert 2009: 2). A broader definition of this phenomenon, offered by McEnery & Hardie (2012: 122–123), is 'the idea that important aspects of the meaning of a word (or another linguistic unit) are not contained within the word itself, considered in isolation, but rather subsist in the characteristic associations that the word participates in, alongside other words or structures with which it frequently co-occurs.' This 'Firthian' notion of collocation was posthumously 'formalised and implemented' by neo-Firthian scholars (most notably John Sinclair),[9] who established its 'application in computational lexicography' (Evert 2009: 2). Collocation in this sense can be determined by calculating the frequency of co-occurring words within a set distance to each other (Xiao 2015: 107), and in its most basic form could be 'considered a methodological elaboration on the concordance' (McEnery & Hardie 2012: 123). While this approach remains a popular method of corpus analysis among neo-Firthian scholars (McEnery & Hardie 2012: 126), it is not universal, and the term 'collocation' has been applied to different notions of co-occurrence (Evert 2009; McEnery & Hardie 2012; Gries 2013; Xiao 2015). Xiao (2015: 107) uses 'neighbourhood collocation' to describe the concordance-based neo-Firthian approach, which is a helpful characterisation when discussing how it differs to alternative measures of collocation.

[9] Sinclair was one of the first scholars to expand Firth's concepts in CL, but many prominent linguists belong to the neo-Firthian tradition, including 'Michael Hoey, Susan Hunston, Bill Louw, Michael Stubbs, Wolfgang Teubert and Elena Tognini-Bonelli' (McEnery & Hardie 2012: 122).

In computational linguistics, 'collocation' is sometimes used to describe 'recurring sequences of two or more words' (McEnery & Hardie 2012: 123), otherwise referred to as 'multiword units' or '*n*-grams' (Xiao 2015: 107). Unlike neighbourhood collocation in the neo-Firthian sense, *n*-grams refer exclusively to adjacent frequency: words that appear next to a particular word (*n*). Within CL, these word sets are referred to as 'lexical bundles' (Biber et al. 2003), 'word clusters' (Scott 2010a) or 'multi-word-expressions' (MWEs) (Rayson 2005), and treated as 'lexicalised word combinations' (Evert 2009: 3) where the order and form of the words is restricted to a particular combination. In this regard, *n-grams* or MWEs differ to neo-Firthian collocation, which is a 'co-occurrence pattern that exists between two items that frequently occur *in proximity* to one another' (McEnery & Hardie 2012: 123), and 'not necessarily adjacently, or, indeed, in any fixed order'. For corpus linguists, the decision to give MWEs 'special treatment' (Evert 2009: 3) in processing or analysis will have methodological implications, as MWEs could be regarded as idiosyncratic items and processed as a set, or additional attributes or markers could be attached to the individual words that form them.[10] Furthermore, the 'compositionality' (Piao et al. 2006: 2) of MWEs makes them harder to classify systematically,[11] and typically requires manual categorisation of any observed collocate frequency (Evert 2009: 2). Thus, while investigating neighbourhood collocation requires selecting the span for measuring co-occurrence and the method for calculating significant collocates within this range, this approach is not sufficient to identify MWEs. To overcome this, it is necessary to identify the semantic properties of MWE parts, which in turn requires a comprehensive 'semantic lexicon' (Piao et al. 2006: 2) and a system for semantic annotation of corpora. Current systems capable of this analysis are discussed in §2.4.3 (USAS) and §2.4.5 (HTST) below. Notably, semantic annotation could also be used to investigate a third approach to collocation, described by Xiao (2015: 107) as 'coherence collocation'.

'Coherence collocation' was used by Xiao (2015: 107) to describe the 'cohesion that results from the co-occurrence of lexical items' that are conceptually similar (Halliday & Hasan (1976: 287) in Xiao 2015: 2). As explained by Halliday & Hasan

[10] See Piao et al. (2005b; 2006) for a thorough discussion of these issues.

[11] An example of 'non-compositional' MWEs by Piao et al. (2006: 2) includes the idioms 'kick the bucket' and 'hot dog', where the meaning cannot be 'predicted' from its parts. In contrast, 'compositional' MWEs like 'traffic light' and 'audio tape' are semantically related to their constituent parts.

(1976: 284), 'lexical reiteration takes place not only through repetition of an identical lexical item but also through occurrence of a different lexical item that is systematically related to the first one, as a synonym or superordinate of it'. Examples of this include words that frequently 'occur in a similar environment' (Xiao 2015: 107), such as 'letter, stamp, and post office' or 'hair, comb, curl, and wave'. Collocation in this sense is difficult to identify through statistical measures (Xiao 2015: 2), as it requires an external marker for classifying lexical items under conceptual headings. Semantic annotation of corpora presents an opportunity for investigating this form of collocation, and some recent attempts at this include the use of 'semantic glosses' for categorising collocates of a lexical node (Rodríguez-Fernández et al. 2016), and the analysis of semantic domains that collocate with different parts of speech (Archer et al. 2006; Culpeper 2009). However, none of these approaches identify the collocation of semantically related items in both the node and collocate position.[12] This gap is addressed in Chapter 7 with the introduction of 'semantic collocation' as a method for identifying conceptually related collocates, and thus establishes a foundation for systematically exploring coherence collocation in the future. Notably, 'semantic collocation' differs from 'neighbourhood collocation' as it uses 'keyness' analysis to identify significant co-occurrence.

2.3.3 Keyness analysis

The keywords of the corpus are those which occur 'with unusual frequency in a given text' when compared to a reference corpus (Scott 1997: 236). A statistical measure, usually Log Likelihood or Chi-squared, is used to determine what qualifies as an unusual occurrence in a given corpus (Rayson et al. 2004a). Words can be key if they occur more or less frequently in the corpus than would be expected by chance, identified as positive and negative keyness respectively (Scott 1998: 71). Crucially, keyword analysis allows researchers to determine distinct features of a specific corpus, and by examining those keywords they can form (and sometimes answer) research questions about the source texts.

Keywords can be used as 'searching tools, in text mining and classification, but also as analytic tools in text interpretation and discourse analysis' (Bondi 2010: 1).

[12] It is worth noting that this was outside the parameters of the authors' research questions.

Bondi focuses on the application of keywords in examining the 'cultural context that informs the text', using them as a proxy for the 'aboutness'[13] of the text and considering how they can be used to determine the cultural influences on the author (Bondi 2010: 1). The ability to carry out this type of investigation on large corpora meant that new ways of investigating texts were garnered from corpus linguistics methodologies. Taylor (2018: 22) notes the popularity of keyness in particular (in the analysis of 'difference') as being the result of the range of software capable of measuring keyness, showing that 'the tools which are available shape and form the type of research which may be carried out'. In this regard, keywords can 'play a role in identifying important elements of the text' (Bondi 2010: 1), and can be used to identify the features of a corpus that could provide insight through further analysis.

The application of keyword analysis in CL has now extended beyond the investigation of words in the corpus, thus expanding the scope of the features that could be examined for keyness (Baker 2004; Rayson 2004; Archer et al. 2006; Koller et al. 2008). A pertinent paper by Culpeper (2009: 30), for example, investigated 'the extension of the notion of keyness to part-of-speech tags and semantic domain tags'. In this paper, the author followed an earlier investigation of keywords in the dialogue of six different characters in Shakespeare's *Romeo and Juliet* (Culpeper 2002) by incorporating grammar and semantics 'explicitly into a keyness analysis' (Culpeper 2009: 41). To achieve this, the dialogue of *Romeo and Juliet* was standardised using the Variant Detector (VARD; Archer et al. 2003) program to resolve spelling variation, then annotated with parts-of-speech using the CLAWS (Constituent Likelihood Automatic Word-tagging System; Garside 1987),[14] the keyness of which was compared to the keywords for each character (Culpeper 2009: 41). To identify the keyness of semantic domains in the dialogue, Culpeper (2009: 46) used UCREL's (University Centre for Computer Research on Language) Semantic Analysis System (USAS; Wilson & Rayson 1993),[15] which assigned 'semantic tags' to 'each lexical item or multiword unit' (Culpeper 2009: 46).

[13] See also Archer et al. (2006).
[14] See §4.3.1.a.
[15] See §2.4.3.

Through their analysis, Culpeper (2009: 54) determined that while 'a straight keyword analysis revealed most of the conclusions' of the extended keyness analysis, 'the part-of-speech and particularly the semantic keyness analyses have much more of a contribution to make, moving the analysis beyond what is revealed by the keywords'. One example of this is the grouping of 'lower frequency words which might not appear as keywords individually' that represent patterns of usage (Culpeper 2009: 54–55), such as the 'general adjectives and (metaphorical) colour terms for Romeo, plural common nouns for Mercutio and items relating to "being" for the Nurse'. These findings suggest that key parts-of-speech and semantic domains could reveal idiosyncratic features of a corpus that would not be identified through a keyword analysis alone. The analysis carried out in §5.3.1 extends this research by investigating key semantic domains in a sample of a diachronic corpus of poetry and expands the scope of the analysis by using an extended semantic taxonomy in the semantic annotation of the corpus.

2.4 Semantic annotation

While semantic annotation of corpora is procedurally similar to grammatical tagging, as both involve assigning labels to words in a corpus, its use is less established in corpus linguistics, in part due to the 'more abstract and/or difficult nature of the phenomena to be analysed' (Garside et al. 1997: ix).[16] The main obstacle preventing a generally established framework for semantic annotation is the difficulty of disambiguating senses in polysemous words (Kennedy 1998: 225). A further concern is the way in which senses are classified, as different approaches to semantic classification will impact on how they are assigned. The final key consideration, therefore, is the process by which the disambiguated meanings are assigned to the words in a corpus. Semantic annotation requires all three of these challenges to be resolved, but there is no universal approach to resolving them.

A further issue arises from the challenge of blending *interpretative* methods with an empirical approach to data. The former has traditionally been associated with subjective analysis, while the latter is grounded in an objective scientific approach and has long been viewed as an advantage of using corpora in research (McEnery

[16] In an earlier work, Leech and Fligelstone (1992: 126) predicted that semantic annotation is 'likely to become a matter of priority in the future'.

& Wilson 1996: 86). Crucially, according to Leech (1997: 2), by acknowledging the interpretative nature of the annotations, we therefore acknowledge that it is 'at least in some degree, the product of the human mind's understanding of the text'. This distinction was key in establishing the relationship between the annotation Leech was discussing and the corpus, as at that time there was 'no purely objective, mechanistic way of deciding what label or labels should be applied to a given linguistic phenomenon.' (Leech 1997: 2). Consequently, any systematic approach to disambiguation is reduced to attaching a 'certain *probability* to each sense' (Leech 1974: 78), with 'the complete ruling-out of a sense being the limiting case of *nil* probability'. That is, while it is possible to definitively rule out a particular sense, the 'correct' sense may in fact be multiple possible definitions of the word in context. Despite substantial growth in corpus semantics, the limits of disambiguation as identified by Leech (1974) remain unchanged, as they are inherent to how meaning is formed in language.

To explicate the difficulty of observing meaning in language, and what this means for corpus semantics, Stubbs (2001: 20) outlined a set of unobservable features that impact the meaning of a word: *expectations* (communicative competence), *real-world inferences* (extra-linguistic knowledge), *linguistic conventions* (familiarity with language patterns), and *text-types* (expectations based on the genre of the text). These concepts correspond to the development of meaning from the perspective of a reader and cannot be explored through computational text analysis. It is, however, possible to use observational methods to identify *probable* meanings in language, as the meaning of a word is also determined by the context in which its used (Stubbs 2001: 20–21). This observable characteristic consequently informs most approaches to systematic disambiguation of meaning in text.

2.4.1 Disambiguating senses

The challenge of accurately identifying the meaning of a word in a corpus is not unique to corpus linguistics but is a shared issue in 'computational linguistics and Natural Language Processing' (NLP) (Alexander et al. 2015b: i19). In computational linguistics, researchers have investigated two different approaches for resolving this issue: the first looks at examining the distinctions between senses and the degree to which each sense can be represented in a text, and the second

looks at what contextual information is available to assist in disambiguating sense (Kennedy 1998: 225). The former commonly involves assigning semantic descriptors to lower-level syntactic annotation, with the view of automating meaning discovery for research or data analysis purposes (Palmer 1990: 1). This approach is not dissimilar to part-of-speech annotation as it relies on 'semantic interpretation rules' to overlay semantic information on a pre-existing syntactic parse based on a set of parameters, in this case, a 'template' containing a limited number of designations and corresponding criteria (Palmer 1990: 34–35). The use of contextual information for attributing semantic markers is less clearly defined.

Inference-driven mapping, for example, extends the template approach by reducing the number of possible markers for a word based on the decomposition of that word (Palmer 1990: 5). That is, by identifying the simple form of the word through semantic decomposition, the constraints of the basic form can be used to assign markers for the complex term (Palmer 1990: 122–123). An example from Palmer (1990: 23) for the 'lexical entry for *attach*' is that 'a contact between an entity, OBJECT1, and another entity, OBJECT2, can be expressed using the verb *attach*'. To determine if '*attach* has been used appropriately', 'contact' is 'set up as a subgoal', representing a possible simple form of *attach* (Palmer 1990: 23). To validate the decomposition of *attach* into 'contact', it must meet three 'subgoals, locpt, locpt and sameplace': 'if a location point on an entity, LOCPT1, and a location point on another entity, LOCPT2, are at the sameplace, then the entities are in contact with each other' (Palmer 1990: 24). If a word is able to meet all of the subgoals of the decomposed domain, its usage is 'proven' and can be attributed to that domain (Palmer 1990: 23). Crucially, by elevating the level at which the constraints are set through decomposition, the process reduces the need to program assignment criteria for complex forms. The drawback of this approach is its reliance on finite domains that are mapped to syntactic roles, which are manually programmed to suit the task.

An alternative method to the template approach reverses the annotation process by grouping words into 'clusters' (Jurafsky & Martin 2000: 640) based on their frequency and context, and manually encoding these clusters with user-determined senses. The process can be further refined by adding additional rules to the clustering algorithm, thus increasing the chance that items will be added to the correct cluster in subsequent iterations. As this approach is popular in

computational linguistics, a range of algorithms have been developed to facilitate this analysis, employing different statistical measures and pre-set restrictions (Jurafsky & Martin 2000: 640). To implement this analysis, however, a degree of specialist knowledge is required, limiting its use outside of computational linguistics. Perhaps the most accessible algorithms to fall within this approach are probabilistic topic models, due to the wide variety of guides, libraries, and tools for topic modelling using popular programming languages (Blei 2012: 78; Jelodar et al. 2019: 15196).[17]

As with other methods of analysing meaning in corpora, topic modelling algorithms are designed to help researchers 'discover and annotate large archives of documents with thematic information' (Blei 2012: 78). The process of extracting meaning through topic modelling, however, differs substantially to the template approach as it does not rely on a pre-existing classification of senses which are attributed to items in a corpus. Instead, words that appear with a significant frequency (such as through keyness analysis) across different texts are iteratively grouped into topic clusters based on the probability that they are thematically related. The researcher is then responsible for defining these clusters based on their interpretation of the topic as in the example of *human, genome, mapping* all being assigned under the topic of 'Genetics' (Blei 2012: 80). In this regard, this approach overcomes the issue of disambiguation, as the topics identified through the model are not attributed any set meaning through the model itself.

Unlike semantic annotation, topic modelling only selects words that appear with a significant frequency in a particular text, which are seen to represent the topics covered in that text (Jelodar et al. 2019: 15172). Words that are not considered significant by the algorithm are not represented, and therefore have no assigned meaning. While it is possible to increase the number of topics modelled by the algorithm, a researcher would still be required to manually classify each topic, thus making the process prohibitively time-consuming (Jurafsky & Martin 2000: 641). Consequently, while topic modelling has uses in exploring meaning in texts, it is

[17] A few examples include *Topic Modeling with Gensim* (*Python*) (Prabhakaran, https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/), *Beginner's Guide to LDA Topic Modelling with R* (tang, https://towardsdatascience.com/beginners-guide-to-lda-topic-modelling-with-r-e57a5a8e7a25), and *jsLDA: In-browser topic modeling* (Mimno, https://mimno.infosci.cornell.edu/jsLDA/) which uses JavaScript to run models within a web browser.

not a viable method of disambiguating meaning across all words in a corpus. Semantic annotation of corpora makes this possible by assigning meaning from a pre-defined taxonomy of senses, thus removing the need for manual classification of topics represented in the text (Rayson et al. 2004b). The drawback of this approach, however, is the need for a viable system of classifying senses at a level of abstraction that makes the annotated data useful (de Andrade et al. 2019: 3). A popular approach to demarcating these conceptual boundaries is by expressing related meanings as belonging to 'the same semantic field' (Wilson & Thomas 1997: 54).

2.4.2 Semantic fields

Broadly, semantic fields represent groupings of concepts that are represented by lexical items, which can be used in as a scheme for semantic annotation (Wilson & Thomas 1997: 54–55).[18] The boundaries represented by semantic fields are not absolute; rather, they are constructs which group 'words that are related by virtue of their being connected – at some level of generality – with the same mental concept' (Wilson & Thomas 1997: 54). This relatedness is not restricted to synonymy (or antonymy), as semantic fields include hypernyms (more general meanings) and hyponyms (more specific meanings), as well as 'words which are associated in other ways with the concept concerned' (Wilson & Thomas 1997: 54). An example this would be the grouping of '*rider, horse, eventing, spurs, saddle, dressage, jump-off* and so on' within the semantic field of 'equestrianism' (Wilson & Thomas 1997: 54).

Grouping concepts into semantic fields allows for a usable framework for semantic classification of texts, albeit one that serves as an idealised version of how relationships between concepts are formed by readers:

> By classifying words according to a category system representing a set of plausible relationships (i.e. semantic fields) we can approximate to a representation of the kinds of relationships which we know to exist in the mind whilst simultaneously presenting these groups of related words in a way which is maximally accessible to end users of an annotated corpus. Classification of words according to semantic field systems seems the best compromise between what we know about the mind, what is useful

[18] A semantic field can also referred to as 'a conceptual field, a semantic domain, a lexical field, or a lexical domain' (Wilson & Thomas 1997: 54).

> for further psycho-linguistically-motivated textual research based upon this knowledge, and what other content-oriented scholars and commercial users will find useful and accessible (Wilson & Thomas 1997: 55).

As such, the advantage of a semantic field taxonomy is that meanings can be classified in a consistent and readable format and disambiguated through contextual semantic information, which is made possible through the UCREL Semantic Analysis System (USAS; Rayson et al. 2004b).

2.4.3 USAS

The USAS tagger was developed by the University Centre for Computer Corpus Research on Language (UCREL; Rayson & Wilson 1996) team at Lancaster University as a resource for semantic annotation of corpora. The distinguishing feature of the USAS tagger is that its 'semantic lexicon employs a semantic field taxonomy and maps words and multiword expression (MWE) templates to their potential semantic categories […] according to their context in use' (Piao et al. 2005a: 1). The tagger utilises the hierarchical structure of the Lancaster semantic lexicon for disambiguating senses into '21 major semantic fields that expand into 232 sub-categories' (Piao et al. 2005a: 3).[19] These semantic fields represent a 'conception of the world that is as general as possible' (Piao et al. 2005a: 3), with top-level fields such as 'Emotion [E]', 'Food and Farming [F]', and 'Life and living things [L]' (Archer et al. 2002: 2).[20] These top-level categories can be decomposed into distinct sub-categories, which reflect narrower conceptual groups, such as 'Food [F1]', 'Drinks [F2]', 'Cigarettes and drugs [F3]', and 'Farming & Horticulture [F4]' as sub-categories of [F] (Piao et al. 2005a: 3–4). The sub-categories allow users to further distinguish the meanings of the semantic tags, but they 'maintain a relatively low level of granularity' (Piao et al. 2005a: 4) to overcome the challenge of intricate word sense disambiguation (Rayson et al. 2004b: 2). This restriction allowed the UCREL team to employ 'a set of context rules' and 'algorithms of

[19] The initial taxonomy was based on Tom McArthur's (1981) *Longman Lexicon of Contemporary English*, but has gone through several revisions to account for the 'practical tagging problems met in the course of ongoing research' (Archer et al. 2004: 816).

[20] Capital letters, such as 'E' for 'Emotion', are used to distinguish the top-level fields in the semantic tag syntax, while digits are used to specify the subcategory the lexeme belongs to, alongside additional optional markers (Archer et al. 2002: 1–2). These additional markers include 'areas of meaning which reflect synonym-antonym, general-specific or meronymy/holonymy' relationships (Piao et al. 2005a: 3).

disambiguation' to assign semantic code tags to each word in the annotated corpus (Rayson et al. 2004b: 3).

USAS assigns these tags over two 'phases' of annotation (Rayson et al. 2004b: 4): in the first phase, 'potential semantic tags' are assigned to each word from the meanings recorded in the semantic lexicon, and in the second phase, 'contextually appropriate' tags are selected from the potential tags identified in phase one. USAS employs multiple parameters to disambiguate potential tags in phase two, which could be further divided into two groups: rules that determine tag priority based on generic lexical features (similar to the template approach discussed in §2.4.1 above), and disambiguation rules that rely on contextual information to identify likely word senses. The former ruleset starts by using the part-of-speech (POS) tag to restrict lexical entries based on type, such as the likely usage of *spring* as '[season]' when used as a temporal noun (Rayson et al. 2004b: 4). In addition to this, 'general likelihood ranking' is used to promote senses based on their 'frequency' (Rayson et al. 2004b: 5), though this 'ranking is derived from limited or unverified sources such as frequency-based dictionaries, past tagging experience and intuition'. A further lexicon-based rule employed by USAS is the preference of multi-word expressions over individual word senses, using a 'set of heuristics' to determine the most likely MWE tag (Rayson et al. 2004b: 5). These rules could reduce the number of potential tags that need to be contextually disambiguated, but it is not clear if they are applied in any order of importance or priority in relation to each other (Rayson & Wilson 1996; Rayson et al. 2004b).[21] This lack of clarity makes the process harder to evaluate, as there is no indication of what rules determined the final tags assigned by the tagger.[22]

The opaqueness of the disambiguation process is even more pronounced in the second, context-based ruleset. The first of these contextual rules uses 'knowledge of the current domain or topic of discourse' to 'raise the likelihood' of a relevant domain 'at the expense' of others (Rayson et al. 2004b: 5). The precise method for making this adjustment, however, is unclear. Similarly, the 'text-based

[21] Earlier work on the system that would become USAS suggested that part-of-speech tagging is a necessary precursor to semantic annotation, but it was unclear if the relationships between semantic tags and grammatical tags would take precedence over subsequent disambiguation stages (Garside & Rayson 1997).

[22] The error-rates of the different disambiguation stages are discussed in Rayson et al. (2004b), but these are not distinguished in the confidence rating included in the final annotation output.

disambiguation' rule declares that words with assigned meanings are likely to have the same meaning elsewhere in the text, but does not specify how the initial sense is determined (Rayson et al. 2004b: 5). Finally, it is not clear to what extent contextual disambiguation relies on pre-defined templates of 'regular contexts in which a word is constrained to occur in a particular sense' and 'local probabilistic disambiguation' (Rayson et al. 2004b: 5), which is a dynamic process that draws from the grammatical and semantic tags assigned during the semantic annotation routine. While these ambiguities are not uncommon in closed source software, they restrict the user's ability to interpret the results or modify the annotation process. The counterbalance to these limitations is that USAS provides an all-inclusive system for semantic annotation, allowing users to investigate meaning in texts without any programming knowledge (c.f. computational approaches discussed in §2.4.1 above).

The advantages of USAS are likely to outweigh the limitations for users looking at present-day English texts, where USAS demonstrated 91.05% accuracy on a test corpus (Rayson et al. 2004b), and carrying out research that benefits from a general semantic classification system (Archer et al. 2004). To extend the use of USAS beyond these parameters, the team investigated implementing a 'modified' semantic lexicon for Early Modern English (Archer et al. 2003), and ultimately extended the tagger to work with 'historical forms of English' and 'annotate deep semantic senses' (Piao et al. 2017: 113) by linking the USAS framework to a semantic taxonomy developed from the *Historical Thesaurus of English* (Kay et al. 2009). The resulting 'Historical Thesaurus Semantic Tagger' (HTST; Piao et al. 2017) employs the rich semantic taxonomy of the *Historical Thesaurus of English* to deliver a tool for systematic semantic annotation of historical texts. While the HTST maintains the restrictions of a closed source system (Alexander et al. 2015b; Piao et al. 2017), its development highlights the potential of the *Historical Thesaurus* as a resource for semantic annotation.

2.4.4 The HTE as a tool for semantic analysis

The release of the *Historical Thesaurus of English* (HTE; Kay 2011: 42) marked a unique contribution to language analysis and, by extension, semantic analysis of texts. First published as the *Historical Thesaurus of the Oxford English Dictionary* (Kay et al. 2009), its release enabled previously impossible research in

the study of English Language. The unique resource, holding just under 800,000 words arranged in 225,131 semantic categories, captures the history of meanings from Old English to the present day. The lexical items are categorised in what amounts to a 'semantic database of the language' (Alexander et al. 2015b: i17). The HTE semantic taxonomy is organised through a hierarchy of senses, ranging from broad to narrow concepts, nested in a tree-like structure under three overarching domains. Thus, unlike the USAS semantic lexicon, which restricts the classification of senses to general semantic domains (Archer et al. 2002), the HTE fulfils the role of a 'computer-readable lexicon containing possible semantic fields for given words' (Wilson & Thomas 1997: 62). The HTE's classification of historical senses further enhances its use in semantic analysis of texts, as it enables diachronic investigation of meaning in text.

One such approach for investigating meaning change with the HTE taxonomy is to examine the 'recategorisation' of words as 'understanding of, or attitude towards' particular concepts changes over time (Alexander & Struan 2013: 233). An example of this application of the HTE taxonomy was a project that focused on the concept of 'incivility' as a conceptual category and examined how related adjectives were applied in cultural documents over time (Alexander & Struan 2013: 234). To examine this, the corpora used for the project included 'OED citation files, the House of Commons recorded debates in *Hansard*, major linguistic corpora and sundry other relevant primary sources.' (Alexander & Struan 2013: 235). The authors' findings included the change in representations of incivility from 'rough, animalistic characteristics […] towards the later significance of the relationship between the person and the state' (Alexander & Struan 2013: 232). This change depicts 'longer-term shifts in attitudes' (Alexander & Struan 2013: 232), which would have eluded identification if only present-day terms for incivility were examined. While this project relied on manual classification of senses with the HTE taxonomy, it demonstrated the importance of diachronic sense classification for interpreting historical texts.

2.4.5 HTST

To extend the use of the HTE beyond manual classification of senses, a collaborative initiative, funded by the Arts and Humanities Research Council (AHRC) and the Economic and Social Research Council (ESRC) in 2014,

attempted to develop a tool for systematic semantic classification with the HTE. The project was titled 'Semantic Annotation and Mark-Up for Enhancing Lexical Searches' (SAMUELS, after Michael Samuels), and the key output was the HTST (Piao et al. 2017).[23]

The SAMUELS project identified the issue of needing a more sophisticated way of determining the concepts in large corpora, which has traditionally been limited by the 'need to search using word forms' (Alexander et al. 2015b: i16). By developing a system of quickly identifying the semantic properties of a corpus, the team hoped to improve the way that users engaged with large textual data. They worked towards developing a system that would allow for 'semantic searches', removing from the user the need to have a pre-determined list of word-forms and instead allowing them to search using an overarching semantic field (Alexander et al. 2015b: i17).

As discussed in §2.4.3 above, the SAMUELS team developed the HTST alongside an existing set of tools already established by the UCREL team at the University Centre for Computer Corpus Research on Language (UCREL) at Lancaster University (Alexander et al. 2015b: i17). The tools include the Variant Detector (VARD) for normalising spelling variation (Baron & Rayson 2008); the Constituent Likelihood Automatic Word-tagging System (CLAWS), the UCREL part-of-speech tagger (Garside & Smith 1997); and UCREL's own semantic tagging system, USAS (Rayson et al. 2004b). By integrating the HTE taxonomy alongside these existing tools, the SAMUELS team could avoid rebuilding the whole tagger from the ground up (Alexander et al. 2015b: i17).

Like USAS, the HTST relies on a hybrid disambiguation model that combines a template approach with further contextual disambiguation phases. [24] However, while the HTST retrieves USAS codes alongside its own tagset in the tagged output, its disambiguation process is mostly independent. An exception to this is the annotation of closed class words and proper nouns, which are assigned HTE sense codes that match the USAS output (Alexander et al. 2015b: i18). The HTE sense codes are represented in two different ways in the annotated output (Alexander et al. 2015: i19): the 'HT [*Historical Thesaurus*] sense code', which

[23] Available online through the University of Glasgow (Alexander et al. 2015c).
[24] As discussed in §2.4.3.

follows the 'highly fine-grained' semantic categories the HTE taxonomy, and the 'Thematic level sense code', which provides a general description of the HTE field. As the taxonomy of the HTE is deeper than the USAS semantic lexicon, the HTST employs additional disambiguation phases to assign HTE senses to corresponding lexical items (Piao et al. 2017). The template phase includes matching 'highly polysemous words' from a curated 'sub-lexicon' of HTE senses (Piao et al. 2017: 119), and the aforementioned mapping of USAS categories to HTE tags for 'function words and proper nouns' (Piao et al. 2017: 120).

Additionally, potential HTE senses of each word are ranked in order of importance using the HTST 'polyseme density' metric (Alexander et al. 2015b; Piao et al. 2017). This metric is unique to the HTST as it is based on how frequently a word form is recorded within a conceptual hierarchy and is used to prioritise popular meanings over more obscure ones. For example, the frequent appearance of 'the word form "wine" in the semantic area of drinking' is used to promote the sense of wine as the drink (Alexander et al. 2015b: i21), instead of, for instance, a religious artefact.[25] It is not clear, however, how much the polyseme density metric impacts on the disambiguation process, and whether all recorded senses are ranked based on their density (Alexander et al. 2015b: i21). The HTST test data combines the sub-lexicon, closed-class word mapping, and polyseme density ranking into one review phase, so it is not clear what the individual contribution of each method is (Piao et al. 2017: 128). These results did, however, reveal that this phase of disambiguation significantly improved the accuracy of the annotation data against a baseline of random HTE categories, [26] achieving an average 69.39% accuracy across the ten sample corpora (Piao et al. 2017: 128). This was a significant improvement to the 15.41% baseline accuracy, but was further improved by two context-based disambiguation phases, reaching an average 79.85% accuracy in testing (Piao et al. 2017: 125–126).

The two context-based disambiguation phases employed by the HTST are unique to the tagger, as both rely on the HTE taxonomy and the 'headwords' that define HTE categories (Alexander et al. 2015b; Piao et al. 2017). The first method

[25] 03.08.05.15.07.02 (n.) Wine. 2021. In *The Historical Thesaurus of English* (2nd ed., version 5.0). University of Glasgow. https://ht.ac.uk/category/?id=181254.
[26] The HTST disambiguation methods were tested against a baseline of randomly selected HTE categories for each word form, and applied to ten sample corpora of different genres and periods (Piao et al. 2017).

retrieves the headwords of the parent categories of each potential sense of a word form, and then compares these headwords against the immediate context of the word form (up to five words in both directions). The result prioritises candidate senses that are closer to neighbouring content words in the HTE hierarchy. So, if applied to the word *osmosis*, the tagger would prioritise the sense 03.07.03|06 (assimilating ideas) if the context included the words *learning* (03.07.03) or *education* (03.07).[27] If the context included *movement* (01.02.03.03.08), *biological processes* (01.02.03.03), or *biology* (01.02.03), then it would prioritise 01.02.03.03.08|06 (diffusion through porous membrane) as the sense of *osmosis* in this context.[28] In testing, this approach increased the average accuracy of the tagger to 79.77%, when implemented after the first template-based disambiguation phase (Piao et al. 2017: 126). The downside of this approach, however, is that it only looks at the lexical context of a word and relies on direct matches based on the category headwords of the target word form. If no match is identified, the tagger refers to a 'manually compiled list' of HTE categories to assess potential senses (Alexander et al. 2015b: i20), which reduces the efficacy of this context-based method.

The second context-based disambiguation phase similarly relies on the HTE taxonomy, but extends the approach by using the connection between the *Thesaurus* senses and the *Oxford English Dictionary* (OED) entries for target word forms (Piao et al. 2017). This process embeds the USAS system by tagging 'word sense definitions in the OED using the USAS tagger' (Piao et al. 2017: 120), and then 'extracting the statistical association metric between the HT [HTE] categories of the headwords and the USAS tags contained in the definition entries'. This process leads the tagger to prioritise HTE senses that correspond to the USAS semantic tag retrieved from the OED word sense definition. In other words, the USAS tags attributed to the OED entry inform the selection of the corresponding HTE tag.[29] In testing, however, this method performed inconsistently across the

[27] 03.07.03|06 (n.) Learning :: assimilating ideas. 2021. In *The Historical Thesaurus of English* (2nd ed., version 5.0). University of Glasgow. https://ht.ac.uk/category/?id=172683.
[28] 01.02.03.03.08|06 (n.) Movement :: diffusion through porous membrane. 2021. In *The Historical Thesaurus of English* (2nd ed., version 5.0). University of Glasgow. https://ht.ac.uk/category/?id=16994.
[29] This association metric was demonstrated by Piao et al. (2017: 120–121) by referencing the connection between the HTE category 03.01.01.03.10 (ancestor) and the USAS category S4 (kinship) through the words *father, mother, progenitor, forefather,* and *grandfather* in the OED entry for *ancestor*.

ten sample corpora: it slightly improved the accuracy for the contemporary samples, but had a 'slightly negative impact on the historical data' (Piao et al. 2017: 126). Piao et al. (2017) attributed these conflicting results to the underlying USAS metadata, as it 'performs less accurately on historical data, thus affecting the performance of the OED data based disambiguation' (Piao et al. 2017: 126). As a result, while the first generation of the HTST addresses some of the limitations of USAS, its application is reduced by its current technical limitations.

2.4.6 Limitations of existing semantic taggers

To make semantic analysis and annotation tools easy to use and accessible outside of specialist research, their design will often have to sacrifice customisation in favour of reliability. This allows their creators to ensure that the tool performs consistently and without requiring the user to troubleshoot any errors (Hardie 2012). The drawback of this is that there is no way for a user to modify the process when it does not fit the requirements of their corpus. While researchers can try to select a tool that would be most compatible with the corpora they intend on studying, this is not always possible; most semantic analysis tools are developed with a particular research goal, and often perform better on certain types of corpora (Piao et al. 2005a). Compatibility becomes even more of a problem when looking for a tool that can annotate 'all lexical units' with intelligible semantic categories (Piao et al. 2017: 114), as the only options currently available for English are the USAS (as part of Wmatrix) [30] and HTST systems (Piao et al. 2005a; 2017).[31]

The USAS tagger was developed to 'undertake the automatic semantic analysis of present-day English' texts (Archer et al. 2003: 22), and its abstracted semantic lexicon is useful for identifying general concepts in a corpus (Piao et al. 2005a). For studies that require a fine-grained semantic taxonomy, the USAS classification scheme might not be sufficiently representative (Koller et al. 2008). The HTST overcomes this limitation by utilising the detailed semantic categorisation of the HTE to classify senses (Alexander et al. 2015b; Piao et al. 2017), which makes it suitable for investigating meaning in corpora at different levels of specificity. However, the HTST was built on top of existing systems and relies on USAS and

[30] http://ucrel.lancs.ac.uk/wmatrix/
[31] http://www.gla.ac.uk/samuels/

pre-defined templates for part of its disambiguation process,[32] including attributing fixed definitions to highly polysemous words (Furkó 2019).

Despite the advantages posed by the HTST over other semantic annotation methods, its lack of customisation makes it difficult to use in text types that require modified disambiguation rules. The SAMUELS team found that the HTST's accuracy varied depending on the text corpora (Alexander et al. 2015b: i20). Samples that contained a lot of 'noise' were tagged less accurately because most of the annotation systems embedded in the tagger (CLAWS and USAS) were 'trained on standard English texts, and hence performed less well on noisy text' (Alexander et al. 2015b: i20). In corpus linguistics, 'noise' refers to features that interfere with the analysis of a corpus, and could include grammatical items (Hughes 2010), irrelevant documents or texts included in the corpus (Gabrielatos 2007), and in general any 'unwanted data' included in the corpus (Desagulier 2019). In semantic annotation, 'noise' can also extend to incorrect or redundant metadata that impacts the accuracy of the tagger, whether as part of the pre-processing POS annotation (Padro & Marquez 1998; Piao et al. 2015), error handling (Alec et al. 2016), or unexplained ambiguity in the assigned semantic tags (Berlanga et al. 2015; Piao et al. 2015; Furkó 2019). To overcome the issue of 'noise' in CL, it is necessary to tailor the corpus to meet the requirements of the tagger; to overcome the issue of 'noise' in the semantic annotation output, the output could be edited in post-processing by the researcher. For corpora that contain texts that cannot be manipulated to meet these requirements, or studies that require adjustments to different stages of the semantic tagging process, a flexible method of annotation is required, which can be tailored to suit the requirements of the corpus or research. Consequently, the limitations of both USAS and HTST can be divided into issues of compatibility with different types of corpora, and issues of compatibility with different types of corpus analysis.

The issue of compatible corpus types impacts on the accuracy of the semantic tagger. While both USAS and the HTST have been applied to a range of different text types, they do not perform as well when used on non-standard texts. Studies that looked at texts that contain technical vocabulary, such as legal documents (Piao et al. 2004), scientific literature (Gábor et al. 2016), or medical texts (Cohen

[32] See §2.4.5.

et al. 2013; Jovanović & Bagheri 2017), and texts that 'infringe orthographic, grammatical and stylistics norms' (Bollmann et al. 2014: 86), such as 'internet data' or 'historical language data' (Bollmann et al. 2014: 86), reported reduced accuracy in the tagged output when compared to the baseline of each tagger. While the developers of both taggers could address this limitation by expanding the semantic lexicon, or adjusting the tagging process for different text types, users are reliant on the developers to make these changes.

For example, the inclusion of VARD (Baron & Rayson 2008) to normalise historical variations in spelling into the pre-processing stage of the HTST could be seen as the developers' attempt to overcome the issue of compatibility. VARD reduces the impact that non-standard spelling has on the annotation process (Rayson et al. 2017), thus improving the accuracy of the tagger on historical language (Piao et al. 2017). However, as VARD was embedded into the existing tagging process, it cannot resolve any issues caused by the other tagging methods that were trained on contemporary texts, namely CLAWS and USAS (Alexander et al. 2015b; Piao et al. 2017). By integrating VARD into the tagging pipeline, the developers of the HTST allowed users to toggle the variant spelling normaliser as a setting of the semantic tagging tool (Alexander et al. 2015c). However, as discussed in §2.4.5, this affects the pre-processing stage of the tagging pipeline, and only impacts the disambiguation process by reducing errors when matching word forms with lexicon items (Piao et al. 2017); it does not change the disambiguation parameters. At the time of writing, the only other configuration option for users of the HTST interface is setting the date range of available HTE senses and the formatting of the tagged output (Alexander et al. 2015c).

The USAS tagger, through the Wmatrix interface (Rayson 2009a), offers more customisation options than the HTST, such as allowing users to set 'preferred domains' so that the tagger 'increases their likelihood' and tries to select corresponding tags 'when an ambiguous word of MWE occurs in the text' (Rayson 2007a). Users can also add to the USAS semantic lexicon by adding their own 'personal dictionaries' for both MWEs and single words (Rayson 2007b). Both features were added as updates to the Wmatrix USAS interface (Rayson 2007c), and serve as a further example of addressing the compatibility issue, as they allow users to adjust the tagging process to suit the needs of a specific corpus. However, while these settings allow users to 'override or extend the system dictionaries' with

their data (Rayson 2009b), and adjust certain variables of the disambiguation process, most of the tagging process is still pre-determined. Furthermore, while these settings are a useful addition to the tagging process, they are only appropriate for studies where the user can identify additional domains they want to add based on their knowledge of the corpus (Desagulier 2019).

The design of the HTST tagger and the expansion of USAS demonstrate their developers' attempts to overcome the issue of compatibility with different text types. However, both the inclusion of VARD into the HTST pipeline, and the configuration settings added to USAS fail to address the second compatibility issue: the impact that the indivisible stages of the tagging process have on corpus analysis. For instance, both taggers assign pre-determined senses to highly polysemous words, which could be 'at odds with the contextual specificities of the occurrence' (Desagulier 2019: 223); neither tagger allows the user to change this tagging setting when it does not suit their needs. Incorrectly tagged and 'unmatched items' (Prentice 2010: 432) similarly pose a problem for users of both taggers, as the 'preferred domain' setting in USAS is applied regardless of context (Rayson 2007a), and the HTST does not allow this customisation option at all (Alexander et al. 2015c). On a larger scale, neither tagger can be used to investigate the meaning in context (Desagulier 2019), as both rely on template-based disambiguation for certain items (Piao et al. 2004; 2017). Overcoming the epistemological issues of current taggers could expand the scope of semantic analysis of different text types and allow different approaches of investigating meaning in corpora.

In summary, while both compatibility issues of current semantic taggers restrict the scope of semantic analysis of corpora, it is not possible to fully account for idiosyncratic features of different corpora if the underlying tagging process is not adjusted as well. Correspondingly, any system that looks to address these limitations must be flexible enough to accommodate different research requirements as well as different text types (Hardie 2012). However, as shown by research that made use of USAS and the HTST, both taggers can be used in research outside of their ideal parameters (Piao et al. 2004; 2017). Any new system of semantic annotation must have a clear advantage over current taggers in areas where current limitations have a significant impact on research. Thus, to establish the extent of these limitations in practice, and to identify the requirements of a

flexible semantic annotation, the following section of the literature review explores a noteworthy use-case of the ambiguity tagger: stylistic analysis of semantically annotated corpora.

2.5 Semantic annotation for corpus stylistics

Corpus stylistics is a particularly useful starting point for exploring different uses of semantic annotation, as it is a growing area of research characterised by innovative blending of 'corpus linguistic techniques with stylistic analysis' (McIntyre & Walker 2019: 1). Research in the field includes application of corpus analysis techniques to 'individual texts alone' (McIntyre & Walker 2019: 12), as well as the use of corpus methods to identify distinctive features in texts through comparison with 'general patterns of a language' (Mahlberg 2013: 7–8). In both cases, the use of corpus linguistics techniques adds a further layer of 'systematicity' to stylistic analysis (Mahlberg 2013: 8), by providing a 'means of checking intuitions and validating (or invalidating) what might otherwise be fairly subjective claims about a text' (McIntyre & Walker 2019: 61). The more significant advantage afforded by corpus methods, however, is the ability to 'supplement insights gained via traditional methods' (McIntyre & Walker 2019: 14), and thus the potential to expand the scope of stylistic analysis in general.

Corpus stylistic research is frequently associated with analysis of literary texts (Mahlberg 2014), particularly when framed as the use of corpus methods to identify 'textual features that are especially characteristic of an author or text' (Biber 2011: 15), though research in the field extends to a range of text types (e.g. non-literary texts in Semino & Short (2004); early news media in Studer (2008)). McIntyre and Walker (2019) reinforce this position by noting that 'the object of study for stylistics is style in all text types, not just literature' (McIntyre & Walker 2019: 309). Notably, while computational and corpus-assisted analysis of literary texts 'has been around for almost as long as corpus linguistics itself' (McIntyre & Walker 2019: 309), the distinguishing feature of emerging corpus stylistics research is the 'focus on linguistic style and the use of stylistic theories and analytical frameworks'. As such, while corpus stylistics is 'fast becoming a recognisable field' (McIntyre & Walker 2019: 1), the methodologies and disciplinary boundaries of corpus stylistics are not yet settled. As the field

continues to grow, new research questions and designs demonstrate the potential of established corpus linguistics approaches for stylistic analysis, while simultaneously uncovering new demands that cannot be met by existing approaches.

2.5.1 Developments in corpus stylistics

While corpus stylistics has been steadily growing since the early 2000s (Semino & Short 2004; Biber 2011; Mahlberg & McIntyre 2011), the use of corpus methods in computational stylistics predates the formation of the field (McIntyre & Walker 2019: 11). One of the earlier influential examples of computer-assisted analysis of style was Burrows' (1987) stylometric investigation of Jane Austen's novels (McIntyre & Walker 2019: 11). Burrows' expository work investigated non-lexical items in the dialogue of Austen's characters compared with the author's narrative style (Burrows 1987), using corpus techniques to explore patterns across several novels. In parallel to Burrows, Louw (1989; 1993) began extending corpus analysis methods to investigate rhetorical features in literary texts, using large reference corpora to show how prosodic inconsistencies could be used to identify 'suasive language' (Louw 1993: 157). These early investigations demonstrated the potential of corpus linguistic approaches in stylistic analysis, and continue to serve as a foundation for different routes of corpus stylistic enquiry (Biber 2011). However, despite these early developments, corpus stylistic analysis of literary texts remained underexplored until recently (McIntyre & Walker 2019).

A possible reason for the slow growth of corpus stylistics is that literary texts are not considered to fulfil the traditional criteria of a corpus; that using CL techniques on a text, or even a large collection of texts by a single or multiple authors does not inherently meet the criteria of a corpus as being 'usually of a size which defies analysis by hand and eye alone within any reasonable timeframe' (McEnery & Hardie 2012: 2). Mahlberg (2013) explains this position by noting that 'a poem, a novel, or a short story provide very limited data' when compared to the large corpora already available to corpus linguists (Mahlberg 2013: 1). This reasoning cannot fully explain the slow development of corpus stylistics, however, as small and specialised corpora are not uncommon in corpus linguistics research (Koester 2010), while the work of Burrows (1987) and Louw (1993) established a

precedent for stylistic analysis of individual authors' work using corpus techniques.

Indeed, instead of corpus size, the more significant barrier to be considered when using corpus techniques for stylistic analysis of literary texts is that tools developed for corpus analysis often decontextualise the texts that make up a corpus (Flowerdew 2005; Baker 2006; Koester 2010). In other words, they do not distinguish between different elements of 'contextual scope' in a corpus (Scott & Tribble 2006: 9); that is, the boundaries drawn within a corpus at different levels, ranging from word, sentence, and paragraph level, to section or individual texts in a corpus, and finally to 'the context of culture' (Scott & Tribble 2006: 9). While the key advantage of integrating computational tools in the study of language and literature is the capacity for analysing large volumes, often millions of words, relatively quickly, they do so by removing structural information and flattening the data into a single textual artefact (Flowerdew 2004; Mautner 2012). For research that includes comparing 'different elements of the data' (McIntyre & Walker 2019: 87), such as the beginning or ending of a chapter in a novel, or 'different textual entities' (McIntyre & Walker 2019: 87), such as dialogue or narrative, a researcher will typically annotate the corpus with structural metadata to identify these features (Baker 2006: 38–42).

Some corpus analysis applications can interpret metadata if it follows general encoding standards,[33] so that the annotated features can be 'quantified and extracted for further analysis' (McIntyre & Walker 2019: 90). The popular CLAWS part-of-speech (POS) tagger, for example, will ignore encoded structural markers if accessed through the Wmatrix interface (Rayson 2009a), [34] and annotate sentence breaks to help researchers interpret the tagged output. This behaviour is helpful if working with annotated corpora, as the external metadata will not be included in the tagged output, but it leaves the researcher unable to filter the annotated results using their original markup. The POS frequency results, for instance, will report tag frequency across the whole corpus, ignoring any structural metadata, limiting

[33] See Baker (2006), Chapter 2, for a general introduction to Standard Generalized Markup Language (SGML) encoding principles, and McIntyre & Walker (2019), Chapter 3, for an overview of common XML (eXtensible Mark-up Language) tags for corpus stylistics.

[34] Weisser's (2014) 'The Simple Corpus Tool' is a useful example of a corpus analysis tool compatible with XML markup, while WordSmith's KeyWords tool (Scott 1998) can identify keywords at paragraph or section level, depending on the markup of the corpus (Scott 2010b: 47).

the scope of the analysis to a 'somewhat atomized, bottom-up type of investigation of the corpus data' (Flowerdew 2005: 324).[35] This process is incompatible with corpus stylistic research that examines the 'style' of a text through 'frequencies of linguistic items in a given context, and thus with *contextual* probabilities' (Enkvist, 1964: 29, cited in Culpeper 2014a: 10). To analyse the style of specific texts or parts of a corpus, the researcher would have to manually extract any sections for further analysis (Mautner 2012). For corpora that are composed of many smaller texts, such as a corpus of poetry, this limitation could significantly impact on what corpus methods could be used to investigate the style of individual poems or works by different authors contained in the corpus.

These concerns are shared by researchers in other areas of language analysis who rely on contextual information to inform their work: research into different language genres, 'where the discourse functions of lexico-grammatical items are examined within different sections of a text' (Flowerdew 2004: 15); corpus studies of specialised language, where 'an important concept may be defined only in the opening paragraph' (Bowker & Pearson 2002: 49), and the location of key 'concepts, terms, patterns and contexts' in a text is significant; and the use of corpora in English for Academic Purposes (EAP) instruction, where beginner students benefit from 'structural level' information, 'such as with audience analysis and/or with the organization of the paper' (Lee & Swales 2006: 57). Corpus analysis tools that allow a researcher to explore different contextual scopes are therefore useful in a variety of research applications beyond corpus stylistics.

2.5.2 The impact of decontextualisation in semantic analysis

For corpus stylistic research that requires semantic annotation, the issue of decontextualisation is even more significant: semantic annotation tools like USAS and the HTST are not trained to identify different contextual boundaries during annotation, and would disambiguate meanings in a corpus of collected texts based on their likelihood across the whole corpus (Piao et al. 2005a; 2017). A corpus stylistics researcher investigating distribution of meaning in literary texts from a specific period, for example, would be able to use these tools to semantically annotate the words in a collection of period literature as if it was part of one

[35] Further information available at http://ucrel.lancs.ac.uk/wmatrix/tutorial/

continuous text (see for example Alexander et al. 2015a), or they would have to have to individually process each text to retrieve semantic information based solely on the context of that work (Alexander et al. 2015b).

As such, corpus stylistic research has so far utilised semantic annotation for individual novels (Alexander et al. 2015b), plays (Culpeper 2014b), or short collections of semantically related plays and poetry (Archer et al. 2006; McIntyre & Walker 2010). These corpus investigations of literature demonstrated the value of semantic annotation tools like USAS and the HTST in stylistic analysis, but they were carried out on small, specialised corpora that could be annotated as a whole, or in smaller sections. Extending these approaches to a corpus that contains a range of varied literary texts would not be possible without either reducing the contextual accuracy of the taggers, or requiring the researcher to manually split the corpus based on the contextual scope they want to investigate.[36] This restricts the application of existing semantic annotation tools in research that investigates stylistic features in collections of shorter texts.

Alexander et al.'s (2015b) use of the HTST tagger in the analysis of metaphor in popular science texts, for example, acknowledged aim of analysing semantic 'metadata about large-scale collections of information in a way that does not require detailed and time-consuming research on individual texts' (Alexander et al. 2015b: i22), but restricted their 'proof of concept' to 'two popular science texts'. This restriction was necessary for the second stage of their investigation, where each text was divided into smaller sections, to allow for 'visual identification of the portions of a text which made extensive use of a particular domain' (Alexander et al. 2015b: i23). This allowed the researchers to investigate the distribution of domains across the corpus (Alexander et al. 2015b), but it required external post-processing of the HTST results by the research team. Furthermore, the sections of the corpus were determined by length, of 'approximately 500 words per unit' (Alexander et al. 2015b: i23), instead of by paragraph or chapter, as these scopes are not identified by the HTST. As such, while the study demonstrated the potential of semantic annotation for investigating domain distribution across a

[36] For instance, to determine whether metaphorical usage decreases in the novel *One Flew over the Cuckoo's Nest* as it progresses, Koller et al. (2008) used Wmatrix to semantically annotate 'the whole of the novel, the novel's first half, and the novel's second half', comparing the results in post-processing (Koller et al. 2008: 155).

corpus, it would be difficult to scale up their approach to corpora that contain multiple different texts.

A further consideration of decontextualised semantic annotation is that the disambiguation approach cannot be altered for different contextual scopes: while researchers are able to cross-examine USAS semantic tags by combining the results of tagged texts using the 'CrossTab' feature in Wmatrix (Rayson 2013), they are not able to assign tags to different scopes within a text using the interface. The planned addition of the HTST tagger into the Wmatrix application will similarly extend the 'corpus indexing and retrieval functions' of the tagger (Alexander et al. 2015b: i20), but not the annotation parameters.[37] This prevents investigation of meaning *development* based on different contexts, such as in a comparison of an author's rhetorical choices in relation to their collected work, or in contrast to the work of their peers or even contemporary texts.[38] Furthermore, the meaning of different words in a text might change as it progresses (O'Halloran 2007), which could only be explored if the semantic tagger can distinguish between different contextual scopes. Similarly, the development of meaning in a text, or what Hoey (2012) refers to as semantic priming, would require a context-aware semantic annotation approach.

The disambiguation methods of the USAS and HTST taggers also restrict their application in corpus stylistic analysis of ambiguity in language, as both systems were trained to identify the most likely meaning of individual words in a text (Piao et al. 2005a; 2017). This does not mean that the taggers are incapable of identifying non-literal meanings, but their capacity for doing so is restricted by their pre-determined disambiguation parameters. USAS 'treats metaphorical items as polysemous words' (Alexander & Bramwell 2014: 3), recording metaphorical senses of a word as separate entries in its general semantic lexicon. Koller et al. (2008) explored the potential of this lexicon for identifying metaphors in corpora of different genres, using the 'first choice and secondary tags' assigned by the USAS tagger (Koller et al. 2008: 144). Their research looked at whether conventional metaphors were annotated with their target domain as the most

[37] It is, however, possible to assign POS-tags to different sections of a corpus using WordSmith Tools (Scott 1998), making it 'well suited' to corpus stylistic enquiry that does not require semantic metadata (Culpeper 2014a: 13).

[38] O'Halloran (2014), for example, uses corpus stylistics methods to explore Robert Frost's poem 'Putting in the Seed' (1916) in relation to 'contemporary knowledge' (O'Halloran 2014: 150).

likely sense, and the source domain as a secondary meaning; *campaign*, for example, was identified with the 'first choice tag X7 ("wanting, planning , choosing")' (Koller et al. 2008: 155), while 'the source domain is represented by the last tag in the string, G3 ("warfare")'. Using this data, the authors hypothesised that 'novel or less conventional metaphoric expressions' would be tagged with the source domain as the 'first choice or even only tag' (Koller et al. 2008: 155), as with *e-campaign*, which was 'allocated only the source domain tag, together with the secondary tag Y2 ("information technology and computing")'. As the researchers were familiar with the corpus data, they used the USAS tagger to investigate metaphorical usage of specific domains, using a purpose-built search function to complement the Wmatrix filter interface (Koller et al. 2008: 154).[39] While the results varied across different corpora, they were able to demonstrate that the USAS semantic lexicon could identify the figurative use of a word 'where the metaphoric meaning is established as predominant' (Koller et al. 2008: 146). In reporting on their process, however, Koller et al. (2008) acknowledged that further work is necessary for supporting a 'corpus-based methodology for the investigation of metaphor in large-scale data sets' (Koller et al. 2008: 158). Additionally, the authors noted that 'not all domains manually identified and labelled by a researcher are actually reflected in the USAS tag set' (Koller et al. 2008: 154), as the lexicon was not designed to account for granular sense distinctions. As such, while their research confirmed the potential benefits of the USAS tagger in metaphor analysis, it also highlighted the restrictions of its semantic lexicon in discovering specialised or atypical metaphorical expressions.

In this regard the HTST is more flexible, as it relies on the HTE taxonomy that includes 'all the meanings recorded in the history of English' (Alexander & Bramwell 2014: 4), and the words that 'have been used to instantiate these meanings'. Using this taxonomy, Alexander et al. (2015b) demonstrated how the HTST could be used to identify metaphorical language in a text through the presence of semantic 'domains which are not directly relevant to the subject matter of the text' (Alexander et al. 2015b: i22). However, the HTST was developed to

[39] This function was described as a 'broad sweep search', which scanned 'the full list of possible semantic tags on each word in the text' (Koller et al. 2008: 154). It was implemented for this project as the Wmatrix interface restricts the search to the most likely semantic tag, requiring manual cross-referencing to investigate secondary tags in the corpus. Unfortunately, this does not appear to be an added function in the general Wmatrix interface, as it is not included in the list of updates for the resource (Rayson 2013).

identify 'correct' senses (Alexander et al. 2015b: i19), and applies a fixed limit to the number of candidate meanings retrieved from the HTE. As it is not possible to alter the disambiguation parameters of the HTST tagger, the user is restricted in their access to the HTE taxonomy in the annotation process. This design limits the HTST's use in the discovery and interpretation of idiosyncratic rhetorical features that could be present in literary texts, such as 'novel metaphors', and 'irony, sarcasm, and allegory' (Castiglione 2017: 106). Similarly, in restricting the meanings of 'highly polysemous' words (Alexander et al. 2015b: i22), both taggers prevent the discovery of atypical usage of those lexical and grammatical items (Furkó 2019).

The way that a semantic tagger handles ambiguity is particularly relevant in literary texts that employ the 'deliberate exploitation of linguistic ambiguity' (Gerbig & Müller-Wood 2002: 76), making them more challenging to annotate with semantic metadata. While ambiguities in general language can typically be resolved, the ambiguities in literature will often allow multiple interpretations of a word. This is particularly true of poetry, where 'ambiguities are frequently brought to the reader's attention, and the simultaneous awareness of more than one interpretation is used for artistic effect' (Leech 1969: 207). Leech (1969) explains that a reader will 'recognize and tolerate more ambiguity in poetry' because they are 'attuned to the acceptance of deviant usages and interpretations' in poetic language (Leech 1969: 207); a semantic tagger for literary language must therefore account for intentional ambiguity in its disambiguation process.

2.5.3 Semantic annotation of a corpus of poetry

A corpus of poetry is therefore a useful starting point for testing a flexible semantic annotation approach, as it can be used to highlight the annotation and disambiguation parameters of figurative language and a context-dependent corpus. Accommodating these parameters within a semantic annotation system would in turn address a key critique of computational analysis of literature: that a restrictive categorisation of meaning ignores the '*transient* nature' of literature (van Peer 1989: 302). While the development of the USAS and HTST semantic taggers have expanded the scope of semantic annotation, van Peer's (1989) assessment that 'there is still no computer program to automatically disambiguate figurative language as such (apart from dead metaphors or other kinds of 'frozen'

figurative meanings)' is still relevant today (van Peer 1989: 303). The interconnected disambiguation process of both taggers restricts their application in exploring figurative language, as they prevent the researcher from altering the way in which they attribute meaning in a corpus.

The decontextualisation of USAS and HTST has similarly restricted the scope of prior analyses of meaning in poetry, as neither tagger can be used to annotate individual poems in a corpus with context-dependent senses. McIntyre and Walker (2010) circumvented this restriction in their corpus stylistic analysis of William Blake's *Songs of Innocence* (SoI) (1789) and *Songs of Experience* (SoE) (1794) by annotating both collections of poems as two separate files (McIntyre & Walker 2010: 517). Using the USAS semantic lexicon and Wmatrix analysis tools, they were able to investigate the key semantic domains in both collections, both in relation to each other, and through the use of a reference corpus. The small number of key domains allowed the authors to assert that 'lexically and semantically the texts are actually quite similar' (McIntyre & Walker 2010: 517), as USAS categorised both collections with similar meaning profiles. They did identify contrasting domains in the collections, with 'HAPPY' being a key domain for SoI (McIntyre & Walker 2010: 517), and 'FEAR/SHOCK' and 'VIOLENT/ANGRY' appearing significantly in SoE. By comparing the words that corresponded to key semantic domains with the keywords in both collections, the authors illustrated why 'key words on their own are not enough to capture certain important differences between texts' (McIntyre & Walker 2010: 519), as the key words analysis failed to identify this semantic contrast through lexical items alone.

However, while the authors acknowledged that the keywords in the text corresponded to 'specific poems in Songs and to specific elements within those poems' (McIntyre & Walker 2010: 521), they did not inspect the key semantic domains for this phenomenon. A possible reason for this omission is the difficulty of attributing the semantic data to individual poems in the text, as there is no way to limit the contextual scope of the results. To enable this type of enquiry, the semantic annotation process must be able to distinguish between different elements, or scopes, within a corpus. A diachronic corpus containing poems by multiple authors can be explored across a range of contextual scopes, at the level of individual sentences, stanzas, and poems, or as part of a broader classification that includes words by one poet, or poems belonging to a specific period, genre, or

school. As such, a corpus of poetry can be used to illustrate the requirements of different contextual scopes when training a semantic tagger. In turn, a semantic annotation approach that can distinguish between different texts within a corpus could be applied to a range of text types, such as internet texts and fragments (Scott 2010b), or collections of shorter literary texts, as with a corpus of poetry, thus expanding the scope of semantic analysis of corpora to a range of texts that are currently incompatible with existing tools.

2.6 Conclusion

The burgeoning research activity in corpus semantics and related disciplinary approaches to analysis of meaning in corpora confirms the demand for methodological and analytical frameworks that support this enquiry. The innovative use of collocation and keyness analysis outlined in sections §2.3.2 and §2.3.3 above shows creative adaptation of existing corpus methods for the purpose of extracting meaning from corpora. The range of new methods for semantic analysis of corpora discussed in §2.4.1 confirms that interest in this field extends beyond corpus linguistics, while also highlighting the broad scope of approaches developed for this purpose. The design of new tools for semantic annotation of corpora allowed non-specialist researchers to explore the semantic properties of their corpora, corresponding to the largest leap forward in corpus semantic research. The development of the USAS (§2.4.3) and HTST (§2.4.5) semantic taggers significantly expanded the corpus linguistics toolkit, allowing new forms of corpus analysis and stimulating research activity beyond corpus semantics. By using the USAS and HTST, researchers demonstrated the benefits of semantic annotation in exploring large collections of textual data and the advantage of using semantic metadata in exploring meaning in a range of textual genres. In exploring these studies, however, it was also possible to identify key limitations of USAS and the HTST (§2.4.6), as the design of the taggers made them less suited to handling non-standard corpora, figurative or ambiguous language, and corpora that contains varied and semantically distinct texts.

Together, these limitations represent a substantial barrier for corpus semantics research, as they restrict how researchers are able to employ semantic annotation in their work. The impact of this can be seen by reviewing the use of both taggers

in another emerging area of corpus linguistics: corpus stylistics. The popularity of semantic annotation in this field establishes the demand for semantic annotation in stylistic analysis of literary corpora, with research demonstrating the benefits of semantic metadata in the analysis of literary style. At the same time, the scope of these experimental projects reifies the restrictions of the two available semantic taggers, USAS and HTST (§2.5.2). As the taggers do not distinguish between different texts in a corpus, they are less suited when working with corpora of collected texts and in analysis of meaning at different contextual scopes, as the researcher must choose between annotating each text individually or decontextualising the corpus. The analysis of figurative or ambiguous meaning is similarly restricted, as it is not possible to alter the disambiguation parameters or the breadth of semantic metadata produced by either tagger, despite research indicating that the different tags attributed to individual lexical items in a corpus can be used to explore abstraction in language (Koller et al. 2008). As a whole, the observed strengths and limitations of current taggers answer the first research question of this thesis (RQ1): What are the barriers to using existing semantic annotation tools and methods in analysing meaning in non-standard corpora?

Concurrently, these limitations represent an opportunity for expansion, as they can be used to define the requirements of an alternative, complementary approach to semantic annotation that overcomes these barriers. Usefully, a diachronic corpus of collected poetry incorporates every element that was identified as posing a challenge to the USAS and HTST taggers (§2.5.3), thus making it a well-suited starting point for designing a flexible approach to semantic annotation. Accordingly, while a flexible semantic annotation tool would expand the scope of corpus semantic analysis in any research areas that similarly face restrictions of existing approaches, a corpus of poetry is useful in guiding the design of this tool. These design considerations are explored in greater detail in the following chapter, which establishes the groundwork for the flexible semantic annotation method proposed by this thesis.

Chapter 3 Groundwork

3.1 Introduction

This chapter builds on the findings of the literature review by discussing the requirements of a semantic annotation method that can address the limitations of existing taggers. Consequently, to bridge the gap between previous semantic annotation approaches (as discussed in Chapter 2) and the alternative method described in the following chapter (Chapter 4), this chapter establishes the evidentiary groundwork for the semantic tagger proposed in this thesis. Thus, while the literature review answered the first research question of this thesis (RQ1) by identifying remaining barriers to semantic annotation, this chapter addresses the second research question (RQ2): What are the practical design parameters for overcoming these barriers?

As identified in the previous chapter, the only currently available tools for semantic annotation that do not require specialist knowledge are the USAS and HTST taggers, which allow users to annotate corpora with general semantic labels (USAS) and more granular descriptions (HTST).[40] The HTST also allows users to retrieve period-appropriate meanings during the annotation process, through the historical semantic taxonomy of the HTE. While both taggers can be used to explore meaning in corpora in a variety of research situations, their application is still curtailed by two key factors: users are not able to alter their disambiguation parameters, reducing their use in the analysis of figurative language, and users are restricted in the type of corpora they can annotate, as the taggers are not able to distinguish between different contextual elements in a corpus, which restricts the analytical scope of semantic annotation.

To demonstrate a viable method of overcoming these restrictions, this chapter will first address the issue of the disambiguation parameters (§3.2), showing how a flexible approach to identifying meaning in a corpus can be used to annotate figurative language. Following this, the significance of annotating corpora at

[40] The USAS and HTST taggers are described in sections §2.4.3 and §2.4.5 respectively.

different contextual scopes is highlighted in reference to key analytical parameters (§3.3). These parameters inform the case study analyses of the diachronic corpus of poetry created for this research. This corpus was used to further assess the requirements of a context-dependent flexible semantic annotation method and informed the development of the tagger. As such, the final set of groundwork parameters discussed in this chapter outline the corpus design considerations (§3.4) of the test corpus created for this project.

3.2 Disambiguation parameters

The USAS and HTST taggers disambiguate senses of words using a hybrid system that combines a template approach for annotating polysemous words and multi-word-expressions, and probability calculations to determine the likelihood of specific senses based on a lexico-grammatical ruleset. Using these disambiguation parameters, USAS reports an average accuracy score of 91.05% when annotating general contemporary texts (Rayson et al. 2004b), while the HTST reported an average accuracy of 81.61% across a range of genres and periods. While these are encouraging results, the developers of both systems concede that the remaining limitation of these taggers is a lack of context-dependent disambiguation.[41] Users of USAS and the HTST have similarly called for a context-aware disambiguation method to complement the hybrid approach already in use by these taggers (Löfberg et al. 2004; Archer 2014; Archer & Malory 2017; Furkó 2019). Consequently, context-dependent disambiguation is the first parameter that must be addressed in an alternative approach to semantic annotation.

3.2.1 Disambiguation with the HTE

The disambiguation process developed for this research builds on Sinclair's (1994) hypothesis that if 'successive meanings can be discerned in the text', then it is possible to use this information to 'associate a meaning or a component of meaning or a shade of meaning with this or that word or phrase that is present in the text' (Sinclair 1994: 22). In other words, by identifying the semantic properties

[41] In evaluating the HTST, Piao et al. (2017) recommend that 'more efficient context-based disambiguation algorithms' are implemented in future iterations of the tagger (Piao et al. 2017: 129). The implementation of 'local probabilistic disambiguation' into the USAS disambiguation process is still in development, as the tagger currently relies on a template of 'contextual rules' (Rayson et al. 2004b: 4).

of the corpus, the specific sense of a word can be determined based on its relationship to the senses identified in the rest of the text. Thus, the process relies on contextual information to determine the most likely meanings of a specific word, based on the contextual scope determined by the researcher. By expanding the range within which 'successive meanings' are recorded, the range of meanings is narrowed through association. This concept can be illustrated by borrowing Stubbs' (2001: 14) example of 'the supermarket is opposite the bank', where the meaning of *bank* is uncertain from the immediate context. Stubbs' argued that by identifying the semantic fields of words that occur in proximity to *bank* ('co-text'), (e.g. '*cashier*, *deposit*', or '*cave*, *cod*'), it could be possible to narrow the specific sense of *bank* (Stubbs 2001: 15). However, this approach is only possible if it makes use of a classification system capable of identifying all possible senses in the text. The only semantic taxonomy that categorises all recorded meanings in a hierarchical structure is the HTE, making it ideally suited to this task.

3.2.2 HTE taxonomy

Any designed system of classification will reflect subjective decisions on behalf of the research team. The HTE began with the goal of developing a 'conceptual thesaurus' of the *Oxford English Dictionary* (OED) (Kay 2011: 45). In the first instance, the categorisation was built on existing research and also quite literally manual, with paper slips used to record and categorise OED items. Categorisation aimed to reflect usage, with the acknowledgement of the distinction between 'folk' categories that develop though usage, and 'scientific' or expert categories that we use to classify the world around us (Alexander & Kay 2019a). With no precedent, the HTE team developed their own classification system, expanding substantially beyond Roget's (1852) classification and encapsulating a hierarchical structure of 'conceptual fields', identified as 'the domain of experience where the word was likely to be used' based on available records (Kay 2011: 46). The editors called this a 'modified folk taxonomy', striving to meet the 'intelligent average individual's view of the world' where possible (Ullmann 1962: 255 in Kay 2011: 51), but deferring to the 'established scientific taxonomy' for categories where an expert taxonomy was preferable (Kay 2011: 51).

The HTE data is not static: the dataset continues to be revised and is maintained by the HTE team at the University of Glasgow, who released the second edition of

the thesaurus in 2020 (Alexander & Kay 2021a). Version 4.2.2 was used for this research, and includes all major revisions to the first edition of the HTE database (Alexander & Kay 2019b). The scope of the resource makes it a clear benefit to semantic analysis of corpora: the live HTE website defines the 'fine-grained conceptual hierarchy' (Alexander & Kay 2019a) as comprised of 'semantic categories', which hold 'almost a quarter of a million concepts'. Furthermore, the unique diachronic taxonomy makes it possible to study texts through the context of 'the options available to a writer to realise their conceptualisations of the world' (Alexander & Struan 2013: 233), while the inclusion of hyponyms in the classification system used by the HTE makes it more applicable for use in disambiguating concepts through contextual information (the appearance of similar themes in close proximity to the text). This is further assisted by the hierarchical structure of the HTE, which 'begins with the most general ways of expressing a concept and moves hierarchically downwards to the most specific' (Alexander & Kay 2019a), allowing for investigation of meaning at different layers of abstraction by selecting different category levels for analysis.

The categories of the HTE are not equal in size: of the three, the first 01 (The World) is the largest, containing 121,032 categories spread to a maximum depth of 12 tiers, reflecting the scope of the category, which contains the words we use 'to describe the physical universe, the creatures living in it, and the operations of the human beings upon it.' (Kay 2011: 47). [42] Further exploration of the HTE taxonomy also reveals societal changes, such as the fact that the third category 03 (Society) contains 'the largest number of categories', and is seen to reflect 'the expanding vocabulary denoting families, government, law, manufacture, trade, communications, and so on.' (Kay 2011: 48).

3.2.3 HTST thematic categories

While substantially expansive for the purpose of annotating senses, the HTE taxonomy is too fine-grained to be used for researcher-led investigation, as it produces a prohibitively large dataset when senses are recorded across all levels of the taxonomy. An illustrative example is the difference between the classification of concrete objects, which 'lend themselves to detailed classification by features

[42] Visual representations of these changes are available at: https://ht.ac.uk/treemaps/

such as “type of” or “part of”’, and abstract concepts, which ‘rarely require the full 12-place taxonomy’ (Kay 2011: 52). A straightforward solution for the issue of interpretation is to truncate the classification at the third level, containing 354 categories, which then expand to ‘a further 236,400 categories and subcategories’ (Kay 2011: 50), and represent a ‘conceptually coherent […] level at which categories are most salient to users of the language’ (Kay 2011: 50). The hierarchical classification makes this possible, as the semantic headings are numbered sequentially, and could be abridged at any level if necessary. However, while this approach is still supported by the tagger’s current output, the tagger also utilises the alternative set of categories, which were purposefully developed for semantic annotation with the HTE: ‘the thematic category set’ (Alexander et al. 2015a: 9). Designed by the team working on HTST, ‘the thematic category set’, also referred to as ‘thematic headings’ (Alexander & Kay 2021b), represent ‘a significantly reduced set of headings for which a researcher may wish to search’ (Alexander et al. 2015a: 9). The rationale for its development was to ‘aid analysis’ (Alexander et al. 2015a: 9) of annotated data, thus solving the problem of how to manually interpret the results of the tagger.

The success of the thematic headings as a ‘human-scale’ version of the HTE taxonomy lies in their descriptive nature; that is, they relate to familiar concepts that are still discrete in relation to each other (Piao et al. 2017: 116). It is not difficult to accept that *AE* (Animals) and *AC* (People) represent different classification groups of the world around us but identifying the degree of separation between the two is less straightforward. The distinct properties of *AR* (The mind) as opposed to *AS* (Attention, judgement, curiosity), *AU* (Emotion), and *AV* (Will) are even harder to differentiate, even if the headings themselves represent recognisable concepts. The broader domains enabled the development of the contextual disambiguation process used for the ambiguity tagger, which took the higher-level conceptual groups as representing different levels of ‘generality in concept relatedness’ (Wilson & Thomas 1997: 57), and allowed the analysis of the results to be carried out at multiple levels of abstraction. In this manner the disambiguation method differs from the ‘polyseme density’ technique proposed by the team working on the HTST, which weighs the likelihood of a particular sense based on the number of recorded senses in the HTE within a sub-category, rather than the related senses in the corpus itself (Alexander et al. 2015b: i21).

3.2.4 Disambiguating figurative language

By utilising the hierarchical taxonomy of the HTE to assign senses to words in the corpus, this method of semantic annotation gives users greater control over the disambiguation process and the semantic metadata produced by the tagger. As the tagger attributes semantic tags by calculating their contextual relevance, the researcher has control over the cut-off point for the number of tags attributed to each item, based on the confidence score reported by the tagger. Thus, the user decides if they want to retrieve a narrow selection of most likely senses, or a broader range of senses that are contextually relevant and could indicate metaphorical usage of a particular word.

The potential of auxiliary semantic metadata in identifying metaphor was illustrated in Koller et al.'s (2008) analysis of USAS semantic metadata, as the tagger identified both source and target domains to highlight popular metaphorical expressions.[43] This research also highlighted two key limitations: the USAS semantic lexicon was not sufficiently granular to capture all instances of metaphorical expression, and the annotation process and semantic output produced by USAS restricted the scale and scope of the analysis. The HTE taxonomy can be used to overcome the first of these limitations, as Alexander & Bramwell (2014) and Alexander et al. (2015b) have demonstrated the superior capability of the HTE in identifying all possible senses of words based on their recorded usage. The issue of scale is addressed in the design of the new semantic tagger, which can annotate a corpus containing multiple different texts along different contextual scopes, making it possible to explore metaphorical expressions across different boundaries in a corpus (see §3.3 below). Finally, by allowing researchers to configure the disambiguation and annotation parameters to suit the requirements of their corpus and type of enquiry, this semantic annotation method increases the scope of analysis beyond what is currently possible, thus addressing the final limitation.

It is still essential, however, that meanings are disambiguated to exclude irrelevant senses from further analysis; to annotate figurative language, a tagger must allow for different interpretations of the text, but it must also systematically remove

[43] Discussed in §2.5.2 above.

irrelevant noise from the semantic metadata for it to be of any value to a researcher. Thus, to annotate the corpus of poetry developed for this project, the tagger must retrieve only those senses that can usefully aid the analysis of meaning in the text. Correspondingly, it must also allow the researcher to determine what is considered useful in relation to their research, based on the interpretative approach they are using to analyse the text. To understand the role of the tagger in this process, it is useful to consider the following advice from Leech (1969):

> A poem offers a vast number of interpretative possibilities; some are simply theoretical possibilities which would rarely, if ever, occur to an actual reader; others are more plausible. The subjective element enters when the reader selects from this array of possibilities that interpretation, or those interpretations, which suit him best. The role of linguistics is to help us to study what possibilities exist; the role of the literary commentator, it may be suggested, is to evaluate the various possibilities, and to arrive at an informed and authoritative interpretation by rejecting some and accepting others (Leech 1969: 215).

In response to Leech (1969), this thesis proposes that the semantic tagger should fulfil the 'role of linguistics' (Leech 1969: 215), providing the researcher with the means to 'study what possibilities exist' for interpreting the text. Moreover, a flexible semantic tagger should enable the researcher to select the scope of 'interpretative possibilities' based on the requirements of the research (Leech 1969: 215). While the USAS and HTST taggers can be seen as fulfilling the first of these requirements, they cannot be used to annotate corpora to suit a particular set of interpretative parameters. The flexible semantic annotation approach demonstrated in this thesis can be used for this purpose, proving a complementary alternative to the USAS and HTST taggers.

3.3 Analytical parameters

However, it is not enough to identify the appearance of senses in a text, and employing computational methods in this task does not absolve the researcher of interpretative responsibility. Traditionally, the researcher is still responsible for identifying 'significance' in the computer's findings (Stubbs 2001: 143). Bringing the above together, these changing paradigms led some linguists to note the blending within CL of 'the use of computational, and consequently algorithmic and statistical, methods on the one hand, and the qualitative change of the

observations that derive from this approach on the other' (Tognini-Bonelli 2001: 1). An attempt by Heuser and Le-Khac (2011; 2012) to reconcile what they refer to as the 'signal' (data) and the 'concept' (what the data represents) placed reasonable emphasis on the need for a robust methodology and comprehensive testing of the results. Yet, in addition to these, the authors note that the 'same careful attention to nuance and complexity that humanists have developed in close reading texts pays dividends when close reading data.' (Heuser & Le-Khac 2012: 48). The challenge of translating quantitative data into meaningful results is at the forefront of the research. Yet in order to produce meaningful discoveries through innovative methodologies, instead of simply supporting existing scholarship, it is not enough to rigorously test the methods; the results must live up to the same degree of scrutiny found in other areas of the humanities.

A clear example of this pursuit for new knowledge is the breadth of corpus linguistics research that re-examines Shakespeare's work, which, despite being extensively studied through more traditional methodologies, remains open for investigation due to the volume of work attributed to the author (Archer et al. 2006: 1). However, Archer et al. (2006) note that Culpeper (2002) found that dominant keywords associated with a character can be skewed by a key event, as in the case of a surprising amount of 'surge features' (Culpeper 2002: 21), or 'outbursts of emotion', associated with Juliet's nurse being the result of the character's one-off reaction to a uniquely traumatising event instead of representing a 'character trait' (Archer et al. 2006: 2). The cautionary advice given by the authors is that keywords should be contextualised manually and reviewed in relation to the original text, though they note that this advice is 'a point often made but not always carried out convincingly.' (Archer et al. 2006: 2).

While statistical measures typically associated with keyness in CL have come under recent criticism due to the high volume of 'significant' results they return, the cause is attributed to using reference corpora that are not themselves randomly sampled (Bestgen 2018: 37). That is, 'the presence of some very specific texts' in even larger corpora could disrupt the measure to the extent that the significance reported could be traced to one single entry (Bestgen 2018: 37–38). This is a valid critique and must be taken in consideration when looking at comparing frequency in corpora, but it assumes that the measure is used to identify salient features in the language that could be replicated if applied to a different sample that meets the

same criteria. In other words, if the observed frequency of a word in a text is seen as significant because it is disproportionate to the expected frequency based on the reference, it should be possible to see the same proportionate frequency in a different text if it has the same features as the first (Bestgen 2018). However, this is only a flaw if the research question extends beyond the text being measured. Consequently, while keyness measures are insufficient for formulating arguments about language use at scale, they are well suited to identifying significant features of a particular text.

3.4 Corpus design parameters

The *Oxford Book of English Verse* (OBEV; Quiller-Couch 1919/1999) was selected as the source material for the corpus created for this research. In addition to being in the public domain, the anthology was selected as it represented a range of literary periods, providing a useful reference for the diachronic element of the analysis. The first edition sold 'over half a million copies' (Bassnett 2001: 255) through a series of regular reprints in the time between the 1900 publication and the 1939 second edition. It remained the seminal reference until Christopher Ricks' (1999) edition, unsurpassed by even Gardner's (1972) revised edition. Indeed, the editions of the anthology reflect changes in literary tradition but also cultural shifts in the same way that the HTE taxonomy became reflective of language change.

Despite the developments in corpora size, we are still limited by the volume of text that is available from pre-digital periods. Thus, while an ideal analysis would employ what Fowler (1979) defined as the 'potential' canon which 'comprises the entire written corpus, together with all surviving oral literature', a more realistic approach is utilising the 'accessible' canon which acknowledges the limits of accessibility and restriction (Fowler 1979: 98). A further limit acknowledged by Fowler is that of the 'official' canon, defined as a 'sizeable subset of the writers and works of the past', delimited by the limits set through 'education, patronage, and journalism' and enforced by tradition (Fowler 1979: 97–98). Otherwise referred to as a 'selective' canon, its members are both recorded in and defined by 'anthologies, syllabi and reviewer's choices' (Harris 1991: 112).

What makes canonical texts particularly suited to this analysis is the wide-ranging critical analysis which has already been conducted on their work. This would allow for the conclusions to be tested in relation to more traditional methods of critical analysis. If this is proven to be viable, then it might be possible to draw relationships and parallels between the works of authors which have not been studied as closely, but which might have still influenced those that followed them. This, in turn, could open up new debates within literary criticism.

3.5 Conclusion

This chapter establishes the justification for the disambiguation (§3.2), analytical (§3.3), and corpus design (§3.4) parameters used in this thesis. By identifying these parameters, it answers the second research question of this thesis (RQ2), showing the design considerations for a semantic annotation method that can overcome the main barriers of existing approaches. In doing this, it examines the extent to which this method can add to our existing knowledge of the source text, showing that interpretation and analysis of semantic metadata are essential to developing semantic annotation further. Consequently, in addition to providing the theoretical groundwork for the flexible semantic annotation method described in the following chapter (Chapter 4), this chapter establishes the importance of the analytical evaluation of the method in Chapter 5, Chapter 6, and Chapter 7.

Chapter 4 Methodology

4.1 Introduction

Previous chapters surveyed recent achievements in developing tools for semantic annotation of corpora, which exemplify the broad range of applications for the process in research and highlight the demand for robust methodologies and applications that enable large-scale annotation. Crucially, these developments are not restricted to one discipline, with semantic annotation initiatives appearing in different forms across a range of fields, creating a rich foundation of experimental approaches that evolves with every advancement. The methodology presented in this chapter marks a further contribution to this field by addressing the gap of semantic annotation approaches for corpora of poetry. This approach builds on existing semantic annotation methodologies, most notably the recent work in utilising the HTE in the annotation process, to enable an 'ambiguous' semantic annotation system that is compatible with corpora of poetry.

Elements of the methodology are technical in nature, so the annotation process is described with enough precision to allow for reproducibility. This chapter is therefore split into the following sections: §4.2 Preparing the corpus, which sets out the approach used for cleaning the OBEV text and splitting the anthology files into four date-delimited corpus groups; §4.3 Preparing the database, which captures the creation of the HTE and corpus group database tables; sections §4.4 Combining the corpus and the HTE, §4.5 Filtering the data and §4.6 Disambiguation, which include the cross-referencing process, the steps taken to filter, sort and otherwise manipulate the result, and the development of a 'readable' version of the tagged output; and §4.7 Accuracy, which sets out the parameters for evaluating the tagger. A summary of the methods is provided in §4.8, concluding the chapter and providing a quick reference for the analysis chapters.

The results of the semantic annotation process introduced in this chapter are evaluated through a macro-level analysis in Chapter 5, which looks at the aggregated semantic metadata for all four corpus groups, followed by a micro-level

analysis in Chapter 6, which examines the results at the level of an individual poem. A further approach to analysing the data is presented in Chapter 7, where the data is further processed to identify frequently co-occurring semantic tags in the corpus, thus enabling the investigation of 'semantic collocation'. To facilitate this approach, Chapter 7 further expands the methods established in this chapter through additional post-processing of the annotated data.

4.2 Preparing the corpus

In contrast to many corpus stylistics projects, which often begin with the corpus and then identify appropriate methods for investigating it, or look towards corpora as a tool for answering research questions that were formed independently or exist as part of disciplinary dialogue, the corpus used in this research was created for the purpose of developing and testing the methodology (McEnery & Wilson 1996: 101–103). This approach was chosen as it is suited to a proof-of-concept for an experimental approach, and the use of a single corpus meant that it was possible to examine the data in more detail during the analysis stage than would have been possible if further corpora were introduced to the study. The drawback of this approach is that it could limit the transferability of the semantic annotation process to other corpora, and so where possible steps were taken to address this limitation. In the first instance, it was necessary to identify the impact of any idiosyncratic features of the source material on the semantic annotation process and address any issues these may cause as part of the pre-processing stage. Subsequent obstacles could then be treated as part of the general configuration of the semantic annotation process, with the goal of creating a standardised approach that could be adapted for use with different corpora.

This objective meant that during the corpus creation process, the features of the corpus were examined in relation to the way they might interact with the semantic annotation process, instead of how they represent the source material. It was not necessary, for example, to consider editorial variation between digital versions of the OBEV, as this would not affect the semantic annotation approach. It was, however, important to determine any textual features that would interfere with the tagger, regardless of the initial source. The corpus had to be sufficiently broad as to encompass enough variation to expose potential issues, but small enough to allow

for different iterations of the process to be tested within a reasonable timeframe. To enable reproducibility, this section looks at the steps taken in preparing the corpus, with a particular focus on ensuring compatibility with the annotation process.

4.2.1 Cleaning the OBEV

A plain text version of the complete *Oxford Book of English Verse* (OBEV; Quiller-Couch 1998) was downloaded from Project Gutenberg for use as the corpus. The text was then edited manually, removing front and end matter, as well as editorial additions to the text.[44] The only extraneous information retained from the text was the number assigned to each poem in OBEV. These served as markers in the corpus, making it easier to select individual poems for further analysis but they were excluded from the disambiguation data and therefore had no impact on the semantic annotation process.[45] The decision was made to keep the titles of the poems in the corpus, which were included in the main semantic analysis and contribute to the overall thematic distribution in the corpus. The justification for this comes from the nature of the poem title, which is typically considered to 'say something *about* the poem' (Ferry 1996: 2–3) and thus offers further contextual information that could be used to disambiguate a poem's content. Crucially, while the title is often considered as a separate element to the poem, rather than part of the poem itself, including it in the analysis is a further methodological distinction that is made to accommodate the distinct features of verse as a source for CL. A further point that had to be considered at this stage was the practice, particularly during the early Renaissance period, of poem titles being 'given by someone other than the author' (Ferry 1996: 12), most frequently during the printing process. For consistency, and to reduce the volume of manual editing required, all titles were retained in the corpus, with the acknowledgement that further testing would be necessary to determine the impact of this decision.

Following the manual cleaning process, the remaining textual data was then submitted to a further series of text-replacements to prepare the corpus for

---

44 To assist the readers of his anthology, Quiller-Couch added glosses for 'archaic and otherwise difficult words' (Quiller-Couch 1998: viii) at the end of poems, which were manually removed at this stage.

45 Western Arabic numerals were used in the edition for poem numbers, making it easier to distinguish them from the text.

lemmatisation and part-of-speech (POS) tagging. It was necessary to lemmatise the corpus before cross-referencing with the HTE to make the data compatible with the headwords recorded in the thesaurus. The lemmatisation process achieves this by reducing 'the words of a corpus to their respective lexemes' and thus producing the headword that 'one would look up if one were looking for the word in a dictionary' (McEnery & Wilson 2001: 53). However, abnormal spelling and punctuation can impact on the accuracy of the lemmatiser, which in turn impacts on the accuracy of the semantic annotation process (Baron & Rayson 2009; Archer 2012). Of course, this presents more of a problem when working with corpora that contain non-standard spellings, as is the case with many of the poems recorded within the OBEV collection (Baron 2011). Earlier poems in the collection were recorded as accurately as possible, with the editor noting that in 'the very earliest poems inflection and spelling are structural, and to modernize is to destroy' (Quiller-Couch 1998: vii). This perspective meant that archaic variants were preserved in the earlier texts, though the editor took some liberties with 'a few small corrections' where an error was deemed 'obvious', and standardised certain spellings in later poems, where 'old spelling becomes less and less vital' (Quiller-Couch 1998: viii).

While, there are a number of options for cleaning textual data, these are often targeted at preparing a corpus for research that does not require the preservation of the original structure of the text, and therefore would strip out features that were to be retained for this analysis (Kübler & Zinsmeister 2015: 6–7). Punctuation, for example, is often removed during the cleaning process, but was to be retained in the creation of this corpus to serve as a marker in the close analysis of the tagged data. Scanning through several hundred rows of vertically annotated text is easier when there are markers separating them, particularly when these are recognisable and serve the same function as in the original text. Consequently, to reach a balance of maintaining as much of the original text as possible while maximising the number of possible matches, the OBEV text was first sanitised in accordance with the CLAWS input guidelines,[46] and then modified with the following text-specific alterations: double-hyphens were removed instead of being converted to *em* dashes, extra spaces before the possessive suffix "'s" were removed, orphan inverted commas were removed where unsuccessfully

[46] Available at http://ucrel.lancs.ac.uk/claws/format.html

tokenised,[47] and extra white space was removed from inside stanza boundaries. All three supplementary alterations were performed as a result of initial issues with the tagging process.

4.2.2 Splitting the OBEV into period groups

The next stage in the analysis involved splitting the OBEV text into groups that would represent selections of poems from distinct date ranges. Dates were used to split the corpus instead of dividing it by word count because of the historical element of the tagger and the need for fine-tuning the date range for cross-referencing with the HTE dataset. The Gutenberg OBEV edition did not contain individual dates for the poems, instead providing the date of birth/death of the authors where available.[48] The complete OBEV text contains poems ranging from c.1250 (Anonymous) to b.1870 (T.Sturge Moore). However, only a small number of poems fell into the earliest group (1250-1500),[49] and so these were omitted from the analysis as the sample would be too small to investigate in relation to the rest of the corpus. A further reason for this omission is the degree of spelling variation in these poems, which would pose an issue during the lemmatisation process. This was unfortunate, as the HTE dataset ranges from Old English to present day; if it was feasible to automate the lemmatisation process for this date range, it would offer further experimental opportunities for the HTE semantic annotation process. The remaining text was divided into four corpus groups, split at 100-year boundaries, with the birth date of the authors as the delimiting factor. A full list of the authors in each corpus is available in Appendix I as Table A1,[50] and a summary of the groups is provided in Table 1 below.

Group	Poem numbers	Number of authors in section	Years	Wordcount
1	34-295	59	1500-1599	43,462
2	296-447	51	1600-1699	38,040
3	448-655	59	1700-1799	46,652
4	656-883	93	1800-1918	55,079

Table 1 Corpus Groups

[47] Examples include ['Dear] and ['Sweetheart], which failed to be separated by the tagger automatically.

[48] Unverified dates were denoted with *?* or c. For instances where the author was still alive at the time of publication, the date was recorded as *b.[DATE]*.

[49] A total of 33 poems was omitted, dated XIII-XIV Century to d.1523 (Stephen Hawes).

[50] A complete list of the poems contained in the OBEV corpus is available in the digital appendix DA1.

Two other splitting methods were considered for this project: dividing the corpus by the literary periods to which the authors belong, or creating groups of equal word counts for simpler comparative analysis. The former approach was dismissed as being too subjective for this stage of the analysis: for boundary-crossing authors a decision would have to be made to attribute them to one particular group to avoid duplication, which conflicted with the objective goal of the methodology. The latter approach would require splitting the text within the boundary of the poem to achieve truly equal corpus groups, undermining the structure of the original text. Thus, while the approach used in this thesis is imperfect as it relies on boundaries that do not directly relate to the nature of language, it was chosen as a compromise between consistency and objectivity. The annotation process was carried out on each group individually, though aggregated semantic data for the full corpus is used in the analysis sections where appropriate. Subsequent reference to the groups will use CG1 for corpus group one, and CG2, CG3, and CG4 for groups two, three, and four, respectively.

4.3 Preparing the database

The annotation process employed MySQL queries for producing semantically annotated versions of the corpus groups.[51] To achieve this, a database was set up for this project with the HTE dataset and corpus data as individual tables, which were first cross-referenced to retrieve HTE data for each word in the corpus. This output is then processed through several additional queries to filter and calculate the distribution of the semantic data and produce the final output. These queries were developed through an iterative process, which is documented in this section, beginning with the setup of the initial data tables.

4.3.1 Corpus tables

To cross-reference the corpus groups with the HTE data it was necessary to convert them into database tables. Importing raw text data, even after it has been sanitised, would not produce a compatible dataset. At a minimum, it is necessary to lemmatise the original text and convert the corpus into a vertical format where

[51] Initial use of the HTE data was through the MS Access database system, but due to the size of the database it was necessary to move the queries over to a MySQL server as this was a free alternative that performed better than the MS system (Bassil 2012).

each lemma becomes a record that can be called on in the query, as established in §4.2.1 above.

4.3.1.a Preparatory tagging of the corpus

To allow for comparison with the semantic annotation results of the HTST, CLAWS was used to lemmatise and POS-tag the corpus. Wmatrix (Rayson 2009a) includes CLAWS as part of its suite of text manipulation resources, and supports larger input files than the publicly available version of CLAWS.[52] However, running the corpus groups through the Wmatrix interface revealed a limitation of the tagger: it was unable to identify lemma forms for early modern variants of words through its LEMMINGS lemmatiser (Rayson 2002: 119). As an example, the Wmatrix results for the following section from the first poem in CG1, [34] 'Forget not yet' by Sir Thomas Wyatt (1503-1542) are given in the Appendix in Table A2:

> FORGET NOT YET
> *The Lover Beseecheth his Mistress not to Forget his*
> *Steadfast Faith and True Intent.*
>
> FORGET not yet the tried intent
> Of such a truth as I have meant;
> My great travail so gladly spent,
> Forget not yet!
>
> *Poem i [34] Sir Thomas Wyatt (1503-1542)*

The output generated by Wmatrix includes the CLAWS POS tag, original word, and the corresponding lemma. It also documents the 'line of the input file [that the] word comes from', here simplified as the 'sentence number', and a 'two digit number to the left of the POS tags is a decision code produced by CLAWS to aid manual postediting', identified as 'CLAWS code' in the table ('CLAWS Input / Output Format Guidelines', n.d.).[53] While mostly correct, the output shows that the lemmatiser did not identify 'beseech' as the lemma of 'beseecheth'. It was, however, able to assign the correct part of speech. The likely reason for this is that the CLAWS tagger 'uses a probabilistic model based on left and right context to guess when it doesn't know a word' (Rayson 2009a), and so relied on the

[52] http://ucrel-api.lancaster.ac.uk/claws/free.html
[53] http://ucrel.lancs.ac.uk/claws/format.html. See (Garside 1987) for a further discussion of the output.

surrounding parts of speech to determine the most likely tag for the word.[54] To increase the number of accurate lemma forms in each corpus group, non-standard spelling had to be standardised, which was possible through the VARD system.

VARD is available as a standalone system for standardising spelling variation in corpora before carrying out additional processing (Baron 2019). Through this tool, it would be possible to edit each corpus group before passing it through the CLAWS tagger, improving the accuracy of the latter. While this process would already cut down on the manual editing when compared to standardising the spelling manually, a quicker alternative was made available through the HTST. The tagger combines the CLAWS and VARD utilities as part of the 'pre-processing' stage before applying the HTE 'knowledge base for a deeper layer of semantic annotation' (Piao et al. 2017: 117). While the tagger was primarily used as a conduit to the CLAWS and VARD utilities, using it to carry out preparatory annotation of the corpus also retrieved corresponding USAS and HTST semantic annotation data, which provided a further reference point for the macro and micro analysis.[55]

Table A3 (HTST output) in Appendix I shows the word, lemma, and POS fields for the HTST output for the same part of Poem i. The table shows no change to the POS tags[56] between the two versions but does correctly display 'beseech' as the lemma of 'Beseecheth'. Despite affecting only one record in this sample, pre-processing the records through VARD increased the compatibility of the corpus groups with the HTE data substantially, as shown by the accuracy tests conducted in the initial cross-referencing query runs. To enable the cross-referencing process, the HTST output for each corpus group was uploaded as four individual tables, referred to as `corpus` tables, using the schema in Appendix CG1_1_CorpusGroupTable.sql (p.317).[57]

---

[54] As noted on the Wmatrix website, the tagger is 'very good at guessing.' https://ucrel-wmatrix4.lancaster.ac.uk/cgi-bin/wmatrix4/help.pl#annotate

[55] The HTST tagger was used for this purpose in the early development stages of the methodology, but the public server was unfortunately offline for maintenance during the later stages. This research is therefore indebted to Dr Scott Piao for his assistance with tagging the corpus through a local version of the HTST.

[56] With the exception of losing the 'rarity marker' (http://ucrel.lancs.ac.uk/claws/) '@' from the adjective (JJ) tag for *tried*. Rarity markers '@' and '%' at the end of any CLAWS 'were added manually during the creation of the CLAWS lexicon to indicate rare tags for words' and have no relevance for the HTE classification system.

[57] The only modification to the original HTST output at this stage is the addition of a blank field for the generation of a primary key that will serve as the identifier for the position of each word in the original corpus.

4.3.2 HTE tables

The HTE data used for this research was obtained at the discretion of the University of Glasgow, which allows the use of the data for research purposes through a licence program. The HTE version used for this research was 4.2.2, but the methodology has forward compatibility with newer editions through the unique identifiers maintained through each revision (Kay et al. 2015). The raw HTE data was split into three files, which were imported as three separate tables into the database: the `lexeme` table, `category` table, and the `thematic headings` table.[58]

4.3.2.a Lexeme table

At 793,736 records, the `lexeme` table is the largest in the HTE dataset, containing every lexeme recorded from Old English to present day. The database table that housed the lexeme data mirrored the structure of the original file, the schema for which is recorded in LexemeTable.sql. While the HTE dataset provided through the University of Glasgow was already set up for use in research projects, it was still necessary to sanitise the data before importing it into the database. This involved adding additional markers to make the data compatible with the database encoding and structure.[59] To use the `lexeme` data, it is necessary to connect it to the `category` and `thematic heading` table before cross-referencing with the `corpus` table. Not all of the fields in each table are relevant to this study, but the full dataset was imported to maintain the integrity of the original file and allow for easier retrospective revisions or modifications to the query in the future. The active fields from the `lexeme` table are listed in Table 2 Lexeme headings below, along with the descriptions of each field.

Field	Description
htid	Unique identifier for each lexeme [Historical Thesaurus Identifier]
catid	Corresponding category identifier [Category Identifier]
word	Lemma/ headword

[58] The scripts LexemeTable.sql, CategoryTable.sql, and ThematicHeadingsTable.sql in Appendix II (MySQL) describe the schema used to set up the three tables for querying the HTE. Where all three HTE tables are mentioned, the database table convention is used (`HTE`).

[59] Examples include the lexeme 'null', which is interpreted by the database as a 'Null value' (a special marker that indicates that no value exists in that field), and the wholly problematic OE records such as '"1324","wylle < wiell(-e,-a)","\"wiell(-e,-a)\"","wylle","'.

apps	First citation (approximate and as recorded in the OED) [Approximate Start]
appe	Last citation (approximate and as recorded in the OED) 2000 is used for lemmas that are still in use [Approximate End]

Table 2 Lexeme headings

A sample record is used to illustrate the connections between the three `HTE` tables used for this research. The word *nightingale* was chosen for this purpose, as it appears only five times in the lexeme table, making it easy to display all corresponding data from the three tables. Table 3 shows the associated data for *nightingale* in the `lexeme` table.

Field Name	htid	catid	word	apps	appe
Record 1	52736	14731	nightingale	1882	2000
Record 2	88092	23001	nightingale	1862	2000
Record 3	128084	36423	nightingale	1250	2000
Record 4	219145	61026	nightingale	1500	2000
Record 5	762192	215050	nightingale	1500	2000

Table 3 Lexeme table with example

All five *nightingale* records in the `lexeme` table show that they are still in use according to the OED data, which is recorded by the *appe* (last citation) field as 2000. The first citation information varies, however, with the earliest citation for a *nightingale* record being 1250 and the latest new citation being 1862. The data shows that two new meanings for *nightingale* were recorded in the 19th century. If used against a corpus that is dated in the 18th century, for example, it would not be appropriate to consider these definitions, as they were not in use at the time the original text was created. The `lexeme` table does not supply information about the meaning of each word, and so it cannot be used on its own to identify the semantic properties of the text; to facilitate this, it must be connected to the `category` table, which holds this information.

4.3.2.b Category table

The `category` table houses all 235,249 records from the HTE category file. The description of the active fields is provided in Table 4 Category headings below, while the full schema can be seen in CategoryTable.sql. The `catid` field is the unique identifier for each semantic category in the `category` table. To show how the `lexeme` table connects to the `category` table, it is possible to take the list of

catids for the *nightingale* records in Table 3 (Lexeme table with example) above and retrieve all corresponding records from the category table.

Field	Description
catid	Unique identifier for each category
t1	Main category tier [e.g. 01 (The world)]
t2	Hierarchical category tier [e.g. 02 of 01.02 (Life and death)]
t3	Hierarchical category tier [e.g. 03 of 01.02.03 (Biology)]
subcat	Subcategory [e.g. 02.01 of 01.02.03.01.04|02.01 (Parts of eukaryote)]
pos	Part of speech
heading	Category name
themid	Thematic heading ID

Table 4 Category headings

Table 5 below lists the corresponding categories for the *nightingale* records. The data shows that all recorded forms of *nightingale* are nouns (n), but the headings for each record do not contain enough information to give a clear description of each sense without requiring additional context. To obtain this context, it is possible to look at the HTE hierarchy, which would reveal the context for each distinct sense, but this process would be time consuming to carry out for each record. Instead, the approach taken for this research was to connect the category records to the thematic headings by referring to the themid provided for each category record.

Field Name	catid	pos	heading	themid
Record 1	14731	n	other garments	343
Record 2	23001	n	names applied to various flowers	597
Record 3	36423	n	luscinia megarhynchos/nightingale	497
Record 4	61026	n	person	928
Record 5	215050	n	sweet singer	3773

Table 5 Category table with example

4.3.2.c Thematic heading table

The thematic heading data (v4) contained 4,033 records, the full schema for which is recorded in ThematicHeadingsTable.sql. The active fields are described in Table 6 below and include the unique identifier tid, which corresponds to the themid field in the category table. The thematic headings have their own hierarchical tiers which extend to five levels and follow the pattern of 'two upper case letters, number, lower case letter, number, lower case letter' to distinguish the system from the category tiers (Alexander & Kay 2021b).

Field	Description
Tid	Unique identifier for each thematic category [links to themid in category table]
s1	First hierarchical thematic tier
s2	Hierarchical thematic tier
s3	Hierarchical thematic tier
thematicheading	Thematic category name

Table 6 Thematic Headings

The value of the thematic headings for this project can be shown by following the *nightingale* example. Matching the `themids` in Table 5 (Category table with example) to the Thematic Headings table returns the `thematic headings` for the records, as shown in Table 7.

Field Name	tid	s1	s2	s3	thematic heading
Record 1	343	AC	2	f	Medical appliances/equipment
Record 2	597	AF	29	Null	Particular cultivated/ornamental plants
Record 3	497	AE	13	o	Order Passeriformes (song-birds)
Record 4	928	AI	15	d	Quality of voice
Record 5	3773	BK	4	g	Singer

Table 7 Thematic Headings with Example

Reviewing this table adds context to the results in Table 5: the »14731 (other garments) `category` attached to `record 1` is now clarified as being part of †343 (Medical appliances/equipment) instead of, for example, being part of the collections of lexemes under †814 (Set/suit of clothes). Three levels of the thematic hierarchy are also shown, as these enabled the aggregation of the semantic metadata for the corpus groups used in the macro analysis. Independently, each `HTE` table reveals a part of the data for the *nightingale* records, but it is necessary to bring them together to understand the distinction between each separate meaning.

4.3.2.d Combined HTE record example

To bring all of this information together, it is possible to join the `catid` from the `lexeme` table to the `catid` in the `category` table, and the `themid` in the `category` table to the `tid` in the `thematic headings` table, pulling through the active fields from each table into one aggregate view. This combined table for *nightingale* is displayed in Table 8.

Merging the tables in this way produces a coherent dataset for every match of the original record from the combined `HTE` dataset. Maintaining the data in individual tables keeps it flexible and allows for different sets of query structures. In this format, it was possible to feed individual words into the query, as with *nightingale,* to retrieve all corresponding fields from the three tables. As a manual process, however, this holds no advantage over the existing search functionality of the HTE website (Kay et al. 2019), and provides less information as it will not place the search result within a visual representation of the taxonomy that the user can then interact with further.[60] Rather, the true advantage of using the `HTE` dataset in a database structure is the ability to carry out bulk queries and retrieve corresponding data for every possible cross-match from the HTE. Using this approach, it was possible to obtain every matching record for every word in the `corpus` groups created from the OBEV.

[60] Credit must be given to the official HTE website, which excels in the design and presentation of the HTE data for individual search queries. It is an excellent portal for engaging with the HTE data.

Field Name	Lexeme table					Category table			Thematic heading table			
	htid	**catid**	**word**	**apps**	**appe**	**pos**	**heading**	**themid**	**s1**	**s2**	**s3**	**thematic heading**
Record 1	52736	14731	nightingale	1882	2000	n	other garments	343	AC	2	f	Medical appliances/equipment
Record 2	88092	23001	nightingale	1862	2000	n	names applied to various flowers	597	AF	29	Null	Particular cultivated/ornamental plants
Record 3	128084	36423	nightingale	1250	2000	n	luscinia megarhynchos / nightingale	497	AE	13	o	Order Passeriformes (song-birds)
Record 4	219145	61026	nightingale	1500	2000	n	person	928	AI	15	d	Quality of voice
Record 5	762192	215050	nightingale	1500	2000	n	sweet singer	3773	BK	4	g	Singer

Table 8 Combined Table

4.4 Combining the corpus and the HTE

This section first describes the initial cross-referencing query that was used to retrieve HTE data that corresponds to the `corpus` groups discussed in §4.3.1. The query used for this research retrieved similar results to those shown in Table 8 above by combining information from the `HTE` tables with the `corpus` table to produce a merged dataset. The query, as recorded in CG1_2_CrossReference.sql, achieves this by joining the `lemma` in the `corpus` table with the corresponding `word` in the `lexeme` table, with subsequent joins connecting the `category` and `thematic heading` tables on key fields. To maintain the word order of the original corpus, an auto-incremented ID was created to act as a reference position, fulfilling the same function as the `record` number in Table 8 (Combined Table). [61] Following this, the process of cleaning the results and resolving any errors in the cross-match process are described, beginning in §4.4.1 below which discusses the accuracy of the cross-reference, and then in §4.5 onwards, which looks at the filtering process that prepared the cross-match dataset for semantic disambiguation.

Returning briefly to the illustrative example of *nightingale*, the output of the query can be demonstrated by using actual instances of the word *nightingale* in the corpus. The word *nightingale* appears in the complete OBEV corpus twenty-seven times, so the first two examples were selected to illustrate the tagging output. These came from [39] *Description of Spring* by Henry Howard, Earl of Surrey (1516–47), and [42] *A True Love* by Nicholas Grimald (1519–62), which were both part of CG1 (1500-1599).

> DESCRIPTION OF SPRING
>
> THE soote season, that bud and bloom forth brings,
> With green hath clad the hill and eke the vale:
> The *nightingale* with feathers new she sings; [62]
> The turtle to her make hath told her tale.
>
> *Poem ii [39] Henry Howard, Earl of Surrey (1516–1547)*
>
> A TRUE LOVE
>
> The oak shall olives bear, the lamb the lion fray,

[61] Recoded in the database as `position_ID`.
[62] Emphasis added.

> The owl shall match the *nightingale* in tuning of her lay,
> Or I my love let slip out of mine entire heart,
> So deep reposèd in my breast is she for her desart!
>
> *Poem iii [42] Nicholas Grimald (1519-1562)*

The `position_ID` numbers for these instances of *nightingale* were #1115 and #1820 respectively. An excerpt from the initial output result for these records can be seen in Table 9 below. As with the earlier example, Table 9 lists all corresponding `HTE` matches for the target record that were retrieved through the cross-referencing query, including those that first appeared long after the death of both authors. The records retrieved through this process are identified in the methodology as *cross-reference matches*, representing *candidate* meanings for the word-forms in the corpus. Of the five HTE cross-reference matches retrieved for the *nightingale* entries, the candidate meanings »14731 (other garments) and »23001 (names applied to various flowers) were least likely to describe the *nightingale* of Henry Howard and Nicholas Grimald, as the senses were first recorded in 1882 and 1862 respectively, long after the poems were written. The filter process described in §4.5.1.a below removed these anachronistic cross-reference matches from data, restricting the list of possible candidate meanings to time-appropriate senses. With the example data, this left behind three candidate meanings for the *nightingale* records. At this stage, a reader may be able to distinguish between these senses manually or consider all three as valid meanings in this context; however, this would not be manageable with highly polysemous words, where dozens of HTE records are returned within the desired time range, and time-prohibitive to carry out manually for the full corpus. Thus, while the output shows the breadth of the lexical data in the HTE, it was not at this stage representative of the language used at the time the original text was published and did not disambiguate between the different candidate senses.

Running the cross-reference query on all four corpus groups returned a complete list of every corresponding lexeme in the HTE dataset. This substantially increased the size of the `corpus` tables: for example, the 43,462-word CG1 originally increased to 55,265 records after being passed through the CLAWS tagger, which tokenised the corpus and created new records for punctuation marks and sentence breaks; however, submitting this data through the cross-referencing query increased this to 1,075,367 records comprised of merged CG1 and HTE cross-

reference matches. As noted above, however, a portion of these records fell outside of the appropriate timeframe for the corpus group. In addition to these anachronistic records, the results also revealed several unmatched words, where no corresponding HTE record could be located. While some of the words in the corpus are understandably absent in the HTE data, either through their absence in the *OED* records or due to their idiosyncrasy, as is the case with many proper nouns, a review of the data was carried out to determine if any records were unmatched due to avoidable issues that could be resolved at this stage of the analysis.

corp_id	corp_word	corp_pos	htid	apps	appe	heading	catid	tid	thematicheading
1115	nightingale	NN1	52736	1882	2000	other garments	14731	343	Medical appliances/equipment
1115	nightingale	NN1	88092	1862	2000	names applied to various flowers	23001	597	Particular cultivated/ornamental plants
1115	nightingale	NN1	128084	1250	2000	luscinia megarhynchos/ nightingale	36423	497	Order Passeriformes (song-birds)
1115	nightingale	NN1	219145	1500	2000	person	61026	928	Quality of voice
1115	nightingale	NN1	762192	1500	2000	sweet singer	215050	3773	Singer
1820	nightingale	NN1	762192	1500	2000	sweet singer	215050	3773	Singer
1820	nightingale	NN1	52736	1882	2000	other garments	14731	343	Medical appliances/equipment
1820	nightingale	NN1	88092	1862	2000	names applied to various flowers	23001	597	Particular cultivated/ornamental plants
1820	nightingale	NN1	128084	1250	2000	luscinia megarhynchos/ nightingale	36423	497	Order Passeriformes (song-birds)
1820	nightingale	NN1	219145	1500	2000	person	61026	928	Quality of voice

Table 9 Example output with Nightingale

4.4.1 Identifying unmatched records

Due to the size of the output files, individually checking each record for potential errors would be prohibitively time-consuming. However, it was possible to quickly find obvious cross-reference errors by filtering for the *Null* values in the output, as these corresponded to records that were not matched to any HTE categories (i.e. unmatched records). Because a right join was used in the query, the output contained all of the records from the `corpus` table, and only the matching records from the corresponding joined `lexeme`, `category`, and `thematic heading` tables.[63] When a match was not possible, a *Null* value was recorded for all of the fields from the linked tables. Of the 1,075,367 records returned for CG1, 14,297 were recorded as *Null* matches. At first glance this number might appear discouraging, but a review of the data shows that most of the *Null* records reference punctuation, sentence boundaries added by CLAWS, and the OBEV numbers used to identify the poems. These elements have no impact on the semantic profile of the corpus and cannot be found in the HTE, so do not represent errors in the tagging process. They were retained through cross-referencing to make it easier to read the output but could be filtered out at any point through the corresponding CLAWS tags.

Of the 14,297 *Null* records for CG1, only 2,708 remained after filtering out punctuation and sentence markers. The remaining *Null* results for the first four poems in CG1 are shown in Table A5 (p.208), with the poem numbers retained to separate the texts. Much like the rest of the cross-reference output, the sample in Table A5 shows grammatical items (*that, your* in the example) and proper nouns (*Vixi*, *Puellis*, *Nuper*, *Indoneus*) resulting in *Null* records. The unmatched proper nouns (CLAWS tag NP1) are easier to explain: they are not routinely recorded in the HTE. There were 257 proper nouns in CG1 that could not be found in the HTE. The lack of matches for *that* and *your* are a bit less straightforward: there are 378 records in the `lexeme` table that contain *that*, but none that match the word exactly, instead listing '*be it that*', '*save that*', '*take it that*', in the word field, much

[63] This was a conscious decision in the design of the query. MySQL allows different `join` types, with the most common being the `inner`, `right`, and `left` joins. The `inner join` retrieves records found in both tables, and so would omit any `corpus` table records if corresponding fields were not found in the `HTE` tables. A `left join` would return all data from the `HTE` tables and only matching records from the `corpus` tables, while the `right join` retrieves the data necessary for this research, which includes all `corpus` data and only the matching `HTE` records.

like the 174 records for variations of *your*. This issue would need to be resolved if grammatical items were considered in the semantic disambiguation process, but this was not implemented in the current methodology as grammatical elements were filtered out using a stop-list.[64] Similarly, hyphenated items like *new-fangleness* presented an issue that could not be resolved within the parameters of this research as it would require a multi-word tagging approach that was not possible within the current process. The full corpus contained 215 hyphenated words, so the impact of their exclusion was minimal, but it would be desirable to incorporate hyphenated and multi-word expressions in the annotation process in the future.

The remaining 2,236 *Null* records for CG1 could be further divided into 294 cardinal numbers, the majority of which were OBEV poem numbers, 219 auxiliary verbs, and 1,340 function words, leaving just 573 words that could not be matched to an HTE record through the cross-referencing query and did not fit the criteria above. These were overwhelmingly made up of historical or regional variants, which either evaded processing through VARD and CLAWS, or were too uncommon to be included in the HTE. A sample of the first ten of these records is shown in Table 10 below.

position_ID	corp_word	corp_lemma	corp_pos
488	*obeyed*	*obeye*	VVN
494	*betrayed*	*betraye*	VVN
504	*unkist*	*unkist*	NN1
843	*affectin*	*affectin*	NN1
1999	*liever*	*liever*	VV0
2027	*leif*	*leif*	NN1
2029	*unmolest*	*unmole*	JJT
2044	*luvis*	*luvis*	NN1
2054	*servit*	*servit*	NN1
2055	*lang*	*lang*	NN1

Table 10 Sample of Lexical Null Records for CG1

The data show that certain forms still presented an issue for VARD and therefore CLAWS, such as *unkist*[unkissed], *servit*[serviette]. Another series of mis-tags were caused by the lemmas produced for *obeyed* and *betrayed*, which appeared to have been stemmed incorrectly. Overall, however, these results reflected a narrow

[64] The rationale and implementation of the stop-list filter is discussed §4.5.1.

margin of error in the pre-processing stage. Furthermore, without VARD, the number of unmatched lexical records for CG1 jumped up to 1,096, representing a loss of 523 possible lexical matches. To reduce the number of lexical *Null* matches, VARD was used in the pre-processing of the remaining corpus groups, leading to 824 unaccounted *Null* records out of 921,310 for CG2, 695 of 1,116,324 for CG3, and 402 of 1,374,740 for CG4. Reducing these numbers further would require manually editing each corpus group, but the small number of remaining *Null* records for each group did not necessitate this step.

4.5 Filtering the data

The large volume of data retrieved from the `HTE` tables required further filtering to narrow the results before carrying out the disambiguation process. Two filters were applied to the cross-referencing output: a stop-list to isolate grammatical items from further semantic analysis, and a date-filter to remove HTE data that fell outside the time frame for each corpus group. The filtering process formed the second stage in the semantic annotation pipeline and was carried out before the semantic disambiguation stage. While both filters have precedent in the literature, with stop-lists being frequently employed in CL research (Baker 2006), and a precedent for filtering the HTE data by date established by the HTST team (Piao et al. 2017), a number of approaches were available for implementing these filters. The parameters chosen for this research are briefly discussed below, complemented by a worked example of the effect these filters have on a small sample of the cross-referenced data.

4.5.1 Stop-list filter

Stop-lists are most frequently used to removed closed-class words from the corpus prior to carrying out keyword analysis, as they are often the most frequently appearing items in natural language texts (Baker 2006). While some CL work focuses on these elements instead of the open-class items, this approach is typically presented as going against the norm (Groom 2010). While the decision was ultimately taken to isolate closed-class (primarily grammatical) words from the semantic analysis, this was done after investigating the cross-referencing output, and presented a methodological choice that could be reversed if required.

As an illustrative example, a section of the cross-reference matches is reproduced in Table A6, showing the output for the first two lines of Poem i, [34] 'Forget not yet': 'FORGET not yet the tried intent/ Of such a truth as I have meant;' (Sir Thomas Wyatt (1503-1542) in Quiller-Couch 1998).

As shown in Table A6, many function words (*not, yet, the*) were recorded in the HTE and could provide semantic data in relation to the corpus. In some cases, as with *the*, these are obscure senses and could be filtered out with minimal impact on the accuracy of the tagger. Elsewhere, however, multiple senses of functional words were recorded in greater detail, as shown with the examples *of, as, a,* and *I* in Table A6. Auxiliary verbs were similarly given fine-grained definitions, and were described by the HTE editors as a particular challenge to categorise (as recorded in Kay (2010: 263) and then again in Kay (2011: 48)), thus presenting a counterpoint to dismissing these items as semantically irrelevant. Retaining grammatical data without diluting the results with highly specific meanings would require manual editing of the cross-reference data or the creation of rule-based assignment of tags for function words. As the goal of the current research was to develop a methodology for semantic annotation that required minimal manual input and processing, this was not attempted at this stage.[65] Consequently, although the HTE contained semantic classification for many function words, the decision was made to remove these from the output to prevent skewing the disambiguation data.

The approach employed in the USAS tagger for filtering grammatical elements, and, by extension, the HTST tagger was to assign a series of default tags to these items and exclude them from further semantic analysis (Archer et al. 2002: 35–36; Piao et al. 2017: 121). To produce the same result with the current data, a stop-list filter was created using the POS tags assigned through CLAWS. POS tags were used to filter the data to circumvent having to use a stop-list capable of handling unusual spelling, as these were already assigned with the help of VARD. The full stop-list created for this process is shown in Table A7 (p.225). The stop-list was employed after the initial cross-reference query, which meant that a record of all filtered words was retained in the initial cross-reference output. These were then re-integrated into the final summarised dataset, preserving the original word order of the poems in the corpus. To streamline the process, the stop-list was integrated

[65] The HTST developed a similar strategy for tagging highly-polysemous words, which could serve as a template for developing these filters in the future (Piao et al. 2017).

into the date-filter query, which eliminated anachronistic HTE records from the data.

4.5.1.a Date filter

Because the corpus groups fell within different ranges, it was not possible to filter the HTE data directly, as this would have required different versions to be created for each corpus group. Instead, the date-filter was applied after the initial cross-reference query, allowing the HTE data to be preserved in its entirety in the `HTE` tables. The query designed to implement the stop-list and date-filter is shown in Appendix CG1_3_filter.sql. Each corpus group was filtered individually, following the ranges identified in §4.2.2. The parameters for each group are listed in Table 11 below, and show the upper and lower date-range boundaries used to match the content of each group. A larger date-range is used for CG1 to reflect the accuracy of the OED data at that time, allowing matches recorded at any point between 1450 and 1700 for the group, acting as a 'buffer' for the dates of the recorded senses. For the remaining groups, a buffer of 50 years was used to reduce the risk of omitting relevant senses, as the first and last recorded appearance are based on available data and cannot be considered fully accurate in all instances.

Group	Poem numbers	Years	Upper year	Lower year
1	34-295	1500-1599	1700	1450
2	296-447	1600-1699	1750	1550
3	448-655	1700-1799	1850	1650
4	656-883	1800-1918	1950	1750

Table 11 Corpus groups with filter numbers

These filters produced a version of the cross-reference output that was more precise but still included a range of senses that were not related to the corpus. For example, using the steps above, the data in Table A6 would be filtered to exclude any HTE records falling outside the 1450-1700 range, and grammatical items included in the CLAWS stop-list filter, reducing the number of records from 336 to 166. However, this left 11 different senses for *forget*, 29 for *yet*, 7 for *tried*, 24 for *intent*, 21 for *truth*, and 73 distinct tags for *meant*. To determine the most likely senses for each word out of the remaining tags, a disambiguation process was developed that limited the number of possible senses to 6 most likely definitions.

4.6 Disambiguation

As discussed in the literature review, different approaches have been developed to identify the semantic properties of corpora, which in turn employ a range of disambiguation parameters. In most cases, however, these parameters are trained towards identifying the primary sense of the word as it is used in the context, with accuracy of the approach determined by how well it can accomplish this task. These approaches are well-suited to extracting surface-level information about a corpus, identifying significant themes, or annotating corpora that contain precise or literal material. When used on corpora that contain literary texts, however, semantic annotation systems that look at identifying specific senses are less effective. The reason for this, as introduced in §3.2.4 above, is the unique challenge of annotating figurative language, where multiple senses can be evoked by single words. To successfully annotate figurative texts, the tagger would have to be capable of attributing multiple senses to each individual word and restrict the output so that implausible senses are excluded. At the time of writing, however, no semantic annotation system was capable of fully fulfilling these parameters, and therefore no precedent existed for the exact disambiguation process used in this thesis. It was therefore necessary to adopt an experimental approach in its design, drawing from related work to explore alternatives for achieving the desired results. The result was the development of the 'ambiguity' semantic tagger.

4.6.1 Contextual disambiguation

Following the rationale that contextual information can be used to disambiguate between likely and unlikely senses of individual words, a system for measuring this relationship through the HTE data was developed for use in the ambiguity tagger. A corpus that is cross-referenced with the HTE data can provide this contextual information at different levels of abstraction by moving through the hierarchical taxonomy.

Using the worked example from Poem i, the candidate meanings for the word *truth,* shown in Table A6, range from the familiar »75959 (reality/quality of being real) to the obscure »114960 (Egyptian) and the antiquated »147387 (Betrothal), with a total of thirty distinct classifications retrieved for the word. The *truth* of

Wyatt's poem could be seen as invoking more than one of the senses recorded in the HTE, but it certainly does not represent all of the recorded meanings simultaneously. To determine which candidate meanings are relevant to this particular use of the word *truth*, the cross-reference matches for the rest of the poem provided the context for disambiguating the candidate meanings of each individual word. To achieve this, the research employed the principles of contextual association (section §3.2), using the hierarchical structure of the HTE as a system of classification for identifying higher-level semantic domains for each candidate meaning. Domains that were activated multiple times in the corpus were seen as having a higher 'semantic relevance' (SR) in relation to the corpus than those which are infrequently activated. The activation was measured by counting the number of times a cross-reference match was retrieved for a word against a higher level in the semantic taxonomy. The semantic relevance of the higher-level domains was then used to disambiguate between fine-grained senses as identified through the cross-referencing process by promoting the senses that correspond to greater SR of high-level domains. As an example, the SR of the category »114960 (Egyptian) in relation to the *truth* of Poem i would be calculated in relation to the cross-matches returned for the rest of the poem.

4.6.2 Category match value

The most straightforward approach for grouping the activated senses in the corpus was to count the number of times an HTE category was retrieved as a candidate meaning for the words in the corpus. Carrying this out on the date and part-of-speech filtered corpus groups showed that CG1 returned matches for a total of 32,008 unique HTE categories, while CG2, CG3, and CG4 returned 33,507, 37,020, and 43,072, respectively. The top ten results for CG1 and CG4 are shown in Table 12 and Table 13 below, representing the earliest and latest parts of the OBEV corpus. Reviewing these results showed a lot of similarity in the most frequently cross-referenced categories: all four corpus groups had »130,547 (Terms of endearment), »130,497 (Loved one), and »130,433 (love in return) in the top ten categories. These results were unsurprising, as they corresponded to words in the corpus that were associated with affection, both common (*sweet, darling, dear, heart, love, treasure*) and uncommon (*turtle, wanton, sparrow, turtle-dove, sun*), and align with the style of lyrical poetry.

Rank	category_id	category_heading	appearance
1	130433	love in return	780
2	123749	consider to be, account as	701
3	130497	Loved one	638
4	90057	pass into state, become	597
5	130547	Terms of endearment	555
6	57520	Sexual desire	509
7	130455	Liking/favourable regard	468
8	128428	take pleasure in/enjoy	467
9	126189	Commend/praise	465
10	77308	cause to be/become	454

Table 12 Top category counts CG1

Rank	category_id	category_heading	appearance
1	90057	pass into state, become	688
2	130547	Terms of endearment	567
3	130433	love in return	562
4	130497	Loved one	525
5	142005	Speak/say/utter	524
6	102861	Go away	452
7	116662	Perceive	442
8	77791	Occur/happen	435
9	128445	source of pleasure	410
10	215908	of part of	409

Table 13 Top category counts CG4

Appearing with similar frequency across all groups are categories that are harder to identify at a glance, such as »123,749 (consider to be, account as) in CG1 and »90,057 (pass into state, become) in both tables. A closer look at the lexemes that fall in these categories revealed them as corresponding to instances of, for example, *see, take, make, fancy, esteem, behold,* and *eye* for »123,749, and *go, get, fall, come, proceed, prove,* and *wax* for 90,057. As with the combined record example in Table 8 above, these categories are easier to interpret alongside their corresponding thematic headings of †1922 (Evaluation, estimation, appraisal) and †1354 (Change) respectively.

When measuring the category matches at the level of an individual poem, a further issue is the dispersion of candidate meanings across the full HTE hierarchy. The category match counts for Poem i, for example, identified a match value of 10 or more for only 12 out of the 928 distinct categories retrieved through the cross-

reference query. In total, only 264 categories reported more than one match, with the majority corresponding to one-off matches. Thus, despite the HTE categories representing conceptual domains, and therefore a broader group than the individual lexemes, they were still too narrow to be used for contextual disambiguation. The decomposition of categories into higher-level categories was initially trialled for calculating the SR of the fine-grained senses, with the third level of the taxonomy chosen as cut off level for abstraction. The release of the thematic heading taxonomy as part of the HTST resource offered an alternative to this, with a curated list of high-level senses, mapped onto the HTE categories, and following a simplified hierarchical structure. Consequently, the thematic headings taxonomy was used as an alternative to the HTE categories for measuring the SR of individual HTE activations.

4.6.3 Thematic heading match value

The thematic headings match value for the retrieved cross-reference matches for Poem i were better suited for contextual disambiguation: of the 467 thematic headings activated by the cross-reference matches for the poem, 273 reported a match value of 2 or more, and 31 reported a match value of 10 or above. However, as discussed in §3.2.3, the size of the thematic headings categories is likely to impact on the match value (MV) metric, with categories that represent broader semantic domains reporting a higher MV because of the large number of categories that sit under them, and not because they are particularly relevant to the corpus. Similarly, thematic headings that correspond to narrow semantic domains were less likely to report a high MV, even if a proportionately higher number of cross-reference matches was retrieved. One of the steps taken to resolve this was removing duplicate candidate meanings for different parts of speech within the same thematic headings by filtering the data to exclude repeats within those parameters.

To further reduce the impact of the thematic heading size on the SR calculation, the match value was normalised against the size of the thematic heading, thus converting the metric into a relative match value (RMV), representing MV as a proportionate frequency. The RMV was calculated using the formula used to calculate the relative frequency of words when comparing corpora of different sizes, adapted to match the parameters of this data (Brezina 2018a: 42–43). The

calculation is represented as $F_n = \frac{F_o}{T_o} \times 10^2$, where the normalised frequency RMV (F_n) of matches for a thematic heading, is calculated by taking the MV for the heading as the observed frequency (F_o), divided by the total number of categories grouped under that heading as the size of the heading, considered to be the *offset* for that heading (T_0), multiplied by 100 as the basis for the normalisation.[66] RMV could therefore be taken as a more accurate measure of the semantic relevance of the thematic headings in relation to each poem, and more effective for contextual disambiguation.

The normalised results for Poem i, for example, reported 291 headings with a relative match value of 2 or more, and 60 with a value of 10 or more. The RMV was used to determine which headings have a greater semantic relevance, and therefore to assign candidate senses from those headings to the words on a scale of confidence. To accomplish this, the candidate meanings for each word were ranked in order of high to low RMV, making it possible to restrict the meaning with an assigned percentile rank. The percentile was calculated at the word level rather than poem level, thus ensuring that word frequency did not impact on the ranking used for assigning semantic tags from the candidate senses. A sample of the query used to calculate this for CG1 is shown in Rank.sql. This approach allows the cut-off of candidate meanings at different points on the percentile scale, depending on the desired confidence for the semantic annotation results. For this research, candidate senses were cut off at the 80th percentile mark, which represented a high-confidence ranking that still allowed flexibility in assigning multiple senses for ambiguous words.

Following this process for the cross-reference results for the first two lines of Poem i shows how the data is reduced from the example in Table A6 to produce the ranked results in Table A9.[67] Here, only the candidate meanings that were found to be semantically relevant are recorded, while still reporting multiple possible meanings for words that are ambiguous in the context. The assigned meanings for *forget*, for example, were all variations within †1789 (Faulty recollection, forgetting), while *intent* was assigned meanings from †1696 (The mind), †1769

[66] An example of the query used to calculate the MV and RMV for CG1 is shown in Appendix RMVbyPoem.sql.
[67] The complete ranked results for all corpus groups are provided in the digital appendix under DA2.

(Meaning), †1901 (Concentration), and †2150 (Intention). Notably, while the meanings assigned to highly polysemous words such as *yet* did contain some errors, in this case the sense †38 (Flood/flooding), overall the results were promising in revealing single or multiple senses for words based on the HTE taxonomy.

4.6.4 Post-processing

As noted in the outline for the project, the goal of the research was to develop a system for semantic annotation of poetry that required minimal input and post-processing, in part to ensure that the tagger is efficient and does not require time-consuming manual configuration, and to ensure consistency in the output available for further investigation. This did not mean, however, that it was not possible to fine-tune the results and eliminate erroneous tags before moving on to the analysis of the data. The parameters for excluding certain senses from the results are liable to change depending on the research goal, the researcher's familiarity with the contents of the corpus, and contextual information that could be used to determine incongruous senses. For this project, the post-processing involved reviewing the top offset thematic headings for each corpus group and reviewing any noteworthy results against the text to determine if the senses were plausible or not. This was of course a subjective exercise, and care was taken to retain any senses that were ambiguous or offered an interpretation of the text that was coherent in relation to the corpus group, even if the sense itself could be considered unexpected.

4.6.5 Summary output

To facilitate the micro-level analysis, an additional view was created that took a maximum of six senses and converted the output to a horizontal display of the data, considered to be the *summary output*. The summary output took the thematic heading code and description, the HTE category code, description, first and last appearance, the RMV, and the rank for the top six assigned senses to create a maximum of six distinct *semantic tags* for each word. In cases of multiple senses having the same rank, the general sense was taken by sorting the category numbers and retrieving the one from higher in the semantic taxonomy. The syntax

for the semantic tags, and an example for one of the tags for *forget* from Poem i is shown below:

> Thematic heading; [category heading, category part-of-speech; apps-appe]; †thematic heading number; [»category number]; #RMV/percentile rank
>
> Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00

The thematic heading was recorded alongside the HTE code to enable analysis of the data to be carried out at either the HTE category or the thematic heading level. The HTST output similarly records both data points in the tagger output to allow investigation at varying levels of abstraction (Piao et al. 2017: 116). The summary output takes the semantic annotation data and presents the information in a horizontal format rather than a vertical list which duplicated each word form by creating a new record for each possible tag.[68] In the summary output, the six top tags are presented in columns, and the words are returned to their condensed vertical structure without being duplicated. This made the semantic metadata easier to interpret when looking at an individual poem, as it converted the results into a readable format. To illustrate this, the summary output for the complete Poem i, [34] 'Forget not yet' by Thomas Wyatt are given in Table A10, and shows the progression from the cross-referenced results in Table A6 to the final annotated output.

4.7 Accuracy

The accuracy of the ambiguity tagger is harder to measure than single-meaning taggers, not least because of the dynamic number of meanings assigned to each word. Due to the size of the tagged output, manually checking the data would be prohibitively time consuming. A further challenge to identifying the accuracy of the output is the 'ambiguous' nature of the tagger, which allows multiple meanings to be simultaneously considered as valid representations of a single word in context. Consequently, while the post-processing stage (§4.6.4) allowed manual correction of the cross-reference data to exclude any matches from domains that

[68] The query for producing this output for CG1 is shown in CG1_5_SummaryView.sql. The complete summary view results for all corpus groups are provided in the digital appendix under DA3.

were noticeably at odds with the corpus, the manual editing was purposefully conservative to allow the tagger to disambiguate candidate meanings with relative autonomy. It is, however, possible to review the results in relation to the HTST tagger, which achieved a reported average accuracy of 81.96% for a sample of Shakespeare's comedies, which was the closest test text to the OBEV corpus (Piao et al. 2017: 128). To this end, the HTST results for Poem i are presented in the appendix under Table A11, which can be reviewed in relation to the ambiguity tagger results in Table A10. The accuracy of the tagger is further explored in the analysis sections, as the semantic metadata is queried at both macro and micro level, then explored in greater detail through a measure of semantic collocation.

4.8 Conclusion

This chapter has outlined the semantic annotation process developed as part of this research. The methods used in this research are experimental and form the first phase in the development of an automated semantic annotation process for figurative language. Despite the constraints of this pilot study, the tagger was able to meet the primary design parameters set out in Chapter 3 in response to RQ2. Furthermore, the analysis of the results, as described in the following sections, revealed multiple applications for the annotation process, aided by the hierarchical structure of the HTE. The macro analysis (Chapter 5) that follows makes use of the high-level thematic headings for exploring salient concepts across all four corpus groups, while the micro analysis (Chapter 6) looks at the annotation results at the level of individual poems and utilises the fine-grained HTE category headings. The final analysis section (Chapter 7) combines a macro and micro approach to investigate frequently co-occurring semantic tags in a sub-section of the corpus, thus presenting an approach for measuring semantic collocation that was made possible through this research. As a reference for the analysis sections, a summary of the key steps and terminology introduced in this chapter is provided below.

First, the corpus is tokenised, lemmatised, and POS-tagged to produce an annotated version of the text. The HTST tagger was used for this purpose, employing the VARD and CLAWS systems. Following this, the cross-reference query *retrieves* all HTE records for every word in each corpus group. This was accomplished by matching the words tokens in the corpus with the lexemes in the

HTE data, which in turn allow the retrieval of the HTE categories and thematic headings that provide the semantic classification for the lexemes. The dataset produced as a result of this process included the original word forms for each corpus group, tokens representing sentence breaks and punctuation, as identified through CLAWS, and all corresponding HTE records, identified as *cross-reference matches*. The MySQL code for the CG1 cross-reference query is shown in CG1_2_CrossReference.sql. With the exception of uncommon or highly specific words, multiple records were retrieved for each word, representing a list of all possible meanings; these were considered to be *candidate* meanings (Wilson & Thomas 1997: 62).[69] These records were filtered to exclude any meanings that fell outside of the date range of the texts in the corpus groups, with a different range used for each group. Records that corresponded to grammatical items were also excluded, as they were not used in the semantic disambiguation process. The *cross-reference matches* represent *candidate* senses and are not referred to as *semantic tags* at this stage, as they do not represent *assigned* senses; they are simply a list of all possible meanings, similar to what would be retrieved if looking up a word in a dictionary.

To determine which *candidate* meanings are more or less likely to represent each word as it is used in the corpus, a *contextual disambiguation* process is employed, using the *retrieved cross-reference matches*. First, the number of matches retrieved from each thematic heading category is counted, producing a *match value* for individual thematic headings in relation to each poem. Thematic headings with a higher *match value* were considered to represent semantic domains that were more relevant to the poem, making the *candidate* meanings that fall within those thematic headings more likely to represent the words in the poem. However, as described in §4.6.3 above, the *match value* was influenced by the size of the thematic heading categories: categories that represented broader semantic fields reported a higher match value, not necessarily because they were relevant to the corpus, but because there was a greater number of candidate meanings distributed under those thematic headings, and therefore a better chance of retrieving a cross-reference match from within those headings.

[69] Further exceptions included words that were not identified in the lexeme data, and therefore returned no possible candidate meanings. This group was mostly comprised of proper nouns, archaic variants, or words that did not return a viable lemma for use in the cross-reference query. See §4.4.1 for the discussion of these records.

Similarly, thematic headings representing narrow or esoteric semantic fields were less likely to report a high match value, even if the number of retrieved matches was proportionately high when taking into account the total number of fine-grained senses encapsulated by the broader thematic heading.

To account for the size of the thematic headings, the *match value* (MV) was converted into a *relative match value* (RMV), which took account of the size of the thematic headings categories to return a proportionate value. The total number of concepts grouped under a thematic heading category were taken to represent the size of that heading, identified as the *offset* value, and the match value reported for the heading was divided by *offset* value to produce the *relative match value*. The RMV was then used to rank *candidate* meanings in order of most relevant (highest RMV) to least relevant (lowest RMV). The senses from the top twenty percent were considered to have a high-enough *semantic relevance* for the poem to represent relevant meanings. At this stage, the ranked *candidate* matches could be considered as *assigned* to individual words, and available for further analysis of the corpus. This informed the macro analysis of the corpus carried out in Chapter 5 below. For the micro analysis in Chapter 6, a summary view was created by limiting the number of semantic tags to a maximum window of six, as demonstrated in the sample output in Table A10.

The semantic annotation process described in this chapter was designed to accommodate figurative language, which has traditionally presented a challenge for semantic annotation systems. The initial aim for the process was a fully automated annotation process, but this would not meet the criteria identified in §3.2, which shows that a flexible approach to semantic annotation is required to overcome the barriers of existing tools and methods. The current process therefore includes a post-processing stage where any errors in the tagging process can be resolved manually, and also allows users to configure their annotation parameters.

Chapter 5 Macro analysis

5.1 Introduction

The macro-level analysis uses the results of the semantic annotation process described in the methodology chapter above to examine the extent to which the ambiguity tagger addresses the barriers to semantic annotation of non-standard corpora (RQ3). For corpus linguistics data to be useful, it must be interpretable. Thus, this chapter also examines the capacity of the ambiguity tagger in analysing the semantic properties of a corpus at a macro level (RQ4). When working with large data emblematic of the field, researchers often look towards visualisation as a conduit to showcasing their data (Allen 2017). Often, this involves selecting key features for presentation, or finding an appropriate measure for aggregating information. The semantic annotation process developed for this thesis enables this by using the hierarchical taxonomy of the HTE, making it possible to review the results at different levels of 'generality' (Wilson & Thomas 1997: 57). This was particularly valuable for handling the volume of data produced by the ambiguity tagger.

The semantic metadata retrieved through the semantic annotation process developed for this thesis is unique in its breadth, as the ambiguity tagger allows multiple meanings to be assigned to individual words based on their semantic relevance in context, but without restricting the number of senses to a set number of tags. While the number of assigned meanings can be restricted by modifying the confidence level for assigning tags, the design of the tagging process purposefully excluded the restriction of meanings to an arbitrary maximum number to allow all 'likely' senses to be considered as valid if they are supported by the context. This disambiguation approach is unique to the semantic annotation process used in this thesis, making a contribution to knowledge by extending the capabilities of semantic taggers to handling multiple meanings in context. The semantic metadata produced by the tagger varied in size, because there was no pre-determined limit to the number of thematic tags that could be assigned to each item. This had to be taken into account when identifying the appropriate method of analysis.

Two different approaches were used to extend the analysis beyond the 1 : 1 ratio of word to semantic tag: for the macro analysis, all assigned meanings for each word were included in the aggregate data, using a ratio of 1 : x with no restriction on the number of senses (x) assigned to each word; for the micro analysis, the summary view was used, thus restricting the number of semantic tags to a ratio of 1 : 6. The latter approach follows the precedent established by the 'portmanteau' tags used in the USAS tagger and the 'merged' codes assigned by the HTST to limit the number of tags assigned to each word (Piao et al. 2005a: 5; 2017: 124). The micro analysis differs from the approach taken in this chapter by focusing on the advantage the semantic data affords to a close reading of the text, in contrast to the aggregated semantic features examined in this chapter. To focus the analysis in this chapter, the core research question RQ4 (How can the tagger facilitate analysis of non-standard corpora using existing corpus analysis methods?) was expanded to include a further sub-question (RQ4a): What kind of insights do these analyses provide?

To answer these questions, this chapter first presents the aggregated semantic data for all four corpus groups, then looks at the results of a keyness analysis of the semantic metadata for each group, and finally looks at the aggregated semantic metadata for the work of a single poet. The division of the corpus into four groups is discussed in §4.2.2 above. Closer examination of specific categories was guided by the aggregate results, with salient features identified first at the macro level. To enable a comparative analysis of the semantic features to be carried out in this chapter, the semantic metadata is aggregated for each corpus group, representing the four different periods represented by the poems in the OBEV.

5.2 Aggregated semantic metadata

While the thematic headings already provided a superordinate form of the semantic categories of the HTE, there were still 4,033 headings that could be activated by the cross-reference matches for each corpus group. The RMV metric (§4.6.3) was used in the disambiguation process to identify the semantic relevance of the headings to each poem, which then guided the assignment of candidate meanings to the corpus. As the corpus was split into four groups for the cross-referencing process, it was possible to query the results for each group

individually, thus combining the assigned senses for each poem using the thematic headings taxonomy. In total, 2,399 thematic headings were identified as semantically relevant for the poems contained in CG1, 2,374 for CG2, 2,574 for CG3, and 2,841 for CG4. Consequently, the granularity of the data, even at the level of thematic headings, made it necessary to condense the semantic data further to enable the interpretation of the results.

To accomplish this, the thematic headings were consolidated to the top tier of the taxonomy, thus flattening the data to restrict the number of datapoints available for analysis. For example, this reduced the 2,841 distinct thematic headings identified in CG4 to 37 top-level headings, while leaving the granular data accessible in the tagged dataset. This made it easier to review the results across all four corpus groups for the initial macro-level analysis, while the granular data enabled the analysis of specific headings at lower levels of the hierarchy in later sections of this chapter. As with the HTE taxonomy, the top level of each thematic category is considered representative of the categories it contains, making it suitable for aggregating senses in this way (Alexander et al. 2015c). As an example, thematic headings beginning with *AU*, ranging from *AU*.01 (Emotions, mood) through to *AU*.47.*d* (Encouragement), were counted under the top-level heading of *AU* (Emotion) in the aggregated data used in this section, while the analysis in §5.2.1 moves down to the second tier of the hierarchy to examine the results in greater detail.

To account for the difference in size of each corpus group, the aggregated results were normalised by following the same process that was used to retrieve the RMV in §4.6.3 above. The aggregated count was normalised against the word count for each group using $F_n = \frac{F_0(10^3)}{C}$, where F_n as the normalised frequency was retrieved by multiplying F_o as the observed frequency of the aggregated thematic headings by 1,000, and then dividing by the corpus size C. While a larger base value, typically a million, is normally used in normalising corpus data, a smaller base of 1,000 (Brezina 2018a: 42–43) is recommended for smaller corpora, and was appropriate in this case. As an example, for CG1, the 2,752 assigned senses falling under *AA* (The world) were normalised against the size of the corpus group ($C =$ 43,462 words), returning $F_n = 633$ as the semantic relevance of *AA* for CG1, while the observed frequency of 2,715 for *AA* in CG2 returned $F_n = 714$ against $C =$

28,040. By normalising the results, the data revealed that despite reporting a smaller observed frequency, senses collected under *AA* (The world) had a higher semantic relevance for CG2 than CG1. Both the original and normalised values for the aggregated counts are available in Table A12, while Figure 1 below shows the normalised distribution of the semantic relevance of each thematic category for the four corpus groups.

The data captured in Figure 1, therefore, shows the number of senses assigned under each top-level thematic heading against a scale of 0 to 3,500, with the largest number of senses assigned under one heading for a corpus group represented by *AU* (Emotion) with 3,095 for CG1. The graph also captures those categories with low semantic relevance to the corpus groups, most notably *AH* (Textiles and clothing), *BE* (Education), and *BC* (Law) with a combined total of 269, 155, and 133 respectively for all four corpus groups. Interestingly, several domains appear to show a gradual trend of semantic relevance from CG1 to CG4, with *AA* (The world), *AB* (Life), *AL* (Space), *AN* (Movement), *BH* (Travel and traveling) and *BJ* (Trade and finance) becoming more semantically relevant as the groups move forward in time, while senses from *AP* (Relative properties), *AS* (Attention, judgement, curiosity), *AW* (Possession/ownership), and *BD* (Morality) were assigned less frequently as the groups became more modern. These shifts could point to cultural shifts or priorities expressed through the poems contained in the collection. The lexical growth in certain semantic fields could also impact on the reported semantic relevance, as with the growth in *BJ* (Trade and finance) and the sharp increase in *BK* (Leisure) for CG4. In this regard, the aggregated results were limited in what they can reveal about the corpus without first accounting for the lexical variety of the HTE taxonomy. They were, however, useful as an entry point into further analysis of the data. To explore this, a domain with one of the most consistently high semantic relevance for all groups was chosen for further analysis, *AR* (The mind).

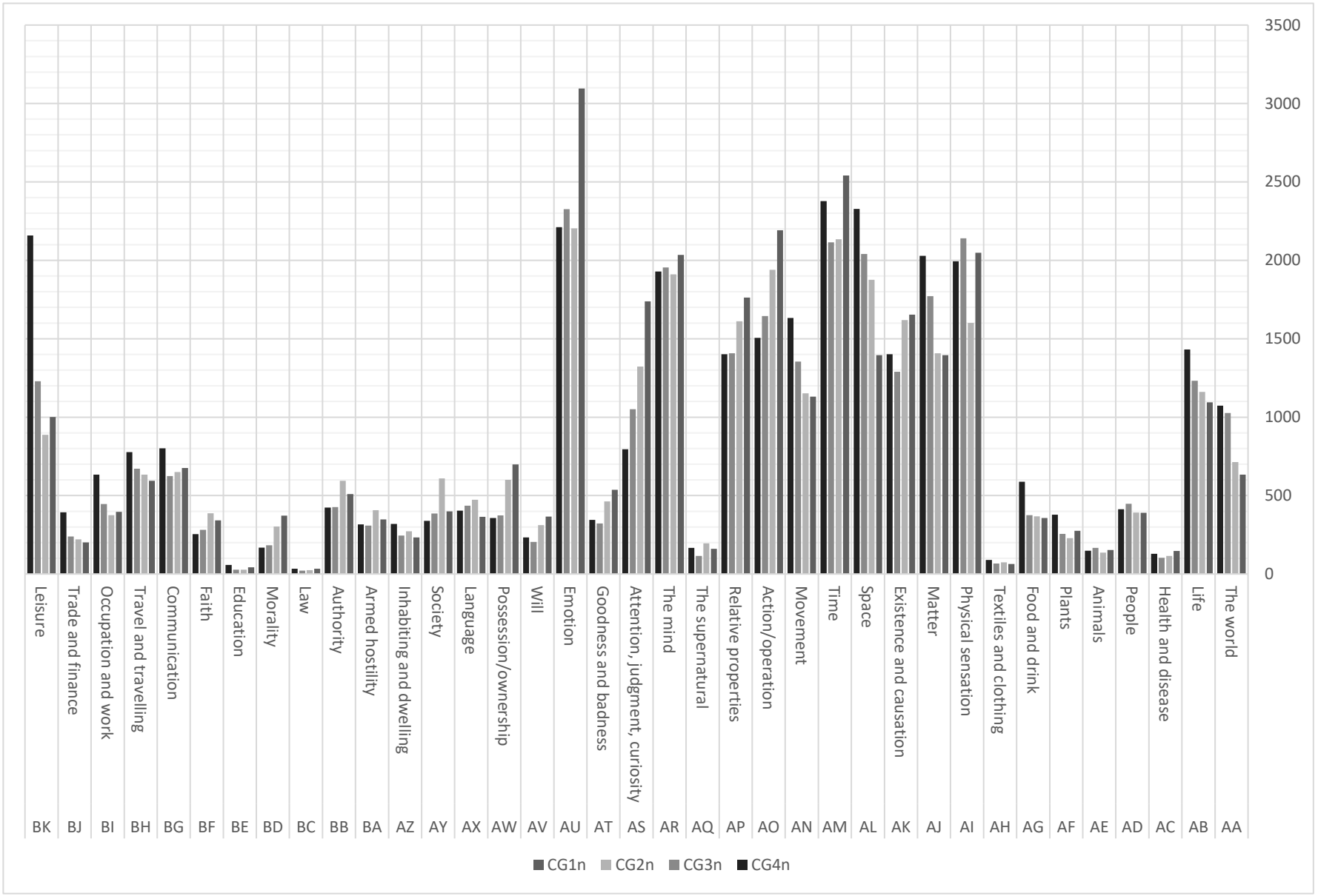

Figure 1 Aggregated Thematic Headings across corpus groups

5.2.1 AR (The mind)

To understand the significance of the high presentation of *AR* (The mind) across the corpus groups, it was necessary to look further down the thematic heading taxonomy. In total, 151 distinct thematic headings from within the sub-tiers of the *AR* domain were activated by the senses assigned to the corpus groups. These were grouped under the 2nd tier of the *AR* domain, of which 40 distinct senses were activated by the corpus. At this level, it is possible to examine the distribution of semantic relevance across the *AR* domain, as shown in Figure 2 below, while Table 14 captures the top ten semantically relevant domains for the corpus.

S1.S2	Thematic Heading	CG1n	CG2n	CG3n	CG4n	Total
AR.16	Understanding	170	171	191	175	707
AR.36	Knowledge	159	144	176	220	699
AR.50	Expectation	217	147	143	152	659
AR.38	Truth, validity, correctness	177	136	114	121	548
AR.45	Belief, opinion	107	106	100	97	410
AR.35	Memory, keeping in mind	130	49	85	126	390
AR.42	Deceit, deception, trickery	121	111	68	54	354
AR.44	Hiding, concealing	60	78	68	60	266
AR.17	Intelligence, cleverness	38	62	83	71	254
AR.12	Imagination	40	33	66	65	204

Table 14 Top 10 activated thematic headings within AR (The mind)

By reviewing the results in Table 14, the most semantically relevant domains could be grouped under distinct thematic headings, which could then be used to form hypotheses about the corpus. While the base activation frequency of semantic domains does not necessarily reveal significant domains in the corpus (see §5.3.1), it is still valuable in reviewing the general 'aboutness' of a text based on the words that 'share the same semantic space' across the different corpus groups (Archer et al. 2006: 2).

The high SR of *AR*.36 (Knowledge), *AR*.17 (Intelligence, cleverness), *AR*.16 (Understanding), and *AR*.38 (Truth, validity, correctness) could serve as an indicator of the priorities in the corpus, whether related to the importance of discerning thought or a universal pursuit of truth and understanding. Antonymous domains represented by *AR*.42 (Deceit, deception, trickery) and *AR*.44 (Hiding,

concealing) have a lower SR than their more positive counterparts, but were assigned with a relatively high frequency compared with the remaining *AR* subdomains. This was particularly clear when comparing the results across the full *AR* domain, as shown in Figure 2. Here, the visual representation of the data across the corpus groups highlighted the variety in SR of subdomains across the groups, which could be concealed by reviewing the total counts alone.

At the second tier of the thematic heading hierarchy, the data obtained through the semantic annotation process could be used to form more specific research questions. The difference between each second-level subcategory was easier to interpret than the difference between the top-level headings, and so it was possible to narrow the research focus to understand why certain senses were more or less represented in the corpus. The multifaceted nature of the data, however, allowed for different approaches to answering these questions. It was possible, for example, to look at the most-referenced headings and trace the relationship between the subheading and the words that trigger the senses during the tagging process. Another approach was to look at the headings that appeared incongruous, whether in relation to the semantic annotation data or to existing knowledge of the reference texts and direct the inquiry towards understanding the cause of the unexpected results. Crucially, each approach offered a new way of engaging with the corpus and the texts contained within it, showing how the ambiguity tagger can facilitate analysis of non-standard corpora (RQ4).

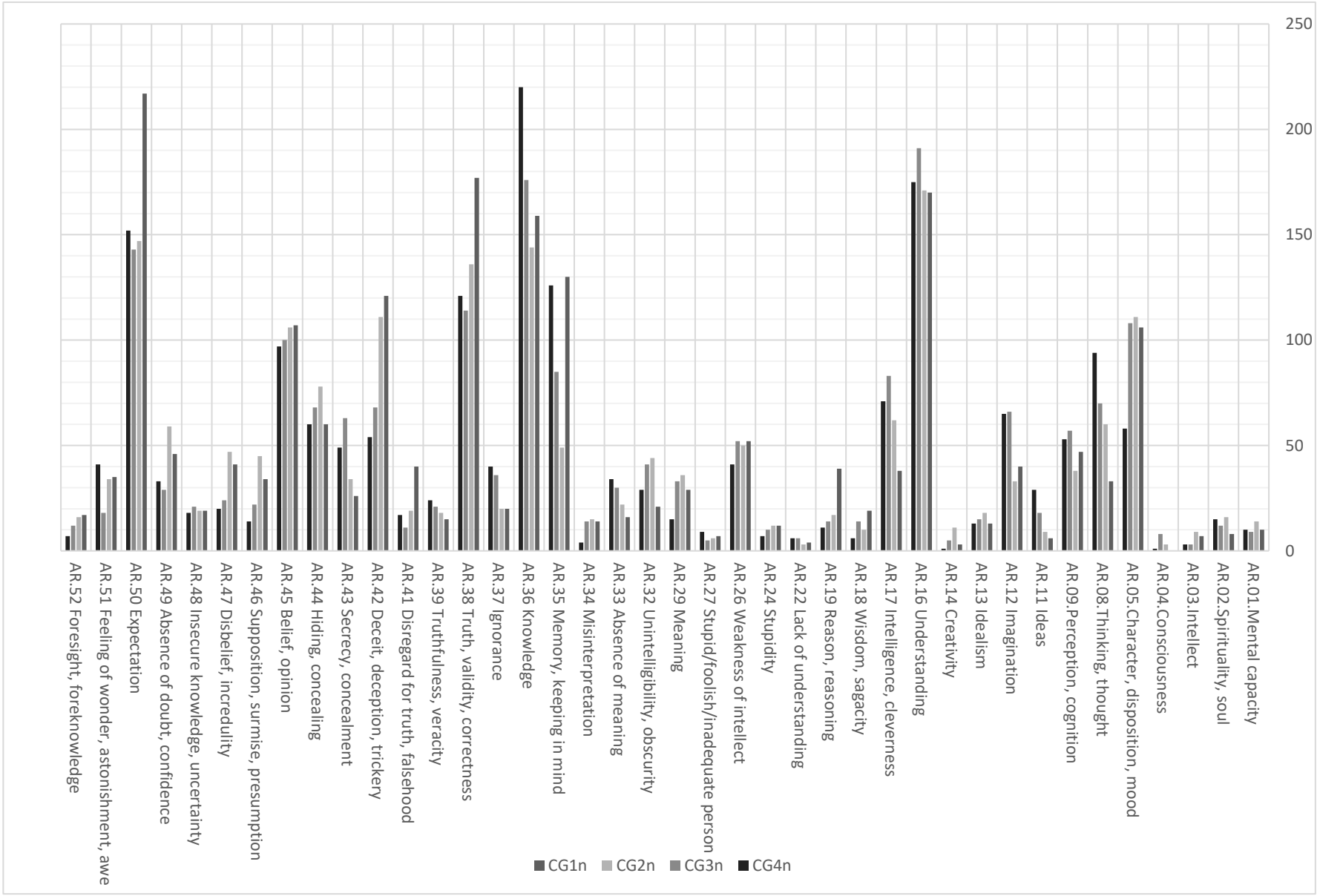

Figure 2 AR Aggregated results across corpus groups

The aggregated semantic metadata analysis in §5.2 highlighted highly relevant domains for all four corpus groups and captured a shift in semantic relevance across the diachronic corpus. This macro perspective allows researchers to examine the distribution of meaning across their corpus at a distance and can be used to select specific semantic fields for further investigation based on their presentation in the text. This serves as a cognate approach to the use of keyness analysis to determine the 'aboutness' of a text (Archer et al. 2006: 1), which can direct 'researchers' attention on aspects of a text (or texts) that deserve further enquiry'. Closer analysis of one highly relevant high-level domain in sections §5.2.1 and §5.2.1.a demonstrated this approach, showing how the results can be explored at different levels of the thematic taxonomy.

These research approaches were made possible by the complexity of the data, which in turn presented the challenge of deciding on the appropriate starting point for enquiry. As shown above, different ways of aggregating the data produced different results, despite relying on the same base dataset. It followed that different approaches to processing the data could in turn inspire different research questions. To explore this further, the sub-domain *AR*.17 (Intelligence, cleverness) was selected for further enquiry, as representative of variation between the groups.

5.2.1.a AR.17 (*Intelligence, cleverness*)

CG1, which contained the works of poets born between 1500 and 1599, had a much lower SR result for *AR.17* (Intelligence, cleverness) than the other three groups, despite reporting average results for the related domains of *AR.16* (Understanding), and *AR.36* (Knowledge). The SR of *AR.17* for CG2, which contained the works of poets born between 1600 and 1699, was almost double that of CG1, while the difference between CG2, CG3, and CG4 was less pronounced. To determine if the data accurately represented a lexical shift between the four corpus groups, the data was queried to reveal the lemmas assigned with *AR*.17 meanings for each group. Table 15 shows the lemmas most frequently assigned meanings from *AR*.17 across the corpus, using the total from all four groups as the sort metric. The results revealed the disparity between the assignment of meanings for *bright* between CG1 and the rest of the corpus, which had a noticeable impact on the aggregated data. This was not caused by the word frequency, as the lemma

bright appeared 175 times in CG1, which was proportionate to the rest of the corpus. In 156 instances, however, the lemma was assigned meanings from *AJ*.08 (Light) instead of *AR*.17. While instances of *bright* were assigned senses from *AJ*.08 in CG2, CG3, and CG4, the words were more likely to be tagged with meanings from both categories in the later corpus groups. Indeed, word *bright* was only associated with *AR*.17 from 1741 onwards, and therefore was not a candidate meaning for the word in CG1. Without the time-based filter, it would not be possible to refine the meanings in this way. The inspection of the results therefore demonstrated a further advantage of the tagger: in addition to including a range of likely meanings, it can inform an analysis of the corpus by restricting inappropriate meanings, thus providing evidence for interpretation.

corp_lemma	CG1	CG2	CG3	CG4	Total
bright	0	22	110	100	232
soul	18	45	54	55	172
deep	11	16	57	53	137
wit	40	36	5	6	87
clear	12	10	18	30	70
strong	12	10	16	26	64
flame	14	23	8	0	45
wide	0	0	24	15	39
strength	6	3	9	14	32
well	10	11	10	0	31

Table 15 AR.17 Top 10 lemmas

The macro analysis in this chapter has so far used the corpus groups as the divisions for comparing the semantic relevance results and has therefore focused on the diachronic variation in semantic relevance. To further investigate the scope of the ambiguity tagger, the second part of the macro analysis chapter looks at the work of an individual author in comparison to the corpus as a whole, using a sample of the aggregated semantic data.

5.3 Macro analysis of single author: John Keats (1795-1821)

The OBEV poem numbers in the corpus made it possible to select the work of a single author for further analysis. John Keats (1795-1821) was chosen for this investigation, as an author who was well-represented in the OBEV collection. An

earlier study by Smith (2006) looked at the language of John Keats as an example for the benefit of using the HTE in stylistic analysis, noting that to fully appreciate the '*inventio*', the practice of 'harnessing materials to hand for creative purposes' of artists, it is 'necessary to have a clear idea of the materials available from which authors could make their choices' (Smith 2006: 169). Here, the historical component of the tagger could be tested to determine if texts annotated with period-appropriate senses could create a unique vantage point for stylistic analysis. Furthermore, while normalising the data removed any restrictions on sample size, the Keats sample opened up more opportunities for comparing the results with existing critical literature, as the corpus captured a number of the poet's most well-known works, with a total of fifteen poems included in the collection.

As a starting point, the semantic metadata for Keats' poems was separated from the rest of the corpus and normalised against the total word count for the sample, using $C = 4664$. As Keats was included in CG3, the aggregated semantic relevance for the group was recalculated to exclude the sample, making it possible to compare the sample with the rest of the corpus. First, to determine if there were significant differences between the semantic relevance of the Keats sample and the rest of the corpus, the average normalised SR for the corpus was identified, using the top tier of the thematic hierarchy. The results, as shown in Figure 3, were able to capture slight differences between the SR of the sample and the corpus, but the distribution of SR across the top tier of the thematic hierarchy seemed to follow a similar pattern for both the sample and the rest of the corpus. For example, both the sample and the corpus had more senses assigned from the domains of *AU* (Emotion), *AM* (Time), *AR* (The mind), *AI* (Physical sensation), and *AL* (space) than any other. The ranking was slightly different for the sample, with *AI* (Physical sensation) identified as the most semantically relevant domain for the sample, despite ranking fourth for the rest of the corpus, but the data failed to reveal any statistically significant differences between the most relevant domains. Curiously, most of the domains with a higher semantic relevance for the sample than the rest of the corpus related to the physical world, with *BF* (Faith) as the only exception to this. However, this data was not granular enough to show if this distribution was typical of the period captured in CG3, or unique to the sample. Consequently, while the average SR reduced the data points for analysis, it was only useful for

making broad observations about the results. While the corpus groups themselves represent notional boundaries, merging the groups reduced the breadth of the data by negating the diachronic elements of the corpus, and was found to be too limited. As a result, the data was reverted to maintain the corpus group boundaries, as with Figure 1 and Figure 2 above.

The expanded results are shown in Figure 4 below, showing the aggregated SR for all four corpus groups alongside the Keats sample. As with the average results, CG3 in Figure 4 excludes the Keats sample from the meanings aggregated within the top tier of the thematic hierarchy. At this level of granularity, the significant domains for the sample could be narrowed down to *AC* (Health and disease), *AE* (Animals), *AF* (Plants) and *AI* (Physical sensation), which all had a higher SR for the sample than any individual corpus group. A closer investigation of the domain with the highest SR for the sample, *AI* (Physical sensation), identified the subdomains *AI*.15 (Hearing/noise) and *AI*.14 (Sight/vision) as the most assigned concepts for the sample. The strong SR of the domains corresponded to the meanings assigned to words that frequently occurred in the sample, including *still, soft, voice, hear* for *AI*.15, and *look, see,* and *eye* for *AI*.14. As the domain was activated by ubiquitous lexical items, it had a high semantic relevance for the corpus as a whole. Again, the data pointed to similarities between the sample and the corpus; while the SR for the domain was higher for the sample, it was not a distinguishing feature of the corpus. It was, however, possible to identify these features by calculating the keyness of the thematic headings assigned to the sample, compared with the rest of the corpus.

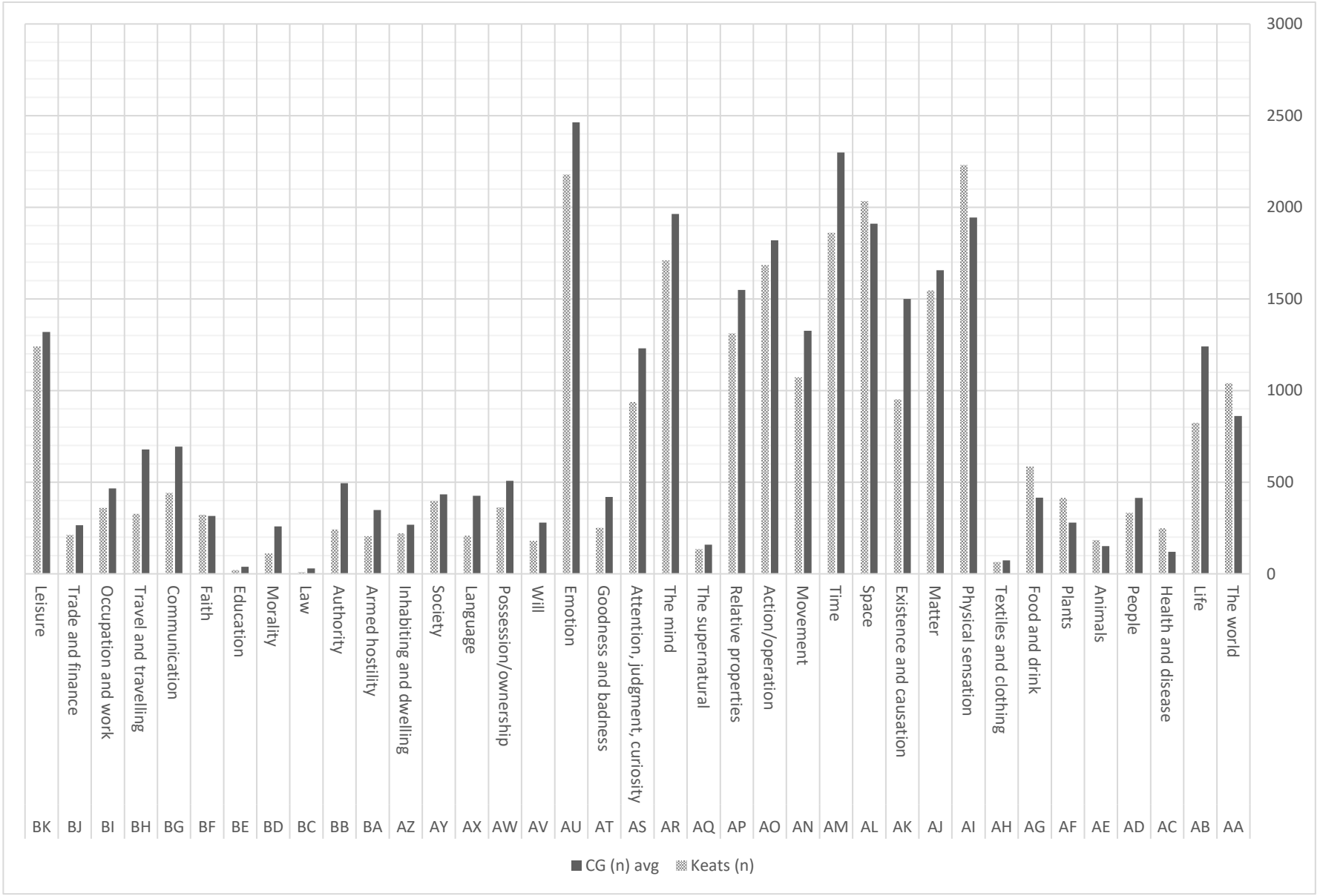

Figure 3 Keats SR cf. Corpus average

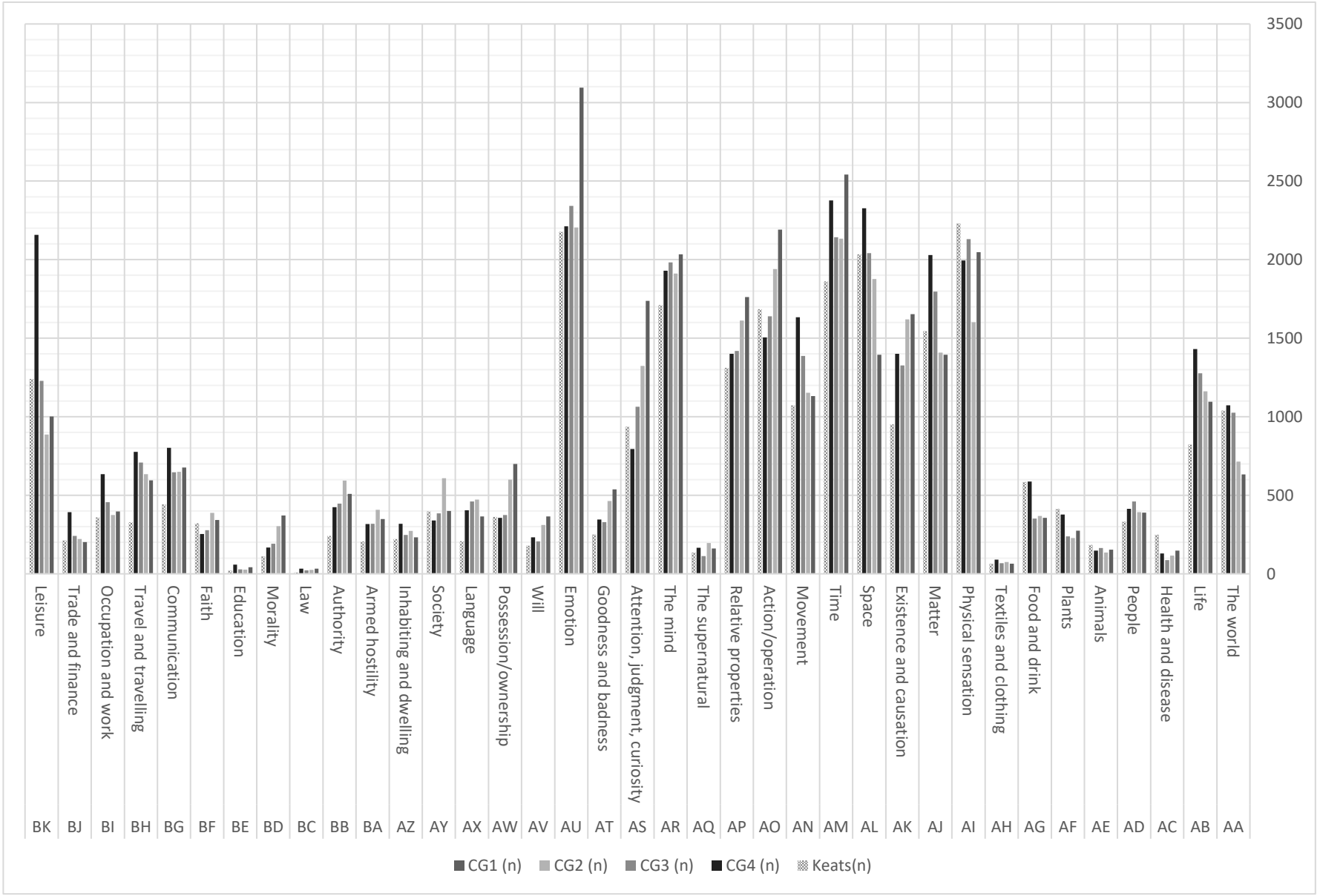

Figure 4 Keats SR cf. Corpus groups

5.3.1 Keyness analysis of semantic relevance

Aggregated semantic data was useful for highlighting prominent semantic domains in the sample, but it was less suited to identifying unique features of the text. Instead, keyness analysis was used to identify the domains with a disproportionate match value for the Keats sample, compared to the rest of the corpus. The keyness of the semantic domains was calculated through another MySQL query, adapted from Lancaster's log-likelihood calculator (Rayson 2016) and modelled on a similar query in Alexander (2011).[70] Using the critical value of 10.83 ($p < 0.001$) as the cut off for significant frequency (Rayson et al. 2004a), the results were split into positive and negative scores; the top thirty positive results for the sample are recorded in Table A12, while the top thirty negative results are listed in Table A13.

Keyness is identified by sorting the table on the log-likelihood value (LL in the appendix), which is indicative of the difference between the observed appearance (here, the number of times meanings were *assigned* from a thematic category, AV), and the expected appearance in a sample of this size (EAV) (Rayson 2016). A larger difference between the AV and the EAV produced a higher LL value, while a smaller difference corresponded to a lower LL score; the positive dataset in Table A12 shows the thematic headings that had a higher AV than expected, meaning that they were assigned more than would be expected, which is reversed for the negative dataset in Table A13. In contrast to the aggregate data, the key thematic tags were incoherent, with disparate concepts ranking highly in both the positive and negative datasets. There was no clear connection between, for example, †18 (Waterlogged/wild land), †1290 (Quality of having sides/being a side), and †700 (Wine), but all three tags reported a significant difference between the AV and the EAV for the sample (Table A12). Neighbouring concepts did appear in the data, with †16 (Poem/piece of poetry) and †24 (Rhyme) for the positive results, and †2079 (Love) and †2087 (Amorous love) in the negative data, but most of the tags corresponded to a seemingly random selection of concepts. Further analysis of the data, however, revealed the true value of the key headings for giving insight into the corpus sample.

[70] The full query is available under Log_Likelihood.sql in the appendix.

The biggest positive difference between the assigned value and the EAV was reported for the thematic category †18 (Waterlogged/wild land), with an LL score of 128.22. The expected AV for the domain was 7.12, based on how frequently meanings from the heading were assigned in the reference corpus, but the tagger assigned 53 meanings from the domain to the Keats sample. This total could be split into 18 tags assigned to the words *natural, bushes, salt, waste, desolate, wild, forest,* and *heath* in [623] 'Song of the Indian Maid' from 'Endymion'; 8 tags assigned to *sunk, wild,* and *waste* in [624] 'Ode to a Nightingale'; 9 to *wild, bare, forest,* and *waste* in [625] 'Ode to a Grecian Urn'; 6 assigned to *forest* and *waste* in [626] 'Ode to Psyche'; and 12 to *wild* in [633] 'Las Belle Dame sans Merci'. What at first appeared to be a curiously specific heading revealed a broader pattern upon inspection, which was further supported by the positive keyness of several concepts pertaining to the land, weather, and nature. The keyness of these domains in relation to the sample gained further importance when considered in relation to existing research that examines the 'human relationship with nature' through Keats' work (Lawrence 1999; Slattery 2005). Notably, while previous studies relied on individual poems as evidence of Keats' blending of nature and narrative, the keyness results reveal a deeper pattern within the author's work.

When looking at the data through the perspective of existing criticism, further patterns began to emerge. The positive key headings of †294 (Wasting disease), †298 (Disordered breathing), †317 (Maiming/mutilation), †322 (Mental illness), †328 (Examination), and †333 (Non-scientific treatments) were easier to understand in relation to existing research into the impact Keats' medical training had on his writing (Smith 2006). A further cluster of key headings falling into the domain of *AF* (Plants) seemed to support the lasting impact of Keats' interest in botany, and provided empirical evidence for this less-explored facet of the author's writing (Evans 2002). The negative keyness of †2079 (Love) and †2087 (Amorous love) in comparison to Keats' contemporaries could be seen as supporting the claim that the author rejected 'old romance' in favour more 'skeptical lyrics' (Stillinger 1968: 605). Indeed, these observations could be drawn together to answer the question of whether the author successfully 'adopt[ed] the "calmness of a Botanist" in order to classify the observed "differences in human character"' (Evans 2002: 37). The semantic metadata produced by the tagger was particularly suited to revealing these connections as it extended the range of analysis beyond

the individual word, thus highlighting significant distribution of concepts in the sample.

The key thematic headings were also useful for investigating how the ambiguity tagger coped with the figurative language of the corpus. As an example, the word *vintage* in [624] 'Ode to a Nightingale' was annotated with three different meanings from the key thematic category †700 (Wine), appearing in the first line of the second stanza: 'O, for a draught of vintage! That hath been/'. The tagger was therefore capable of identifying the use of *vintage* to mean *wine* in context: 'the speaker is literally asking for a glass of wine, but this is never explicitly mentioned. Instead, the poem presents us with figurative terms for wine, calling it "a draught of *vintage*"' (Furniss & Bath 2013: 13). An example from later in the poem demonstrates how the breadth of the HTE data could be used to advance existing criticism, with the assignment of »130614 (inspire with affection/kindly feelings) and »128256 (Impetuous) as two likely meanings of *warm* in 'a beaker of the full warm South'. Here, the tagger offers alternatives to Furniss & Bath's (2013) interpretation of the metaphor as drawing on 'some of the connotations of southern Europe [...] used in holiday and wine advertisements' (Furniss & Bath 2013: 13), instead positioning the ameliorating and destructive qualities of wine in direct conflict, with the speaker abandoning both in favour of the pursuit of poetry by the end of the. In working down from key headings identified at macro level to a close reading of the text and corresponding semantic metadata, this example highlights one of the ways that the tagger could be used to explore the corpus. The use of the semantic metadata in a close reading of the text is explored in further detail in the following chapter, which reverses the analytical process by starting the enquiry at the level of the poem, further expanding the range of applications for this tagging process.

The macro analysis of a single author was carried out to demonstrate the potential of the ambiguity tagger for producing useful semantic metadata, which could be used to inform an analysis of a smaller sample of poetry. While the observations drawn from the Keats sample data aligned with existing criticism, and served to highlight further avenues for discussion, they were directed towards demonstrating the scope of the data, and not intended as a definitive discussion of Keats' work. The keyness analysis of the sample returned the thematic tags that

were key when compared to the rest of the corpus, which limited their application; the sample could not be considered as representative of Keats' style as it contained an unbalanced sample of the poet's work, and a larger reference corpus would be necessary to obtain more robust results. It was clear from reviewing the data, however, that the tagger produced metadata that could be used in macro analysis of poetry.

5.4 Conclusion

This chapter examined different approaches for utilising the semantic metadata produced by the tagger in a macro-level analysis of poetry. In exploring the data, the aggregated semantic metadata analysis in §5.2 was found to be better suited to exploring broader conceptual trends between the corpus groups, and less effective in providing evidence to support existing research. The divisions between the corpus groups impacted on this, making it more difficult to connect the results to scholarship. In this respect, the sample analysis of a single author in §5.3 was better suited for examining how the results could be used to reflect existing knowledge, as it was possible to review the semantic metadata within existing critical parameters. While the sample analysis successfully demonstrated a range of applications for the ambiguity tagger, it was limited by the selection of poems contained in the OBEV corpus. Consequently, while the data lends itself to a macro-level analysis, this type of investigation would be even more effective if working with a larger corpus.

Chapter 6 Micro analysis

6.1 Introduction

When working with corpora, it can be difficult to separate the whole from the constituent parts, and it can seem counterintuitive to engage qualitatively with smaller sections of the corpus when working within a quantitative methodological framework. Indeed, much of the debate surrounding CL practice looks at reconciling the subjective approach to analysis with empirical methods (McEnery & Hardie 2012). However, as shown in Chapter 5 above, quantitative analysis of corpus data can involve subjective input in both selecting the entry point into data analysis and in the decision to pursue further inquiry into specific features. There is, therefore, no arbitrary line that can be drawn between objective and subjective analysis of the annotated corpus; rather, the parameters that guide the enquiry at either level should reflect the appropriate approach for critical engagement. In the previous chapter, the case studies were selected by reviewing the general semantic features of the corpus at a distance. Correspondingly, the analysis centred on the different approaches to preparing and analysing the data, and how this impacted on what the semantic metadata could reveal about the corpus at a macro-level. In this respect, this chapter differs from the previous section by focusing on the extent to which the ambiguity tagger can be used in a micro-level analysis of the text.

To determine if the tagger presents any advantage to a closer reading of the text, this chapter narrows the focus of the analysis by examining the semantic metadata produced for individual poems. The poems were selected to evaluate different forms of micro analysis, and this analysis is split into two distinct case studies: the first case study looked at the semantic metadata for poems by two authors from different corpus groups, Alexander Pope and Lord Byron, to evaluate the efficacy of a comparative analysis, while the second case study examined the results for a selection of poems by a single author, William Blake, and considered whether they could be used to meaningfully engage with the author's work. In this regard, this

approach differed from the analysis of John Keats in §5.3 above, which looked at the work of one author in the OBEV corpus as a single coherent sample.

A particular challenge of this approach was drawing the boundary between investigating the potential of the semantic metadata in a micro-level analysis and carrying out a full-scale analysis, which is not within the scope of this thesis. To this end, this chapter presents possible avenues for further investigation using the semantic metadata to explore the value of further enquiry (RQ4), and displays the extent to which the annotation can contribute to further analysis of poetry from different literary periods (RQ4a).

6.2 Comparison: Alexander Pope (1688-1744) and Lord Byron (1788-1824)

Though a frequent practice in criticism, the comparison of texts from different literary periods is not without controversy: critics acknowledge that writers of any period are rarely immune to the 'dominant presuppositions and attitudes' of their time, but often speak of the influence that past and contemporary poets have on each other's work (Stephens & Waterhouse 1990: 96–97). To understand the work of one author in relation to another, it is necessary to look beyond the literary periods or formal classifications either may belong to, and to focus on these elements alone would miss out on a range of influencing factors that separate one author from another, and indeed one poem from the next. Because of this, a comparative analysis of all creative artefacts must allow for factors of influence that extend beyond the zeitgeist.

This informed the choice of poets for the comparative case study, as both simultaneously embody the dominant attitudes of their time and stand alone as masters of their own unique poetic style. The first, Alexander Pope (1688-1732), was active during the Augustan period in English literature, an age that saw a civilised and measured tone replace the 'conflicts and enthusiasms' associated with Puritanism and the Restoration (Preminger et al. 1974: 230–231; Daiches 1979: 590). As with all periods in literature, the Augustan age serves as a boundary that covers a range of voices and contradicting styles, and Pope, seen as the 'dominant poetic figure' of the age, still stands apart from Swift and Dryden, his

contemporaries and dominant literary figures of the time (Preminger et al. 1974: 230; Daiches 1979: 621). Pope's distinctiveness, wit, and the depth and variety he was able to draw from within the Pastoral constraints of the time made him an interesting candidate for further investigation (Daiches 1979: 324).

The second poet chosen for the case study holds a more divisive position in literary history: Lord Byron (George Gordon Byron; 1788-1824), who played an active role during the Romantic period. While Pope is often lauded for his carefully constructed verse, Lord Byron remains a more controversial literary figure, depicted on the one hand as 'arrogant, passionate, and wayward' and overrated when compared to other Romantic poets due to his 'lack of verbal distinction' (Preminger et al. 1974: 233), and on the other as misunderstood, commanding a range of 'vocabularies and registers' unmatched by his contemporaries (Stephens & Waterhouse 1990: 120). These competing valuations made Lord Byron a curious case for further study, and existing comparisons to Pope and Augustan registers, which stand at odds with the shift from 'Society to Self and Nature' that defines the movement from Augustan to Romantic verse, made Byron a suitable choice for this case study (Stephens & Waterhouse 1990: 97;120).

In selecting these authors, however, the first limitation of comparative analysis was highlighted: the poems available for the micro analysis were restricted by the OBEV corpus, which contained only three poems by Pope and five poems by Byron, and therefore made only a small sample of each poet's work available for analysis (Quiller-Couch 1919/1999). Indeed, later editions of the OBEV included a broader sample of both poets' work, most notably through the inclusion of their mock-epics, and both editors acknowledged that Quiller-Couch's edition under-represented both Pope and Byron in this regard (Gardner 1972; Ricks 1999). This limited the scope of the analysis as it was not possible to relate it directly to existing criticism, which juxtaposed the authors' mock-epics in their analysis (England 1975; Beatty 1990; Rawson 1990). Consequently, the focus was shifted to examine any salient semantic patterns present in the authors' annotated poems, beginning with a survey of the dominant themes in each individual text.

6.2.1 Aggregated results by poem

Eight poems in total were available for the comparative analysis: Pope was included in CG2 and tagged with senses available between 1550-1750, while Byron's poems were included in CG3 and tagged with meanings from 1650-1850. As a starting point, the aggregate measure used in Chapter 5 was used to identify the overarching domains with the most semantic relevance for each poem. The results are summarised in Table 16 for Pope and Table 17 for Byron, which show the top three thematic domains for each poem, a list of the unique words that were assigned meanings from that domain, with the number of assigned senses in brackets next to each word. The 'Total' column shows the total number of meanings assigned for each poem, and the total for each listed domain, thus providing a reference point for the proportion of meanings attributed to the top two domains. Taking as an example the first poem in Table 16, [440] 'On a certain Lady at Court' by Pope, the data shows that of the 252 conceptual tags assigned to the poem, the largest concentration fell within the domains AR (The mind) and AU (Emotion), with 45 distinct meanings assigned from the former domain and 31 from the latter. Surveying the results in this way displays the most frequently referenced semantic domains in each poem,[71] and the lexical variety in the text that contributed to the high semantic relevance for the domains. Consequently, the results provide an alternative starting point for a corpus-led analysis of the text than the one provided by traditional techniques, including wordcount and keyword analysis.

When comparing the results for Pope and Byron's top thematic headings, it would seem that senses ranked most highly for Pope's work were triggered by a broader range of lexical items than in Byron's work, when taking into account the length of each poem. To verify this, the type-token ratio measure was adapted to calculate lexical variety within the top three semantic domains for the poems. The type-token ratio (TTR) is typically used to determine the richness of vocabulary in a text (Scott 1998; Xiao & Yue 2012), but it has previously been adapted to investigate the lexical realisation of semantic tags in a corpus as the 'tag-lemma ratio' (Semino et al. 2005: 3). Here, the TTR for the top three senses was calculated between the

[71] See Archer (2006) for an analogous study of Shakespeare's tragedies and comedies.

unique corresponding lemmas (type) and the total number of activations for each heading (token). As this analysis considered higher level headings rather than specific tags in calculating lexical variety within a domain, it is referred to as the 'domain richness measure' (DRM) to distinguish it from other TTRs. Furthermore, while TTR is sensitive to text length, and often calculated by standardising the results to every n words in a corpus,[72] the DRM is calculated to determine lexical variability of individual domains and can be used to highlight differences in how the domains are realised in the text. The DRM for the top three categories is shown in the final column of Table 16 and Table 17 below, with a higher ratio indicating more lexical variety, up to a maximum of 1.

The average DRM for the top three domains of Pope's poems is 0.28, while for Byron the average was 0.23. While this confirms the initial observation of higher lexical variety in Pope's work, the difference in the average is minor. More significantly, domains with a higher DRM would be harder to identify through lexical corpus analysis methods, such as keyness (Semino et al. 2005). As shown in Table 16 and Table 17 below, domains that correspond to more varied lexical items include multiple words that appear only once or twice in the poem, which for longer poems might not be significant enough to highlight the prevalence of a particular semantic field in the text. Thus, the aggregated results can provide a unique perspective of the semantic properties of the text.

[72] The default for standardising the type-token ratio in WordSmith tools is $n = 1{,}000$ (Scott 1998).

Top Headings	Corresponding words	Total	DRM
[440] 'On a certain Lady at Court'		252	
AR (The mind)	*know*(20), *reasonable*(5), *folly*(5), *sensible*(4), *certain*(3), *witty*(2), *warp*(2), *fault*(2), *soft*(1), *aver*(1)	45	0.22
AU (Emotion)	*melancholy*(9), passion(6), *pride*(6), *envy*(4), *grave*(2), *gay*(2), *friend*(1), *good-humour*(1)	31	0.26
AP (Relative properties)	*equal*(10), *mixture*(6), *most*(5), *attend*(3), *conspire*(3), *uncommon*(1), *handsome*(1), *rumour*(1)	30	0.27
[441] 'Elegy to the memory of an Unfortunate Lady'		2027	
AU (Emotion)	*proud*(16), *mourn*(14), *love*(12), *warm*(7), *humble*(7), *melt*(6), *heart*(6), *woe*(6), *burn*(6), *state*(5), *mean*(5), *stone*(4), *low*(4), *fury*(4), *flow*(4), *fall*(4), *grave*(4), *tear*(4), *mournful*(4), *light*(3), *tender*(3), *grieve*(3), *sullen*(3), *air*(2), *blood*(2), *trembling*(2), *beloved*(2), *dull*(2), *complaint*(2), *glow*(2), *desire*(2), *weeping*(2), *rise*(2), *pang*(2), *fool*(2), *unlamented*(1), *lady*(1), *unpitied*(1), *wife*(1), *kind*(1), *cold*(1), *congenial*(1), *pitying*(1), *out*(1), *peep*(1), *sable*(1), *roman*(1), *midnight*(1), *public*(1), *lover*(1), *marble*(1), *face*(1), *firm*(1), *lightly*(1), *blow*(1), *ground*(1), *make*(1), *well*(1), *mute*(1), *want*(1), *bleeding*(1), *muse*(1), *yield*(1)	186	0.34
AB (Life)	*life*(20), *breast*(16), *hand*(15), *grave*(8), *breath*(7), *heart*(6), *tomb*(5), *rise*(4), *die*(4), *earth*(4), *fall*(4), *sepulchre*(4), *lay*(3), *lie*(3), *green*(3), *bosom*(3), *age*(3), *hearse*(3), *death*(3), *ball*(2), *ear*(2), *rest*(2), *race*(2), *ghost*(2), *flower*(2), *sleep*(2), *part*(2), *state*(2), *cold*(2), *roll*(2), *marble*(2), *child*(2), *glow*(2), *pass*(2), *stand*(1), *yield*(1), *light*(1), *day*(1), *bier*(1), *body*(1), *dying*(1), *compose*(1), *burn*(1), *sable*(1), *nature*(1), *dirge*(1), *low*(1), *blast*(1), *see*(1), *blood*(1), *memory*(1), *blow*(1), *ground*(1), *relic*(1), *make*(1), *leave*(1), *stone*(1), *dust*(1), *air*(1), *elegy*(1)	174	0.34
AO (Action opetation)	*how*(18), *so*(9), *grace*(8), *fall*(8), *art*(6), *idle*(5), *avail*(5), *kind*(5), *stand*(5), *unfortunate*(4), *melt*(4), *vulgar*(4), *way*(4), *keep*(4), *thus*(4), *guardian*(4), *love*(3), *rest*(3), *tender*(3), *lazy*(3), *sleep*(3), *mean*(3), *leave*(2), *air*(2), *form*(2), *confine*(2), *lay*(2), *friendly*(2), *forget*(2), *well*(2), *earth*(2), *bear*(2), *dress*(2), *power*(2), *pageant*(1), *rite*(1), *unpaid*(1), *act*(1), *firm*(1), *compose*(1), *grieve*(1), *dance*(1), *public*(1), *pass*(1), *lie*(1), *bestow*(1), *blow*(1), *silver*(1), *peaceful*(1), *fault*(1), *wealth*(1), *useless*(1), *wait*(1), *line*(1), *rule*(1), *bleeding*(1), *gate*(1)	162	0.35

Top Headings	Corresponding words	Total	DRM
[442] 'The Dying Christian to his Soul'		414	
AN (Movement)	*fly*(18), *wing*(8), *mount*(7), *tremble*(5), *come*(4), *open*(4), *recede*(3), *away*(2), *draw*(1)	52	0.17
AR (The mind)	*steal*(11), *sense*(9), *spirit*(6), *sound*(6), *fond*(5), *frame*(2), *nature*(2), *open*(2), *flame*(1), *say*(1), *draw*(1), *tell*(1), *heaven*(1), *eye*(1), *ring*(1)	50	0.30
AB (Life)	*vital(10), life(10), breath(7), mortal(4), dying(2), nature(2), sight(2), death(2), spark(1), away(1), draw(1), ear(1), grave(1)*	44	0.30

Table 16 Pope top thematic headings

Top Headings	Corresponding words	Total	DRM
[597] 'When we Two parted'		433	
AJ (Matter)	*cold*(20), *light*(16), *chill*(6), *pale*(5), *sunk*(2), *shudder*(1), *spirit*(1)	51	0.14
AM (Time)	*long*(27), *year*(14), *morning*(5), *hour*(2), *now*(1), *light*(1)	50	0.12
AR (The mind)	*know(18), secret(6), truly(4), tell(4), light(3), deeply(3), spirit(3), feel(2), heart(2), deceive(2), foretold(1)*	48	0.23
[598] 'For Music'		275	
AI (Physical sensation)	*voice*(8), *sweet*(7), *soft*(7), *sound*(6), *still*(6), *wind*(3), *deep*(3), *listen*(3), *seem*(2), *summer*(2), *water*(1), *lie*(1)	49	0.24
AP (Relative properties)	*like*(8), *full*(7), *as*(3), *sound*(1), *so*(1)	20	0.25
AL (Space)	*bow(5), wind(3), heave(3), ocean(2), wave(2), lie(2), weave(2)*	19	0.37
[599] 'We'll go no more a-roving'		276	
AM (Time)	*night*(12), *go*(9), *return*(5), *soon*(5), *late*(3), *wear*(3), *make*(3), *day*(3), *yet*(3), *outwear*(1), *out*(1), *rest*(1)	49	0.24

Top Headings	Corresponding words	Total	DRM
AO (Action/operation)	*rest*(10), *still*(8), *so*(6), *return*(4), *pause*(3), *moon*(2), *breathe*(2)	35	0.20
AK (Existence and causation)	*go(9), make(8), heart(2), soul(2), out(2), light(2), wear(1)*	26	0.27
[600] 'She walks in Beauty'		389	
AJ (Matter)	*light*(16), *dark*(8), *shade*(7), *ray*(6), *bright*(5), *mellow*(4), *lighten*(4), *soft*(4), *tender*(3), *pure*(3), *starry*(2), *day*(2), *sky*(1), *wave*(1), *softly*(1), *serenely*(1), *glow*(1)	69	0.25
AO (Action/operation)	*how*(12), *so*(6), *soft*(5), *grace*(3), *softly*(3), *calm*(3), *tender*(2), *light*(2), *peace*(2), *thus*(1), *goodness*(1), *heart*(1)	41	0.29
BH (Travel and travelling)	*walk*(28), *win*(3)	31	0.06
[601] 'The Isles of Greece'		1947	
AL (Space)	*fill*(28), *where*(25), *set*(22), *high*(20), *place*(14), *head*(13), *stand*(9), *face*(9), *sit*(8), *back*(8), *lie*(5), *lay*(5), *strike*(5), *fall*(5), *rise*(4), *bore*(4), *line*(4), *even*(4), *leave*(4), *there*(3), *here*(2), *breast*(2), *at*(2), *arise*(2), *come*(2), *in*(2), *break*(2), *rising*(2), *mountain*(2), *broad*(2), *see*(2), *out*(1), *hand*(1), *beat*(1), *ship*(1), *phalanx*(1), *present*(1), *chain*(1), *bind*(1), *rock*(1), *spring*(1), *rank*(1), *shield*(1), *steep*(1)	234	0.19
AI (Physical sensation)	*still*(21), *sound*(14), *look*(12), *silent*(10), *see*(6), *virgin*(6), *voice*(6), *dead*(6), *feel*(6), *sweep*(5), *call*(4), *mute*(4), *break*(4), *sing*(3), *face*(3), *echo*(3), *eye*(3), *brow*(2), *blush*(2), *living*(2), *strike*(2), *set*(2), *go*(2), *line*(2), *gaze*(2), *maid*(2), *murmur*(2), *lie*(1), *voiceless*(1), *beat*(1), *grave*(1), *ship*(1), *out*(1), *find*(1), *harp*(1), *fall*(1), *arise*(1), *dumb*(1), *give*(1), *rise*(1), *peace*(1), *render*(1)	151	0.28
AM (Time)	*set(18), new(14), long(10), day(8), yet(6), come(6), now(5), summer(4), present(4), hour(4), go(4), high(4), fall(4), stand(3), strike(3), save(2), eternal(2), sun(2), sound(2), break(2), at(2), living(2), back(1), distant(1), face(1), even(1), arise(1), hand(1), beat(1), below(1), sit(1), give(1), then(1), further(1), sure(1), rock(1), exist(1), dwell(1), see(1), spring(1), wave(1), ne'er(1), down(1)*	132	0.33

Table 17 Byron top thematic headings

At the aggregated semantic level, the results for individual poems could be used to identify differences in semantic relevance between multiple poets, as suggested by the data above. The distribution of semantic relevance across the poems of Pope and Byron in the OBEV corpus, for example, could be used to further investigate whether the apparent division between abstract and physical domains is supported in the authors' other work, or if the trend extends to their contemporaries. The small number of poems included for each poet, however, meant that it was not possible to use the semantic metadata as a valid starting point for a quantitative comparison, as there was not enough data to support an investigation into the significance of these observed differences. Consequently, the conclusion drawn from this stage of the analysis pointed to the need for a larger sample of an author's work to allow direct comparison between multiple authors on the basis of SR distribution across a selection of their poems. As a potential to overcoming this issue, an alternative approach to comparative analysis between two authors was trialled, using the semantic metadata produced for a single poem by each author as a potential entryway into critical analysis.

6.2.2 Alexander Pope

One of the salient features of the SR distribution across all three Pope poems was the trend towards lexical representations of abstract ideas, such as those falling within the domains of *AR* (The mind) and *AU* (Emotion), with different aspects of those domains evoked in each poem. The high SR of *AN* (Movement) in [442] 'The Dying Christian to his Soul' appeared as an outlier to this, particularly when contrasted with the high SR of *AB* (Life) in [441] 'Elegy to the memory of an Unfortunate Lady'. While both domains relate more to the physical world and its circumstances, the meanings assigned within *AB* cover a broader conceptual range, particularly when considering the figurative use of words such as *life, heart,* and *rest,* and made the comparatively narrow focus on movement and flight within *AN* in [442] even more conspicuous, and the SR results appear uncharacteristically high for a physical domain (see §3.2.2). As the second most prominent domain for [442] was AR (The mind), which represented the abstract ideas present in the other two poems, the poem was chosen for further inquiry as an example of typical and atypical SR distribution, and to determine if these results could be explained in relation to the text.

Returning first to the aggregated results, in addition to the high SR for *AN* (Movement) and *AR* (The mind), the third most prominent domain for the poem was *AB* (Life), with 44 meanings assigned from the field to words in the poem. This corresponded sensibly to the poem's central theme of the acceptance of death, with words such as *vital, life, breath, mortal, dying, nature, sight,* and *death* assigned meanings from *AB*. Indeed, the fact that this was not the top domain for the poem was curious in its own right, and further highlighted the significance of a high SR for *AN*. To explore these results further, the poem was cross-referenced with the SR results to determine if the distribution of the meanings revealed anything in relation to the text. For reference, the full text is shown as Poem iv below, edited to highlight the words falling within *AN* (Movement) in bold, and *AR* (The mind) as underlined, and *AB* (Life) in italics.

THE *DYING* CHRISTIAN TO HIS SOUL

VITAL spark of heav'nly flame!
Quit, O quit this *mortal* frame:
Trembling, hoping, ling'ring, **flying**,
O the pain, the bliss of *dying*!
Cease, fond *Nature*, cease thy strife,
And let me languish into *life*.

Hark! they whisper; angels say,
Sister Spirit, **come *away***!
What is this absorbs me quite?
Steals my senses, shuts my *sight*,
Drowns my spirits, ***draws*** my *breath*?
Tell me, my soul, can this be *death*?

The world **recedes**; it disappears!
Heav'n **opens** on my eyes! my *ears*
With sounds seraphic ring!
Lend, lend your **wings**! I **mount**! I **fly**!
O *Grave*! where is thy victory?
O *Death*! where is thy sting?

Poem iv [442] Alexander Pope (1688-1744)

Cross-referencing the results with the original text reveals that the meanings from all three domains were distributed across the poem, and in some instances overlapped on a single word. The most notable example of the latter was the meanings assigned to *draws*, which included †26102 (Inhale) from *AB*, †102731 (draw towards/attract) from *AN*, and †118197 (Pervert/distort) from *AR*. In this

case, the ambiguity tagger pointed towards both literal and figurative interpretations for *draws*, based on the context it was used in.

When considering the rest of the meanings that fall within the top domains in relation to the poem, the tagger seemed to reveal the contradictory feelings of the speaker. The speaker, knowing that he is dying, is energetic and erratic, calling out to death as if challenging it to a fight: 'Quit, O quit this mortal frame:/[..]/ O the pain, the bliss of dying!'. This energy is maintained to the very end, with the final stanza shown in Poem iv above, where the speaker continues taunting death as the world disappears around him. The tagger also picked up the contrast between this manic energy within *AB* and the quiet of *AR*, which picks up when the speaker seems to temporarily lose confidence and starts to sound almost afraid: 'What is this absorbs me quite?/ Steals my senses, shuts my sight,/Drowns my spirits, draws my breath?'. The meanings within *AN*, when considered in relation to *AR* and *AB*, appear to move between resignation and defiance, both accepting the inevitability of moving on, and at the same time refusing to go quietly: 'Lend, lend your wings! I mount! I fly!'. Indeed, while at the aggregated level the results seemed to deviate from Pope's other works, a closer inspection of the data was in-keeping with existing valuations of his work:

> he had a subtle ear for variety within unity, as well as the kind of wit which sought and achieved most effective expression in those verbal devices which, by varying delicately the balance or progression of the thought to which the verse had been leading, at the same time demonstrated technical virtuosity and created new overtones of meaning (Daiches 1979: 324).

A larger sample of poems would be necessary to take this analysis further, but the results indicated that the ambiguity tagger could be used to carry out a closer reading of the text at this level. In this example, the aggregated results formed the basis for the closer examination of the text, which could be a useful technique when selecting texts for closer examination and serve as an entry point into corpus-driven analysis of poetry.

6.2.3 Lord Byron

Following on from the closer examination of Pope's work, the poem chosen for closer examination of Byron's work was [599] 'We'll go no more a-roving', which

similarly deals with the concept of mortality, though from the perspective of growing older. The aggregated results for the poem, as shown in Table 17 above, appeared to be more straightforward than those for Pope. Indeed, one of the reasons for the selection of this poem for the attempt at comparative analysis based on its thematic relationship to [442] was the straightforwardness of the top domains in Byron's work. In short, the results did indicate a reliance on more concrete concepts in the poet's tagged poems, and therefore did not show a particular benefit to using an ambiguity tagger. The closer analysis was carried out to determine if this was the case, or if summarising the data concealed anything curious in the semantic metadata for the poems. Again, for reference, the poem is reproduced below, as Poem v. As with Pope's poem, the text was marked up to reflect the words that corresponded to the top domains, with *AM* (Time) shown in bold, *AO* (Action/operation) underlined, and *AK* (Existence and causation) in italics.

WE'LL ***GO*** NO MORE A-ROVING

SO, we'll ***go*** no more a-roving
So **late** into the **night**,
Though the heart be still as loving,
And the moon be still as bright.

For the sword **outwears** its sheath,
And the *soul* ***wears out*** the breast,
And the *heart* must pause to breathe,
And love itself have **rest**.

Though the **night** was **made** for loving,
And the **day returns** too **soon**,
Yet we'll ***go*** no more a-roving
By the *light* of the moon.

Poem v [599] George Gordon Byron, Lord Byron (1788-1824)

Byron's poem differs in SR distribution to Pope's as it is less divisive, with both high-ranking fields corresponding to different elements of ageing, effectively conceding that time is running out through the dual presentation of *AM* and *AO* alongside *AK*. Individual meanings assigned from those domains similarly reflect this theme, even when they offer multiple interpretations of a word, as in the case of *night*, which was tagged with †88678 (Night), †88679 (as a division/period of time), †88680 (marking lapse of time), and †88698 (the kind of night one has

had), thus offering different perspectives on the word's use in 'So late into the night'.

The word *go* presented more of a problem, as it was tagged under †90381 (in specified state) under *AM* and †77626 (in a specific manner) within *AK*. Here, the tagger could be configured to prioritise a more general definition of *go*, such as †189127 (Departure/leaving/going away) or †100829 (Motion in a certain direction), both of which fell just outside the RMV cut-off value for this test. The unexpected appearance of *moon* as one of the lemmas in *AO* corresponds to a fairly specific sense of the word, †80181 (typically), in †1453 (Practical impossibility). Within this category, the specific sense of *moon* sits under the sub-category of †80179 (condition of being unattainable), which helps to explain why the sense was picked up as relevant in context. As with *go*, users of the tagger can choose to remove these highly granular senses of words from the aggregated analysis, or manually correct mis-identified tags to configure the tagger for their corpus parameters.

Unfortunately, the tagger struggled with *a-roving*, failing to identify any matches because the word was not tokenised during the initial pre-processing stage and was therefore retained in the corpus data as one item. It is possible that the RMV of the candidate meanings would have been different if the word was tagged successfully, but the analysis proceeded with the returned data as this issue corresponded to the CLAWS lemmatisation process and not the annotation parameters. Resolving this issue consistently would require another pre-processing stage before POS-tagging the corpus, to resolve non-standard hyphenation that falls outside of the CLAWS ruleset. It is also possible to resolve lemmatisation errors during the annotation process and re-tag incorrectly tokenised items. The present analysis retained the error to better present the strengths and limitations of the annotation process with minimal interference. However, as the HTE has thirteen recorded meanings for *roving,* eleven of which fall within the date ranged used for CG3, it is possible that multiple meanings would have been assigned to the word if tagged, thus impacting on the summary ranking in Table 17.

The semantic metadata for Byron's poem was more precise, which was useful for exploring the text at macro level and identifying the dominant themes in the poem without needing to engage with the text directly. At the micro level, the results reinforced the interpretation of the poem as a lament of ageing, and therefore aligned with existing readings of the poem (Daiches 1969: 925). This was an encouraging result, as it supported the use of the tagger in validating existing interpretations, which was a key desired outcome for the annotation process. Overall, however, the lack of ambiguity in the poem lessened the scope of the annotated analysis, as there were fewer possible interpretations made available by the ambiguity tagger. This is not inherently an issue, as the high-confidence RMV cut-off was chosen precisely to exclude meanings that were unlikely based on context, and in this case it highlighted the straightforward nature of the poem. It did, however, constrain the scope of the analysis, as the primary finding was a lack of ambiguity in the text. One notable exception to this was the annotation of *soul* and *breast* in 'And the soul wears out the breast', which revealed an interesting interpretation of the line: one of the thematic tags assigned to *soul* was †117146 (High intelligence, genius), while *breast* was assigned †115311 (Mind, soul, spirit, heart) as one of the likely meanings of the word. While this does not contradict the accepted interpretation of the lyric as a reflection of Byron's concern over 'suddenly facing the loss of his youth and of his emotional venturesomeness', it adds a possible interpretation that the author was concerned about the intense impact of his intellectual pursuits, and perhaps afraid that the better part of his creative output was behind him (Daiches 1969: 925). Byron wrote the poem after a period of illness in 1817, and it is possible that his concern over his perceived weakness extended beyond the loss of his youthful energy (Abrams & Greenblatt 2000: 560).

As with the previous discussion of Pope's poem, this enquiry was restricted to a discussion of how the tagger could be used to explore the use of ambiguous meanings in the text. While Pope's poem provided a broader scope for this enquiry, the results for Byron's poem were still useful in revealing the dominant themes in the text and as a starting point for further investigation of a broader interpretation of the concerns embedded in the text. As a comparative analysis, however, this case study highlighted one of the challenges of using an anthology-based corpus to investigate poetry. While the OBEV corpus provided enough scope

for a macro discussion of the changing SR distribution across the centuries represented in the anthology, it was harder to adapt it to a comparative analysis as there was no control over the content from each author that could be used in the discussion. There was enough evidence in the limited data for each author, however, to justify extending the analysis to a larger sample of each author's work, which would enable further exploration through the ambiguity tagger. To further investigate the scope of the tagger, an additional comparative analysis was carried out, focusing on the work of a single author from the OBEV to determine if the restrictions of the corpus could be overcome by focusing on multiple poems by the same author.

6.3 Comparison: William Blake (1757-1827)

William Blake's (1757-1827) highly regarded *Songs of Innocence* (SoI) and *Songs of Experience* (SoE) were chosen for the second comparative case study.[73] Blake was a prolific author and cannot be represented through the *Songs* alone, but they undoubtedly remain his most examined work; indeed, despite the 'apparent simplicity' of the lyrics, much has been written on the 'depth of meaning' achieved by the poet in these 'seemingly direct little poems' (Bottrall 1970: 11–12). The layered interpretations of the Songs made them a useful brief case study for testing the ambiguity tagger, and the duality of the series lends the poems to a comparative investigation (Bottrall 1970: 13).

Indeed, McIntyre and Walker's (2010) use of corpus methods in the analysis of Blake's *Songs of Innocence and of Experience* stands as a rare example of CL approaches to verse, and provided a reference point for the comparative case study that was unique to the OBEV corpus. The authors initially looked towards identifying the key semantic domains in both series, using the Wmatrix system for annotating the texts (McIntyre & Walker 2010: 517). Their key finding was the positive keyness of 'HAPPY' for SoI and the contrasting positive keyness of 'FEAR/SHOCK' and 'VIOLENT/ANGRY' for SoE, which validated the 'semantic contrast' typically associated with the Songs (McIntyre & Walker 2010: 517). However, as

[73] Songs is used when referring to both Songs of Innocence and of Experience in the text, while SoI and SoE are used when discussing either collection independently.

Wmatrix does not use a hierarchical structure to organise the semantic domains, the authors turned towards the words that corresponded to the key domains and focused the remainder of their discussion on a keyword analysis of the texts, therefore examining the significant presentation of lexical items in the series as a way of exploring the poems (McIntyre & Walker 2010: 517–519). Their approach was similar to the discussion in the first micro-level case study, which began with the higher-level domains and then looked at the individual words that correspond to the semantic fields. To this end, this case study looks at the intermediate layer of the ambiguity tagger that is made available through the fine-grained taxonomy of the HTE, to further highlight the distinguishing elements of the ambiguity tagger in the analysis of poetry.

The OBEV corpus contained only six poems from the *Songs* collection, four from SoI and two from SoE (Quiller-Couch 1919/1999). As a result, it was not possible to carry out a full comparison of the songs, but with at least two poems from each collection, it was possible to compare the results from the semantic tagger to see if they corresponded to existing scholarship and McIntyre & Walker's research (2010). Furthermore, the corpus contained two poems that were published under the title of *Introduction* when the *Songs* were first released, though they were listed under different titles in the OBEV: [486] 'Reeds of Innocence' and [488] 'Hear the Voice', which introduce the SoI and SoE respectively (Quiller-Couch 1919/1999; Blake 1757/2007). Both poems are reproduced below as Poem vi and Poem vii below, and formed the basis of the comparative analysis as representative of the contrasting elements of the *Songs*, with the SoE lyric viewed as the 'corresponding poem' to the SoI entry (Hughes 2011: 33). The poems were tagged as part of CG3, with senses recorded between 1650 and 1850. The full summary views for both poems are available in Table A17 and Table A18, showing the top 6 thematic tags assigned to each word.

REEDS OF INNOCENCE

PIPING down the valleys wild,
  Piping songs of pleasant glee,
On a cloud I saw a child,
  And he laughing said to me:
'Pipe a song about a Lamb!'
  So I piped with merry cheer.
'Piper, pipe that song again;'
  So I piped: he wept to hear.

'Drop thy pipe, thy happy pipe;
  Sing thy songs of happy cheer!'
So I sung the same again,
  While he wept with joy to hear.
'Piper, sit thee down and write
  In a book that all may read.'
So he vanish'd from my sight;
  And I pluck'd a hollow reed,
And I made a rural pen,
  And I stain'd the water clear,
And I wrote my happy songs
  Every child may joy to hear.

Poem vi [486] William Blake (1757-1827)

HEAR THE VOICE

HEAR the voice of the Bard,
Who present, past, and future, sees;
Whose ears have heard
The Holy Word
That walk'd among the ancient trees;
Calling the lapsed soul,
And weeping in the evening dew;
That might control
The starry pole,
And fallen, fallen light renew!
'O Earth, O Earth, return!
Arise from out the dewy grass!
Night is worn,
And the morn
Rises from the slumbrous mass.
'Turn away no more;
Why wilt thou turn away?
The starry floor,
The watery shore,
Is given thee till the break of day.'

Poem vii [488] William Blake (1757-1827)

As an overview, the most frequently assigned thematic tags for the SoI poem were: »3762 (Singing), »2057 (Weeping), »903 (Hearing/noise), »481 (Sound/bird defined by), and »1574 (Commotion/disturbance/disorder). For the SoE poem, these were »1354 (Change), »903 (Hearing/noise), »1294 (Inclination),[74] »3049 (Conversion), and »1497 (Adversity). These senses are not split into binary positive and negative connotations, but rather blend into each other through the repeated emphasis on »903 (Hearing/noise), which in SoI seems to relate more to music and emotion while in SoE is linked to different stages of change and resignation.

[74] More specifically, †97019 (downwards).

Music here is associated with 'inspiration and origination', as the 'piper inspires the child' through song (Hughes 2011: 31). Consequently, its absence in SoE is all the more pronounced as a result: 'the Bard's injunction to 'Hear' is not musical, [..] he speaks with the voice of moralising obligation and omniscience' (Hughes 2011: 33). This distinction is visible in the thematic tags assigned to the word *hear* in the SoI and SoE poems, which for the former includes »183467 (be informed), »59882 (Hear), and »59898 (listen), and for the latter only shows variants of »122297 (listen attentively) and »183467 (be informed).

The granularity of the HTE tagset used by the ambiguity tagger was able distinguish the meanings in the poems with more precision than the Wmatrix analysis, thus expanding the analysis started in McIntyre & Walker's (2010) paper without contradicting the authors' findings. The HTST, which similarly uses the HTE taxonomy as the thematic dataset, assigned the same meanings to *hear* in both poems, and thus did not pick up on the distinction that was identified by the ambiguity tagger. Furthermore, as both taggers are restricted in the number of senses they assign, the potential for highlighting multiple interpretations is lessened as a result. For Blake, this could be seen as a significant limitation when considering the following observation of the poet:

> His poetry can at times work with ambiguity, stimulating a kind of reflection that so far from resulting in very exact pinpointing of meaning, leaves the mind itself at play. [...] The wealth of ambiguities involved in this text make it difficult to hold it in the mind as a clear entity (Beer 2005: 63–64).

In this regard, the ambiguity tagger is expertly suited for the task, and the results for both poems reinforce this, with multiple avenues for comparative discussion made available through the assigned senses, as shown by the example used in this case study. Thus, despite the restriction of the OBEV corpus on the scope of this investigation, it served to further reinforce the potential benefits of the tagger in a critical reading of poetry.

6.4 Conclusion

The micro analysis chapter looked at the possible applications of the ambiguity tagger in a close reading of the poems. Through two comparative case studies, the chapter outlined several approaches for exploring the semantic metadata in a critical reading. However, one of the challenges identified early on in this process was the limited selection of poems by single authors in the OBEV collection, which impacted on the scope of the analysis. Consequently, while the chapter demonstrated the potential for the ambiguity tagger in a micro analysis, it also highlighted the need for a wider sample of each author's work to be able to use the tagger in a comparative analysis.

Chapter 7 Semantic collocation

7.1 Introduction

Previous sections of this thesis demonstrated the process of applying the semantic hierarchy of the HTE to a corpus of poetry compiled from the OBEV and the use of the semantic metadata in a macro- and micro-level analysis of the corpus. This chapter extends the scope of the tagger by presenting an approach to measuring the co-occurrence of semantic domains in the corpus for the purpose of stylistic analysis. Building on existing work in identifying key semantic domains in corpora (Rayson 2008; McIntyre & Archer 2010), and prior applications of the HTE in exploring sequences of meaning (Archer & Malory 2015), this chapter employs keyness measures in identifying significant co-occurrences of semantic fields within corpora. The first part of this chapter outlines the research that inspired this project and provides the background to the methodology. As this approach expands the core methods utilised in earlier sections of the thesis, §7.3 of this chapter summarises these changes and the rationale for making them. The remainder of the chapter presents case-study analyses of the semantic collocation data, showing how this approach can address the final research question of this thesis (RQ5): What new research opportunities are opened by the ambiguity tagger?

7.2 Semantic collocation proposal

Collocation is viewed traditionally as 'a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text' (Stubbs 2001: 24). Collocation is therefore typically measured by studying the frequency with which individual words appear (co-occur) in a corpus within a pre-defined distance of a 'node' word (the word that is being studied). The co-occurring words are the 'collocates', and frequent co-occurrence is viewed as evidence of 'collocation' (Stubbs 2001: 29). The systematic analysis of significantly co-occurring words began as a foundational practice in corpus linguistics, and

early research in the field showcased this innovative approach to analysing language (Sinclair 1991). However, while research into collocation is primarily concerned with the co-occurrence of words in a corpus, recent work has begun expanding the scope of collocation analysis by investigating significant co-occurrence between lexical and grammatical items through collocation networks (Culpeper 2009; Brezina 2016; 2018b), and, crucially, the co-occurrence of semantic fields in corpora as identified through semantic annotation (Alexander et al. 2015a; Archer & Malory 2015; 2017). Building on this work, this thesis proposes a systematic approach to identifying 'semantic collocation' using the results from the ambiguity tagger.

As the nature of this enquiry was experimental, with no existing precedent for calculating semantic collocation on this scale, the study focused on a narrow span of 1: 1, or $N + 1$ and $N - 1$, where N is the node being studied and the range covers the corpus items that are directly before and after the node. The term *corpus items* is used instead of *tags* as the collocation extended to words and corpus items that were not assigned a meaning by the tagger or had been excluded from the semantic annotation process. While the 1: 1 range is narrow for measuring collocation, it is not without precedent (Semino 2010: 209). In addition to selecting the narrow span for practical reasons, as 1:1 restricts the calculation range, the data being measured has already undergone a process of abstraction, where an external signifier has been attributed to each node, and in the case of tagged lemmas, up to six external signifiers for one base node, and therefore represents a broader range of concepts than a single word-form.

To carry out this experimental analysis, collocation was calculated using the 'ranking' approach rather than a 'threshold' method (Evert 2009: 6). In lexical collocation, the ranking approach sorts 'words pairs on a scale of collocation strength without strict separation into collocations and non-collocations' (Evert 2009: 6), while the latter approach uses a threshold to find 'true collocations' in pairs 'whose association score exceeds a (more or less arbitrary) threshold value specified by the researcher'. Furthermore, while the association strength of collocates is calculated by considering the 'co-occurrence frequency' of words within a set span alongside the 'marginal' frequency of individual words in the corpus (Evert 2009: 12), these measures do not account for the variable number of

the semantic tags assigned to each word in the corpus. The measure would typically take the frequencies of individual words in a pair (f_1 and f_2), and calculate a separate 'frequency signature' (w_1, w_2) for 'every recurrent word pair' (Evert 2009: 12). Due to the number of variables that could be included in the marginal frequency count in this corpus (the frequency of individual words, tag frequency, tag frequency in relation to individual words), the initial analysis of semantic collocation was simplified to just the observed frequency of tag co-occurrence within the 1:1 span.

7.3 Methods

7.3.1 Pre-processing

The semantic metadata produced by the ambiguity tagger was not compatible with existing tools for identifying collocates as each word is assigned a variable number of semantic tags. These were limited for the collocation analysis to a maximum of six possible meanings, using the summary view data discussed in §4.6.5 above. While this reduced the maximum number of meanings that had to be calculated for each word, the semantic data for each word could still vary by having less than six tags, or by providing a *Null* value where it was not possible to match the lemmatised form of the word with the lexeme table (§4.4.1). Additionally, any attempt to match semantic tags with surrounding elements had to account for the stop-listed items (§4.5.1), which also reported a *Null* value after the candidate meanings for these words were removed to enable the contextual disambiguation process (§4.6.1). The six position fields were also *Null* for punctuation marks and sentence markers added to the data during the CLAWS POS-tagging process (§4.3.1.a). To determine node and collocate pairings across the corpus, these untagged fields would either have to be excluded, which would distort the results by producing pairs from a span that does not correspond to the original text, left blank, which would produce inaccurate results as every *Null* field would be counted as a single entry without distinguishing between the different reasons for the result, or they could be replaced with an alternative signifier, which could then be used to demarcate different *Null* entries. The latter approach was chosen as it

would provide an additional variable for the collocation measure and therefore expand the scope of the analysis.

The parameters for separating the *Null* records were based on the USAS and HTST tagging systems, which similarly disambiguate between non-lexical items in their tagsets. As the corpus was lemmatised and POS-tagged through the HTST tagger, and therefore annotated with both HTST and USAS codes at that stage, it was possible to use these results as the first stage in identifying groups of *Null* fields that represented distinct features. The process was less straight-forward than anticipated, however, because the thematic heading hierarchy document available through the SAMUELS website does not include any grammatical distinctions (Alexander & Kay 2021b). For example, while the document includes all headings ranging from *AA* (The world) to *BK*.09.*o* (Modern dance), the HTST output includes additional headings that range from *ZA*.01 (Personal Name) to *ZZ* (Unrecognised). Reviewing the results from the HTST annotation of the corpus indicated that the HTE codes within the *Z* range corresponded to matching designations in the USAS tagset, but the relationship was not always a direct one-for-one match: *ZA*.01 (Personal Name) was attributed to words tagged with USAS codes *Z*1 (Personal names), *Z*1 *mf* (Personal names, gender neutral), *Z*1f (Personal names, female), *Z*1m (Personal names, male), and in some cases ambiguous annotations like *Z*1m *S*9/*O*1.2 (Personal names, male, religion and the supernatural/ Substances and materials generally: Liquid) (Archer et al. 2002). This made a direct cross-match between the descriptions of the USAS fields and HTE codes not possible, as the documentation for the USAS codes could not be seen as corresponding directly to the HTE senses. Using the USAS codes in place of HTE categories for these items would introduce another disambiguation approach into the methodology, and research into the taggers indicated that the USAS system was less accurate when used on 'historical data' (Piao et al. 2017: 126). Instead, the *Null* fields were disambiguated by referring to the CLAWS part-of-speech tags,[75] which were part of the pre-processing annotation and already embedded in the semantic annotation process through the stop-list, along with the S_BEGIN and S_END designations for sentence markers. The query used to match *Null* values with their determiner is shown in Collmatch.sql (p.327), with

[75] The C8 tagset was used for this stage, which can be found at: http://ucrel.lancs.ac.uk/claws/

the result being a list of reference values for each item that could be used for identifying collocation.

7.3.2 Identifying pairs

The next stage in the process involved identifying pairs of tags within a 1: 1 radius. While the pre-processing replaced the *Null* values without semantic tags, it still left an irregular dataset with one to six possible tags for each individual item in the corpus. To illustrate this, an example of the first ten items from CG1, with their corresponding determiner or CATID is shown in Table 18 below.[76]

ID	Lemma	1st tag	2nd tag	3rd tag	4th tag	5th tag	6th tag
1	*34*	MC					
2	PUNC	YSTP					
3	S_END	S_END					
4	S_BEGIN	S_BEGIN					
5	*forget*	118408	118411	118414			
6	*not*	XX					
7	*yet*	90369	2205	89116	89117	87971	
8	*the*	AT					
9	*lover*	130460	131049				
10	*beseech*	142758	142814	142823	142828	142833	142834

Table 18 Tags for first 10 items in CG1

While calculating collocate pairs for the lemma forms would allow for simple $N + 1$ pairings (such as *34;*PUNC, or *forget;not*), calculating pairs from a range of assigned semantic tags requires taking into account multiple pairing sets for each item. For the data in Table 18, it is possible to calculate $N + 1$ for nodes 1+2 (*34*;PUNC), 2+3 (PUNC;S_END), and 3+4 (S_END;S_BEGIN) as direct pairs. However, there are three possible pairs for node 4+5 (S_BEGIN;»118408, S_BEGIN;»118411, S_BEGIN;»118414), and 30 possible pairs for node 9 + 10 (»130460;»142758, »130460;»142814, etc.), while the maximum number of pairs would be 36 if all six positions were occupied. The calculation must therefore allow for both $N + 1$ and $N - 1$ to represent between 2 and 36 possible unique pairs. Another query was written for this purpose, Collpairs.sql, which took an ordered list of up to six possible identifiers for each node and matched them to six possible

[76] See appendix Table A9 (p.184) for heading descriptions.

collocates in the following row.[77] It was not necessary to repeat the calculation for the previous row, as the field was limited to *N*+/-1 and each pair represented both the node and collocate for each subsequent row as the query iterated through the data.

The results of the query for $N + 1(1)$ and $N + 1(2)$ for the CG1 sample above is displayed in Table 19 below. The table was cropped to show only the pairs for the first and second position tags and identifiers, while the full 36 possible variations are available in the accompanying digital appendix DA4 (CG_collpairs.csv). This decision to match all possible pairs based on the CATID instead of the TID, as shown in Table 19, was made to ensure that no duplicate pairs were created by the query where multiple distinct HTE senses were subsumed under one thematic heading. This reverses the approach used to calculate the relevance of the sense, where the thematic heading was used for disambiguating between possible HTE senses through the RMV calculation (§4.6.3). By pairing the HTE category codes and, where necessary, alternative signifiers, the output produced by the query was more precise, while leaving the option of grouping the pairs again under higher level categories or thematic headings at a later stage of the analysis. The $N + 1$ pair for rows 6 and 7, for example, produced three distinct pairs within the 1+ range (Table 19), but would have returned two unique pairs and one duplicate if the TID codes were used instead, as #7:»89116 (as formerly/still/to this day) and #7:»89117 (even now (though not until now)) both fall within the thematic heading category of 1340/AM.08.b (The present (time)). If duplicates were allowed at this stage, it could impact on later investigation into frequently occurring and key pairs, and therefore impede the process of identifying semantic collocates in the corpus. Instead, the query returns positive and negative pairs for each unique item, as identified by the `position_ID`, and null values (shown as blank fields in the tables) in cases where no further identifiers were associated with an item.

[77] The only modification to the data at this stage was replacing the items without a count_value, which were filtered out in the stop-listing stage with a count of '1' to use as an index for picking out the identifiers assigned to each item.

ID	1-1+words	1-1+	1-2+	1-3+	1-4+	1-5+	1-6+	2-1+	2-2+	2-3+	2-4+	2-5+	2-6+
1	34; PUNC	MC; YSTP											
2	PUNC; S_END	YSTP; S_END											
3	S_END; S_BEGIN	S_END; S_BEGIN											
4	S_BEGIN; *forget*	S_BEGIN; 118408	S_BEGIN; 118411	S_BEGIN; 118414									
5	*forget*; *not*	118408; XX						118411;XX					
6	*not; yet*	XX; 90369	XX; 2205	XX; 89116	XX; 89117	XX; 87971							
7	*yet; the*	90369; AT						2205; AT					
8	*the; lover*	AT; 130460	AT; 131049										
9	*lover; beseech*	130460; 142758	130460; 142814	130460; 142823	130460; 142828	130460; 142833	130460; 142834	131049;142758	131049; 142814	131049; 142823	131049; 142828	131049; 142833	131049; 142834
10	*beseech; his*	142758; APPGE						142814;APPGE					

Table 19 Positive pairs sample

7.3.3 Pair frequency

In total, the Collpairs.sql query returned 623,240 unique pairs for N+/-1. In the first instance, a simple count was used to identify how frequently each pair appeared in the corpus. To achieve this for the dataset, it was first necessary to group the pairs from all 36 possible variations for each item into one list of unique pairs, and then cross reference this with the Collpairs.sql table to count the number of times each tag appeared across all columns.[78] A further query, Colldescs.sql then split the pairs into two columns, to allow for the descriptions for each tag to be retrieved from either the `category` or `CLAWS_filter` table depending on the type of signifier. The pairs were sorted by appearance frequency, and the descriptions added crucial context for interpreting the data.

As expected, the top frequent pairs in the corpus as a whole were sentence breaks and complementary grammatical items. The top ten frequent pairs for the whole corpus are shown in Table 20 below, which lists the pair, the total count, and a concatenated version of the two corresponding descriptions. The top two results are understandably the start and end of each sentence (S_END;S_BEGIN), and the combination of a full stop (YSTP) and the end of a sentence (S_END). Similarly, the combination of a comma (YCOM) and coordinating (CC) or subordinating conjunctions (CS) is unsurprisingly common, though the high appearance of exclamation marks (YEX) at the end of a sentence is likely higher than one would expect of a non-poetry corpus. Confirming the latter would require comparing the results against a reference corpus that was tagged and processed in the same way, and the previously covered challenges in identifying an appropriate reference for verse corpora prevented this from being carried out in the present research. Despite this limitation, it was still possible to carry out different forms of analysis on the tag pairs, both in terms of frequency of appearance and in a comparative analysis of smaller samples of the corpus against the remainder in an effort to identify significant pairs and explore the possibility of calculating 'semantic collocation'.

[78] Using the query Collcount.sql.

Pair	Count	Descriptions
S_END;S_BEGIN	7587	S_END;S_BEGIN
YSTP;S_END	5150	punctuation tag - full-stop;S_END
YCOM;CC	3704	punctuation tag - comma;coordinating conjunction (e.g. and, or)
II;AT	2727	general preposition;article (e.g. the, no)
II;APPGE	1544	general preposition;possessive pronoun, pre-nominal (e.g. my, your, our)
YEX;S_END	1333	punctuation tag - exclamation mark;S_END
YCOM;CS	1068	punctuation tag - comma;subordinating conjunction (e.g. if, because, unless, so, for)
YCOM;II	1058	punctuation tag - comma;general preposition
YCOM;AT	933	punctuation tag - comma;article (e.g. the, no)
MC;YSTP	872	cardinal number, neutral for number (two, three..);punctuation tag - full-stop

Table 20 Top 10 pairs

Returning to Table 20 above, while the previously mentioned tags corresponded to, in most part, the same word-forms and structural markers, certain tag pairs related to a broader range of items, even within the restrictions of closed-class grammatical items. As an example of the former, the YCOM and CC pair, which occurred 3,704 times in the tagged corpus data, corresponded to 3,173 instances of the *comma* and, 317 cases of *comma or*, and 214 uses of *comma nor*.[79] As evidence for the latter, the next pair down in the table, listed as II and AT, or 'general preposition' and 'article', represented 61 distinct pairings, while the combination of general preposition (II) and possessive, pre-nominal pronoun (APPGE) covered 192 unique pairs. The corresponding word-forms for these two tag pairs are listed in Table 21 below, with the frequency of appearance in brackets next to each pair. This approach extends the scope of both part-of-speech tagging and wordcount analysis by providing a different frame of reference, suitable for both syntagmatic and paradigmatic analysis.

[79] Note that the CLAWS C8 tagset lists *but* under CCB for 'adversative coordinating conjunction', and *so* and *for* as subordinating conjunctions (CS), which explains their absence here, and reinforces the close connection between subjective parameters and objective methods in CL.

II;AT
in;the(545), *to;the*(345), *on;the*(322), *from;the*(233), *by;the*(131), *through;the*(118), *upon;the*(109), *at;the*(105), *like;the*(69), *o'er;the*(62), *as;the*(61), *into;the*(50), *among;the*(38), *about;the*(38), *under;the*(37), *beneath;the*(35), *over;the*(32), *above;the*(28), *round;the*(28), *along;the*(26), *unto;the*(25), *within;the*(24), *before;the*(22), *down;the*(21), *across;the*(19), *amid;the*(19), *up;the*(17), *beyond;the*(17), *against;the*(16), *beside;the*(14), *till;the*(11), *below;the*(10), *out;the*(9), *around;the*(8), *amidst;the*(7), *behind;the*(7), *between;the*(6), *near;the*(6), *past;the*(5), *towards;the*(5), *toward;the*(5), *underneath;the*(5), *after;the*(4), *off;the*(4), *by;no*(4), *to;no*(3), *twixt;the*(3), *betwixt;the*(3), *save;the*(2), *in;no*(2), *amongst;the*(2), *through;no*(1), *nigh;the*(1), *from;no*(1), *throughout;the*(1), *ere;the*(1), *outside;the*(1), *at;no*(1), *worth;the*(1), *but;the*(1), *besides;the*(1)
II; APPGE
in;my(102), *in;her*(88), *in;thy*(59), *to;my*(56), *to;his*(52), *in;his*(50), *in;their*(43), *on;thy*(40), *on;my*(38), *on;his*(35), *in;your*(32), *to;her*(31), *in;our*(31), *on;her*(31), *on;your*(27), *from;her*(27), *to;thy*(26), *at;my*(26), *upon;her*(25), *from;thy*(25), *from;his*(25), *at;her*(24), *from;my*(22), *to;their*(21), *upon;my*(20), *from;their*(19), *to;your*(18), *to;our*(17), *about;her*(16), *in;its*(16), *by;his*(14), *at;his*(14), *by;thy*(14), *within;my*(12), *at;thy*(12), *at;their*(12), *by;my*(10), *on;their*(10), *at;your*(10), *to;its*(9), *on;our*(9), *into;my*(9), *through;my*(9), *upon;his*(8), *from;our*(8), *within;his*(8), *into;her*(8), *by;their*(8), *upon;thy*(7), *unto;my*(7), *at;our*(7), *round;her*(6), *against;their*(6), *by;your*(6), *upon;our*(6), *on;its*(6), *unto;her*(5), *upon;their*(5), *against;his*(5), *through;his*(5), *by;her*(5), *through;thy*(5), *o'er;thy*(5), *from;its*(5), *about;my*(5), *against;my*(4), *between;her*(4), *within;her*(4), *over;her*(4), *before;his*(4), *about;thy*(4), *at;its*(4), *through;your*(4), *by;its*(4), *into;thy*(4), *upon;your*(4), *into;his*(4), *beneath;thy*(4), *beneath;their*(4), *beneath;her*(4), *from;your*(3), *round;his*(3), *o'er;his*(3), *round;my*(3), *about;his*(3), *over;my*(3), *off;my*(3), *beneath;my*(3), *across;her*(3), *within;thy*(3), *o'er;their*(3), *unto;your*(2), *unto;our*(2), *by;our*(2), *among;his*(2), *against;thy*(2), *behind;her*(2), *before;thy*(2), *within;their*(2), *like;my*(2), *into;our*(2), *beside;thy*(2), *like;its*(2), *above;thy*(2), *into;their*(2), *through;their*(2), *like;thy*(2), *upon;its*(2), *about;your*(2), *over;thy*(2), *like;your*(2), *near;her*(2), *under;her*(2), *above;her*(2), *into;your*(2), *beside;her*(2), *above;my*(2), *unto;thy*(2), *beneath;your*(2), *amidst;my*(1), *less;our*(1), *after;their*(1), *ere;thy*(1), *before;her*(1), *near;his*(1), *amid;its*(1), *worth;its*(1), *before;my*(1), *ere;your*(1), *about;our*(1), *off;his*(1), *past;my*(1), *amongst;his*(1), *up;his*(1), *beyond;thy*(1), *o'er;her*(1), *within;its*(1), *through;her*(1), *like;her*(1), *twixt;his*(1), *down;their*(1), *unto;their*(1), *underneath;thy*(1), *amidst;their*(1), *round;thy*(1), *throughout;his*(1), *toward;his*(1), *betwixt;their*(1), *near;my*(1), *unto;his*(1), *amongst;her*(1), *o'er;my*(1), *near;thy*(1), *between;my*(1), *beside;their*(1), *amid;their*(1), *twixt;thy*(1), *near;our*(1), *under;their*(1), *underneath;their*(1), *worth;my*(1), *beneath;our*(1), *round;their*(1), *amid;your*(1), *underneath;my*(1), *amongst;our*(1), *throughout;your*(1), *beyond;my*(1), *among;thy*(1), *round;its*(1), *unlike;your*(1), *across;thy*(1), *under;his*(1), *below;her*(1), *like;our*(1), *amid;thy*(1), *below;its*(1), *over;your*(1), *beside;his*(1), *around;my*(1), *behind;his*(1), *under;our*(1), *over;its*(1), *ere;his*(1), *into;its*(1), *till;my*(1), *o'er;its*(1), *among;her*(1), *unlike;our*(1), *through;its*(1), *along;her*(1), *below;his*(1)

Table 21 Pair examples

The degree of variation captured in Table 21, even within the comparatively narrow parameters of grammatical word-forms, exemplifies the diverse nature of the corpus. The frequent use of archaic forms is particularly conspicuous, while the preference of certain constructions over others invites further investigation. In II;AT, for example, *o'er the* appears almost twice as frequently as *over the*, with the highest concentration of use in CG3. Furthermore, while *over the* appears only once in CG1 and three times in CG2, increasing in frequency to ten times in CG3 and fourteen in CG4, *o'er the* is the preferred form in all four corpus groups. Curiously, while most instances of *o'er the* and *over the* appear in separate poems, a closer look at the data revealed that in two instances both forms were used in the same poem.

The first example appeared in [514] 'Kilmeny' by James Hogg (1770-1835), a Scottish poet now recognised for his contribution to the Romantic period after being overlooked during his time (Mack 2004). The decision to use both forms of 'over the' lends itself to existing critical interpretations of Hogg's work, though more frequently in relation to his prose (Pittock 2008). His use of 'different forms of Scottish speech' is seen as both a measure of character integrity (Pittock 2008: 219), and as a conscious presentation of his 'authorial identity' (Alker & Nelson 2009: 7). Furthermore, while no reference to this particular example was found in critical work on Hogg, a curious observation was made in a paper on Wordsworth's about a revision the poet made to replace 'For' with 'O'er' in his poem, 'Extempore Effusion Upon the Death of James Hogg', further reinforcing the significance of Hogg's original (Currie 2005: 11). Further analysis of this appearance is beyond the remit of this research, as its goal is to demonstrate the potential of the tagger rather than carry out extended stylistic analyses of individual features in the corpus. It does, however, show that by allowing users to annotate and explore the corpus at different contextual scopes, the tagger can highlight unusual semantic and lexical pairs, thus pointing to a potential stylistic marker.

The second appearance of both *o'er the* and *over the* in the same poem was Samuel Taylor Coleridge's (1772-1834) [549] 'The Rime of the Ancient Mariner', in which *over the* is used by the third-person narrative voice in the first part of the poem, while *o'er the* was chosen by the Mariner in 'Part the Third', and then again in 'Part the Sixth'. Again, the investigation of the tag pairs could be traced back to a

corresponding feature of the text, though in this case it forms part of a broader division between the third and first person narratives taking part in the poem.

7.3.3.a Beginnings and endings

The decision to include all corpus items in the calculation, while expanding the range of data available for analysis, did introduce a significant limitation: punctuation marks featured heavily as possible collocates, making it possible to identify tags that most frequently preceded or followed a particular mark; however, they separated the semantic tags with the sentence marker signifiers, meaning that it was not possible to determine which tags appeared more or less frequently at the end of a sentence. One possible solution would be to omit punctuation from the Collmatch.sql query, though further comparison of the data would be necessary to determine the consequence this would have on the validity of the results. One of the key features of the ambiguity tagger is that it retains contextual parameters in annotating the corpus, which includes the use of sentence boundary markers as possible scopes for analysis. Removing these from the corpus data would require a further processing step to allow users to cross reference the collocation results with the original markers in the corpus. To determine whether the approach was worth pursuing at this time, the analysis was first carried out on the data that pertained to the beginnings of sentences, where punctuation marks were less likely to interfere.

It was possible to investigate the pairs at the beginning of sentences by examining the possible $N + 1$ collocates for the S_BEGIN node, which returned 3,286 unique combinations of S_BEGIN+1. Of these, however, only 208 appeared ten or more times in the whole corpus, and 18 combinations appeared more than 100 times. As shown in Table 22 below, which depicts these 18 S_BEGIN pairs, the majority point to sentences that start with grammatical elements. A further feature of the corpus highlighted here is the S_BEGIN and MC pair, which primarily corresponds to the OBEV numbers that were retained to identify individual poems in the corpus.

Pair	Count	Descriptions
S_BEGIN;MC	584	S_BEGIN;cardinal number, neutral for number (two, three..)
S_BEGIN;AT	574	S_BEGIN;article (e.g. the, no)
S_BEGIN;CC	416	S_BEGIN;coordinating conjunction (e.g. and, or)
S_BEGIN;UH	385	S_BEGIN;interjection (e.g. oh, yes, um)
S_BEGIN;II	362	S_BEGIN;general preposition
S_BEGIN;CS	272	S_BEGIN;subordinating conjunction (e.g. if, because, unless, so, for)
S_BEGIN;APPGE	251	S_BEGIN;possessive pronoun, pre-nominal (e.g. my, your, our)
S_BEGIN;PPIS1	247	S_BEGIN;1st person sing. subjective personal pronoun (I)
S_BEGIN;CCB	226	S_BEGIN;adversative coordinating conjunction (but)
S_BEGIN;PPHS1	217	S_BEGIN;3rd person sing. subjective personal pronoun (he, she)
S_BEGIN;DDQ	134	S_BEGIN;wh-determiner, interrogative (which, what).
S_BEGIN;PPY	130	S_BEGIN;2nd person personal pronoun (you)
S_BEGIN;AT1	125	S_BEGIN;singular article (e.g. a, an, every)
S_BEGIN;IF	112	S_BEGIN;for (as preposition)
S_BEGIN;NP1	105	S_BEGIN;singular proper noun (e.g. London, Jane, Frederick)
S_BEGIN;89757	101	S_BEGIN;Different time
S_BEGIN;89758	101	S_BEGIN;(by) that time/(since) that time
S_BEGIN;89759	101	S_BEGIN;at that time

Table 22 Top S_BEGIN pairs

The data in Table 22 includes HTE semantic tags among the top pairs, which reference items not excluded through the stop-list query, but these are adverbial or introductory elements, and function similarly to grammatical elements. While the range here does imply a degree of variation in the corpus and could be explored further, the data referring to sentence beginnings did not provide a substantial insight into the corpus. It did, however, provide justification for taking a closer look at sentence endings, if only to determine whether there was a substantial difference between the collocates.

To overcome the punctuation issue, it was necessary to go back to the results of the Collmatch.sql query and filter out any records corresponding to punctuation marks, which was done by removing any rows that contained 'PUNC' as the lemma. These missing rows would have created an issue for the Collpairs.sql

query, which references the unique key for each node sequentially, so a temporary ID was created for the filtered data. By pointing the Collpairs.sql and Colldescs.sql queries to the temporary ID, the calculation skipped punctuation marks when matching nodes to neighbouring collocates, which increased the number of unique pairs for the corpus from 623,240 to 700,144. This was an encouraging result, validating the decision to adapt the methodology to ignore punctuation.[80] Furthermore, while a quick examination of the S_BEGIN $N + 1$ collocates in the filtered data returned only 32 additional unique pairings, which confirmed the previous assumption that punctuation is less likely to interfere at the beginning of sentences, the number of unique combinations of S_END for $N - 1$ grew from 8 to 7,413.

Exploring the S_END N-1 collocates revealed a greater number of unique pairs appearing ten or more times than the S_BEGIN N+1 group, with 732 pairs compared to the 284 now identified for S_BEGIN. Additionally, while the S_BEGIN pairs continued to weigh heavily for grammatical items, with only slight changes to the number of matches captured in Table 22 above,[81] the S_END data was more varied. Table 23 below displays the top ten S_END pairs, along with the count and descriptions.

Pair	Count	Descriptions
MC;S_END	877	cardinal number, neutral for number (two, three…);S_END
NN1;S_END	159	singular common noun (e.g. book, girl);S_END
PPY;S_END	140	2nd person personal pronoun (you);S_END
PPIO1;S_END	132	1st person sing. objective personal pronoun (me);S_END
NP1;S_END	127	singular proper noun (e.g. London, Jane, Frederick);S_END
130497;S_END	120	Loved one;S_END
15487;S_END	109	Die;S_END
130547;S_END	106	Terms of endearment;S_END
189127;S_END	101	Depart/leave/go away;S_END
117117;S_END	95	Understand;S_END

Table 23 Top ten S_END pairs

[80] DA5 (CG_collpairs_nopunc.csv) lists all tag pairs for the four corpus groups after removing punctuation rows from the analysis.
[81] For example, excluding punctuation returned two extra matches for S_BEGIN;AT, bringing the total to 576.

A notable feature of the S_END pairs is the smaller number of variants that exceed 100, again contrasting with the front-loaded data for S_BEGIN. The only similar feature was the high frequency of MC pairings, again explained by the OBEV numbers in the corpus. Following this, the NN1 group appeared to contain primarily personal names, presumably mis-tagged with the common noun POS-tag instead of NP1 (proper nouns). Examples include *Rosaleen*, *Rosaline, Leucippe, Alciphron, Samela*, each one appearing multiple times at the end of a sentence. Different names appeared in the NP1;S_END list, but otherwise the top four most frequent collocates for S_END offered no surprises. The results for the semantic tag pairings, however, were considerably more interesting.

The first noticeable feature of the top semantic collocates for S_END−1 was that each tag represented a familiar concept, with even the more abstract 117117 (Understand) being recognisable as an idea. This meant that it was not necessary to refer to the thematic headings to begin the inquiry into the data, as the HTE category tags were salient enough. Another noticeable feature was the appearance of related concepts as frequent collocates: »130497 (Loved one) and »130547 (Terms of endearment), »15478 (Die) and »189127 (Depart/leave/go away). The presence of »117117 (Understand), by comparison, seems to stand out in the data because it shares no obvious thematic connection with either group. It was also unclear from the data if the frequency of these semantic collocates was the result of a small number of common words, and if the appearance of related HTE categories in the high-ranking data originated from both tags occupying one of the six positions in the tagged data as possible senses, or if the collocates referred to a broader range of lexical items and indicated common themes employed by a range of poets in the corpus. As with each case study analysis undertaken for this research, these questions were used to further explore the semantically annotated corpus.

The first step in understanding the composition of the high-frequency collocates in Table 23 involved querying the data to retrieve the corresponding word-forms for the collocate tags. The lemmatised forms of the words for each of the five top semantic collocates for S_END−1 are shown in Table 24 below, sorted in descending order by appearance frequency, which is given in brackets after each collocate pair. The frequency here refers to the number of times each word tagged with the HTE code appeared as a collocate of S_END, rather than the raw

collocation count. For example, the lemma *sweet* appears as a collocate of S_END seven times as »130497 (Loved one), and fifteen times as »130547 (Terms of endearment). This highlights a key distinguishing element of semantic collocation analysis as implemented in this thesis: collocation is identified through the co-occurrence of senses with the target node rather than individual word forms. This allows for a different entry point into investigating collocation in the corpus when compared to more traditional approaches for measuring collocation, which rely on frequency counts of individual word forms. A word-form based approach would consider each appearance of *sweet* in the $N - 1$ range for S_END, returning the total appearance count as fifteen. This would place the pairing in the 71st frequency position for S_END collocates out of 2,782 total combinations. To determine related lexical collocates in that list, it would be necessary to carry out further post-filtering of the collocate data to group the results into conceptual groups. Here, the collocation calculation provides this information directly, promoting less frequent lexical collocates as variations of common semantic collocates in the corpus. The single appearance of *turtle-dove*, for example, as a lexical collocate within »130547 (Terms of endearment), becomes significant precisely because it deviates from the more-common terms used elsewhere in the corpus. This observation could be pursued to a further stage of enquiry, by investigating whether different authors featured in the corpus have preferred forms or use uncommon constructions to express the same concepts as their peers.

Pair	Count	Lemmas
130497;S_END	120	*love*;S_END(85), *sun*;S_END(19), *sweet*;S_END(7), *treasure*;S_END(5), *passion*;S_END(2), *beloved*;S_END(1), *darling*;S_END(1)
15487;S_END	109	*die*;S_END(71), *end*;S_END(11), *part*;S_END(9), *depart*;S_END(6), *buy*;S_END(4), *sink*;S_END(3), *expire*;S_END(2), *drop*;S_END(1), *ghost*;S_END(1), *starve*;S_END(1)
130547;S_END	106	*heart*;S_END(28), *soul*;S_END(19), *joy*;S_END(16), *sweet*;S_END(15), *dear*;S_END(14), *beautiful*;S_END(4), *toy*;S_END(3), *treasure*;S_END(2), *lover*;S_END(2), *lamb*;S_END(1), *darling*;S_END(1), *turtle-dove*;S_END(1)
189127;S_END	101	*away*;S_END(62), *part*;S_END(9), *depart*;S_END(8), *ago*;S_END(7), *hence*;S_END(7), *return*;S_END(3), *remove*;S_END(2), *off*;S_END(2), *leave*;S_END(1)

117117;S_END	95	*see*;S_END(38), *know*;S_END(27), *in*;S_END(7), *take*;S_END(6), *sound*;S_END(6), *skill*;S_END(3), *espy*;S_END(2), *tend*;S_END(2), *comprise*;S_END(1), *ken*;S_END(1), *receive*;S_END(1), *understand*;S_END(1)

Table 24 Semantic S_End pairs

Taking the top two high-frequency semantic collocates, Table A19 and Table A20 show the individual lemma collocates separated by the authors in the corpus. With this vantage point, a further series of patterns become clear. Taking the results for »130497 (Loved one) in Table A19 first, an initial observation could be made of the authors that frequently end sentences in reference to a loved one: Robert Herrick (1591-1674) is recorded with the most frequent use at four separate references, while Richard Crashaw (1613?-1649), Elizabeth Barrett Browning (1806-1861), and Robert Bridges (b.1844) use the refrain three separate times. Furthermore, in each of the above cases the authors use more than one word-form to express the idea, thus revealing a pattern that could be missed if relying on lexical collocation analysis alone.

This variation is further visible when reviewing the data for each individual corpus group. Curiously, poets represented in CG1, CG2, and CG3 rely on a narrow variety of words to represent »130497 (Loved one) at the end of a sentence, with *love*, *sun*, and *treasure* used in CG1, and only *love* and *sweet* in CG2, and *love* and *sun* in CG3. In CG4, however, authors employ *love, sun, sweet, treasure, passion, beloved,* and *darling* to express the same sentiment when closing a sentence. The larger size of CG4 compared to the other three corpus groups could be the reason for this variation, but if that was the only cause it would follow that the same pattern could be observed for other prominent semantic collocates, such as »15487 (Die). However, as shown in Table A20, earlier corpus groups report more variation, with CG3 having the most individual words representing the idea, followed by CG4, then CG1 and finally CG2 with the least unique word-forms. This would suggest that the size of the corpus group is not enough to account for variety in words used to express the same concept, and further enquiry would be necessary to understand this phenomenon in relation to the source material.

Indeed, every observation made when reviewing this data could be further explored in relation to the source poems. This is not further pursued in this research as it extends beyond the scope of the work, which primarily looks at how

this approach could be used to inspire alternative routes into critical engagement with corpora of poetry (RQ5). Furthermore, the size of the corpus limits the scope of the analysis as it is possible only to form observations regarding the poems included in the Quiller-Couch OBEV edition (1919/1999), and therefore cannot be considered as representative of the full range of the author's work. As mentioned previously, this does not prevent the analysis of the data, but does impact on the scope of the analysis outwith the source material. Indeed, the editor's preference for ballads is likely to influence the frequent appearance of this collocate as a stylistic trait of the genre. While these limitations are inherent in small corpus research of this nature, they are further enhanced by the absence of a suitable reference corpus to use as a baseline for identifying significant features. However, when working with the semantic collocation data, the limitation of reviewing frequency alone is that it can only point to areas of active engagement; the data cannot show where authors have chosen alternatives, and where the trend was broken. To overcome this, a keyness analysis was used to compare samples of the corpus against the rest as a reference, to identify significantly appearing pairs in two genres of poetry captured in the corpus.

7.4 Key semantic collocates

The use of keyness analysis to identify semantic collocates has been attempted in the past, most notably by Archer et al. (2006) in their analysis of co-occurring domains with the semantic field of *love* in Shakespeare's tragedies as compared with his comedies. The authors' goal was to 'discover which semantic tags collocate significantly with a small number of key semantic tags' that were identified by carrying out a keyness analysis on a semantically annotated corpus (Archer et al. 2006: 3). This work has clear parallels to the research carried out for this thesis, particularly within this chapter as the analysis moved towards semantic collocation. However, key differences in research design and methodology separate the endeavours, with both approaches acting as complementary examples of different routes into investigating semantic collocation. In particular, the authors relied on the USAS system for semantically annotating the corpus, this providing them with a different dataset to the HTE ambiguity tagger used in this research, as established in Chapter 4. Similarly, while the authors were also unable to use an existing tool for calculating semantic collocation, they were able to carry out their

analysis by calculating keyness with the Wmatrix software (Rayson 2009a) as it was not necessary to account for multiple possible tags for each individual word, as was the case with this research (Archer et al. 2006: 3). Finally, while the authors were similarly looking at 'the extent to which important collocate information can be discovered at the domain level' as opposed to 'the word level', they looked towards identifying collocates for a pre-determined list of semantic domains, rather than all possible collocate combinations (Archer et al. 2006: 10). However, while these methodological differences are substantial enough to distinguish this study, the goal for enquiry is to further build on the existing tools and methodologies for measuring semantic collocation, thus adding to the existing research.

To calculate significant semantic collocates from data produced through the Collpairs.sql query, the keyness measure that was used to identify key semantic tags in the corpus groups was adapted to identify key semantic pairs,[82] enabling a measure of significant semantic collocates within the N+/-1 range. While it was possible to use the corpus groups as samples for calculating keyness, a different approach was chosen for investigating key semantic collocates, as this would allow an additional route to investigation to be explored in analysis.

7.5 Samples

Two samples were chosen to allow for some initial comparative analysis to be carried out with the results. The samples had to be selected from within the timeframe selected for the reference, and as such, two sub-collections were identified within the corpus that could serve for the initial stage of the experiment. These were the works of a collection of writers commonly referred to as *Metaphysical* poets (c1600-c1690) and the *Cavalier* poets (c1640-c1660). These samples were advantageous as both groups were contemporaries, their work was well represented in the OBEV corpus, and neither sample exceeded 15,000 words and grammatical items, and 100,000 individual collocate pairs, making the remainder of the corpus suitable for acting as a reference. A further advantage, and one that inspired the selection, was that both groups were seen as embodying

[82] See Log_Likelihood.sql.

diverse poetic traditions, and often referred to as antecedent movements of the period (Miner 1971: 12).

7.5.1 Metaphysical poets (*S*1)

The label of the 'metaphysical poets' is often attributed to a group of 17th century poets, and most notably John Donne, who is often seen as a leading influence among the metaphysics (Alvarez 1961). The concept of metaphysical poetry existed prior to its modern use, and was initially applied 'to witty, conceited poetry, but in the vaguest possible way' (Hammond 1974: 11). The term was first used to describe the group in Johnson's (1779) biography of Abraham Cowley, but it was Dryden's dismissive remark, that John Donne 'affected the metaphysics, not only in his satires, but in his amorous verses where nature only should reign', which popularised the term as referring to Donne and 'his followers' (Hammond 1974: 12).

The particular traits that mark the 'metaphysical' poets are loosely defined, and there are varying accounts of the authors that belong to this tradition. Initially used to describe the work of Donne, Abraham Cowley, and John Cleveland (Hammond 1974: 13), it was later expanded to include Andrew Marvell, George Herbert, Henry Vaughan, and Richard Crashaw (Willy 1971; Bennett 1989; Austin 1992; Burrow 2006). Edited collections of 'metaphysical poetry' still range substantially in scope, but rarely through the omission of these names. Of these authors, the OBEV corpus lacked only the work of John Cleveland, thus making it possible to select a sample of 11,262 words and grammatical items that included the work of several metaphysical poets.

7.5.2 Cavalier Poets (*S*2)

The second sample corpus (S2) contained the work of the Cavalier poets who were included in the OBEV collection. Like the Metaphysical poets, the Cavaliers were prolific writers of the 17th century, though their main period of activity begins and ends a few years after the core group of metaphysical poets. Joined by their Royalist ideals and their 'use of direct and colloquial language', the Cavaliers were often seen as writing in opposition to the high conceits of the Metaphysicals (Skelton 1969: 7). The Cavalier poets include Robert Herrick, Thomas Carew, Sir

John Suckling, Richard Lovelace (Skelton 1969), and Edmund Waller (Miner 1971; Clayton 1978). Often seen as taking influence from the work of Ben Jonson, the poets are sometimes referred to as the 'Sons of Ben' (as well as the variant 'Tribe of Ben'), Jonson's impact is felt unequally across the work of the Cavaliers (Clayton 1978: xiv–xv).

It was possible to include all five key Cavalier poets in the sample, since a selection of each poet's work was included in OBEV. However, the sample was substantially smaller than S1, at 7,078 items compared to the 11,262 for S1. While the difference in size of S1 and S2 was normalised as part of the keyness measure, the small size of the corpus limited the scope of analysis to simply the features arising in that particular collection of poems. A larger sample could allow for a discussion of emerging semantic collocation patterns in relation to the schools, but any connections found in the active sample are limited to a discussion of recorded words only. A summary of both samples is provided in Table 25 and Table 26 below, while the full list of poems included in each sample is available in Table A21 and Table A22 of the appendix.

S1 Poets	Poem numbers	CG
John Donne. 1573–1631	195-202	CG1
George Herbert. 1593–1632	281-286	CG1
Richard Crashaw. 1613?–1649	336-342	CG2
Abraham Cowley. 1618–1667	349-353	CG2
Andrew Marvell. 1621–1678	355-361	CG2
Henry Vaughan. 1621–1695	362-365	CG2

Table 25 S1 poets summary

S2 Poets	Poem numbers	CG
Robert Herrick. 1591–1674	247-275	CG1
Thomas Carew. 1595?–1639?	289-295	CG1
Edmund Waller. 1606–1687	304-306	CG2
Sir John Suckling. 1609–1642	325-328	CG2
Richard Lovelace. 1618–1658	343-348	CG2

Table 26 S2 poets summary

7.6 Results

One particular challenge when interpreting quantitative data of this volume is that is it difficult to identify a starting point for analysis. There are several ways of ordering the results to bring salient features to the front, but each had a particular drawback when applied to this research: ordering by keyness meant that comparing semantic collocations across both samples was only possible by looking at those most significant, rather than the presentation of specific themes; alphabetical ordering meant that it was difficult to pick out key collocates; ordering by frequency ultimately produced a list of the most frequent refrains in the poems, where lexical pairs were used multiple times in a particular sample. However, the analysis of key semantic collocates in the samples provided a starting point for investigating the data.

Using a cut-off value of 15.13 ($p<0.0001$),[83] the calculation identified 60 distinct key positive semantic collocates for the Metaphysical (S1) sample, and 167 key positive collocates for the Cavalier (S2) sample. The first 50 results for each sample are shown in Table A23 and Table A24, while the full list of key positive collocates for both samples is available in DA6. Positive collocates were chosen for further analysis as they represent pairs that appear more frequently than would be expected by chance, and are therefore a useful starting point for discussing the key themes that collocate in each sample and how they relate to existing criticism. A comparative overview of the two samples side-by-side highlighted the differences between the significant co-occurrences of semantic fields. A small selection of these is discussed below, showing both the advantages and drawbacks of the approach.

7.6.1 Metaphysical concern with the soul

The most pronounced result for the S1 collocates when compared against S2 was the frequent reference to the soul collocating with grammatical elements. The first, most frequent, and most direct is 'APPGE;174916', (possessive pronoun, pre-

[83] See Rayson (2008).

nominal (e.g. my, your, our);Soul). The lexical results for this semantic collocate are depicted in Table 27 below, split by author and poem.

words	author	CG	poemID	totals
Our;soul	John Donne. 1573–1631	CG1	198	2
my;soul	George Herbert. 1593–1632	CG1	284	1
my;soul	George Herbert. 1593–1632	CG1	286	1
my;soul	Abraham Cowley. 1618–1667	CG2	352	1
Thy;soul	Abraham Cowley. 1618–1667	CG2	352	1
his;soul	Abraham Cowley. 1618–1667	CG2	352	2
My;soul	Andrew Marvell. 1621–1678	CG2	359	1
her;soul	Andrew Marvell. 1621–1678	CG2	361	1
my;soul	Henry Vaughan. 1621–1695	CG2	362	2
MY;soul	Henry Vaughan. 1621–1695	CG2	363	2
Thy;spirit	Henry Vaughan. 1621–1695	CG2	365	1

Table 27 'APPGE;174916' S1 collocates

In this case, it was clear why the semantic collocate pair was significant for the sample: all but one of the authors employed the pairing, often several times across different poems. However, aside from *thy spirit*, the authors all use the word *soul* to follow the possessive pronoun, and though this does indicate a commonality in their work, the frequency of *soul* would have likely been observed with a keyword analysis of the sample and therefore cannot be considered a significant finding through the key semantic collocate analysis. Similar results were found for 'APPGE;174929' (possessive pronoun, pre-nominal (e.g. my, your, our);with regard to moral aspect) which was composed of a pronoun and *soul*, while 'VM;174949' (modal auxiliary (can, will, would, etc.);of soul: die) and *die*. The results for the 'APPGE;76262' (possessive pronoun, pre-nominal (e.g. my, your, our);Essence/intrinsic nature) semantic collocate were more varied, however, and reported a match for each one of the S1 poets, as shown in Table 28 below.

words	author	CG	poemID	totals
Our;soul	John Donne. 1573–1631	CG1	198	2
their;bone	John Donne. 1573–1631	CG1	202	1
Thy;root	George Herbert. 1593–1632	CG1	281	1
my;flower	George Herbert. 1593–1632	CG1	282	1
my;soul	George Herbert. 1593–1632	CG1	284	1
my;savour	George Herbert. 1593–1632	CG1	284	1
my;soul	George Herbert. 1593–1632	CG1	286	1
my;soul	Abraham Cowley. 1618–1667	CG2	352	1

Thy;soul	Abraham Cowley. 1618–1667	CG2	352	1
his;soul	Abraham Cowley. 1618–1667	CG2	352	2
My;soul	Andrew Marvell. 1621–1678	CG2	359	1
her;soul	Andrew Marvell. 1621–1678	CG2	361	1
my;soul	Henry Vaughan. 1621–1695	CG2	362	2
MY;soul	Henry Vaughan. 1621–1695	CG2	363	2
their;root	Henry Vaughan. 1621–1695	CG2	364	1
Thy;spirit	Henry Vaughan. 1621–1695	CG2	365	1

Table 28 'APPGE;76262' S1 collocates

The significance of this collocation is further reinforced when comparing these results with the Cavalier sample (S2), which returned four results for the 'APPGE;174916' pair and four for 'APPGE;76262', all in reference to '*my;soul',* and each appearing only once in a poem (Table 29 below).

words	author	CG	poemID	totals
My;soul	Robert Herrick. 1591–1674	CG1	262	1
my;soul	Thomas Carew. 1595?–1639?	CG1	291	1
my;soul	Sir John Suckling. 1609–1642	CG2	328	1
my;soul	Richard Lovelace. 1618–1658	CG2	348	1

Table 29 'APPGE;174916' and 'APPGE;76262' S2 collocates

A comparison of the results in Table 28 and Table 29 would suggest that the collocating pair is significant for S1 in part due to a broader range of figurative expressions were identified through the 'APPGE;76262' (possessive pronoun, pre-nominal (e.g. my, your, our);Essence/intrinsic nature) pairing. The lemma *bone*, as an example, was identified by the tagger as »76262 (Essence/intrinsic nature) by the semantic tagger, thus contributing to the significance of the semantic collocate pair. In context, *their;bone* is the lemmatised pairing of *their* and *bones* in Donne's [202] 'Death':

> DEATH
> *Death, be not proud*
>
> And soonest our best men with thee do go—
> Rest of their bones and souls' delivery!
> Thou'rt slave to fate, chance, kings, and
> desperate men,
> And dost with poison, war, and sickness
> dwell;

Poem viii [202] John Donne (1573–1631)

The reading of *bones* as the essence left behind when the *best men* depart with death has precedent; Abdulla & Lutfi (2019) argued that *rest of their bones* and

soul's delivery represents the 'underlying conceptual metaphor' of 'BODY IS A CONTAINER' (Abdulla & Fadhil 2019: 85). They suggest that '*bones* metonymically stand for the whole body', which can be interpreted as the 'soul is contained in the body' (Abdulla & Fadhil 2019: 85). Elsewhere, Sperry (2019) proposes that 'hair and bone enter into Donne's lyrics as emblems of extreme decay, providing an image through which to examine the possibilities of existence at the further reaches of corporeality' (Sperry 2019: 47). They argue for the significance of *bones* in Donne's work, suggesting that 'future readers of both the bones and the lyric describing them' should 'look to these decayed remains as objects of divine significance' (Sperry 2019: 50), as they represent a state of existence that is abandoned in death. These readings support the interpretation of *bone* identified by the ambiguity tagger, which was in turn highlighted through the collocate analysis as a potentially significant pairing for further investigation. Notably, both USAS and the HTST identified only the literal senses of *bone* in [202] 'Death', tagging the lemma with 'Anatomy and Physiology [B1]', *AB.17.f* (Bodily substance, tissue) and *AB.17.g* (Bone/bones), and thus missing the figurative use of the word in context.

Although the small samples cannot truly represent stylistic differences between the works of the Metaphysical and Cavalier poets, they were useful in showing the potential of keyness analysis in identifying significant semantic collocates across multiple annotated poems. Furthermore, as the corpus was annotated by using the contextual scope of individual poems, key semantic collocates represent pairs that appear in multiple poems but correspond to a variety of lexical items. This approach could be used to examine collocating patterns across multiple texts in a corpus or aid a comparative analysis of two different texts instead of the collected samples used in this study. Finally, while poetic language will often invite different interpretations, the significant paring in Donne's [202] 'Death' supports existing readings of the text, but it would have been missed if the corpus was only tagged with the USAS and HTST systems.

7.7 Conclusion

This chapter presented an experimental method for calculating semantic collocation, using the results of the ambiguity tagger. This method was developed

in response to RQ5 (What new research opportunities are opened by the ambiguity tagger?) and sought to extend the scope of lexical collocation analysis, continuing earlier work in this area (§7.1). Consequently, this chapter serves as a proof of concept for this new approach, outlining the rationale for investigating collocating domains in a corpus (§7.2) and the method developed for calculating collocating semantic pairs using the ambiguity tagger output (§7.3).

By testing this method on the OBEV corpus, sections §7.3.3 and §7.3.3.a demonstrate how it can be used to explore frequently co-occurring semantic pairs, highlighting patterns that might be missed in an analysis of lexical collocation. However, while the results described in these sections captured certain distinctive features of the corpus, the frequency of appearance was not enough to identify significant semantic collocation in the corpus. Adapting the keyness measure to identify key semantic pairs (§7.4) in two samples of the OBEV corpus (§7.5) proved more valuable, as the results highlighted significant collocates for the samples in comparison to the rest of the OBEV corpus (§7.6). These findings suggest that further development of this novel method should prioritise identifying significant semantic collocates, allowing researchers to engage with their corpus in a new way.

Chapter 8 Conclusion

8.1 Introduction

This thesis presents a new, unique method for semantic annotation of corpora: the ambiguity tagger. This tagger was developed to address key limitations of existing semantic taggers and represents the first semantic tagging system designed to disambiguate senses in a diachronic corpus of poetry. To meet the requirements of this corpus, the tagger includes several features that are absent from the two main semantic taggers available at present, USAS (§2.4.3) and HTST (§2.4.5). These features include flexible disambiguation parameters, which can be adjusted to different contextual scopes, and greater control over the semantic metadata produced by the tagger, which can be tailored to show a narrow or broader range of candidate senses based on their relevance in context. These flexible parameters enabled annotation of individual poems in the corpus based on the contextual information of those poems rather than the corpus as a whole, while the configurable metadata parameters were used to explore semantic ambiguities inherent to poetic language.

To conclude this thesis, this chapter surveys the scope of the ambiguity tagger (§8.2), highlighting the key contributions this work makes to existing research (§8.2.1), and discussing the current limitations of the semantic annotation method and the requirements for overcoming these limitations (§8.2.2). The next section shows how the scope of the ambiguity tagger can be further expanded through future development (§8.2.3). The envoi in §8.3 brings the thesis to a close with a final reflection on the work as a whole.

8.2 The scope of the ambiguity tagger

As stated in Chapter 1, this thesis aimed to build on the work being carried out in two related areas of research: semantic annotation (§1.2) and semantic analysis of corpora (§1.3). This thesis therefore introduced the 'ambiguity tagger', developed to expand both the methodological scope of semantic annotation and the analytical

opportunities afforded by semantic annotation of corpora. While the work described in this thesis contributes to both areas, its advances in method are easier to define as it was grounded in an attempt to remedy some of the limitations of existing systems. The analytical chapters demonstrated the potential of the tagger in assisting both a macro- and micro-level analysis, and in identifying semantic collocation in a corpus, but within the scope of a thesis of this length, the OBEV corpus could not fully support a detailed investigation of the stylistic features highlighted by the tagger. The OBEV corpus was chosen for this research as it reflected the design parameters for the ambiguity tagger, but prioritising these considerations had the natural consequence of not giving the case studies in the thesis a wide scope to engage with broader issues in stylistics. Consequently, this thesis had as its main focus a contribution to corpus semantics research and methods rather than corpus stylistics practice.

8.2.1 Key contributions

This thesis makes the following contributions to research:

1. It delivers a new method for semantic annotation, the ambiguity tagger, which was trained to annotate figurative texts:
    a. This method is unique, as no other semantic annotation system offers the same degree of flexibility in attributing senses to all words in a corpus (§2.4);
    b. It relies on custom disambiguation parameters for annotating figurative language, allowing for a broader range of candidate senses in the annotated output (§4.6);
    c. It responds to the demand for context-based disambiguation methods by attributing senses based on their relevance within a set contextual scope (§4.6.1, cf. §2.5.2);
    d. It gives users the option to configure the annotated summary output to include additional metadata for further research, such as higher-level categories, or simplify the results by removing any unneeded data, which is not currently possible through the USAS or HTST taggers (§4.6.5).
2. The ambiguity tagger enables semantic annotation of a corpus of poetry with period-appropriate senses:

    a. While a diachronic tagger was already available in the HTST, no system has been previously designed specifically for handling poetry (§4.4);
    b. Furthermore, as the results have indicated, existing semantic annotation approaches struggle with annotating the semantic ambiguity inherent in verse, thus making this approach unique amongst its counterparts (§3.2, §4.6, cf. §2.5.3).
3. It demonstrates the use of the HTE semantic taxonomy in contextual disambiguation of senses, expanding on prior work by presenting a systematic approach for attributing HTE senses to words in a corpus (§4.4, §4.6):
    a. This is the first project to directly employ the taxonomy to identify relevant candidate senses based on the appearance of related semantic domains in the corpus, as the HTST does not employ the HTE taxonomy in this manner (§4.6.2, §4.6.3, cf. §2.4.5);
    b. The promising results of this disambiguation method further highlight the value of the HTE resource, alongside existing projects that utilise the taxonomy in innovative ways (§4.7, cf. §2.4.4).[84]
4. It provides specialist case studies for utilising the ambiguity tagger across different levels of corpus analysis, showing how the tagger can be used to explore semantic data at a macro and micro level:
    a. At the macro level, this includes aggregating the metadata produced by the tagger to identify the relevance of different HTE semantic domains across a corpus (§5.2), and using these results to inform the analysis of specific texts in a corpus (§5.3);
    b. At the micro level, the comparative analysis of the ambiguity tagger's results for different authors (§6.2) and poems (§6.3) in the corpus showcased the use of the HTE taxonomy in investigating highly granular semantic patterns in text.
5. Finally, it presents an innovative new method for investigating semantic collocation, which utilises the ambiguity tagger results to identify frequently co-occurring semantic domains in a corpus:

[84] See https://ht.ac.uk/bibliography/

a. This is the first semantic collocation approach that calculates co-occurrence of multiple candidate tags across all items in a corpus (§7.2);
b. This approach allows users to identify frequently co-occurring semantic domains across the whole corpus or specific contextual scopes (§7.3), and identify significant semantic collocates by calculating the keyness of co-occurring semantic domains against a reference corpus (§7.4).

8.2.2 Current limitations

The analysis of the results revealed some of the issues with the current disambiguation parameters used by the tagger, which could be resolved in future iterations of the system. While the nature of the ambiguity tagger places less emphasis on precise disambiguation, particularly when compared to semantic taggers that are trained to pick up a limited number of senses for each word, the need for a post-processing stage in the current disambiguation process could be reduced by making future modifications to the current pipeline. One such change could be the implementation of a set number of meanings for highly polysemous words, such as the one used in the HTST system (Piao et al. 2017). To maintain the flexibility of the tagger, this would be implemented as an editable dataset, which could be called on for annotating specific highly polysemous words.

Similarly, while the ambiguity tagger does not currently support MWEs, these could be added in future iterations of the method by embedding relevant entries from the HTST or USAS tagset, as both systems implement specific sub-lexicons to annotate MWEs (Piao et al. 2005a; 2017). This approach was not pursued in the current design of the tagger as these sub-lexicons were designed for the disambiguation methods of USAS and HTST, which do not currently support context-based annotation as described in this thesis.

The technical issues identified in analysis, such as the incorrect tokenisation of hyphenated words (§6.2.3), are expected in a first-generation tool. Further testing of the tagger is necessary to identify and resolve these issues, as different text types might produce different errors in pre-processing. While the OBEV corpus was useful in highlighting the requirements of a diachronic corpus of poetry, it was

typographically and editorially consistent. Thus, it is necessary to test the tagger on a broader range of corpora to determine the extent of the technical limitations and their impact on the annotation method. To improve the ambiguity tagger, these issues should be addressed in relation to different text types as part of the future development of the tagger.

8.2.3 Future development

The current scope of the ambiguity tagger has been defined in relation to its methodological and analytical contribution to research, but future development of the method must also overcome the restrictions inherent to a research thesis; namely, the siloed development and implementation of the tagger. To reach the standards established by the USAS and HTST, subsequent iterations of the tagger must allow for significant user testing and collaborative development. As the tagger was designed to function independently of any existing software, its development can continue as both an extension of the research presented in this thesis, and through third-party involvement with the release of the code that was written for this purpose. The tagger's reliance on the HTE hierarchy does restrict public use as it requires access to the HTE data, but any researcher with access to the dataset (or extracts of the HTE, or other similarly structured thesauri) can use the ambiguity tagger in their work.

8.3 Conclusion

This thesis aimed to expand the scope of semantic annotation and analysis of corpora by presenting a new method for semantic annotation, the ambiguity tagger, and showcasing its use in the analysis of a diachronic corpus of poetry. Although the project had ambitious aims for a thesis-length outcome, the work described here successfully extends the boundaries of semantic annotation beyond the capabilities of existing taggers by delivering a unique context-based disambiguation method that can be used to annotate figurative language, and expands the scope of semantic analysis by presenting an innovative approach for measuring semantic collocation. Thus, to echo the opening lines of this thesis, this research expands the horizon of what we can see when we look at a lot of language at once.

References

Abdulla, Ismail & Abbas Fadhil. 2019. Conceptual Metaphors in Donne's 'Death, Be Not Proud'. Presented at the International Conference on English Language and Culture (ICELC 2019), Koya University, Iraq. DOI: http://dx.doi.org/10.14500/icelc2019.lin187.

Abrams, M.H. & Stephen Greenblatt (eds.). 2000. *The Norton Anthology of English Literature*. New York: Norton. 7th ed.

Alec, Céline, Chantal Reynaud-Delaître & Brigitte Safar. 2016. An Ontology-Driven Approach for Semantic Annotation of Documents with Specific Concepts. In Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, & Christoph Lange (eds.), *The Semantic Web: Latest Advances and New Domains*. Switzerland: Springer International Publishing. 609–624. DOI: 10.1007/978-3-319-34129-3_37.

Alexander, Marc, Alistair Baron, Fraser Dallachy, Scott Piao, Paul Rayson & Stephen Wattam. 2015a. Semantic Tagging and Early Modern Collocates. In Federica Formato & Andrew Hardie (eds.), *Corpus Linguistics 2015 Abstract Book*. Lancaster, United Kingdom: Lancaster University. 8–10. Retrieved from http://ucrel.lancs.ac.uk/cl2015/.

Alexander, Marc, Fraser Dallachy, Scott Piao, Alistair Baron & Paul Rayson. 2015b. Metaphor, Popular Science, and Semantic Tagging: Distant Reading with the Historical Thesaurus of English. *Digital Scholarship in the Humanities*. 30(suppl_1). i16–i27. DOI: 10.1093/llc/fqv045.

Alexander, Marc, Alistair Baron, Fraser Dallachy, Scott Piao, Paul Rayson & Steven Wattam. 2015c. The Historical Thesaurus Semantic Tagger. [Online]

Available at: https://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/ [Accessed: 23 March 2019].

Alexander, Marc & Ellen Bramwell. 2014. Mapping Metaphors of Wealth and Want: A Digital Approach. In Clare Mills, Michael Pidd, & Esther Ward (eds.), *Studies in the Digital Humanities*. The University of Sheffield. Retrieved from http://www.hrionline.ac.uk/openbook/chapter/dhc2012-alexander.

Alexander, Marc Gabriel. 2011. *Meaning Construction in Popular Science*. PhD Thesis. University of Glasgow.

Alexander, Marc & Christian Kay. 2019a. Classification. [Online] Available at: https://ht.ac.uk/classification/ [Accessed: 15 July 2019].

Alexander, Marc & Christian Kay. 2019b. Versions and Updates. [Online] Available at: https://ht.ac.uk/versions-and-changes/ [Accessed: 21 April 2019].

Alexander, Marc & Christian Kay. 2021a. Our Second Edition. [Online] Available at: https://ht.ac.uk/second-edition/ [Accessed: 27 March 2021].

Alexander, Marc & Christian Kay. 2021b. Thematic Categories. [Online] Available at: https://ht.ac.uk/thematic/ [Accessed: 27 March 2021].

Alexander, Marc & Andrew Struan. 2013. 'In Countries so Unciviliz'd as Those?': The Language of Incivility and the British Experience of the World. In Martin Farr & Xavier Guégan (eds.), *The British Abroad Since the Eighteenth Century, Volume 2: Experiencing Imperialism*. London: Palgrave Macmillan UK. 232–249. DOI: 10.1057/9781137304186_13.

Alker, Sharon-Ruth & Holly Faith Nelson. 2009. *James Hogg and the Literary Marketplace: Scottish Romanticism and the Working-Class Author*. Surrey: Ashgate Publishing Limited.

Allen, William. 2017. Making Corpus Data Visible: Visualising Text with Research Intermediaries. *Corpora*. 12(3). 459–482. DOI: 10.3366/cor.2017.0128.

Alvarez, Alfred. 1961. *The School of Donne*. London: Chatto and Windus.

Archer, Dawn, Andrew Wilson & Paul Rayson. 2002. Introduction to the USAS Category System. [Online]. Retrieved from http://ucrel.lancs.ac.uk/usas/usas_guide.pdf.

Archer, Dawn, Tony McEnery, Paul Rayson & Andrew Hardie. 2003. Developing an Automated Semantic Analysis System for Early Modern English. In *Proceedings of the Corpus Linguistics 2003 Conference*. Centre for Computer Corpus Research on Language Technical Papers, University of Lancaster, Lancaster. 16. 22–31.

Archer, Dawn, Paul Rayson, Scott Piao & Tony McEnery. 2004. Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies. In *Proceedings of the EURALEX-2004 Conference*. European Association for Lexicography. 817–827.

Archer, Dawn, Jonathan Culpeper & Paul Rayson. 2006. Love-a Familiar or a Devil? An Exploration of Key Domains in Shakespeare's Comedies and Tragedies. In *Word Frequency and Keyword Extraction*. Lancaster University. 136–157.

Archer, Dawn. 2012. Corpus Annotation: A Welcome *Addition* or an *Interpretation* Too Far? *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*. 10. [Online].

Archer, Dawn. 2014. Exploring Verbal Aggression in English Using USAS. In Irma Taavitsainen, Andreas H. Jucker, & Jukka Tuominen (eds.), *Diachronic Corpus Pragmatics*. Amsterdam: John Benjamins Publishing Company. 277–302.

Archer, Dawn & Bethan Malory. 2015. Tracing Verbal *Aggression* over Time, Using the Historical Thesaurus of English. In Federica Formato & Andrew Hardie (eds.), *Corpus Linguistics 2015 Abstract Book*. Lancaster, United Kingdom: Lancaster University. 27–28.

Archer, Dawn & Bethan Malory. 2017. Tracing Facework over Time Using Semi-Automated Methods. *International Journal of Corpus Linguistics*. 22(1). 27–56. DOI: 10.1075/ijcl.22.1.02arc.

Arthur, Paul Longley & Katherine Bode (eds.). 2014. *Advancing Digital Humanities*. London: Palgrave Macmillan UK. DOI: 10.1057/9781137337016.

Austin, Frances. 1992. *The Language of the Metaphysical Poets*. London: MacMillan Press.

Baker, Paul. 2004. Querying Keywords: Questions of Difference, Frequency, and Sense in Keywords Analysis. *Journal of English Linguistics*. 32(4). 346–359. DOI: 10.1177/0075424204269894.

Baker Paul. 2006. *Using Corpora in Discourse Analysis*. Continuum. Retrieved from http://ci.nii.ac.jp/ncid/BA78013638.

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

Baron, Alistair. 2019. VARD - About. [Online] Available at: http://ucrel.lancs.ac.uk/vard/about/ [Accessed: 4 July 2019].

Baron, Alistair & Paul Rayson. 2008. VARD 2: A Tool for Dealing with Spelling Variation in Historical Corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Birmingham, UK: Aston University.

Baron, Alistair & Paul Rayson. 2009. Automatic Standardisation of Texts Containing Spelling Variation: How Much Training Data Do You Need? In Michaela Mahlberg, Victorina González-Díaz, & Catherine Smith (eds.), *Proceedings of the Corpus Linguistics Conference*. Lancaster, United Kingdom: Lancaster University. Retrieved from https://eprints.lancs.ac.uk/id/eprint/42529/.

Bassil, Youssef. 2012. A Comparative Study on the Performance of the Top DBMS Systems. *Journal of Computer Science & Research*. 1(1). 20–31.

Bassnett, Susan. 2001. A Century of Editing: *The Oxford Book of English Verse*, 1900-1999. In Ulrich Broich & Susan Bassnett (eds.), *Britain at the Turn of the Twenty-First Century*. New York: Brill Rodopi. 251–264.

Beatty, Bernard. 1990. Continuities and Discontinuities of Language and Voice in Dryden, Pope, and Byron. In Andrew Rutherford (ed.), *Byron: Augustan and Romantic*. London: Palgrave Macmillan UK. 117–135. DOI: 10.1007/978-1-349-21060-2_6.

Beer, John B. 2005. *William Blake: A Literary Life*. New York: Palgrave Macmillan.

Bennett, Joan. 1989. *Five Metaphysical Poets: Donne, Herbert, Vaughan, Crashaw, Marvell*. Cambridge: University Press.

Berlanga, Rafael, Victoria Nebot & María Pérez. 2015. Tailored Semantic Annotation for Semantic Search. *Journal of Web Semantics*. 30. 69–81. DOI: 10.1016/j.websem.2014.07.007.

Bestgen, Yves. 2018. Getting Rid of the Chi-Square and Log-Likelihood Tests for Analysing Vocabulary Differences between Corpora. *Quaderns de Filologia - Estudis Lingüístics*. 22. 33–56. DOI: 10.7203/qf.22.11299.

Biber, Douglas, Susan Conrad & Viviana Cortes. 2003. Lexical Bundles in Speech and Writing: An Initial Taxonomy. In Andrew Wilson, Paul Rayson, & Tony McEnery (eds.), *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*. Frankfurt: Peter Lang. 71–93.

Biber, Douglas. 2011. Corpus Linguistics and the Study of Literature: Back to the Future? *Scientific Study of Literature*. 1(1). 15–23. DOI: 10.1075/ssol.1.1.02bib.

Biber, Douglas & Randi Reppen. 2015. Introduction. In Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press. 1–8.

Blake, William. 2007. *Songs of Innocence and of Experience: Shewing the Two Contrary States of the Human Soul*. London: Tate. Repr. (Original work published 1757).

Blei, David M. 2012. Probabilistic Topic Models. *Communications of the ACM*. 55(4). 77–84. DOI: 10.1145/2133806.2133826.

Bollmann, Marcel, Florian Petran, Stefanie Dipper & Julia Krasselt. 2014. CorA: A Web-Based Annotation Tool for Historical and Other Non-Standard Language Data. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Gothenburg, Sweden: Association for Computational Linguistics. 86–90. DOI: 10.3115/v1/W14-0612.

Bondi, Marina. 2010. Perspectives on Keywords and Keyness. In Marina Bondi & Mike Scott (eds.), *Keyness in Texts*. Philadelphia: John Benjamins Publishing Company. 1–18.

Bottrall, Margaret (ed.). 1970. *William Blake, 'Songs of Innocence and Experience': A Casebook*. London: Macmillan.

Bowker, Lynne & Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=85474.

Brezina, Vaclav. 2016. Collocation Networks: Exploring Associations in Discourse. In Paul Baker & Jesse Egbert (eds.), *Triangulating Methodological Approaches in Corpus Linguistic Research*. London: Routledge. 90–107. Retrieved from http://eprints.lancs.ac.uk/84111/.

Brezina, Vaclav. 2018a. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press. 1st ed. DOI: 10.1017/9781316410899.

Brezina, Vaclav. 2018b. Collocation Graphs and Networks: Selected Applications. In Pascual Cantos-Gómez & Moisés Almela-Sánchez (eds.), *Lexical Collocation Analysis: Advances and Applications*. Germany: Springer International Publishing. 59–83. DOI: 10.1007/978-3-319-92582-0_4.

Burrow, Colin. 2006. *Metaphysical Poetry*. London: Penguin Books.

Burrows, J.F. 1987. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press, Oxford University Press.

Castiglione, Davide. 2017. Difficult Poetry Processing: Reading Times and the Narrativity Hypothesis. *Language and Literature*. 26(2). 99–121. DOI: 10.1177/0963947017704726.

Clark, Alexander, Chris Fox & Shalom Lappin. 2010. Introduction. In Alexander Clark, Chris Fox, & Shalom Lappin (eds.), *The Handbook of Computational Linguistics and Natural Language Processing*. West Sussex: John Wiley & Sons. 1–8.

CLAWS Input / Output Format Guidelines. n.d. [Online] Available at: http://ucrel.lancs.ac.uk/claws/format.html [Accessed: 2 June 2019].

Clayton, Thomas. 1978. *Cavalier Poets: Selected Poems*. Oxford: Oxford University Press.

Cohen, Raphael, Michael Elhadad & Noémie Elhadad. 2013. Redundancy in Electronic Health Record Corpora: Analysis, Impact on Text Mining Performance and Mitigation Strategies. *BMC Bioinformatics*. 14(1). 10. DOI: 10.1186/1471-2105-14-10.

Culpeper, Jonathan. 2002. Computers, Language and Characterisation: An Analysis of Six Characters in Romeo and Juliet. In Ulla Melander-Marttala, Carin Ostman, & Merja Kytö (eds.), *Conversation in Life and in Literature: Papers from the ASLA Symposium*. Uppsala: Association Suedoise de Linguistique Appliquee. 15. 11–30.

Culpeper, Jonathan. 2009. Keyness: Words, Parts-of-Speech and Semantic Categories in the Character-Talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics*. 14(1). 29–59. DOI: 10.1075/ijcl.14.1.03cul.

Culpeper, Jonathan. 2014a. Keywords and Characterization: An Analysis of Six Characters in Romeo and Juliet. In David L. Hoover, Jonathan Culpeper, & Kieran O'Halloran (eds.), *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama*. Abingdon: Routledge. 16. 9–34.

Culpeper, Jonathan. 2014b. Developing Keyness and Characterization: Annotation. In D. L. Hoover, J. Culpeper, & Kieran O'Halloran (eds.), *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama*. Abingdon: Routledge. 16. 35–63.

Currie, Janette. 2005. Re-Visioning James Hogg: The Return of the Subject to Wordsworth's 'Extempore Effusion'. *Romantic Textualities: Literature and Print Culture, 1780–1840*. 15(Winter 2005). 7–68.

Daiches, David. 1969. *A Critical History of English Literature, Vol. 4*. India: Allied Publishers.

Daiches, David. 1979. *A Critical History of English Literature, Vol. 3*. India: Allied Publishers.

de Andrade, Guidson Coelho, Alcione de Paiva Oliveira & Alexandra Moreira. 2019. Ontological Semantic Annotation of an English Corpus Through Condition Random Fields. *Information*. 10(5). 171. DOI: 10.3390/info10050171.

Desagulier, Guillaume. 2019. Can Word Vectors Help Corpus Linguists? *Studia Neophilologica*. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/00393274.2019.1616220.

England, A.B. 1975. The Style of *Don Juan* and Augustan Poetry. In John Davies Jump (ed.), *Byron: A Symposium*. London: MacMillan Press. 94–112.

Evans, Gareth. 2002. Poison Wine — John Keats and the Botanic Pharmacy. *The Keats-Shelley Review*. 16(1). 31–55. DOI: 10.1179/ksr.2002.16.1.31.

Evert, Stefan. 2009. Corpora and Collocations. In *Corpus Linguistics. An International Handbook*. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110213881.2.1212.

Evison, Jane. 2010. What Are the Basics of Analysing a Corpus? In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge. 122–135. DOI: 10.4324/9780203856949.ch1.

Ferry, Anne. 1996. *The Title to the Poem*. Stanford: Stanford University Press.

Firth, John Rupert. 1957. *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Flowerdew, Lynne. 2004. The Argument for Using English Specialized Corpora to Understand Academic and Professional Language. In Ulla Connor & Thomas A Upton (eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins Publishing Company. 16. 11–33. Retrieved from https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=622472.

Flowerdew, Lynne. 2005. An Integration of Corpus-Based and Genre-Based Approaches to Text Analysis in EAP/ESP: Countering Criticisms against Corpus-Based Methodologies. *English for Specific Purposes*. 24(3). 321–332. DOI: 10.1016/j.esp.2004.09.002.

Fowler, Alastair. 1979. Genre and the Literary Canon. *New Literary History*. 11(1). 97–119.

Furkó, Péter B. 2019. Exploring the Fuzzy Boundaries of Discourse Markers Through Manual and Automatic Annotation. In Péter B. Furkó, Ildikó Vaskó, Csilla Ilona Dér, & Dorte Madsen (eds.), *Fuzzy Boundaries in Discourse Studies: Theoretical, Methodological, and Lexico-Grammatical Fuzziness*. Germany: Springer International Publishing. 215–238. DOI: 10.1007/978-3-030-27573-0_10.

Furniss, Tom & Michael Bath. 2013. *Reading Poetry: An Introduction*. London: Routledge.

Gábor, Kata, Haïfa Zargayouna, Davide Buscaldi, Isabelle Tellier & Thierry Charnois. 2016. Semantic Annotation of the ACL Anthology Corpus for the Automatic Analysis of Scientific Literature. In *LREC 2016*. Portoroz, Slovenia. Retrieved from https://hal.archives-ouvertes.fr/hal-01360407.

Gabrielatos, Costas. 2007. Selecting Query Terms to Build a Specialised Corpus from a Restricted-Access Database. *ICAME Journal*. 31. 5–44.

Gardner, Helen Louise. 1972. *The New Oxford Book of English Verse, 1250-1950*. New York, Oxford University Press. Retrieved from http://archive.org/details/newoxfordbookofe00gard.

Garside, Roger. 1987. The CLAWS Word-Tagging System. In Roger Garside, Geoffrey Leech, & Geoffrey Sampson (eds.), *The Computational Analysis of English: A Corpus-Based Approach*. London: Longman. 30–41.

Garside, Roger, Geoffrey Leech & Tony McEnery (eds.). 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Abingdon: Longman.

Garside, Roger & Paul Rayson. 1997. Higher-Level Annotation Tools. In Roger Garside, Geoffrey Leech, & Tony McEnery (eds.), *Corpus Annotation:*

Linguistic Information from Computer Text Corpora. Abingdon: Longman. 179–193.

Garside, Roger & Nicholas Ross Smith. 1997. A Hybrid Grammatical Tagger. In Roger Garside, Geoffrey Leech, & Tony McEnery (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Abingdon: Routledge. 102–121.

Gerbig, Andrea & Anja Müller-Wood. 2002. Trapped in Language: Aspects of Ambiguity and Intertextuality in Selected Poetry and Prose by Sylvia Plath. *Style*. 36(1). 76–92.

Gries, Stefan Th. 2013. 50-Something Years of Work on Collocations. *International Journal of Corpus Linguistics*. 18(1). 137–166. DOI: 10.1075/ijcl.18.1.09gri.

Groom, Nicholas. 2010. Closed-Class Keywords and Corpus-Driven Discourse Analysis. In Marina Bondi & Mike Scott (eds.), *Keyness in Texts*. Philadelphia: John Benjamins Publishing Company. 59–78.

Halliday, Michael Alexander Kirkwood & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Hammond, Gerald. 1974. *The Metaphysical Poets*. London: Macmillan.

Hardie, Andrew. 2012. CQPweb — Combining Power, Flexibility and Usability in a Corpus Analysis Tool. *International Journal of Corpus Linguistics*. 17(3). 380–409. DOI: 10.1075/ijcl.17.3.04har.

Harris, Wendell V. 1991. Canonicity. *Publications of the Modern Language Association of America*. 106(1). 110–121. DOI: 10.2307/462827.

Heuser, Ryan & Long Le-Khac. 2011. Learning to Read Data: Bringing out the Humanistic in the Digital Humanities. *Victorian Studies*. 54(1). 79. DOI: 10.2979/victorianstudies.54.1.79.

Heuser, Ryan & Long Le-Khac. 2012. *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*. Stanford: Stanford Literary Lab. 4.

Hoey, Michael. 2012. *Lexical Priming: A New Theory of Words and Language*. London: Routledge. DOI: 10.4324/9780203327630.

Hughes, John. 2011. *Affective Worlds: Writing, Feeling & Nineteenth-Century Literature*. Brighton: Sussex Academic Press.

Hughes, Rebecca. 2010. What a Corpus Tells Us about Grammar Teaching Materials. In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge. 401–412.

Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li & Liang Zhao. 2019. Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey. *Multimedia Tools and Applications*. 78(11). 15169–15211. DOI: 10.1007/s11042-018-6894-4.

Jovanović, Jelena & Ebrahim Bagheri. 2017. Semantic Annotation in Biomedicine: The Current Landscape. *Journal of Biomedical Semantics*. 8(1). 44. DOI: 10.1186/s13326-017-0153-x.

Jurafsky, Dan & James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, N.J: Prentice Hall.

Kay, Christian, Jane Roberts, Michael Samuels & Irené Wotherspoon (eds.). 2009. *Historical Thesaurus of the Oxford English Dictionary: With Additional Material from A Thesaurus of Old English*. OUP Oxford.

Kay, Christian. 2010. Classification: Principles and Practice. In Michael Adams (ed.), *Cunning Passages, Contrived Corridors: Unexpected Essays in the History of Lexicography*. Monza: Polimetrica. 255–270.

Kay, Christian. 2011. Developing *The Historical Thesaurus of the OED*. In Kathryn Allan & Justyna A. Robinson (eds.), *Current Methods in Historical Semantics*. Berlin: De Gruyter Mouton. 41–58.

Kay, Christian, Jane Roberts, Michael Samuels, Irené Wotherspoon & Marc Alexander (eds.). 2015. *The Historical Thesaurus of English*. Glasgow: University of Glasgow. 4.2.2. Retrieved from http://ht.ac.uk/.

Kay, Christian, Jane Roberts, Michael Samuels, Irené Wotherspoon & Marc Alexander (eds.). 2019. *The Historical Thesaurus of English*. Glasgow: University of Glasgow. 4.21. Retrieved from http://ht.ac.uk/.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. London: Routledge. DOI: 10.4324/9781315843674.

Koester, Almut. 2010. Building Small Specialised Corpora. In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge. 66–79. DOI: 10.4324/9780203856949.ch1.

Koller, Veronika, Andrew Hardie, Paul Rayson & Elena Semino. 2008. Using a Semantic Annotation Tool for the Analysis of Metaphor in Discourse. *Metaphorik.De*. [Online].

Kreimeyer, Kory, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F. Jones, Richard Forshee, Mark Walderhaug & Taxiarchis

Botsis. 2017. Natural Language Processing Systems for Capturing and Standardizing Unstructured Clinical Information: A Systematic Review. *Journal of Biomedical Informatics*. 73. 14–29. DOI: 10.1016/j.jbi.2017.07.012.

Kübler, Sandra & Heike Zinsmeister. 2015. *Corpus Linguistics and Linguistically Annotated Corpora*. London: Bloomsbury Academic. DOI: 10.5040/9781472593573.

Lawrence, Elizabeth A. 1999. Melodius Truth: Keats, a Nightingale, and the Human/Nature Boundary. *Interdisciplinary Studies in Literature and Environment*. 6(2). 21–30.

Lee, David & John Swales. 2006. A Corpus-Based EAP Course for NNS Doctoral Students: Moving from Available Specialized Corpora to Self-Compiled Corpora. *English for Specific Purposes*. 25(1). 56–75. DOI: 10.1016/j.esp.2005.02.010.

Leech, Geoffrey. 1974. *Semantics*. England: Penguin UK.

Leech, Geoffrey. 1991. The State of the Art in Corpus Linguistics. In K Aijmer & B Altenberg (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman. 8–29.

Leech, Geoffrey. 1997. Introducing Corpus Annotation. In Roger Garside, Geoffrey Leech, & Tony McEnery (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Abingdon: Routledge. 1–18.

Leech, Geoffrey & Steven Fligelstone. 1992. Computers and Corpus Analysis. In Christopher Butler (ed.), *Computers and Written Texts*. Oxford: B. Blackwell. 115–140.

Leech, Geoffrey N. 1969. *A Linguistic Guide to English Poetry*. London: Longman.

Löfberg, Laura, Jukka-Pekka Juntunen, Asko Nykänen, Paul Rayson & Dawn Archer. 2004. Using a Semantic Tagger as a Dictionary Search Tool. Presented at the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004), Lorient, France. 127–134.

Louw, Bill. 1989. Sub-Routines in the Integration of Language and Literature. In Ronald Carter & Richard Walker (eds.), *Literature and the Learner: Methodological Approaches*. London: Modern English Publications. 47–54. Retrieved from https://www.academia.edu/844190/Sub_routines_in_the_integration_of_language_and_literature.

Louw, Bill. 1993. Irony in the Text or Insincerity in the Writer? — The Diagnostic Potential of Semantic Prosodies. In Mona Baker, Gill Francis, & Elena Tognini-Bonelli (eds.), *Text and Technology*. Amsterdam: John Benjamins Publishing Company. 157. DOI: 10.1075/z.64.11lou.

Mack, Douglas S. 2004. *Hogg, James (Bap. 1770, d. 1835), Poet and Novelist*. Oxford: Oxford University Press. Retrieved from https://www.oxforddnb.com/view/10.1093/ref:odnb/9780198614128.001.0001/odnb-9780198614128-e-13470.

Mahlberg, Michaela. 2013. *Corpus Stylistics and Dickens's Fiction*. Abingdon: Routledge.

Mahlberg, Michaela. 2014. Corpus Stylistics. In M. Burke (ed.), *The Routledge Handbook of Stylistics*. London: Routledge. 378–392.

Mahlberg, Michaela & Dan McIntyre. 2011. A Case for Corpus Stylistics: Ian Fleming's *Casino Royale*. *English Text Construction*. 4(2). 204–227. DOI: 10.1075/etc.4.2.03mah.

Marchi, Anna & Charlotte Taylor. 2018. Introduction: Partiality and Reflexivity. In Charlotte Taylor & Anna Marchi (eds.), *Corpus Approaches to Discourse: A Critical Review*. London: Routledge. 1–15.

Martínez-Cámara, Eugenio, Maria Teresa Martín-Valdivia, L. Alfonso Ureña López & Arturo Montejo-Ráez. 2014. Sentiment Analysis in Twitter. *Natural Language Engineering*. 20(01). 1–28. DOI: 10.1017/S1351324912000332.

Mautner, Gerlinde. 2012. Corpora and Critical Discourse Analysis. In Paul Baker (ed.), *Contemporary Corpus Linguistics*. London: Continuum International Publishing Group. 32–46.

McCarthy, Michael & Anne O'Keeffe. 2010. Historical Perspective: What Are Corpora and How Have They Evolved? In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge. 3–13. DOI: 10.4324/9780203856949.ch1.

McEnery, Tony & Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, Tony & Andrew Wilson. 1996. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

McEnery, Tony & Andrew Wilson. 2001. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press. 2. ed., repr.

McIntyre, Dan & Dawn Archer. 2010. A Corpus-Based Approach to Mind Style. *Journal of Literary Semantics*. 39(2). 167–182. DOI: 10.1515/jlse.2010.009.

McIntyre, Dan & Brian Walker. 2010. How Can Corpora Be Used to Explore the Language of Poetry and Drama? In Anne O'Keeffe & Michael McCarthy

(eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge. 516–530. DOI: 10.4324/9780203856949.ch1.

McIntyre, Dan & Brian Walker. 2019. *Corpus Stylistics: Theory and Practice*. Edinburgh: Edinburgh University Press.

Miner, Earl. 1971. *The Cavalier Mode from Jonson to Cotton*. Princeton, NJ: Princeton University Press. Retrieved from https://archive.org/details/cavaliermodefrom00mine.

O'Halloran, Kieran. 2007. Corpus-Assisted Literary Evaluation. *Corpora*. 2(1). 33–63. DOI: 10.3366/cor.2007.2.1.33.

O'Halloran, Kieran. 2014. Performance Stylistics: Deleuze and Guattari, Poetry, and (Corpus) Linguistics. In D. L. Hoover, J. Culpeper, & Kieran O'Halloran (eds.), *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama*. Abingdon: Routledge. 16.

Padro, Lluís & Lluís Marquez. 1998. On the Evaluation and Comparison of Taggers: The Effect of Noise in Testing Corpora. In *Joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, COLING-ACL*. Montréal, Canada. 997–1002. Retrieved from http://arxiv.org/abs/cs/9809112.

Palmer, Martha Stone. 1990. *Semantic Processing for Finite Domains*. Cambridge: Cambridge University Press.

Piao, Scott, Paul Rayson, Dawn Archer & Tony McEnery. 2004. Evaluating Lexical Resources for A Semantic Tagger. In *Proceedings of 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal. II. 499–502.

Piao, Scott, Dawn Archer, Olga Mudraya, Paul Rayson, Roger Garside, Tony McEnery & Andrew Wilson. 2005a. A Large Semantic Lexicon for Corpus Annotation. In *Proceedings from the Corpus Linguistics Conference Series*. Birmingham, UK. 1. [Online].

Piao, Scott, Paul Rayson, Dawn Archer & Tony McEnery. 2005b. Comparing and Combining a Semantic Tagger and a Statistical Tool for MWE Extraction. *Computer Speech & Language*. 19(4). 378–397. DOI: 10.1016/j.csl.2004.11.002.

Piao, Scott, Paul Rayson, Olga Mudraya, Andrew Wilson & Roger Garside. 2006. Measuring MWE Compositionality Using Semantic Annotation. In *Proceedings of the Workshop on Multiword Expressions Identifying and Exploiting Underlying Properties - MWE '06*. Sydney, Australia: Association for Computational Linguistics. 2. DOI: 10.3115/1613692.1613695.

Piao, Scott, Francesca Bianchi, Carmen Dayrell, Angela D'Egidio & Paul Rayson. 2015. Development of the Multilingual Semantic Annotation System. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics. 1268–1274. DOI: 10.3115/v1/N15-1137.

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. DOI: 10.1016/j.csl.2017.04.010.

Pittock, Murray. 2008. *Hogg, Maturin, and the Gothic National Tale*. Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199232796.001.0001/acprof-9780199232796-chapter-9.

Preminger, Alex, Frank J. Warnke & O.B. Hardison (eds.). 1974. *Princeton Encyclopedia of Poetry and Poetics*. Princeton, N.J: Princeton University Press.

Prentice, Sheryl. 2010. Using Automated Semantic Tagging in Critical Discourse Analysis: A Case Study on Scottish Independence from a Scottish Nationalist Perspective. *Discourse & Society*. 21(4). 405–437. DOI: 10.1177/0957926510366198.

Quiller-Couch, Arthur (ed.). 1998. *Bulchevy's Book of English Verse*. Retrieved from http://www.gutenberg.org/ebooks/1304.

Quiller-Couch, Arthur T., Sir. 1999. *The Oxford Book of English Verse*. Oxford: Clarendon. Retrieved from https://www.bartleby.com/br/101.html (Original work published 1919).

Rawson, Claude. 1990. Byron Augustan: Mutations of the Mock-Heroic in Don Juan and Shelley's Peter Bell the Third. In Andrew Rutherford (ed.), *Byron: Augustan and Romantic*. London: Palgrave Macmillan UK. 82–116. DOI: 10.1007/978-1-349-21060-2_5.

Rayson, Paul. 2002. *Matrix: A Statistical Method and Software Tool for Linguistic Analysis through Corpus Comparison*. PhD Thesis. Lancaster University.

Rayson, Paul, Damon Berridge & Brian Francis. 2004a. Extending the Cochran Rule for the Comparison of Word Frequencies between Corpora. In *7th International Conference on Statistical Analysis of Textual Data*.

Rayson, Paul. 2004. Keywords Are Not Enough. Presented at the JAECS (Japan Association for English Corpus Studies) at Chuo University, Tokyo, Japan.

Rayson, Paul, Dawn Archer, Scott Piao & Tony Mcenery. 2004b. The UCREL Semantic Analysis System. In *Proceedings of the Workshop on Beyond Named Entity Recognition Semantic Labelling for NLP Tasks in Association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal. 7–12.

Rayson, Paul. 2005. *Right from the Word Go: Identifying MWE for Semantic Tagging*. [Lecture], UCREL, Computing Department, Lancaster University.

Rayson, Paul. 2007a. Wmatrix Domain Tag Wizard [Software]. [Online] Available at: https://ucrel-wmatrix4.lancaster.ac.uk/cgi-bin/wmatrix4/prompt_lite_domain.pl.

Rayson, Paul. 2007b. Wmatrix My Tag Wizard [Software]. [Online] Available at: https://ucrel-wmatrix4.lancaster.ac.uk/wmatrix4.html.

Rayson, Paul. 2007c. Updates (My Tag Wizard) [Software]. [Online] Available at: https://ucrel-wmatrix4.lancaster.ac.uk/wm4/help/update.html.

Rayson, Paul. 2008. From Key Words to Key Semantic Domains. *International Journal of Corpus Linguistics*. 13(4). 519–549. DOI: 10.1075/ijcl.13.4.06ray.

Rayson, Paul. 2009a. *Wmatrix: A Web-Based Corpus Processing Environment*. Lancaster University: Computing Department. Retrieved from http://ucrel.lancs.ac.uk/wmatrix/.

Rayson, Paul. 2009b. Introduction to Wmatrix [Software]. [Online] Available at: http://ucrel.lancs.ac.uk/wmatrix/index.html#screen.

Rayson, Paul. 2013. Updates (CrossTab) [Software]. [Online] Available at: https://ucrel-wmatrix4.lancaster.ac.uk/wm4/help/update.html.

Rayson, Paul. 2016. Log-Likelihood and Effect Size Calculator. [Online] Available at: http://ucrel.lancs.ac.uk/llwizard.html.

Rayson, Paul, Alistair Baron & Andrew Moore. 2017. *VARDing to Modernise Spellings in Historical Texts for Improved Corpus Analysis*. Presented at the VARDsourcing challenge: Text Hackathon, CTS, De Montford University. Retrieved from http://cts.dmu.ac.uk/events/hackathon/slides/PR-handout.pdf.

Rayson, Paul & Andrew Wilson. 1996. The ACAMRIT 1 Semantic Tagging System: Progress Report. In *AISB96 Workshop Proceedings* 13–20.

Ricks, Christopher B. 1999. *The Oxford Book of English Verse*. Oxford: Oxford University Press. Retrieved from http://archive.org/details/oxfordbookofengl00chri.

Rodríguez-Fernández, S., Luis Espinosa-Anke, Roberto Carlini & Leo Wanner. 2016. Semantics-Driven Recognition of Collocations Using Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany. 2. 499–505.

Roget, Peter Mark. 1852. *Thesaurus of English Words and Phrases: Classified and Arranged So as to Facilitate the Expression of Ideas and Assist in Literary Composition*. London: Longman.

Scott, Mike. 1997. PC Analysis of Key Words — And Key Key Words. *System*. 25(2). 233–245. DOI: 10.1016/S0346-251X(97)00011-0.

Scott, Mike. 1998. *WordSmith Tools Manual Version 3.0*. Oxford University Press.

Scott, Mike. 2010a. What Can Corpus Software Do? In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge. 136–151. DOI: 10.4324/9780203856949.ch1.

Scott, Mike. 2010b. Problems in Investigating Keyness, or Clearing the Undergrowth and Marking out Trails…. In Marina Bondi & Mike Scott (eds.), *Keyness in Texts*. Amsterdam: John Benjamins Publishing Company. 43–58.

Scott, Mike & Christopher Tribble. 2006. *Textual Patterns*. Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/scl.22.

Semino, Elena, Andrew Hardie, Veronika Koller & Paul Rayson. 2005. A Computer-Assisted Approach to the Analysis of Metaphor Variation across Genres. Presented at the Corpus-Based Approaches to Figurative Language: A Corpus Linguistics 2009 Colloquium, University of Birmingham School of Computer Science.

Semino, Elena. 2010. Descriptions of Pain, Metaphor, and Embodied Simulation. *Metaphor and Symbol*. 25(4). 205–226. DOI: 10.1080/10926488.2010.510926.

Semino, Elena & Mick Short. 2004. *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. London: Routledge.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John. 1994. Trust the Text. In Malcolm Coulthard (ed.), *Advances in Written Text Analysis*. London: Routledge. 12–25.

Sinclair, John. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge. 1st ed. DOI: 10.4324/9780203594070.

Skelton, Robin. 1969. *Cavalier Poets*. Harlow, England: Longmans.

Slattery, Dennis Patrick. 2005. The Myth of Nature and the Nature of Myth: Becoming Transparent to Transcendence. *International Journal of Transpersonal Studies*. 24(1). 29–36. DOI: 10.24972/ijts.2005.24.1.29.

Smith, Jeremy J. 2006. Notes on the Medical Vocabulary of John Keats. In Graham D. Caie, Carole Hough, & Irené Wotherspoon (eds.), *The Power of Words: Essays in Lexicography, Lexicology and Semantics in Honour of Christian J. Kay*. Amsterdam: Rodopi. 159–169.

Sperry, Eileen M. 2019. Decay, Intimacy, and the Lyric Metaphor in John Donne. *SEL Studies in English Literature 1500-1900*. 59(1). 45–66. DOI: 10.1353/sel.2019.0002.

Stephens, John & Ruth Waterhouse. 1990. *Literature, Language, and Change: From Chaucer to the Present*. London: Routledge.

Stillinger, Jack. 1968. Keats and Romance. *Studies in English Literature, 1500-1900*. 8(4). 593. DOI: 10.2307/449467.

Stubbs, Michael. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford, UK: Blackwell Publishers.

Stubbs, Michael. 2010. Three Concepts of Keywords. In Marina Bondi & Mike Scott (eds.), *Keyness in Texts*. Philadelphia: John Benjamins Publishing Company. 21–42.

Studer, Patrick. 2008. *Historical Corpus Stylistics: Media, Technology and Change*. London: Continuum.

Taylor, Charlotte. 2013. Searching for Similarity Using Corpus-Assisted Discourse Studies. *Corpora*. 8(1). 81–113. DOI: 10.3366/cor.2013.0035.

Taylor, Charlotte. 2018. Similarity. In Charlotte Taylor & Anna Marchi (eds.), *Corpus Approaches to Discourse: A Critical Review*. New York: Routledge. 19–37.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins Publishing Company.

Tognini-Bonelli, Elena. 2010. Theoretical Overview of the Evolution of Corpus Linguistics. In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge. 14–28. DOI: 10.4324/9780203856949.ch1.

van Peer, W. 1989. Quantitative Studies of Literature. A Critique and an Outlook. *Computers and the Humanities*. 23(4). 301–307. DOI: 10.1007/BF02176635.

Weisser, Martin. 2014. The Simple Corpus Tool (SCT). [Online] Available at: http://martinweisser.org/ling_soft.html#tagger [Accessed: 1 March 2021].

Willy, Margaret (ed.). 1971. *The Metaphysical Poets*. London: Edward Arnold.

Wilson, Andrew & Paul Rayson. 1993. Automatic Content Analysis of Spoken Discourse: A Report on Work in Progress. In Clive Souter & Eric Atwell (eds.), *Corpus-Based Computational Linguistics*. Amsterdam: Rodopi. 215–226.

Wilson, Andrew & Jenny Thomas. 1997. Semantic Annotation. In Roger Garside, Geoffrey Leech, & Tony McEnery (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Abingdon: Routledge. 54–65.

Xiao, Richard. 2015. Collocation. In Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press. 106–124.

Xiao, Richard & Ming Yue. 2012. Using Corpora in Translation Studies: The State of the Art. In Paul Baker (ed.), *Contemporary Corpus Linguistics*. London: Continuum International Publishing Group. 237–261.

Appendix I Tables

Table A1 Table with authors and number of poems

Excluded group Author's name	Number of entries
Anonymous. XIII–XIV Century	7
Robert Mannyng of Brunne. 1269–1340	1
John Barbour. d. 1395	1
Geoffrey Chaucer. 1340?–1400	3
Thomas Hoccleve. 1368–9?–1450?	1
John Lydgate. 1370?–1450?	1
King James I of Scotland. 1394–1437	1
Robert Henryson. 1425–1500	2
William Dunbar. 1465–1520?	4
Anonymous. XV–XVI Century	8
John Skelton. 1460?–1529	2
Stephen Hawes. d. 1523	2

Group 1 Author's name	Number of entries
Sir Thomas Wyatt. 1503–1542	5
Henry Howard, Earl of Surrey. 1516–47	3
Nicholas Grimald. 1519–62	1
Alexander Scott. 1520?–158–	2
Robert Wever. c. 1550	1
Richard Edwardes. 1523–66	1
George Gascoigne. 1525?–77	1
Alexander Montgomerie. 1540?–1610?	1
William Stevenson. 1530?–1575	1
Anonymous. XVI–XVII Century	23
Nicholas Breton. 1542–1626	2
Sir Walter Raleigh. 1552–1618	4
Edmund Spenser. 1552–1599	6
John Lyly. 1553–1606	2
Anthony Munday. 1553–1633	1
Sir Philip Sidney. 1554–86	8
Fulke Greville, Lord Brooke. 1554–1628	1
Thomas Lodge. 1556?–1625	4
George Peele. 1558?–97	2

Group 1 Author's name	Number of entries
Robert Greene. 1560–92	3
Alexander Hume. 1560–1609	1
George Chapman. 1560–1634	1
Robert Southwell. 1561–95	2
Henry Constable. 1562?–1613?	1
Samuel Daniel. 1562–1619	3
Mark Alexander Boyd. 1563–1601	1
Joshua Sylvester. 1563–1618	1
Michael Drayton. 1563–1631	5
Christopher Marlowe. 1564–93	2
William Shakespeare. 1564–1616	42
Richard Rowlands. 1565–1630?	1
Thomas Nashe. 1567–1601	2
Thomas Campion. 1567?–1619	9
John Reynolds. 16th Cent.	1
Sir Henry Wotton. 1568–1639	3
Sir John Davies. 1569–1626	1
Sir Robert Ayton. 1570–1638	2
Ben Jonson. 1573–1637	11
John Donne. 1573–1631	8
Richard Barnefield. 1574–1627	1
Thomas Dekker. 1575–1641	1
Thomas Heywood. 157?–1650	2
John Fletcher. 1579–1625	11
John Webster. ?–1630?	3
William Alexander, Earl of Stirling. 1580?–1640	1
Phineas Fletcher. 1580–1650	1
Sir John Beaumont. 1583–1627	1
William Drummond, of Hawthornden. 1585–1649	9
Giles Fletcher. 158?–1623	1
Francis Beaumont. 1586–1616	1
John Ford. 1586–1639	1
George Wither. 1588–1667	4
William Browne, of Tavistock. 1588–1643	7
Robert Herrick. 1591–1674	29
Francis Quarles. 1592–1644	2
Henry King, Bishop of Chichester. 1592–1669	3
George Herbert. 1593–1632	6
James Shirley. 1596–1666	2
Thomas Carew. 1595?–1639?	7

Group 2 Author's name	Number of entries
Jasper Mayne. 1604–1672	1
William Habington. 1605–1654	2
Thomas Randolph. 1605–1635	2
Sir William Davenant. 1606–1668	3
Edmund Waller. 1606–1687	3
John Milton. 1608–1674	18
Sir John Suckling. 1609–1642	4
Sir Richard Fanshawe. 1608–1666	1
William Cartwright. 1611–1643	4
James Graham, Marquis of Montrose. 1612–1650	1
Thomas Jordan. 1612?–1685	1
Richard Crashaw. 1613?–1649	7
Richard Lovelace. 1618–1658	6
Abraham Cowley. 1618–1667	5
Alexander Brome. 1620–1666	1
Andrew Marvell. 1621–1678	7
Henry Vaughan. 1621–1695	4
John Bunyan. 1628–1688	1
Anonymous: Ballads.	26
William Strode. 1602–1645	1
Thomas Stanley. 1625–1678	1
Thomas D'Urfey. 1653–1723	1
Charles Cotton. 1630–1687	1
Katherine Philips ('Orinda'). 1631–1664	1
John Dryden. 1631–1700	5
Charles Webbe. c. 1678	1
Sir George Etherege. 1635–1691	2
Thomas Traherne. 1637?–1674	1
Thomas Flatman. 1637–1688	1
Charles Sackville, Earl of Dorset. 1638–1706	1
Sir Charles Sedley. 1639–1701	2
Aphra Behn. 1640–1689	2
John Wilmot, Earl of Rochester. 1647–1680	4
John Sheffield, Duke of Buckinghamshire. 1649–1720	2
Thomas Otway. 1652–1685	1
John Oldham. 1653–1683	1
John Cutts, Lord Cutts. 1661–1707	1
Matthew Prior. 1664–1721	7
William Walsh. 1663–1708	1
Lady Grisel Baillie. 1665–1746	1
William Congreve. 1670–1729	2
Joseph Addison. 1672–1719	1
Isaac Watts. 1674–1748	2

Group 2 Author's name	Number of entries
Thomas Parnell. 1670–1718	1
Allan Ramsay. 1686–1758	1
William Oldys. 1687–1761	1
John Gay. 1688–1732	1
Alexander Pope. 1688–1744	3
George Bubb Dodington, Lord Melcombe. 1691?–1762	1
Henry Carey. 1693?–1743	2
William Broome. ?–1745	2

Group 3 Author's name	Number of entries
James Thomson. 1700–1748	1
George Lyttelton, Lord Lyttelton. 1709–1773	1
Samuel Johnson. 1709–1784	2
Richard Jago. 1715–1781	1
Thomas Gray. 1716–1771	4
William Collins. 1721–1759	4
Mark Akenside. 1721–1770	3
Tobias George Smollett. 1721–1771	1
Christopher Smart. 1722–1770	1
Jane Elliot. 1727–1805	1
Oliver Goldsmith. 1728–1774	2
Robert Cunninghame-Graham of Gartmore. 1735–1797	1
William Cowper. 1731–1800	2
James Beattie. 1735–1803	1
Isobel Pagan. 1740–1821	1
Anna Lætitia Barbauld. 1743–1825	1
Fanny Greville. 18th Cent.	1
John Logan. 1748–1788	1
Lady Anne Lindsay. 1750–1825	1
Sir William Jones. 1746–1794	1
Thomas Chatterton. 1752–1770	1
George Crabbe. 1754–1832	3
William Blake. 1757–1827	10
Robert Burns. 1759–1796	14
Henry Rowe. 1750–1819	2
William Lisle Bowles. 1762–1850	1
Joanna Baillie. 1762–1851	1
Mary Lamb. 1765–1847	1
Carolina, Lady Nairne. 1766–1845	1
James Hogg. 1770–1835	2

Group 3 Author's name	Number of entries
William Wordsworth. 1770–1850	27
Sir Walter Scott. 1771–1832	7
Samuel Taylor Coleridge. 1772–1834	7
Robert Southey. 1774–1843	1
Walter Savage Landor. 1775–1864	20
Charles Lamb. 1775–1834	3
Thomas Campbell. 1774–1844	2
Thomas Moore. 1779–1852	4
Edward Thurlow, Lord Thurlow. 1781–1829	1
Ebenezer Elliott. 1781–1849	2
Allan Cunningham. 1784–1842	3
Leigh Hunt. 1784–1859	1
Thomas Love Peacock. 1785–1866	3
Caroline Southey. 1787–1854	1
George Gordon Byron, Lord Byron. 1788–1824	5
Sir Aubrey De Vere. 1788–1846	1
Charles Wolfe. 1791–1823	2
Percy Bysshe Shelley. 1792–1822	14
Hew Ainslie. 1792–1878	1
John Keble. 1792–1866	1
John Clare. 1793–1864	1
Felicia Dorothea Hemans. 1793–1835	1
John Keats. 1795–1821	15
Jeremiah Joseph Callanan. 1795–1839	1
William Sidney Walker. 1795–1846	1
George Darley. 1795–1846	3
Hartley Coleridge. 1796–1849	4
Thomas Hood. 1798–1845	8
William Thom. 1798–1848	1

Group 4 Author's name	Number of entries
Sir Henry Taylor. 1800–1866	1
Thomas Babington Macaulay, Lord Macaulay. 1800–1859	1
William Barnes. 1801–1886	2
Winthrop Mackworth Praed. 1802–1839	1
Sara Coleridge. 1802–1850	2
Gerald Griffin. 1803–1840	1
James Clarence Mangan. 1803–1849	2
Thomas Lovell Beddoes. 1803–1849	3
Ralph Waldo Emerson. 1803–1882	4

Group 4 Author's name	Number of entries
Richard Henry Horne. 1803–1884	1
Robert Stephen Hawker. 1804–1875	2
Thomas Wade. 1805–1875	1
Francis Mahony. 1805–1866	1
Elizabeth Barrett Browning. 1806–1861	10
Frederick Tennyson. 1807–1898	1
Henry Wadsworth Longfellow. 1807–1882	1
John Greenleaf Whittier. 1807–1892	1
Helen Selina, Lady Dufferin. 1807–1867	1
Caroline Elizabeth Sarah Norton. 1808–1876	1
Charles Tennyson Turner. 1808–1879	1
Edgar Allan Poe. 1809–1849	3
Edward Fitzgerald. 1809–1883	2
Alfred Tennyson, Lord Tennyson. 1809–1892	11
Richard Monckton Milnes, Lord Houghton. 1809–1885	1
Henry Alford. 1810–1871	1
Sir Samuel Ferguson. 1810–1886	3
Robert Browning. 1812–1889	16
William Bell Scott. 1812–1890	1
Aubrey De Vere. 1814–1902	2
George Fox. 1815–?	1
Emily Brontë. 1818–1848	4
Charles Kingsley. 1819–1875	2
Arthur Hugh Clough. 1819–1861	1
Walt Whitman. 1819–1892	2
John Ruskin. 1819–1900	1
Ebenezer Jones. 1820–1860	1
Frederick Locker-Lampson. 1821–1895	1
Matthew Arnold. 1822–1888	8
William Brighty Rands. 1823–1880	2
William Philpot. 1823–1889	1
William (Johnson) Cory. 1823–1892	2
Coventry Patmore. 1823–1896	5
Sydney Dobell. 1824–1874	4
William Allingham. 1824–1889	1
George MacDonald. 1824–1905	1
Dante Gabriel Rossetti. 1828–1882	1
George Meredith. 1828–1909	5
Alexander Smith. 1829–1867	2
Christina Georgina Rossetti. 1830–1894	11
Thomas Edward Brown. 1830–1897	4
Edward Robert Bulwer Lytton, Earl of Lytton. 1831–1892	2
James Thomson. 1834–1882	4

Group 4 Author's name	Number of entries
William Morris. 1834–1896	3
Roden Berkeley Wriothesley Noel. 1834–1894	2
Thomas Ashe. 1836–1889	2
Theodore Watts-Dunton. 1836–1914	1
Algernon Charles Swinburne. 1837–1909	4
William Dean Howells. b. 1837	1
Bret Harte. 1839–1902	1
John Todhunter. 1839–1916	2
Wilfrid Scawen Blunt. b. 1840	8
Henry Austin Dobson. b. 1840	3
Henry Clarence Kendall. 1841–1882	1
Arthur William Edgar O'Shaughnessy. 1844–1881	3
John Boyle O'Reilly. 1844–1890	1
Robert Bridges. b. 1844	9
Andrew Lang. 1844–1912	1
William Ernest Henley. 1849–1903	3
Edmund Gosse. b. 1849	1
Robert Louis Stevenson. 1850–1894	3
T. W. Rolleston. b. 1857	1
John Davidson. 1857–1909	2
William Watson. b. 1858	3
Henry Charles Beeching. 1859–1919	2
Bliss Carman. b. 1861	1
Douglas Hyde. b. 1861	1
Arthur Christopher Benson. b. 1862	1
Henry Newbolt. b. 1862	1
Gilbert Parker. b. 1862	1
William Butler Yeats. b. 1865	3
Rudyard Kipling. b. 1865	3
Richard Le Gallienne. b. 1866	2
Laurence Binyon. b. 1869	2
George William Russell ('A. E.'). b. 1853	2
T. Sturge Moore. b. 1870	1
Francis Thompson. 1859–1907	1
Henry Cust. 1861–1917	1
Katharine Tynan Hinkson. b. 1861	1
Frances Bannerman.	1
Alice Meynell. b. 1850	2
Dora Sigerson. d. 1918	1
Margaret L. Woods. b. 1856	1
R. D. Blackmore. 1825–1900.	1

Table A2 Wmatrix output

Sentence number	CLAWS code	CLAWS POS	original word	lemma
0000001	002	-----	-----	PUNC
0000002	010	MC	34	34
0000002	011	.	.	PUNC
0000002	020	VV0	Forget	forget
0000002	030	XX	not	not
0000002	040	RR	yet	yet
0000003	010	AT	The	the
0000003	020	NN1	Lover	lover
0000003	030	VVZ	Beseecheth	beseecheth
0000003	040	APPGE	his	his
0000003	050	NN1	Mistress	mistress
0000003	060	XX	not	not
0000003	070	TO	to	to
0000003	080	VV0	Forget	forget
0000003	090	APPGE	his	his
0000004	010	JJ	Steadfast	steadfast
0000004	020	NN1	Faith	faith
0000004	030	CC	and	and
0000004	040	JJ	True	true
0000004	050	NN1	Intent	intent
0000006	010	VV0	FORGET	forget
0000006	020	XX	not	not
0000006	030	RR	yet	yet
0000006	040	AT	the	the
0000006	050	JJ@	tried	tried
0000006	060	NN1	intent	intent
0000007	010	IO	Of	of
0000007	020	DA	such	such
0000007	030	AT1	a	a
0000007	040	NN1	truth	truth
0000007	050	CSA	as	as
0000007	060	PPIS1	I	i
0000007	070	VH0	have	have
0000007	080	VVN	meant	mean
0000007	081	;	;	PUNC
0000008	010	APPGE	My	my
0000008	020	JJ	great	great
0000008	030	NN1	travail	travail
0000008	040	RG	so	so
0000008	050	RR	gladly	gladly
0000008	060	VVN	spent	spend
0000008	061	,	,	PUNC
0000009	010	VV0	Forget	forget
0000009	020	XX	not	not
0000009	030	RR	yet	yet

0000009	031	!	!	PUNC

Table A3 HTST output

Word	Lemma	POS
34	34	MC
.	PUNC	YSTP
S_END	NULL	NULL
S_BEGIN	NULL	NULL
Forget	forget	VV0
not	not	XX
yet	yet	RR
The	the	AT
Lover	lover	NN1
Beseeches	beseech	VVZ
his	his	APPGE
Mistress	mistress	NN1
not	not	XX
to	to	TO
Forget	forget	VV0
his	his	APPGE
Steadfast	steadfast	JJ
Faith	faith	NN1
and	and	CC
True	true	JJ
Intent	intent	NN1
FORGET	forget	VV0
not	not	XX
yet	yet	RR
the	the	AT
tried	tried	JJ
intent	intent	NN1
Of	of	IO
such	such	DA
a	a	AT1
truth	truth	NN1
as	as	CSA
I	i	PPIS1
have	have	VH0
meant	mean	VVN
;	PUNC	YSCOL
My	my	APPGE
great	great	JJ
travail	travail	NN1
so	so	RG
gladly	gladly	RR
spent	spend	VVN
,	PUNC	YCOM
Forget	forget	VV0

not	not	XX
yet	yet	RR

Table A4 Null lexical items for Corpus Group One before VARD

corp_id	corp_word	corp_lemma	corp_pos
10	Beseecheth	beseecheth	VVZ
62	whan	whan	NN1
315	pitye	pitye	NN1
319	loveth	loveth	VVZ
460	obeyed	obeye	VVN
465	betrayed	betraye	VVN
474	unkist	unkist	NN1
479	Vixi	vixi	NP1@
480	Puellis	puellis	NP1
481	Nuper	nuper	NP1
482	Idoneus	idoneus	NP1
664	fangleness	fangleness	NN1
806	affectin	affectin	NN1
1050	soote	soote	NN1
1101	hart	hart	NN1
1123	flete	flete	NN1
1142	pursueth	pursueth	VVZ
1145	smale	smale	NN1
1217	woful	woful	JJ
1254	rememberance	rememberance	NN1
1439	drencheth	drencheth	VVZ
1787	desart	desart	NN1
1868	Apollo	apollo	NP1
1920	hald	hald	VVI
1924	soverane	soverane	NN1
1930	liever	liever	VV0

Table A5 Null results corpus group one

position_ID	corp_word	corp_lemma	corp_pos
1	34	34	MC
123	hath	hath	VHZ
131	that	that	CST
152	hath	hath	VHZ
157	Whose	whose	DDQGE
168	35	35	MC
174	An	an	AT1
245	That	that	DD1
246	hath	hath	VHZ
287	That	that	DD1
288	hath	hath	VHZ
334	that	that	CST
362	36	36	MC
431	Your	your	APPGE
480	that	that	DD1
483	That	that	CST
488	obeyed	obeye	VVN
494	betrayed	betraye	VVN
496	that	that	CST
504	unkist	unkist	NN1
506	37	37	MC
510	Vixi	vixi	NP1
511	Puellis	puellis	NP1
512	Nuper	nuper	NP1
513	Idoneus	idoneus	NP1
521	that	that	CST
547	That	that	DD1
557	That	that	CST
562	themselves	themselves	PPX2
590	hath	hath	VHZ
696	new-fangleness	new-fangleness	NN1
702	that	that	CST
716	hath	hath	VHZ

Table A6 Pre-filter output

ID	Word	POS	apps	appe	heading	catid	Cat pos	tid	thematicheading
22	FORGET	VV0	1385	1390	Be/become mad	12184	vi	322	Mental illness
22	FORGET	VV0	1400	1670	cease the practice/observation of something	79165	vt	1437	Ceasing
22	FORGET	VV0	1300	2000	leave undone/fail to perform/carry out	79339	vt	1441	Not doing
22	FORGET	VV0	1297	1797	Be careless/heedless of	85336	vt	1511	Carelessness
22	FORGET	VV0	1538	2000	overlook inadvertently	85338	vt	1511	Carelessness
22	FORGET	VV0	1200	1225	Omit to care for oneself	85354	vr	1511	Carelessness
22	FORGET	VV0	1861	2000	act of forgetting	118383	n	1789	Faulty recollection, forgetting
22	FORGET	VV0	1382	2000	forget, fail to remember	118408	vi	1789	Faulty recollection, forgetting
22	FORGET	VV0	1000	2000	forget, cease to know	118411	vt	1789	Faulty recollection, forgetting
22	FORGET	VV0	1787	2000	fail to recollect	118413	vt	1789	Faulty recollection, forgetting
22	FORGET	VV0	1300	2000	leave behind	118414	vt	1789	Faulty recollection, forgetting
22	FORGET	VV0	1000	2000	Ignore, disregard	122409	vt	1902	Inattention
22	FORGET	VV0	1000	1842	omit, pass over	122417	vt	1902	Inattention
22	FORGET	VV0	1000	1841	through inadvertance	125146	vi	2013	Undutifulness
22	FORGET	VV0	1582	1582	be/get lost	188998	vr	3270	Travel in specific course/direction
23	not	XX	1602	1680	awnless/beardless	22495	aj	584	Cereal/corn/grain
23	not	XX	1837	1875	hornless sheep	38722	n	520	Genus Ovis (sheep)
23	not	XX	1587	2000	having small/no horns	38787	aj	520	Genus Ovis (sheep)

23	not	XX	1587	2000	having no horns	38935	aj	523	Bos taurus (ox)
23	not	XX	1380	1508	nothing	75904	n	1181	Non-existence
23	not	XX	1362	2000	not	75937	av	1181	Non-existence
23	not	XX	1959	2000	particular operations	109235	n	1618	Computing/information technology
23	not	XX	1530	1674	cut	135671	vt	1984	Hairdressing
23	not	XX	1601	1866	utterance/instance of	145934	n	2241	Denial, dissent
24	yet	RR	1513	1536	cause to flow in flood	2205	vt	38	Flood/flooding
24	yet	RR	1200	1300	draw/drain of blood	11834	vt	316	Injury
24	yet	RR	1000	1513	Excrete	25804	vt	247	Organs of excretion
24	yet	RR	1000	1533	Be emitted	68032	vi	1039	Liquid which has been emitted
24	yet	RR	1000	1533	in (a) stream(s)	68168	vi	1040	Action/process of flowing
24	yet	RR	1000	2000	Pour	68222	vt	1041	Action/fact of pouring/being poured
24	yet	RR	1560	1560	into a cavity	68224	vt	1041	Action/fact of pouring/being poured
24	yet	RR	1374	1501	Shine	73597	vt	1126	Light
24	yet	RR	1606	2000	still continuing/enduring	87971	aj	1316	Duration
24	yet	RR	1000	2000	as formerly/still/to this day	89116	av	1340	The present (time)
24	yet	RR	1000	1879	even now (though not until now)	89117	av	1340	The present (time)
24	yet	RR	1000	1460	once/at some former time	89193	av	1341	The past
24	yet	RR	1250	2000	in advance/beforehand	89297	av	1343	Antecedence/being earlier
24	yet	RR	1000	2000	yet/still/hitherto	89305	av	1343	Antecedence/being earlier
24	yet	RR	1000	1849	from now/henceforth	89525	av	1346	The future/time to come
24	yet	RR	1250	1352	at the latest	89539	av	1346	The future/time to come
24	yet	RR	1000	2000	at some future time/one day	89762	av	1338	Relative time

24	yet	RR	1300	1565	anew/again	89962	av	1352	Frequency
24	yet	RR	1606	2000	remaining in specified condition	90369	aj	1361	Lasting quality, permanence
24	yet	RR	1000	1563	copiously	102567	vt	1402	Letting/sending out
24	yet	RR	1000	1866	in/as in a stream	102568	vt	1402	Letting/sending out
24	yet	RR	1300	1565	for a second time, again	107332	av	1589	Two
24	yet	RR	1000	2000	in addition/more	111396	av	1640	Addition/supplementation
24	yet	RR	1000	2000	however, nevertheless, notwithstanding	123250	av	1915	Qualification, condition, reservation
24	yet	RR	1300	1450	emphasizing an extreme case	145595	av	2239	Statement, declaration
24	yet	RR	1000	2000	strengthening/emphasising comparative	145671	av	2239	Statement, declaration
24	yet	RR	1300	1884	used for emphasis after nor	145672	av	2239	Statement, declaration
24	yet	RR	1382	1570	melt	199063	vt	3415	Working with metal
24	yet	RR	1000	1533	found/cast (object)	199073	vt	3415	Working with metal
24	yet	RR	1387	1552	found/cast (metal)	199075	vt	3415	Working with metal
24	yet	RR	1387	1808	set/fasten with specific material	199097	vt	3415	Working with metal
25	the	AT	1788	1827	tea-party	43468	n	675	Feast
25	the	AT	1788	1827	tea-/coffee-party	212942	n	3727	Party
26	tried	JJ	1382	1382	sieved	43740	aj	680	Preparation of grain
26	tried	JJ	1627	1639	rendered (of fat)	43954	aj	684	Preparation for table/cooking
26	tried	JJ	1382	1639	cleared of refuse	56049	aj	950	Clearing of refuse matter
26	tried	JJ	1412	2000	tried/tested	79791	aj	1448	Trial/test/testing
26	tried	JJ	1724	2000	skilled/experienced	87392	aj	1535	Skill/skilfulness
26	tried	JJ	1412	2000	certified, verified	121201	aj	1863	Absence of doubt, confidence

26	tried	JJ	1412	2000	Tested	123103	aj	1913	Testing
26	tried	JJ	1400	1581	choice/excellent	124238	aj	1991	Superiority, excellence, perfection
26	tried	JJ	1400	1400	Famous/eminent person	126443	n	1945	Famous/eminent person
26	tried	JJ	1400	1611	refined	205371	aj	3557	Metal
27	intent	NN1	1340	1483	Endeavour	79594	n	1446	Endeavour
27	intent	NN1	1400	1400	attempt to obtain/attain	79656	vt	1446	Endeavour
27	intent	NN1	1650	1650	vigorous/intense in operation	84541	aj	1502	Vigour/energy
27	intent	NN1	1500	1500	see to/about	85186	vt	1509	Care/carefulness/attention
27	intent	NN1	1300	1623	The mind	115259	n	1696	The mind
27	intent	NN1	1300	1623	Intellect	115362	n	1699	Intellect
27	intent	NN1	1300	1623	state of mind	115503	n	1703	Character, disposition, mood
27	intent	NN1	1300	1623	Perception/cognition	116626	n	1728	Perception, cognition
27	intent	NN1	1460	1670	Topic, subject-matter	116810	n	1732	Topic, subject, concern
27	intent	NN1	1303	1676	Drift, tenor, purport	117845	n	1769	Meaning
27	intent	NN1	1320	1704	Attention	122237	n	1900	Attention
27	intent	NN1	1606	2000	of gaze, etc.	122278	aj	1900	Attention
27	intent	NN1	1610	2000	Intent	122310	aj	1901	Concentration
27	intent	NN1	1400	1611	be intent	122324	vi	1901	Concentration
27	intent	NN1	1613	1613	Accuse	123544	vt	1919	Accusation, charge
27	intent	NN1	1695	1695	Accuse	123544	vt	1919	Accusation, charge
27	intent	NN1	1225	1225	Wish/inclination	136526	n	2145	Wish/inclination
27	intent	NN1	1225	2000	intention/purpose	136856	n	2150	Intention
27	intent	NN1	1340	2000	end/purpose/object	136862	n	2150	Intention

27	intent	NN1	1300	1587	Intend	136927	vi	2150	Intention
27	intent	NN1	1300	1587	Intend	136933	vt	2150	Intention
27	intent	NN1	1386	1830	a plan	136957	n	2151	Planning
27	intent	NN1	1610	2000	Resolute/determined	137182	aj	2154	Resolution, determination
27	intent	NN1	1574	1767	construction put on something by the law	166021	n	2734	Jurisprudence
27	intent	NN1	1575	1575	one's case	167763	n	2764	Action of courts in claims/grievances
27	intent	NN1	1600	1737	Carry on/institute (an action)	167859	vt	2764	Action of courts in claims/grievances
28	Of	IO	1000	2000	Native	40759	p	398	Native people
28	Of	IO	1220	2000	Of/belonging to a thing as a quality	76258	p	1187	Intrinsicality/inherence
28	Of	IO	1200	2000	of/belonging to a thing as something related	76259	p	1187	Intrinsicality/inherence
28	Of	IO	1923	2000	characteristic of	76260	p	1187	Intrinsicality/inherence
28	Of	IO	1200	2000	characterized by	76261	p	1187	Intrinsicality/inherence
28	Of	IO	1450	2000	in the form of	76411	p	1191	Extrinsicality/externality
28	Of	IO	1000	1586	out of/from	76477	p	1193	State/condition
28	Of	IO	1382	2000	created by	76770	p	1197	Creation
28	Of	IO	1000	2000	because of	77370	p	1210	Cause/reason
28	Of	IO	1000	1894	from/out of	77452	p	1211	Source/origin
28	Of	IO	1000	1569	whence action is directed	77970	p	1421	Action/operation
28	Of	IO	1000	2000	indicating the agent/doer	78186	p	1423	Doing
28	Of	IO	1000	1824	By the instrumentality of	81713	p	1468	Use (made of things)
28	Of	IO	1369	1833	during	87890	p	1314	Time
28	Of	IO	1000	2000	at some time during	87891	p	1314	Time

28	Of	IO	1526	2000	of/belonging to a time	88204	p	1320	Particular time
28	Of	IO	1000	1625	from the beginning of a period	88259	p	1322	Period
28	Of	IO	1000	2000	Change	90072	p	1354	Change
28	Of	IO	1000	2000	Belonging to/localized in a place	93480	p	1256	Place
28	Of	IO	1000	2000	as deriving a title from it	93481	p	1256	Place
28	Of	IO	1000	1350	away from/out of	93874	p	1261	Absence
28	Of	IO	1000	1613	away from (denoting departure)	102924	p	1406	Going away
28	Of	IO	1000	2000	respecting/concerning	104478	p	1540	Relation/relationship
28	Of	IO	1225	2000	of/in respect of	104479	p	1540	Relation/relationship
28	Of	IO	1470	2000	in respect of being	104480	p	1540	Relation/relationship
28	Of	IO	1470	1820	in the person of	105733	p	1559	Individual character/quality
28	Of	IO	1440	2000	A member/part of	112080	p	1649	Part of whole
28	Of	IO	1000	2000	of which (something consists)	112081	p	1649	Part of whole
28	Of	IO	1000	2000	a portion of	112082	p	1649	Part of whole
28	Of	IO	1000	2000	in partitive expressions	112083	p	1649	Part of whole
28	Of	IO	1382	2000	Pre-eminent	124305	p	1991	Superiority, excellence, perfection
28	Of	IO	1200	2000	As one's possession	137954	p	2161	Possession/ownership
28	Of	IO	1000	2000	expressing origin of name	145291	p	2237	Naming
28	Of	IO	1000	2000	Expressing descent	146978	p	2269	Descendant
28	Of	IO	1000	2000	Related to as ruler	159201	p	2605	Ruler/governor
28	Of	IO	1000	2000	Of liberation	164621	p	2713	Liberation
29	such	DA	1000	2000	the same thing as mentioned before	104586	n	1544	Identity
29	such	DA	1000	2000	persons/things before mentioned	104588	n	1544	Identity

29	such	DA	1375	2000	the same as already mentioned	104599	aj	1544	Identity
29	such	DA	1000	2000	such persons/things	104871	n	1549	Similarity
29	such	DA	1823	2000	people of the same kind as	104875	n	1549	Similarity
29	such	DA	1200	2000	such	104897	aj	1549	Similarity
29	such	DA	1000	1509	so/in such a manner	104924	av	1549	Similarity
29	such	DA	1000	1390	such a thing	105375	n	1556	Kind/sort
29	such	DA	1000	2000	of this/that kind	105379	aj	1556	Kind/sort
29	such	DA	1297	2000	so many of that kind	105380	aj	1556	Kind/sort
29	such	DA	1000	2000	of the kind that	105381	aj	1556	Kind/sort
29	such	DA	1460	2000	and indescribable	105444	n	1557	Generality
29	such	DA	1000	2000	anyone	105454	n	1557	Generality
29	such	DA	1000	2000	so many/much	110833	aj	1633	Quantity
29	such	DA	1553	2000	as absolute intensive	110834	aj	1633	Quantity
29	such	DA	1420	2000	of that kind/degree	110867	aj	1633	Quantity
29	such	DA	1776	1776	to such an extent that	110875	av	1633	Quantity
30	a	AT1	1927	2000	specific	26281	n	260	Blood
30	a	AT1	1500	2000	of/in respect of	104479	p	1540	Relation/relationship
30	a	AT1	1305	1485	specific cry of grief	129379	in	2055	Expression of grief
30	a	AT1	1866	2000	signs and symbols	134388	n	1896	Logic
30	a	AT1	1350	2000	particular words in specific dialects	140927	vi	2203	Dialect
30	a	AT1	1889	2000	highest/lowest	149286	n	2327	Social class
30	a	AT1	1936	2000	member of	150045	n	2343	Specific classes of common people
30	a	AT1	1932	2000	designating international standard paper size	185296	aj	3187	Paper

30	a	AT1	1921	2000	road of specific class	191064	n	3302	Road
30	a	AT1	1609	2000	notes of specific scales	213556	n	3742	Pitch
31	truth	NN1	1977	2000	top/truth	72779	n	1110	Quark
31	truth	NN1	1380	2000	Reality/real existence/actuality	75958	n	1182	Reality/real existence/actuality
31	truth	NN1	1599	1844	reality/quality of being real	75959	n	1182	Reality/real existence/actuality
31	truth	NN1	1531	1774	the reality as opposed to what is apparent	75962	n	1182	Reality/real existence/actuality
31	truth	NN1	1552	1552	true nature	76296	n	1189	Character/nature
31	truth	NN1	1881	1881	adjust for accuracy	106503	vt	1569	Adaptation/adjustment
31	truth	NN1	1858	2000	Egyptian	114960	n	1690	Other deities
31	truth	NN1	1644	1843	what is true	118524	n	1792	Knowledge
31	truth	NN1	1380	2000	Self-evident truth, axiom	118867	n	1799	Saying, maxim, proverb
31	truth	NN1	1570	2000	Conformity with what is known, truth	119010	n	1806	Truth, validity, correctness
31	truth	NN1	1362	2000	personified	119012	n	1806	Truth, validity, correctness
31	truth	NN1	1380	2000	truth known by observation, fact	119013	n	1806	Truth, validity, correctness
31	truth	NN1	1340	2000	true facts/circumstances	119021	n	1806	Truth, validity, correctness
31	truth	NN1	1380	2000	reality	119023	n	1806	Truth, validity, correctness
31	truth	NN1	1534	1854	In truth	119054	in	1806	Truth, validity, correctness
31	truth	NN1	1638	1638	Describe truly	119092	vt	1806	Truth, validity, correctness
31	truth	NN1	1669	1862	of tools, materials, etc.	119118	n	1807	Accuracy, precision
31	truth	NN1	1400	2000	Truthfulness, veracity	119147	n	1808	Truthfulness, veracity
31	truth	NN1	1300	1300	Accept as true, believe	120427	vt	1842	Belief, opinion
31	truth	NN1	1300	1677	Belief, trust, confidence	120469	n	1844	Belief, trust, confidence

31	truth	NN1	1000	1700	(good) faith	143018	n	2225	Promise, pledge
31	truth	NN1	1275	1440	Betrothal	147387	n	2281	Betrothal
31	truth	NN1	1315	1315	Engage oneself to marry	147407	vi	2281	Betrothal
31	truth	NN1	1330	1412	Betroth	147411	vt	2281	Betrothal
31	truth	NN1	1000	2000	Faithfulness/trustworthiness	170326	n	2800	Faithfulness/trustworthiness
31	truth	NN1	1000	1860	Fidelity/loyalty	170338	n	2801	Fidelity/loyalty
31	truth	NN1	1400	1520	Faith	173497	n	2873	Faith
31	truth	NN1	1382	1611	Piety	174799	n	2926	Piety
31	truth	NN1	1868	2000	question and answer games	212237	n	3702	Parlour and party games
31	truth	NN1	1828	2000	accurate	215900	n	3805	Representation in art
32	as	CSA	1400	2000	because	77330	av	1210	Cause/reason
32	as	CSA	1230	1816	in which way	84454	av	1501	Manner of action
32	as	CSA	1220	1885	during/in the course of (a certain time)	87868	av	1314	Time
32	as	CSA	1297	1420	at the place which	93578	av	1258	Here/there, etc.
32	as	CSA	1200	2000	so/in such a manner	104924	av	1549	Similarity
32	as	CSA	1000	1800	as if/as though	104926	av	1549	Similarity
32	as	CSA	1225	2000	as/like	104978	cj	1549	Similarity
32	as	CSA	1340	1705	For instance	105782	av	1560	An individual case/instance
32	as	CSA	1460	1824	Linking comparatives	122963	cj	1911	Comparison, contrast
32	as	CSA	1601	2000	ancient Roman	207914	n	3615	Coins collectively
32	as	CSA	1847	2000	in the character of	221556	av	3913	Acting
33	I	PPIS1	1953	1962	symbol of quantum number of	72531	n	1106	Particle physics

33	I	PPIS1	1000	2000	I	105656	n	1559	Individual character/quality
33	I	PPIS1	1599	2000	as a word	105657	n	1559	Individual character/quality
33	I	PPIS1	1265	1641	namely/that is to say	105960	av	1562	The quality of being specific
33	I	PPIS1	1450	2000	symbol denoting	107119	n	1586	One
33	I	PPIS1	1710	2000	subjective being, self	115447	n	1702	Self-consciousness
33	I	PPIS1	1552	2000	signs and symbols	134388	n	1896	Logic
33	I	PPIS1	1946	2000	character assumed by author	220584	n	3889	Fiction
34	have	VH0	1382	2000	give birth	8445	vt	143	Birth
34	have	VH0	1000	2000	Apprehend by sensuous perception	56702	vt	821	Physical sensation
34	have	VH0	1594	2000	Have sexual intercourse with	57697	vt	848	Sexual intercourse
34	have	VH0	1205	2000	bring (a person/thing) into a state/condition	77287	vt	1209	Cause
34	have	VH0	1390	2000	cause to be done (to someone/something)	77312	vt	1209	Cause
34	have	VH0	1000	2000	be affected by some action	78057	vt	1422	Operation upon something
34	have	VH0	1000	2000	keep up (a proceeding/performance)	78831	vt	1432	Continuing
34	have	VH0	1175	1175	give effect to/show in action	78998	vt	1434	Carrying out
34	have	VH0	1386	1556	Behave oneself	85465	vr	1513	Behaviour
34	have	VH0	1000	2000	Relate to	104452	vt	1540	Relation/relationship
34	have	VH0	1000	2000	contain as a constituent part	112176	vt	1651	Mutual relation of parts to whole
34	have	VH0	1000	2000	Have in the mind	115295	vt	1697	Mental capacity

34	have	VH0	1591	2000	Understand	117117	vt	1743	Understanding
34	have	VH0	1591	2000	Have knowledge, know	118568	vt	1792	Knowledge
34	have	VH0	1000	2000	Experience	118592	vt	1793	Experience
34	have	VH0	1805	2000	Deceive	119595	vt	1817	Deceit, deception, trickery
34	have	VH0	1816	2000	nonplus	120978	vt	1857	Perplexity, bewilderment
34	have	VH0	1816	2000	Refute, disprove	123602	vt	1920	Disproof
34	have	VH0	1000	1728	consider to be, account as	123749	vt	1922	Evaluation, estimation, appraisal
34	have	VH0	1000	2000	Hold/entertain/cherish (a feeling)	127642	vt	2024	Emotions, mood
34	have	VH0	1000	2000	entertain (an intention)	136934	vt	2150	Intention
34	have	VH0	1200	1605	Possession	137889	n	2161	Possession/ownership
34	have	VH0	1000	2000	Have/possess	137926	vt	2161	Possession/ownership
34	have	VH0	1000	2000	possess a condition/position	137931	vt	2161	Possession/ownership
34	have	VH0	1836	2000	Possessor	138027	n	2162	Owning
34	have	VH0	1836	2000	well-off person/people	138336	n	2165	Wealth
34	have	VH0	1000	2000	Acquire	138697	vt	2169	Acquisition
34	have	VH0	1000	2000	Retain	138997	vt	2174	Retaining
34	have	VH0	1377	2000	take this	139666	vi	2184	Offering
34	have	VH0	1000	2000	Take	139707	vt	2185	Taking
34	have	VH0	1449	2000	Express in phrases	145098	vt	2235	Words and phrases
34	have	VH0	1449	2000	Maintain/uphold as true	145659	vt	2239	Statement, declaration
34	have	VH0	1000	2000	hold relationship with	146359	vt	2247	Kinship/relationship
34	have	VH0	1596	2000	have authority over	158002	vt	2580	Authority
34	have	VH0	1000	2000	Have/had duty/obligation	170037	vi	2791	Duty/obligation

34	have	VH0	1000	2000	have (a duty)	170043	vt	2791	Duty/obligation
34	have	VH0	1175	2000	by one's action/behaviour	181533	vt	3093	Manifestation
34	have	VH0	1420	1849	Travel/proceed/make one's way	188558	vi	3266	Travel and travelling
35	meant	VVN	1841	2000	bring about as a consequence/entail	77303	vt	1209	Cause
35	meant	VVN	1432	1565	Intercession/influence on someone's behalf	80946	n	1462	Intercession/influence on someone's behalf
35	meant	VVN	1455	1606	one who	80947	n	1462	Intercession/influence on someone's behalf
35	meant	VVN	1449	1654	bring about by mediation	80961	vt	1462	Intercession/influence on someone's behalf
35	meant	VVN	1347	2000	(a) means	81626	n	1468	Use (made of things)
35	meant	VVN	1374	1612	person as	81668	n	1468	Use (made of things)
35	meant	VVN	1377	1615	acting as intermediate agent	81677	aj	1468	Use (made of things)
35	meant	VVN	1439	1707	specifically of person	81678	aj	1468	Use (made of things)
35	meant	VVN	1449	1654	be intermediate means in	81705	vt	1468	Use (made of things)
35	meant	VVN	1590	1613	An opportunity	81809	n	1472	An opportunity
35	meant	VVN	1545	1718	Restrained/moderate behaviour	86310	n	1520	Restrained/moderate behaviour
35	meant	VVN	1425	1425	Restrained/moderate	86318	aj	1520	Restrained/moderate behaviour
35	meant	VVN	1848	2000	of a horse, etc.	86526	aj	1523	Unkindness
35	meant	VVN	1920	2000	skilful/adroit	87412	aj	1535	Skill/skilfulness
35	meant	VVN	1387	1738	unable/incompetent/ineffectual	87621	aj	1537	Inability
35	meant	VVN	1464	1772	intervening	87862	aj	1314	Time
35	meant	VVN	1439	1707	at a time between two dates	87863	aj	1314	Time
35	meant	VVN	1548	1642	in the midst/middle of a period of time	87869	av	1314	Time

35	meant	VVN	1420	1688	middle/centre	94504	n	1269	Central condition/position
35	meant	VVN	1435	1541	situated in the centre/middle	94538	aj	1269	Central condition/position
35	meant	VVN	1340	1593	that which is interjacent	96193	n	1282	Condition/fact of being interjacent
35	meant	VVN	1541	1541	Interjacent	96207	aj	1282	Condition/fact of being interjacent
35	meant	VVN	1633	1633	aim at	98644	vt	1312	Direction
35	meant	VVN	1374	2000	mean	105302	n	1554	Condition of being mean/average
35	meant	VVN	1803	2000	average	105304	n	1554	Condition of being mean/average
35	meant	VVN	1340	2000	Mean	105307	aj	1554	Condition of being mean/average
35	meant	VVN	1374	1697	average	105308	aj	1554	Condition of being mean/average
35	meant	VVN	1340	1822	Middle	106995	aj	1579	Order/sequence/succession
35	meant	VVN	1571	2000	mean	108416	n	1607	Arithmetic/algebraic operations
35	meant	VVN	1391	2000	mean	108444	aj	1607	Arithmetic/algebraic operations
35	meant	VVN	1882	2000	calculate mean	108476	vt	1607	Arithmetic/algebraic operations
35	meant	VVN	1571	2000	measures of central tendency	109077	n	1617	Probability/statistics
35	meant	VVN	1420	1679	having some quality in a moderate degree	111449	aj	1633	Quantity
35	meant	VVN	1535	1612	To a moderate extent/degree	111451	av	1633	Quantity
35	meant	VVN	1599	2000	small/trifling in amount/degree	111502	aj	1641	Smallness of quantity/amount/degree
35	meant	VVN	1398	1398	comparatively less	111530	av	1641	Smallness of quantity/amount/degree
35	meant	VVN	1362	1597	mediator between God and man	114460	n	1683	Christian God
35	meant	VVN	1000	2000	Mean	117780	vt	1769	Meaning
35	meant	VVN	1000	2000	Mean, signify, express	117816	vt	1769	Meaning
35	meant	VVN	1513	2000	make reference to	117818	vt	1769	Meaning

35	meant	VVN	1300	1513	Have in one's mind, remember	118222	vi	1784	Memory, keeping in mind
35	meant	VVN	1303	1440	Call to mind, recollect	118225	vt	1784	Memory, keeping in mind
35	meant	VVN	1300	1637	Hold an opinion, opine	120563	vi	1847	Expressed belief, opinion
35	meant	VVN	1225	1250	Accuse	123542	vi	1919	Accusation, charge
35	meant	VVN	1561	1561	in a middle way	123965	av	1926	Absence of prejudice
35	meant	VVN	1460	1628	Mediocre	124109	aj	1989	Mediocrity
35	meant	VVN	1377	1770	Inferior thing	124692	aj	1998	Inferiority/baseness
35	meant	VVN	1387	1738	in ability	124730	aj	1998	Inferiority/baseness
35	meant	VVN	1817	2000	Wretched	124857	aj	2001	Wretchedness/baseness
35	meant	VVN	1626	1861	Basely	124868	av	2001	Wretchedness/baseness
35	meant	VVN	1848	2000	Corrupt	125094	aj	2011	Corruption
35	meant	VVN	1888	2000	Be important	125644	vt	1928	Importance, influence
35	meant	VVN	1585	1807	of little importance/trivial	125768	aj	1929	Unimportance, triviality
35	meant	VVN	1610	1823	low/subordinate	125784	aj	1929	Unimportance, triviality
35	meant	VVN	1600	2000	paltry/mean/contemptible	125811	aj	1929	Unimportance, triviality
35	meant	VVN	1719	1719	in a paltry/mean/contemptible manner	125831	av	1929	Unimportance, triviality
35	meant	VVN	1610	2000	Base/ignoble	127455	aj	1970	Ignobleness/baseness
35	meant	VVN	1665	2000	specifically of character	127456	aj	1970	Ignobleness/baseness
35	meant	VVN	1626	1626	Basely/ignobly	127460	av	1970	Ignobleness/baseness
35	meant	VVN	1300	2000	instance/act of lamenting	129303	n	2055	Expression of grief
35	meant	VVN	1000	1800	Lament/express grief	129323	vi	2055	Expression of grief
35	meant	VVN	1000	2000	Lament/express grief for	129328	vt	2055	Expression of grief
35	meant	VVN	1175	1790	Lament/express grief	129334	vr	2055	Expression of grief

35	meant	VVN	1513	1599	Lament the death of	129345	vt	2055	Expression of grief
35	meant	VVN	1440	1603	Feel pity for	131893	vt	2100	Compassion
35	meant	VVN	1605	1605	middle term	134501	n	1896	Logic
35	meant	VVN	1610	1823	undignified	134884	aj	1976	Bad taste
35	meant	VVN	1841	2000	Necessitate	136374	vt	2143	Necessity
35	meant	VVN	1000	2000	Intend	136933	vt	2150	Intention
35	meant	VVN	1400	2000	intend/be intended for a purpose	136942	vt	2150	Intention
35	meant	VVN	1908	2000	mean what one says	137199	vt	2154	Resolution, determination
35	meant	VVN	1362	1776	poor	138479	aj	2166	Poverty
35	meant	VVN	1200	2000	shared	138594	aj	2168	Sharing
35	meant	VVN	1938	1938	niggard/mean person	139051	n	2175	Frugality, meanness
35	meant	VVN	1860	2000	Niggardly/mean	139057	aj	2175	Frugality, meanness
35	meant	VVN	1530	1530	reflexive verb	143367	n	2228	Part of speech
35	meant	VVN	1530	1583	middle	143993	aj	2231	Grammatical categories
35	meant	VVN	1300	2000	Of low rank/condition	150151	aj	2347	Low rank/condition
35	meant	VVN	1440	1440	mediate between	150704	vt	2361	Bringing about concord/peace
35	meant	VVN	1535	1670	intermediate/intervening	159542	aj	2615	Lord
35	meant	VVN	1665	2000	Not magnanimous	171486	aj	2829	Lack of magnanimity/noble-mindedness
35	meant	VVN	1626	1626	In a manner lacking magnanimity	171496	av	2829	Lack of magnanimity/noble-mindedness
35	meant	VVN	1300	1625	Give information	183464	vi	3138	Action of informing
35	meant	VVN	1000	1494	inform (a person)	183476	vt	3138	Action of informing
35	meant	VVN	1400	1706	be bound for/head for	188970	vt	3270	Travel in specific course/direction

35	meant	VVN	1330	1698	middle parts	213728	n	3745	Harmony/sounds in combination
35	meant	VVN	1597	1721	middle parts	213751	aj	3745	Harmony/sounds in combination
35	meant	VVN	1330	1698	performer of middle part	215017	n	3772	Musician
35	meant	VVN	1330	1698	instrument for middle part	215257	n	3784	Musical instrument
35	meant	VVN	1879	2000	specific strings	215418	n	3791	Lute-/viol-type parts
35	meant	VVN	1400	1586	Plain/simple	218899	aj	3865	Plainness
35	meant	VVN	1610	2000	low in style	219116	aj	3868	Inelegance
36	;	YSC OL							

Table A7 CLAWS stop-list filter

Tag	Description
APPGE	possessive pronoun, pre-nominal (e.g. my, your, our)
AT	article (e.g. the, no)
AT1	singular article (e.g. a, an, every)
BCL	before-clause marker (e.g. in order (that),in order (to))
CC	coordinating conjunction (e.g. and, or)
CCB	adversative coordinating conjunction (but)
CS	subordinating conjunction (e.g. if, because, unless, so, for)
CSA	as (as conjunction)
CSN	than (as conjunction)
CST	that (as conjunction)
CSW	whether (as conjunction)
DA	after-determiner or post-determiner capable of pronominal function (e.g. such, former, same)
DA1	singular after-determiner (e.g. little, much)
DA2	plural after-determiner (e.g. few, several, many)
DAR	comparative after-determiner (e.g. more, less, fewer)
DAT	superlative after-determiner (e.g. most, least, fewest)
DB	before determiner or pre-determiner capable of pronominal function (all, half)
DB2	plural before-determiner (both)
DD	determiner (capable of pronominal function) (e.g any, some)
DD1	singular determiner (e.g. this, that, another)
DD2	plural determiner (these, those)
DDQ	wh-determiner (which, what)
DDQGE	wh-determiner, genitive (whose)
DDQV	wh-ever determiner, (whichever, whatever)
EX	existential there
FO	formula
FU	unclassified word
FW	foreign word
GE	germanic genitive marker - (' or's)
IF	for (as preposition)
II	general preposition
IO	of (as preposition)
IW	with, without (as prepositions)
MC	cardinal number, neutral for number (two, three...)
MC1	singular cardinal number (one)

MC2	plural cardinal number (e.g. sixes, sevens)
MCGE	genitive cardinal number, neutral for number (two's, 100's)
MCMC	hyphenated number (40-50, 1770-1827)
MD	ordinal number (e.g. first, second, next, last)
MF	fraction, neutral for number (e.g. quarters, two-thirds)
PN	indefinite pronoun, neutral for number (none)
PN1	indefinite pronoun, singular (e.g. anyone, everything, nobody, one)
PNQO	objective wh-pronoun (whom)
PNQS	subjective wh-pronoun (who)
PNQV	wh-ever pronoun (whoever)
PNX1	reflexive indefinite pronoun (oneself)
PPGE	nominal possessive personal pronoun (e.g. mine, yours)
PPH1	3rd person sing. neuter personal pronoun (it)
PPHO1	3rd person sing. objective personal pronoun (him, her)
PPHO2	3rd person plural objective personal pronoun (them)
PPHS1	3rd person sing. subjective personal pronoun (he, she)
PPHS2	3rd person plural subjective personal pronoun (they)
PPIO1	1st person sing. objective personal pronoun (me)
PPIO2	1st person plural objective personal pronoun (us)
PPIS1	1st person sing. subjective personal pronoun (I)
PPIS2	1st person plural subjective personal pronoun (we)
PPX1	singular reflexive personal pronoun (e.g. yourself, itself)
PPX2	plural reflexive personal pronoun (e.g. yourselves, themselves)
PPY	2nd person personal pronoun (you)
TO	infinitive marker (to)
UH	interjection (e.g. oh, yes, um)
VB0	be, base form (finite i.e. imperative, subjunctive)
VBDR	were
VBDZ	was
VBG	being
VBI	be, infinitive (To be or not... It will be ...)
VBM	am
VBN	been
VBR	are
VBZ	is
VD0	do, base form (finite)
VDD	did
VDG	doing
VDI	do, infinitive (I may do... To do...)
VDN	done

VDZ	does
VH0	have, base form (finite)
VHD	had (past tense)
VHG	having
VHI	have, infinitive
VHN	had (past participle)
VHZ	has
VM	modal auxiliary (can, will, would, etc.)
VMK	modal catenative (ought, used)
XX	not, n't
ZZ1	singular letter of the alphabet (e.g. A,b)
ZZ2	plural letter of the alphabet (e.g. A's, b's)

Table A8 Post-lexical filter output

corp_id	corp_word	corp_pos	apps	appe	heading	pos	thematicheading	tid
22	FORGET	VV0	1382	2000	forget, fail to remember	vi	Faulty recollection, forgetting	1437
22	FORGET	VV0	1400	1670	cease the practice/observation of something	vt	Ceasing	3270
22	FORGET	VV0	1300	2000	leave behind	vt	Faulty recollection, forgetting	1441
22	FORGET	VV0	1300	2000	leave undone/fail to perform/carry out	vt	Not doing	1789
22	FORGET	VV0	1582	1582	be/get lost	vr	Travel in specific course/direction	1789
22	FORGET	VV0	1297	1797	Be careless/heedless of	vt	Carelessness	1511
22	FORGET	VV0	1538	2000	overlook inadvertently	vt	Carelessness	1511
24	yet	RR	1513	1536	cause to flow in flood	vt	Flood/flooding	1361
24	yet	RR	1606	2000	remaining in specified condition	aj	Lasting quality, permanence	1316
24	yet	RR	1300	1565	for a second time, again	av	Two	1352
24	yet	RR	1560	1560	into a cavity	vt	Action/fact of pouring/being poured	38
24	yet	RR	1300	1450	emphasizing an extreme case	av	Statement, declaration	1343
24	yet	RR	1374	1501	Shine	vt	Light	1126
24	yet	RR	1300	1884	used for emphasis after nor	av	Statement, declaration	1589
24	yet	RR	1606	2000	still continuing/enduring	aj	Duration	2239
24	yet	RR	1382	1570	melt	vt	Working with metal	2239
24	yet	RR	1250	2000	in advance/beforehand	av	Antecedence/being earlier	1041

24	yet	RR	1387	1552	found/cast (metal)	vt	Working with metal	3415
24	yet	RR	1387	1808	set/fasten with specific material	vt	Working with metal	3415
24	yet	RR	1300	1565	anew/again	av	Frequency	3415
26	tried	JJ	1400	1581	choice/excellent	aj	Superiority, excellence, perfection	1991
26	tried	JJ	1627	1639	rendered (of fat)	aj	Preparation for table/cooking	1863
26	tried	JJ	1400	1611	refined	aj	Metal	1448
26	tried	JJ	1382	1639	cleared of refuse	aj	Clearing of refuse matter	684
26	tried	JJ	1412	2000	tried/tested	aj	Trial/test/testing	1913
26	tried	JJ	1412	2000	certified, verified	aj	Absence of doubt, confidence	3557
26	tried	JJ	1412	2000	Tested	aj	Testing	950
27	intent	NN1	1300	1623	Intellect	n	Intellect	1703
27	intent	NN1	1400	1611	be intent	vi	Concentration	1900
27	intent	NN1	1386	1830	a plan	n	Planning	1900
27	intent	NN1	1300	1623	state of mind	n	Character, disposition, mood	1728
27	intent	NN1	1613	1613	Accuse	vt	Accusation, charge	1769
27	intent	NN1	1610	2000	Resolute/determined	aj	Resolution, determination	1919
27	intent	NN1	1300	1623	Perception/cognition	n	Perception, cognition	1919
27	intent	NN1	1695	1695	Accuse	vt	Accusation, charge	1446
27	intent	NN1	1574	1767	construction put on something by the law	n	Jurisprudence	2150
27	intent	NN1	1340	1483	Endeavour	n	Endeavour	2150
27	intent	NN1	1460	1670	Topic, subject-matter	n	Topic, subject, concern	2150
27	intent	NN1	1575	1575	one's case	n	Action of courts in claims/grievances	2150

27	intent	NN1	1303	1676	Drift, tenor, purport	n	Meaning	1509
27	intent	NN1	1225	2000	intention/purpose	n	Intention	1502
27	intent	NN1	1600	1737	Carry on/institute (an action)	vt	Action of courts in claims/grievances	2151
27	intent	NN1	1650	1650	vigorous/intense in operation	aj	Vigour/energy	2154
27	intent	NN1	1320	1704	Attention	n	Attention	1732
27	intent	NN1	1340	2000	end/purpose/object	n	Intention	1696
27	intent	NN1	1500	1500	see to/about	vt	Care/carefulness/attention	2764
27	intent	NN1	1606	2000	of gaze, etc.	aj	Attention	2764
27	intent	NN1	1300	1587	Intend	vi	Intention	1901
27	intent	NN1	1300	1623	The mind	n	The mind	1901
27	intent	NN1	1610	2000	Intent	aj	Concentration	2734
27	intent	NN1	1300	1587	Intend	vt	Intention	1699

Table A9 1_4 sample group CG1

ID	Word	POS	apps	appe	heading	catid	Cat pos	tid	thematicheading	MV	RMV
22	FORGET	VV0	1382	2000	forget, fail to remember	118408	vi	1789	Faulty recollection, forgetting	39	92.86
22	FORGET	VV0	1000	2000	forget, cease to know	118411	vt	1789	Faulty recollection, forgetting	39	92.86
22	FORGET	VV0	1300	2000	leave behind	118414	vt	1789	Faulty recollection, forgetting	39	92.86
23	not	XX									
24	yet	RR	1606	2000	remaining in specified condition	90369	aj	1361	Lasting quality, permanence	11	42.31
24	yet	RR	1513	1536	cause to flow in flood	2205	vt	38	Flood/flooding	9	29.03
24	yet	RR	1000	2000	as formerly/still/to this day	89116	av	1340	The present (time)	18	26.47
24	yet	RR	1000	1879	even now (though not until now)	89117	av	1340	The present (time)	18	26.47
24	yet	RR	1606	2000	still continuing/enduring	87971	aj	1316	Duration	30	25.42
25	the	AT									
26	tried	JJ	1412	2000	Tested	123103	aj	1913	Testing	1	7.69
26	tried	JJ	1412	2000	certified, verified	121201	aj	1863	Absence of doubt, confidence	6	5.41
27	intent	NN1	1300	1623	The mind	115259	n	1696	The mind	3	20
27	intent	NN1	1303	1676	Drift, tenor, purport	117845	n	1769	Meaning	11	13.75
27	intent	NN1	1610	2000	Intent	122310	aj	1901	Concentration	4	11.76
27	intent	NN1	1400	1611	be intent	122324	vi	1901	Concentration	4	11.76

27	intent	NN1	1225	2000	intention/purpose	136856	n	2150	Intention	15	11.11
27	intent	NN1	1340	2000	end/purpose/object	136862	n	2150	Intention	15	11.11
27	intent	NN1	1300	1587	Intend	136927	vi	2150	Intention	15	11.11
27	intent	NN1	1300	1587	Intend	136933	vt	2150	Intention	15	11.11
28	Of	IO									
29	such	DA									
30	a	AT1									
31	truth	NN1	1000	2000	Faithfulness/trustworthiness	170326	n	2800	Faithfulness/trustworthiness	5	41.67
31	truth	NN1	1570	2000	Conformity with what is known, truth	119010	n	1806	Truth, validity, correctness	14	13.73
31	truth	NN1	1362	2000	personified	119012	n	1806	Truth, validity, correctness	14	13.73
31	truth	NN1	1380	2000	truth known by observation, fact	119013	n	1806	Truth, validity, correctness	14	13.73
31	truth	NN1	1340	2000	true facts/circumstances	119021	n	1806	Truth, validity, correctness	14	13.73
31	truth	NN1	1380	2000	reality	119023	n	1806	Truth, validity, correctness	14	13.73
31	truth	NN1	1534	1854	In truth	119054	in	1806	Truth, validity, correctness	14	13.73
31	truth	NN1	1638	1638	Describe truly	119092	vt	1806	Truth, validity, correctness	14	13.73
32	as	CSA									
33	I	PPIS1									
34	have	VH0									
35	meant	VVN	1610	2000	Base/ignoble	127455	aj	1970	Ignobleness/baseness	6	46.15

35	meant	VVN	1665	2000	specifically of character	127456	aj	1970	Ignobleness/baseness	6	46.15
35	meant	VVN	1626	1626	Basely/ignobly	127460	av	1970	Ignobleness/baseness	6	46.15
35	meant	VVN	1432	1565	Intercession/influence on someone's behalf	80946	n	1462	Intercession/influence on someone's behalf	6	37.5
35	meant	VVN	1455	1606	one who	80947	n	1462	Intercession/influence on someone's behalf	6	37.5
35	meant	VVN	1449	1654	bring about by mediation	80961	vt	1462	Intercession/influence on someone's behalf	6	37.5
35	meant	VVN	1374	2000	mean	105302	n	1554	Condition of being mean/average	6	27.27
35	meant	VVN	1340	2000	Mean	105307	aj	1554	Condition of being mean/average	6	27.27
35	meant	VVN	1374	1697	average	105308	aj	1554	Condition of being mean/average	6	27.27
35	meant	VVN	1000	2000	Mean	117780	vt	1769	Meaning	11	13.75
35	meant	VVN	1000	2000	Mean, signify, express	117816	vt	1769	Meaning	11	13.75
35	meant	VVN	1513	2000	make reference to	117818	vt	1769	Meaning	11	13.75
35	meant	VVN	1400	1706	be bound for/head for	188970	vt	3270	Travel in specific course/direction	17	13.6
35	meant	VVN	1300	1513	Have in one's mind, remember	118222	vi	1784	Memory, keeping in mind	8	12.31
35	meant	VVN	1300	2000	instance/act of lamenting	129303	n	2055	Expression of grief	11	11.7
35	meant	VVN	1000	1800	Lament/express grief	129323	vi	2055	Expression of grief	11	11.7
35	meant	VVN	1000	2000	Lament/express grief for	129328	vt	2055	Expression of grief	11	11.7
35	meant	VVN	1175	1790	Lament/express grief	129334	vr	2055	Expression of grief	11	11.7
35	meant	VVN	1513	1599	Lament the death of	129345	vt	2055	Expression of grief	11	11.7

Table A10 Summary result [34]

ID	Word	Position_1	Position_2	Position_3	Position_4	Position_5	Position_6
1	34						
2	.						
3	S_END						
4	S_BEGIN						
5	Forget	Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00	Faulty recollection, forgetting; [forget, cease to know; vt; 1000-2000]; †1789; [»118411]; #92.86/0.00	Faulty recollection, forgetting; [leave behind; vt; 1300-2000]; †1789; [»118414]; #92.86/0.00			
6	not						
7	yet	Lasting quality, permanence; [remaining in specified condition; aj; 1606-2000]; †1361; [»90369]; #42.31/0.00	Flood/flooding; [cause to flow in flood; vt; 1513-1536]; †38; [»2205]; #29.03/0.04	The present (time); [as formerly/still/to this day; av; 1000-2000]; †1340; [»89116]; #26.47/0.09	The present (time); [even now (though not until now); av; 1000-1879]; †1340; [»89117]; #26.47/0.09	Duration; [still continuing/enduring; aj; 1606-2000]; †1316; [»87971]; #25.42/0.17	
8	The						
9	Lover	Liking; [one who likes/favours; n; 1340-2000]; †2080; [»130460]; #9.30/0.00	A lover; [One who loves/a lover; n; 1225-2000]; †2089; [»131049]; #7.84/0.14				

10	Beseeches	Requesting; [earnest request/entreaty; n; 1606-1625]; †2222; [»142758]; #8.76/0.00	Requesting; [beseech/implore; vi; 1225-1655]; †2222; [»142814]; #8.76/0.00	Requesting; [Request/ask for; vt; 1175-2000]; †2222; [»142823]; #8.76/0.00	Requesting; [beseech/implore; vt; 1175-2000]; †2222; [»142828]; #8.76/0.00	Requesting; [a person a thing; vt; 1205-1588]; †2222; [»142833]; #8.76/0.00	Requesting; [a person of a thing; vt; 1300-1604]; †2222; [»142834]; #8.76/0.00
11	his						
12	Mistress	A lover; [specifically a female sweetheart/girlfriend; n; 1509-2000]; †2089; [»131070]; #7.84/0.00	Unchastity; [a mistress; n; 1430-2000]; †2839; [»171934]; #5.00/0.04	Ruler/governor; [thing personified as; n; 1369-1677]; †2605; [»159142]; #4.69/0.08	Ruler/governor; [female; n; 1366-1785]; †2605; [»159144]; #4.69/0.08	Ruler/governor; [person/state ruling over another; n; 1375-2000]; †2605; [»159191]; #4.69/0.08	
13	not						
14	to						
15	Forget	Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00	Faulty recollection, forgetting; [forget, cease to know; vt; 1000-2000]; †1789; [»118411]; #92.86/0.00	Faulty recollection, forgetting; [leave behind; vt; 1300-2000]; †1789; [»118414]; #92.86/0.00			
16	his						
17	Steadfast	Constancy; [Constant/steadfast; aj; 1000-2000]; †2155; [»137210]; #14.29/0.00	Constancy; [Constantly/steadfastly; av; 1300-1887]; †2155; [»137215]; #14.29/0.00	A look/glance; [fixed/steady; aj; 1300-1817]; †884; [»58902]; #5.41/0.12	Warrior; [qualities/attributes; aj; 1000-1821]; †2493; [»155199]; #5.41/0.12		

18	Faith	Faithfulness/trustworthiness; [Faithfulness/trustworthiness; n; 1250-1863]; †2800; [»170326]; #41.67/0.00	Truth, validity, correctness; [In truth; in; 1586-1840]; †1806; [»119054]; #13.73/0.08	Belief, trust, confidence; [Belief, trust, confidence; n; 1300-2000]; †1844; [»120469]; #13.64/0.15			
19	and						
20	TRUE	Faithfulness/trustworthiness; [person; n; 1400-1470]; †2800; [»170330]; #41.67/0.00	Faithfulness/trustworthiness; [Faithful/trustworthy; aj; 1000-2000]; †2800; [»170332]; #41.67/0.00	Truth, validity, correctness; [In conformity with truth, true; aj; 1205-2000]; †1806; [»119029]; #13.73/0.12	Truth, validity, correctness; [In accordance with truth, truly; av; 1300-1883]; †1806; [»119041]; #13.73/0.12	Truth, validity, correctness; [genuine, real; aj; 1398-2000]; †1806; [»119071]; #13.73/0.12	Truth, validity, correctness; [of blood, breeding; aj; 1578-2000]; †1806; [»119074]; #13.73/0.12
21	Intent	The mind; [The mind; n; 1300-1623]; †1696; [»115259]; #20.00/0.00	Meaning; [Drift, tenor, purport; n; 1303-1676]; †1769; [»117845]; #13.75/0.05	Concentration; [Intent; aj; 1610-2000]; †1901; [»122310]; #11.76/0.09	Concentration; [be intent; vi; 1400-1611]; †1901; [»122324]; #11.76/0.09	Intention; [intention/purpose; n; 1225-2000]; †2150; [»136856]; #11.11/0.18	Intention; [end/purpose/object; n; 1340-2000]; †2150; [»136862]; #11.11/0.18
22	FORGET	Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00	Faulty recollection, forgetting; [forget, cease to know; vt; 1000-2000]; †1789; [»118411]; #92.86/0.00	Faulty recollection, forgetting; [leave behind; vt; 1300-2000]; †1789; [»118414]; #92.86/0.00			
23	not						

24	yet	Lasting quality, permanence; [remaining in specified condition; aj; 1606-2000]; †1361; [»90369]; #42.31/0.00	Flood/flooding; [cause to flow in flood; vt; 1513-1536]; †38; [»2205]; #29.03/0.04	The present (time); [as formerly/still/to this day; av; 1000-2000]; †1340; [»89116]; #26.47/0.09	The present (time); [even now (though not until now); av; 1000-1879]; †1340; [»89117]; #26.47/0.09	Duration; [still continuing/enduring; aj; 1606-2000]; †1316; [»87971]; #25.42/0.17	
25	the						
26	tried	Testing; [Tested; aj; 1412-2000]; †1913; [»123103]; #7.69/0.00	Absence of doubt, confidence; [certified, verified; aj; 1412-2000]; †1863; [»121201]; #5.41/0.17				
27	intent	The mind; [The mind; n; 1300-1623]; †1696; [»115259]; #20.00/0.00	Meaning; [Drift, tenor, purport; n; 1303-1676]; †1769; [»117845]; #13.75/0.05	Concentration; [Intent; aj; 1610-2000]; †1901; [»122310]; #11.76/0.09	Concentration; [be intent; vi; 1400-1611]; †1901; [»122324]; #11.76/0.09	Intention; [intention/purpose; n; 1225-2000]; †2150; [»136856]; #11.11/0.18	Intention; [end/purpose/object; n; 1340-2000]; †2150; [»136862]; #11.11/0.18
28	Of						
29	such						
30	a						
31	truth	Faithfulness/trustworthiness; [Faithfulness/trustworthiness; n; 1000-2000]; †2800; [»170326]; #41.67/0.00	Truth, validity, correctness; [Conformity with what is known, truth; n; 1570-2000]; †1806; [»119010]; #13.73/0.05	Truth, validity, correctness; [personified; n; 1362-2000]; †1806; [»119012]; #13.73/0.05	Truth, validity, correctness; [truth known by observation, fact; n; 1380-2000]; †1806; [»119013]; #13.73/0.05	Truth, validity, correctness; [true facts/circumstances; n; 1340-2000]; †1806; [»119021]; #13.73/0.05	Truth, validity, correctness; [reality; n; 1380-2000]; †1806; [»119023]; #13.73/0.05
32	as						

33	I						
34	have						
35	meant	Ignobleness/baseness; [Base/ignoble; aj; 1610-2000]; †1970; [»127455]; #46.15/0.00	Ignobleness/baseness; [specifically of character; aj; 1665-2000]; †1970; [»127456]; #46.15/0.00	Ignobleness/baseness; [Basely/ignobly; av; 1626-1626]; †1970; [»127460]; #46.15/0.00	Intercession/influence on someone's behalf; [Intercession/influence on someone's behalf; n; 1432-1565]; †1462; [»80946]; #37.50/0.04	Intercession/influence on someone's behalf; [one who; n; 1455-1606]; †1462; [»80947]; #37.50/0.04	Intercession/influence on someone's behalf; [bring about by mediation; vt; 1449-1654]; †1462; [»80961]; #37.50/0.04
36	;						
37	My						
38	great	High reputation, honour; [Majestically/exaltedly; av; 1698-1698]; †1946; [»126513]; #15.38/0.00	Proper pride/self-respect; [dignified; aj; 1585-1697]; †2108; [»132209]; #15.38/0.00	Meaning; [gist; n; 1369-1450]; †1769; [»117846]; #13.75/0.06	Greatness of quantity/amount/degree; [(a) great quantity/amount; n; 1557-1557]; †1634; [»110900]; #10.00/0.09	Greatness of quantity/amount/degree; [Great in quantity/amount/degree; aj; 1000-2000]; †1634; [»110919]; #10.00/0.09	Greatness of quantity/amount/degree; [great (of quantity/amount); aj; 1000-2000]; †1634; [»110920]; #10.00/0.09
39	travail	Weariness/exhaustion; [Weary/exhaust; vt; 1483-1568]; †841; [»57474]; #12.90/0.00	Use (made of things); [bring/put into use; vt; 1390-1630]; †1468; [»81512]; #6.83/0.07	Harassment, oppression; [Harass; vt; 1303-1695]; †2065; [»129632]; #6.25/0.13			

40	so	Manner of action; [in this way; av; 1250-2000]; †1501; [»84451]; #23.88/0.00	Manner of action; [in this very manner; av; 1598-2000]; †1501; [»84452]; #23.88/0.00	Manner of action; [in that way; av; 1000-2000]; †1501; [»84453]; #23.88/0.00			
41	gladly	Cheerfulness; [Cheerfully; av; 1000-2000]; †2047; [»128871]; #2.86/0.00					
42	spent	Ceasing; [Cease activity; vr; 1663-2000]; †1437; [»79176]; #22.58/0.00	Weariness/exhaustion; [Weary/exhaust; vt; 1582-1674]; †841; [»57474]; #12.90/0.03	Weariness/exhaustion; [Weary/exhaust; vr; 1593-2000]; †841; [»57479]; #12.90/0.03	Liquid which has been emitted; [Emit copiously; vt; 1602-2000]; †1039; [»68036]; #11.24/0.08	Letting/sending out; [in/as in a stream; vt; 1602-1820]; †1402; [»102568]; #10.22/0.11	Time; [Pass/elapse/go (of time); vi; 1607-1681]; †1314; [»87876]; #10.00/0.14
43	,						
44	Forget	Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00	Faulty recollection, forgetting; [forget, cease to know; vt; 1000-2000]; †1789; [»118411]; #92.86/0.00	Faulty recollection, forgetting; [leave behind; vt; 1300-2000]; †1789; [»118414]; #92.86/0.00			
45	not						
46	yet	Lasting quality, permanence; [remaining in specified condition; aj; 1606-2000]; †1361; [»90369]; #42.31/0.00	Flood/flooding; [cause to flow in flood; vt; 1513-1536]; †38; [»2205]; #29.03/0.04	The present (time); [as formerly/still/to this day; av; 1000-2000]; †1340; [»89116]; #26.47/0.09	The present (time); [even now (though not until now); av; 1000-1879]; †1340; [»89117]; #26.47/0.09	Duration; [still continuing/enduring; aj; 1606-2000]; †1316; [»87971]; #25.42/0.17	
47	!						
48	S_END						

49	S_BEGIN						
50	Forget	Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00	Faulty recollection, forgetting; [forget, cease to know; vt; 1000-2000]; †1789; [»118411]; #92.86/0.00	Faulty recollection, forgetting; [leave behind; vt; 1300-2000]; †1789; [»118414]; #92.86/0.00			
51	not						
52	yet	Lasting quality, permanence; [remaining in specified condition; aj; 1606-2000]; †1361; [»90369]; #42.31/0.00	Flood/flooding; [cause to flow in flood; vt; 1513-1536]; †38; [»2205]; #29.03/0.04	The present (time); [as formerly/still/to this day; av; 1000-2000]; †1340; [»89116]; #26.47/0.09	The present (time); [even now (though not until now); av; 1000-1879]; †1340; [»89117]; #26.47/0.09	Duration; [still continuing/enduring; aj; 1606-2000]; †1316; [»87971]; #25.42/0.17	
53	when						
54	first						
55	began	Beginning action/activity; [Begin action/activity; vi; 1000-2000]; †1430; [»78699]; #6.00/0.00	Beginning action/activity; [make a beginning in some enterprise; vi; 1200-2000]; †1430; [»78712]; #6.00/0.00	Beginning action/activity; [Begin/enter upon (an action); vt; 1000-2000]; †1430; [»78725]; #6.00/0.00			
56	The						

57	weary	Weariness/exhaustion; [Weary/exhausted; aj; 1000-2000]; †841; [»57437]; #12.90/0.00	Weariness/exhaustion; [showing signs of weariness/exhaustion; aj; 1000-2000]; †841; [»57456]; #12.90/0.00	Weariness/exhaustion; [causing weariness/exhaustion; aj; 1315-2000]; †841; [»57458]; #12.90/0.00	Weariness/exhaustion; [Become weary/exhausted; vi; 1000-2000]; †841; [»57467]; #12.90/0.00	Weariness/exhaustion; [Weary/exhaust; vt; 1000-2000]; †841; [»57474]; #12.90/0.00	
58	life	Record; [life/case history; n; 1000-2000]; †3152; [»183955]; #8.70/0.00	Life, process of living; [Life; n; 1000-2000]; †142; [»8195]; #6.59/0.05	Life, process of living; [condition/state of being alive; n; 1000-2000]; †142; [»8196]; #6.59/0.05	Life, process of living; [as property of living things; n; 1567-2000]; †142; [»8198]; #6.59/0.05	Life, process of living; [as possession; n; 1000-2000]; †142; [»8199]; #6.59/0.05	Life, process of living; [as dependent on sustenance; n; 1000-2000]; †142; [»8200]; #6.59/0.05
59	ye						
60	know	Understanding; [Understand; vi; 1000-2000]; †1743; [»117114]; #11.54/0.00	Understanding; [Understand; vt; 1000-2000]; †1743; [»117117]; #11.54/0.00	Statement, declaration; [Acknowledge/avow; vr; 1225-1478]; †2239; [»145550]; #9.91/0.08	Knowledge; [Know, be aware of; vi; 1200-2000]; †1792; [»118532]; #9.73/0.12	Knowledge; [Know, be aware of; vt; 1000-2000]; †1792; [»118533]; #9.73/0.12	Knowledge; [Have knowledge of; vi; 1000-2000]; †1792; [»118564]; #9.73/0.12
61	,						
62	since						
63	when	Qualification, condition, reservation; [although; cj; 1489-1489]; †1915; [»123272]; #15.38/0.00	Particular time; [time of occurrence; n; 1616-2000]; †1320; [»88146]; #13.25/0.06	Particular time; [when/at the time that; av; 1000-2000]; †1320; [»88171]; #13.25/0.06	Particular time; [when?/at what time?; av; 1000-2000]; †1320; [»88173]; #13.25/0.06	Particular time; [since when; av; 1300-2000]; †1320; [»88174]; #13.25/0.06	Particular time; [at what time/in what circumstances; av; 1000-2000]; †1320; [»88175]; #13.25/0.06
64	The						

65	suit	Intention; [have as purpose/object; vt; 1559-1686]; †2150; [»136937]; #11.11/0.00	Endeavour; [attempt to obtain; n; 1568-1627]; †1446; [»79610]; #10.96/0.02	Requesting; [beseeching/importuning; n; 1449-2000]; †2222; [»142760]; #8.76/0.04	Requesting; [to/of/upon someone; vi; 1526-1719]; †2222; [»142807]; #8.76/0.04	Seeking marriage; [seeking hand in marriage; n; 1590-2000]; †2278; [»147327]; #7.32/0.09	Seeking marriage; [be a suitor; vi; 1590-1749]; †2278; [»147345]; #7.32/0.09
66	,						
67	the						
68	service	Pear; [fruit of service-tree; n; 1530-1796]; †626; [»42233]; #14.29/0.00	Course; [Course; n; 1601-1765]; †674; [»43446]; #11.11/0.02	Piety; [condition of being; n; 1230-1549]; †2926; [»174808]; #11.11/0.02	Piety; [serving of God by; n; 1175-2000]; †2926; [»174813]; #11.11/0.02	Civil service; [Civil service; n; 1297-2000]; †2620; [»159700]; #9.09/0.08	A lover; [condition of being the servant (of love); n; 1374-1700]; †2089; [»131079]; #7.84/0.10
69	,						
70	none						
71	tell	Understanding; [reach understanding of; vt; 1400-2000]; †1743; [»117119]; #11.54/0.00	Statement, declaration; [State/declare/set forth; vt; 1200-2000]; †2239; [»145465]; #9.91/0.03	Perception, cognition; [Perceive; vt; 1370-2000]; †1728; [»116662]; #9.09/0.06	Requesting; [Request/ask for; vt; 1393-1500]; †2222; [»142823]; #8.76/0.09	Act of convincing, conviction; [by assertion; vt; 1440-2000]; †1845; [»120523]; #8.57/0.12	Action of informing; [Give information; vi; 1000-1558]; †3138; [»183464]; #8.54/0.15
72	can	Knowledge; [Have knowledge of; vi; 1250-2000]; †1792; [»118564]; #9.73/0.00	Knowledge; [Have knowledge, know; vt; 1000-1649]; †1792; [»118568]; #9.73/0.00	Knowledge; [Know, be conversant with; vt; 1000-1649]; †1792; [»118613]; #9.73/0.00	Knowledge; [know how to; vt; 1000-1726]; †1792; [»118614]; #9.73/0.00		
73	;						

74	Forget	Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00	Faulty recollection, forgetting; [forget, cease to know; vt; 1000-2000]; †1789; [»118411]; #92.86/0.00	Faulty recollection, forgetting; [leave behind; vt; 1300-2000]; †1789; [»118414]; #92.86/0.00			
75	not						
76	yet	Lasting quality, permanence; [remaining in specified condition; aj; 1606-2000]; †1361; [»90369]; #42.31/0.00	Flood/flooding; [cause to flow in flood; vt; 1513-1536]; †38; [»2205]; #29.03/0.04	The present (time); [as formerly/still/to this day; av; 1000-2000]; †1340; [»89116]; #26.47/0.09	The present (time); [even now (though not until now); av; 1000-1879]; †1340; [»89117]; #26.47/0.09	Duration; [still continuing/enduring; aj; 1606-2000]; †1316; [»87971]; #25.42/0.17	
77	!						
78	S_END						
79	S_BEGIN						
80	Forget	Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00	Faulty recollection, forgetting; [forget, cease to know; vt; 1000-2000]; †1789; [»118411]; #92.86/0.00	Faulty recollection, forgetting; [leave behind; vt; 1300-2000]; †1789; [»118414]; #92.86/0.00			
81	not						

82	yet	Lasting quality, permanence; [remaining in specified condition; aj; 1606-2000]; †1361; [»90369]; #42.31/0.00	Flood/flooding; [cause to flow in flood; vt; 1513-1536]; †38; [»2205]; #29.03/0.04	The present (time); [as formerly/still/to this day; av; 1000-2000]; †1340; [»89116]; #26.47/0.09	The present (time); [even now (though not until now); av; 1000-1879]; †1340; [»89117]; #26.47/0.09	Duration; [still continuing/enduring; aj; 1606-2000]; †1316; [»87971]; #25.42/0.17	
83	the						
84	great	High reputation, honour; [Majestically/exaltedly; av; 1698-1698]; †1946; [»126513]; #15.38/0.00	Proper pride/self-respect; [dignified; aj; 1585-1697]; †2108; [»132209]; #15.38/0.00	Meaning; [gist; n; 1369-1450]; †1769; [»117846]; #13.75/0.06	Greatness of quantity/amount/degree; [(a) great quantity/amount; n; 1557-1557]; †1634; [»110900]; #10.00/0.09	Greatness of quantity/amount/degree; [Great in quantity/amount/degree; aj; 1000-2000]; †1634; [»110919]; #10.00/0.09	Greatness of quantity/amount/degree; [great (of quantity/amount); aj; 1000-2000]; †1634; [»110920]; #10.00/0.09
85	assays	Endeavour; [an attempt; n; 1386-2000]; †1446; [»79599]; #10.96/0.00	Endeavour; [tentative; n; 1560-1677]; †1446; [»79601]; #10.96/0.00	Endeavour; [Make an attempt/endeavour; vi; 1370-2000]; †1446; [»79634]; #10.96/0.00	Endeavour; [Attempt; vt; 1300-2000]; †1446; [»79651]; #10.96/0.00	Endeavour; [attempt to obtain/attain; vt; 1597-1597]; †1446; [»79656]; #10.96/0.00	Character, disposition, mood; [Disposition/character; n; 1393-1579]; †1703; [»115493]; #10.42/0.12
86	,						
87	The						
88	cruel	High/intense degree; [excessively; av; 1573-2000]; †1635; [»111034]; #7.37/0.00	Ill-treatment; [person; n; 1420-1725]; †1526; [»86637]; #5.36/0.12	Ill-treatment; [Cruel; aj; 1297-2000]; †1526; [»86653]; #5.36/0.12	Ill-treatment; [specifically of thing/action; aj; 1300-2000]; †1526; [»86654]; #5.36/0.12		

89	wrong	Unjustness; [Unjustness; n; 1000-2000]; †2014; [»125165]; #50.00/0.00	Unjustness; [Unfair, unjust; aj; 1275-2000]; †2014; [»125166]; #50.00/0.00	Unjustness; [Unjustly; av; 1250-2000]; †2014; [»125171]; #50.00/0.00	Unjustness; [Do wrong; vi; 1390-1676]; †2014; [»125173]; #50.00/0.00	Unjustness; [treat; vt; 1330-2000]; †2014; [»125175]; #50.00/0.00	Unjustness; [and disrespectfully; vt; 1449-2000]; †2014; [»125176]; #50.00/0.00
90	,						
91	the						
92	scornful	Contempt; [contemptuous; aj; 1400-2000]; †1950; [»126646]; #2.13/0.00	Contempt; [contemptible; aj; 1570-1624]; †1950; [»126707]; #2.13/0.00				
93	ways	Manner of action; [Manner of action; n; 1000-2000]; †1501; [»84418]; #23.88/0.00	Travel in specific course/direction; [Travel in specific course/direction; n; 1000-2000]; †3270; [»188881]; #13.60/0.03	Vascular system; [vessel; n; 1425-1615]; †256; [»26138]; #7.69/0.06	Use (made of things); [course adopted to achieve an end; n; 1175-2000]; †1468; [»81620]; #6.83/0.09	State/condition; [good or bad condition/order; n; 1467-2000]; †1193; [»76434]; #6.25/0.12	State/condition; [that way/condition; n; 1598-2000]; †1193; [»76449]; #6.25/0.12
94	,						
95	The						
96	painful	Care/carefulness/attention; [careful/painstaking; aj; 1549-1877]; †1509; [»85138]; #7.37/0.00	Care/carefulness/attention; [characterized by painstaking care; aj; 1380-1894]; †1509; [»85139]; #7.37/0.00				

97	patience	Patience; [Patience; n; 1225-2000]; †2038; [»128349]; #20.00/0.00	Patience; [personified; n; 1377-2000]; †2038; [»128350]; #20.00/0.00	Patience; [forbearance/tolerance; n; 1377-2000]; †2038; [»128355]; #20.00/0.00	Patience; [patience in waiting; n; 1375-2000]; †2038; [»128357]; #20.00/0.00	Patience; [Be patient; vi; 1596-1835]; †2038; [»128374]; #20.00/0.00	Patience; [Be patient; vr; 1605-1605]; †2038; [»128385]; #20.00/0.00
98	in						
99	delays	Delay/postponement; [Delay/postponement; n; 1297-2000]; †1351; [»89880]; #7.84/0.00	Delay/postponement; [Delay; vi; 1509-2000]; †1351; [»89903]; #7.84/0.00	Delay/postponement; [Delay; vt; 1290-2000]; †1351; [»89917]; #7.84/0.00	Delay/postponement; [a person; vt; 1388-1768]; †1351; [»89920]; #7.84/0.00		
100	,						
101	Forget	Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00	Faulty recollection, forgetting; [forget, cease to know; vt; 1000-2000]; †1789; [»118411]; #92.86/0.00	Faulty recollection, forgetting; [leave behind; vt; 1300-2000]; †1789; [»118414]; #92.86/0.00			
102	not						
103	yet	Lasting quality, permanence; [remaining in specified condition; aj; 1606-2000]; †1361; [»90369]; #42.31/0.00	Flood/flooding; [cause to flow in flood; vt; 1513-1536]; †38; [»2205]; #29.03/0.04	The present (time); [as formerly/still/to this day; av; 1000-2000]; †1340; [»89116]; #26.47/0.09	The present (time); [even now (though not until now); av; 1000-1879]; †1340; [»89117]; #26.47/0.09	Duration; [still continuing/enduring; aj; 1606-2000]; †1316; [»87971]; #25.42/0.17	
104	!						

105	S_END						
106	S_BEGIN						
107	Forget	Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00	Faulty recollection, forgetting; [forget, cease to know; vt; 1000-2000]; †1789; [»118411]; #92.86/0.00	Faulty recollection, forgetting; [leave behind; vt; 1300-2000]; †1789; [»118414]; #92.86/0.00			
108	not						
109	!						
110	S_END						
111	S_BEGIN						
112	O						
113	,						
114	forget	Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00	Faulty recollection, forgetting; [forget, cease to know; vt; 1000-2000]; †1789; [»118411]; #92.86/0.00	Faulty recollection, forgetting; [leave behind; vt; 1300-2000]; †1789; [»118414]; #92.86/0.00			
115	not						
116	this						
117	!						
118	S_END						

119	S_BEGIN						
120	How	Duration; [for the whole time/duration; av; 1639-1639]; †1316; [»87989]; #25.42/0.00	Manner of action; [Manner of action; n; 1551-2000]; †1501; [»84418]; #23.88/0.04	Manner of action; [In (some/any) way; av; 1000-2000]; †1501; [»84444]; #23.88/0.04	Manner of action; [in the way that; av; 1400-2000]; †1501; [»84448]; #23.88/0.04	Manner of action; [in what way; av; 1000-2000]; †1501; [»84449]; #23.88/0.04	Manner of action; [in whatever way; av; 1400-1869]; †1501; [»84450]; #23.88/0.04
121	long	Lasting quality, permanence; [Permanent; aj; 1000-2000]; †1361; [»90367]; #42.31/0.00	Duration; [a long time; n; 1000-2000]; †1316; [»87940]; #25.42/0.03	Duration; [Having/pertaining to duration; aj; 1300-2000]; †1316; [»87956]; #25.42/0.03	Duration; [long-lasting/enduring; aj; 1000-2000]; †1316; [»87957]; #25.42/0.03	Duration; [long-seeming; aj; 1592-2000]; †1316; [»87962]; #25.42/0.03	Duration; [of long standing; aj; 1220-2000]; †1316; [»87963]; #25.42/0.03
122	ago	The past; [ago; aj; 1314-2000]; †1341; [»89162]; #12.20/0.00					
123	hath						
124	been						
125	,						
126	and						
127	is						
128	,						
129	The						
130	mind	The mind; [The mind; n; 1000-1784]; †1696; [»115259]; #20.00/0.00	State of feeling, mood; [State of feeling/mood; n; 1500-2000]; †2025; [»127644]; #15.38/0.02	State of feeling, mood; [towards another/others; n; 1470-1611]; †2025; [»127647]; #15.38/0.02	Memory, keeping in mind; [something remembered; n; 1000-1489]; †1784; [»118209]; #12.31/0.06	Memory, keeping in mind; [Have in one's mind, remember; vi; 1422-2000]; †1784; [»118222]; #12.31/0.06	Memory, keeping in mind; [Call to mind, recollect; vt; 1382-2000]; †1784; [»118225]; #12.31/0.06

131	that						
132	never	Infrequency; [never; av; 1000-2000]; †1353; [»89999]; #6.90/0.00					
133	meant	Ignobleness/baseness; [Base/ignoble; aj; 1610-2000]; †1970; [»127455]; #46.15/0.00	Ignobleness/baseness; [specifically of character; aj; 1665-2000]; †1970; [»127456]; #46.15/0.00	Ignobleness/baseness; [Basely/ignobly; av; 1626-1626]; †1970; [»127460]; #46.15/0.00	Intercession/influence on someone's behalf; [Intercession/influence on someone's behalf; n; 1432-1565]; †1462; [»80946]; #37.50/0.04	Intercession/influence on someone's behalf; [one who; n; 1455-1606]; †1462; [»80947]; #37.50/0.04	Intercession/influence on someone's behalf; [bring about by mediation; vt; 1449-1654]; †1462; [»80961]; #37.50/0.04
134	amiss	Lack of truth, falsity, error; [amiss, out of order; aj; 1315-2000]; †1812; [»119286]; #15.25/0.00	Lack of truth, falsity, error; [in a wrong way, amiss; av; 1250-2000]; †1812; [»119295]; #15.25/0.00	Lack of truth, falsity, error; [An error, mistake; n; 1477-1700]; †1812; [»119308]; #15.25/0.00	Lack of truth, falsity, error; [Wrongly, erroneously; av; 1380-1833]; †1812; [»119322]; #15.25/0.00		
135	Forget	Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00	Faulty recollection, forgetting; [forget, cease to know; vt; 1000-2000]; †1789; [»118411]; #92.86/0.00	Faulty recollection, forgetting; [leave behind; vt; 1300-2000]; †1789; [»118414]; #92.86/0.00			
136	not						

137	yet	Lasting quality, permanence; [remaining in specified condition; aj; 1606-2000]; †1361; [»90369]; #42.31/0.00	Flood/flooding; [cause to flow in flood; vt; 1513-1536]; †38; [»2205]; #29.03/0.04	The present (time); [as formerly/still/to this day; av; 1000-2000]; †1340; [»89116]; #26.47/0.09	The present (time); [even now (though not until now); av; 1000-1879]; †1340; [»89117]; #26.47/0.09	Duration; [still continuing/enduring; aj; 1606-2000]; †1316; [»87971]; #25.42/0.17	
138	!						
139	S_END						
140	S_BEGIN						
141	Forget	Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00	Faulty recollection, forgetting; [forget, cease to know; vt; 1000-2000]; †1789; [»118411]; #92.86/0.00	Faulty recollection, forgetting; [leave behind; vt; 1300-2000]; †1789; [»118414]; #92.86/0.00			
142	not						
143	then	Antecedence/being earlier; [Antecedent/earlier; aj; 1584-2000]; †1343; [»89273]; #23.75/0.00	Relative time; [Different time; n; 1549-2000]; †1338; [»89757]; #19.35/0.17	Relative time; [(by) that time/(since) that time; n; 1300-2000]; †1338; [»89758]; #19.35/0.17	Relative time; [at that time; av; 1000-2000]; †1338; [»89759]; #19.35/0.17		
144	thine						
145	own						
146	approved	Approval/sanction; [approved/accepted; aj; 1667-2000]; †1935; [»126093]; #6.25/0.00					

147	,						
148	The						
149	which						
150	so	Manner of action; [in this way; av; 1250-2000]; †1501; [»84451]; #23.88/0.00	Manner of action; [in this very manner; av; 1598-2000]; †1501; [»84452]; #23.88/0.00	Manner of action; [in that way; av; 1000-2000]; †1501; [»84453]; #23.88/0.00			
151	long	Lasting quality, permanence; [Permanent; aj; 1000-2000]; †1361; [»90367]; #42.31/0.00	Duration; [a long time; n; 1000-2000]; †1316; [»87940]; #25.42/0.03	Duration; [Having/pertaining to duration; aj; 1300-2000]; †1316; [»87956]; #25.42/0.03	Duration; [long-lasting/enduring; aj; 1000-2000]; †1316; [»87957]; #25.42/0.03	Duration; [long-seeming; aj; 1592-2000]; †1316; [»87962]; #25.42/0.03	Duration; [of long standing; aj; 1220-2000]; †1316; [»87963]; #25.42/0.03
152	hath						
153	thee						
154	so	Manner of action; [in this way; av; 1250-2000]; †1501; [»84451]; #23.88/0.00	Manner of action; [in this very manner; av; 1598-2000]; †1501; [»84452]; #23.88/0.00	Manner of action; [in that way; av; 1000-2000]; †1501; [»84453]; #23.88/0.00			
155	loved	Liking; [Liking/favourable regard; n; 1000-2000]; †2080; [»130455]; #9.30/0.00	Liking; [Have liking for; vt; 1200-2000]; †2080; [»130485]; #9.30/0.00	Liking; [like very much; vt; 1000-2000]; †2080; [»130488]; #9.30/0.00	A lover; [one who is loved/a sweetheart; n; 1225-2000]; †2089; [»131069]; #7.84/0.09	Love; [Love; n; 1000-2000]; †2079; [»130324]; #5.11/0.12	Love; [as an abstract principle; n; 1000-2000]; †2079; [»130325]; #5.11/0.12

156	,						
157	Whose						
158	steadfast	Constancy; [Constant/steadfast; aj; 1000-2000]; †2155; [»137210]; #14.29/0.00	Constancy; [Constantly/steadfastly; av; 1300-1887]; †2155; [»137215]; #14.29/0.00	A look/glance; [fixed/steady; aj; 1300-1817]; †884; [»58902]; #5.41/0.12	Warrior; [qualities/attributes; aj; 1000-1821]; †2493; [»155199]; #5.41/0.12		
159	faith	Faithfulness/trustworthiness; [Faithfulness/trustworthiness; n; 1250-1863]; †2800; [»170326]; #41.67/0.00	Truth, validity, correctness; [In truth; in; 1586-1840]; †1806; [»119054]; #13.73/0.08	Belief, trust, confidence; [Belief, trust, confidence; n; 1300-2000]; †1844; [»120469]; #13.64/0.15			
160	yet	Lasting quality, permanence; [remaining in specified condition; aj; 1606-2000]; †1361; [»90369]; #42.31/0.00	Flood/flooding; [cause to flow in flood; vt; 1513-1536]; †38; [»2205]; #29.03/0.04	The present (time); [as formerly/still/to this day; av; 1000-2000]; †1340; [»89116]; #26.47/0.09	The present (time); [even now (though not until now); av; 1000-1879]; †1340; [»89117]; #26.47/0.09	Duration; [still continuing/enduring; aj; 1606-2000]; †1316; [»87971]; #25.42/0.17	
161	never	Infrequency; [never; av; 1000-2000]; †1353; [»89999]; #6.90/0.00					
162	moved	Organs of excretion; [provoke excretion; vt; 1597-1605]; †247; [»25806]; #12.35/0.00	Requesting; [appeal to/invoke; vt; 1399-1768]; †2222; [»142844]; #8.76/0.02	Requesting; [petition; vt; 1660-2000]; †2222; [»142851]; #8.76/0.02	Bodily movement; [Move the body/a member; vi; 1330-2000]; †1363; [»98936]; #7.69/0.06	Bodily movement; [move as a living being; vi; 1400-2000]; †1363; [»98937]; #7.69/0.06	Bodily movement; [move (of part of body); vi; 1535-2000]; †1363; [»98939]; #7.69/0.06
163	:						

164	Forget	Faulty recollection, forgetting; [forget, fail to remember; vi; 1382-2000]; †1789; [»118408]; #92.86/0.00	Faulty recollection, forgetting; [forget, cease to know; vt; 1000-2000]; †1789; [»118411]; #92.86/0.00	Faulty recollection, forgetting; [leave behind; vt; 1300-2000]; †1789; [»118414]; #92.86/0.00			
165	not						
166	this						
167	!						

Table A11 [34] HTST result

ID	Word	USAS	HTE Category	Thematic Heading
1	34	N1 T1.2 T3	01.16.04 [Number]	AP.04 [Number];
2	.	PUNC		
3	S_END	Z99		
4	S_BEGIN	Z99		
5	Forget	X2.2-	02.01.11.04-01 [0.92307692] [forget, cease to know]; 02.01.11.04-02 [0.92307692] [forget, fail to remember]; 01.15.09.02-03 [0.95454545] [leave undone/fail to perform/carry out];	AR.35.e [Faulty recollection, forgetting]; AR.35.e [Faulty recollection, forgetting]; AO.10.b [Not doing];
6	not	Z6	04.04 [];	ZD [Negative];

7	yet	T1.1.2	01.13.08.03-06 [0021.8008747062] [as formerly/still/to this day]; 02.02.06.01.03-08 [0007.4483575891] [however, nevertheless, notwithstanding]; 01.13.08.06-04 [0000.2112507544] [at some future time/one day];	AM.08.b [The present (time)]; AS.05.b [Qualification, condition, reservation]; AM.08 [Relative time];
8	The	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
9	Lover	S3.2/S2mf E2+/S2mf	01.17.04.01-12 [0.93750000] [as lover/spouse]; 02.04.13.12 [0.94117647] [One who loves/a lover]; 03.06.05.07.02-05.02.02 [0.94444444] [illicit male lover];	AQ.04.a [Christian God]; AU.29.b [A lover]; BD.05.g.02 [Unchastity];
10	Beseeches	Q2.2 X7+	02.07.03.12-03 [0.93333333] [beseech/implore]; 02.07.03.12-06 [1.00000000] [a person a thing]; 02.07.03.12-07 [1.00000000] [a person of a thing];	AX.12 [Requesting]; AX.12 [Requesting]; AX.12 [Requesting];
11	his	Z8m	02.06.01-03.03.01 [his/her]; 04.06 [];	AW.01 [Owning]; ZF [Pronoun];
12	Mistress	S7.1+/S2.1f S3.2/S2.1f P1/S2.1f	03.06.05.07.02-05.02.01 [0.94117647] [a mistress]; 01.15.22.01-19.04 [0.94444444] [a master/mistress]; 02.04.13.12-14.01 [0.95454545] [specifically a female sweetheart/girlfriend];	BD.05.g.02 [Unchastity]; AO.23.a [Skill/skilfulness]; AU.29.b [A lover];
13	not	Z6	04.04 [];	ZD [Negative];
14	to	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
15	Forget	X2.2-	02.01.11.04-01 [0.93750000] [forget, cease to know]; 02.01.11.04-02 [0.93750000] [forget, fail to remember]; 01.15.09.02-03 [0.96000000] [leave undone/fail to perform/carry out];	AR.35.e [Faulty recollection, forgetting]; AR.35.e [Faulty recollection, forgetting]; AO.10.b [Not doing];
16	his	Z8m	02.06.01-03.03.01 [his/her]; 04.06 [];	AW.01 [Owning]; ZF [Pronoun];
17	Steadfast	A2.1-	01.14.09-11.04 [0.93750000] [firmly fixed]; 02.04.13.05-09 [0.94117647] [steadfast/constant in affection]; 01.17.04.01.02-13 [0.94444444] [steadfast];	AN.09 [Absence of movement]; AU.27.d [Affection, tenderness]; AQ.04.a.02 [Nature/attributes of God];
18	Faith	E6+ S9 X2.1	03.08 [0.88888889] [Faith]; 03.08.01.05.01 [0.92857143] [A religion/church]; 02.01.13.02 [0.94117647] [Belief, trust, confidence];	BF [Faith]; BF.03.a [A religion/church]; AR.45.b [Belief, trust, confidence];
19	and	Z5	04.03 [];	ZC [Grammatical Item];

20	True	A5.2+ A5.4+	03.06.01.03.01 [0013.8825610437] [Faithful/trustworthy]; 03.06.02 [0008.1201675347] [Due/morally fitting/proper]; 01.13.11.05.01 [0007.9184800463] [Stable, fixed];	BD.01.c.01 [Faithfulness/trustworthiness]; BD.02 [Dueness/propriety]; AM.11.e [Absence of change, changelessness];
21	Intent	X7+	02.05.04-03 [0042.3321434697] [intention/purpose]; 02.05.04-03.06 [0042.3321434697] [end/purpose/object]; 02.05.04.01-01 [0024.1284648987] [a plan];	AV.03 [Intention]; AV.03 [Intention]; AV.03.a [Planning];
22	FORGET	X2.2-	02.01.11.04-01 [0.94117647] [forget, cease to know]; 02.01.11.04-02 [0.94117647] [forget, fail to remember]; 01.15.09.02-03 [0.96153846] [leave undone/fail to perform/carry out];	AR.35.e [Faulty recollection, forgetting]; AR.35.e [Faulty recollection, forgetting]; AO.10.b [Not doing];
23	not	Z6	04.04 [];	ZD [Negative];
24	yet	T1.1.2	01.13.08.03-06 [0021.8008747063] [as formerly/still/to this day]; 02.02.06.01.03-08 [0007.4483575891] [however, nevertheless, notwithstanding]; 01.13.08.06-04 [0000.2112507544] [at some future time/one day];	AM.08.b [The present (time)]; AS.05.b [Qualification, condition, reservation]; AM.08 [Relative time];
25	the	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
26	tried	X2.4	01.09.10.04-01 [0.94444444] [cleared of refuse]; 01.15.10.01.01-03 [0.95454545] [tried/tested]; 01.07.01.22.06.01-11 [1.00000000] [rendered (of fat)];	AI.16.d [Clearing of refuse matter]; AO.11.a.01 [Trial/test/testing]; AG.01.t.05 [Preparation for table/cooking];
27	intent	X7+	02.05.04.01-01 [0.92857143] [a plan]; 02.01.07.04.03 [0.94736842] [Topic, subject-matter]; 03.05.06-08 [0.96153846] [construction put on something by the law];	AV.03.a [Planning]; AR.11.a [Topic, subject, concern]; BC.03 [Jurisprudence];
28	Of	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
29	such	A13.3	01.16.01.09-03 [0.93750000] [such]; 01.16.02-02.02 [0.94117647] [of the kind that]; 01.16.02-02 [0.94117647] [of this/that kind];	AP.01.i [Similarity]; AP.02 [Kind/sort]; AP.02 [Kind/sort];
30	a	A13.3	04.03 [Grammatical]	ZC [Grammatical Item];
31	truth	A5.2+	02.01.12.08-02 [0.88888889] [personified]; 02.01.12.08-04 [0.90000000] [true facts/circumstances]; 02.01.12.08.03 [0.94444444] [Truthfulness, veracity];	AR.38 [Truth, validity, correctness]; AR.38 [Truth, validity, correctness]; AR.39 [Truthfulness, veracity];
32	as	Z5	04.03 [Grammatical]; 01.16.01.09-01 [.so/in such a manner]	ZC [Grammatical Item]; AP.01.i [Similarity];
33	I	Z8mf	04.06 [Pronoun];	ZF [Pronoun];

34	have	Z5	04.03 [Grammatical];	ZC [Grammatical Item];
35	meant	Q1.1 X7+ S6+	02.01.10.02 [0.92857143] [Mean]; 02.01.11 [1.00000000] [Have in one's mind, remember]; 02.01.13.03 [1.00000000] [Hold an opinion, opine];	AR.29 [Meaning]; AR.35 [Memory, keeping in mind]; AR.45.e [Expressed belief, opinion];
36	;	PUNC		
37	My	Z8	02.06.01-03.02 [denationalize]; 04.06 [];	AW.01 [Owning]; ZF [Pronoun];
38	great	A5.1+ A11.1+ N3.2+	02.02.08-06 [0.94736842] [of high/great importance]; 01.12.02.04.01 [1.00000000] [Large];	AS.10 [Importance, influence]; AL.02.d.01 [Largeness];
39	travail	A12-	01.15.20.02-10 [0002.4821849797] [labour/toil]; 01.15.20.02-10.06 [0002.4821849797] [a piece of hard work]; 03.13.03.04.01-02 [0001.7507317969] [regarded as the result of labour];	AO.21.b [Effort/exertion]; AO.21.b [Effort/exertion]; BK.06.a [A written composition];
40	so	A13.3	01.16.01.09-01 [0.95238095] [so/in such a manner]; 01.11.03.01 [1.00000000] [For that reason/therefore]; 01.13.08.05-10 [1.00000000] [after/afterwards/later];	AP.01.i [Similarity]; AK.03.a [Cause/reason]; AM.08.d [The future/time to come];
41	gladly	X5.2+	02.04.10.09 [0002.0935753078] [Cheerfully]; 02.04.10.08 [0001.3433328753] [Joyfully];	AU.12 [Cheerfulness]; AU.11 [Happiness];
42	spent	I1.2 A1.5.1 T1.3	01.13.01 [0.92857143] [Spend time/allow time to pass]; 03.12.20.02 [0.93333333] [Spend/incur expense]; 01.09.02.03 [1.00000000] [Weary/exhaust];	AM.01 [Spending time]; BJ.01.y.02 [Expenditure]; AI.08.c [Weariness/exhaustion];
43	,	PUNC		
44	Forget	X2.2-	02.01.11.04-01 [0.93750000] [forget, cease to know]; 02.01.11.04-02 [0.93750000] [forget, fail to remember]; 01.15.09.02-03 [0.96000000] [leave undone/fail to perform/carry out];	AR.35.e [Faulty recollection, forgetting]; AR.35.e [Faulty recollection, forgetting]; AO.10.b [Not doing];
45	not	Z6	04.04 [];	ZD [Negative];
46	yet	T1.1.2	01.13.08.03-06 [0021.8008747062] [as formerly/still/to this day]; 02.02.06.01.03-08 [0007.4483575891] [however, nevertheless, notwithstanding]; 01.13.08.06-04 [0000.2112507544] [at some future time/one day];	AM.08.b [The present (time)]; AS.05.b [Qualification, condition, reservation]; AM.08 [Relative time];
47	!	PUNC		

48	S_END	Z99		
49	S_BEGIN	Z99		
50	Forget	X2.2-	02.01.11.04-01 [0.92307692] [forget, cease to know]; 02.01.11.04-02 [0.92307692] [forget, fail to remember]; 01.15.09.02-03 [0.95454545] [leave undone/fail to perform/carry out];	AR.35.e [Faulty recollection, forgetting]; AR.35.e [Faulty recollection, forgetting]; AO.10.b [Not doing];
51	not	Z6	04.04 [];	ZD [Negative];
52	yet	T1.1.2	01.13.08.03-06 [0022.2555140213] [as formerly/still/to this day]; 02.02.06.01.03-08 [0004.1999527488] [however, nevertheless, notwithstanding]; 01.13.08.06-04 [0000.4885441875] [at some future time/one day];	AM.08.b [The present (time)]; AS.05.b [Qualification, condition, reservation]; AM.08 [Relative time];
53	when	Z5	01.13.03-04 [when/at the time that]; 04.03 [];	AM.03 [Particular time]; ZC [Grammatical Item];
54	first	N4	01.16.04 [Number]	AP.04 [Number];
55	began	T2+	02.07.03-18 [0.94117647] [begin to speak]; 01.15.03.02 [0.94444444] [Begin action/activity]; 01.16.03.03.02 [0.94736842] [Begin];	AX.03 [Speech]; AO.04 [Beginning action/activity]; AP.03.c.02 [Beginning];
56	The	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
57	weary	B2- X5.2-	01.09.02.03 [0.95000000] [Weary/exhausted]; 02.04.11.11-01 [0.95454545] [weary of person/thing]; 02.04.11.11-01.02 [0.96428571] [anxious to be rid of someone];	AI.08.c [Weariness/exhaustion]; AU.24 [Weariness, tedium]; AU.24 [Weariness, tedium];
58	life	L1+ A3+ T1.3	01.02 [0.88888889] [Life]; 01.02-09 [0.92857143] [course/span of life];	AB.01 [Life, process of living]; AB.01 [Life, process of living];
59	ye	Z8mf	04.06 [];	ZF [Pronoun];
60	know	X2.2+ S3.2/B1%	02.01.12 [0.91666667] [Know, be aware of]; 02.01.12.02.03 [0.93333333] [Have mastery of]; 02.01.08 [1.00000000] [Understand];	AR.36 [Knowledge]; AR.36.b [Scholarly knowledge, erudition]; AR.16 [Understanding];
61	,	PUNC		
62	since	Z5	04.03 [Grammatical]	ZC [Grammatical Item];

63	when	Z5	01.13.03-12 [when/at which time]; 04.03 [];	AM.03 [Particular time]; ZC [Grammatical Item];
64	The	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
65	suit	B5 G2.1 K5.2	01.08.02.02.12 [0.93333333] [Set/suit of clothes]; 03.13.01.05.02.07.01-14 [0.95454545] [suit]; 03.12.19.03.02-04.02 [0.95833333] [in lieu of attendance at court];	AH.02.b.08 [Set/suit of clothes]; BK.01.d.04.e [Card-game]; BJ.01.s.02 [Payment/service to feudal superior];
66	,	PUNC		
67	the	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
68	service	S8+ S7.1-I2.2	03.04.09.02 [0.95454545] [Service]; 03.04.09.02.01-02 [0.96666667] [performance of duties of]; 01.15.14-07.01 [1.00000000] [beneficial action];	BB.09.b [Service]; BB.09.b.01 [Servant]; AO.15 [Advantage];
69	,	PUNC		
70	none	Z6/Z8c	04.04 [];	ZD [Negative];
71	tell	Q2.2 Q2.1 X3	03.13.03.04.07.04 [0.95454545] [Tell (story)]; 02.07.03-17 [1.00000000] [announce/make known]; 02.07.03.03-01 [1.00000000] [be narrated];	BK.06.g.03 [Narrative/story]; AX.03 [Speech]; AX.04 [Narration, description];
72	can	O2 F2%	01.16.05.03.01.01-03.18 [0.96296296] [tin/can]; 01.07.02.20.01 [1.00000000] [Drinking vessel]; 03.11.11.40.06.10 [1.00000000] [Other specific vessels for holding liquids];	AP.05.c.01.a [Amount defined by capacity]; AG.01.aj.01 [Drinking vessel]; BI.11.x.04 [Vessel];
73	;	PUNC		
74	Forget	X2.2-	02.01.11.04-01 [0.93750000] [forget, cease to know]; 02.01.11.04-02 [0.93750000] [forget, fail to remember]; 01.15.09.02-03 [0.96000000] [leave undone/fail to perform/carry out];	AR.35.e [Faulty recollection, forgetting]; AR.35.e [Faulty recollection, forgetting]; AO.10.b [Not doing];
75	not	Z6	04.04 [];	ZD [Negative];
76	yet	T1.1.2	01.13.08.03-06 [0022.2555140213] [as formerly/still/to this day]; 02.02.06.01.03-08 [0004.1999527488] [however, nevertheless, notwithstanding]; 01.13.08.06-04 [0000.4885441875] [at some future time/one day];	AM.08.b [The present (time)]; AS.05.b [Qualification, condition, reservation]; AM.08 [Relative time];
77	!	PUNC		

78	S_END	Z99		
79	S_BEGIN	Z99		
80	Forget	X2.2-	02.01.11.04-01 [0.92307692] [forget, cease to know]; 02.01.11.04-02 [0.92307692] [forget, fail to remember]; 01.15.09.02-03 [0.95454545] [leave undone/fail to perform/carry out];	AR.35.e [Faulty recollection, forgetting]; AR.35.e [Faulty recollection, forgetting]; AO.10.b [Not doing];
81	not	Z6	04.04 [];	ZD [Negative];
82	yet	T1.1.2	01.13.08.03-06 [0025.3883943543] [as formerly/still/to this day]; 02.02.06.01.03-08 [0001.9354741379] [however, nevertheless, notwithstanding]; 01.13.08.06-04 [0000.1415240514] [at some future time/one day];	AM.08.b [The present (time)]; AS.05.b [Qualification, condition, reservation]; AM.08 [Relative time];
83	the	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
84	great	A5.1+ A11.1+ N3.2+	02.02.08-06 [0.93750000] [of high/great importance]; 01.12.02.04.01 [1.00000000] [Large];	AS.10 [Importance, influence]; AL.02.d.01 [Largeness];
85	assays	X2.4/Y1	03.03.05 [0055.4025616050] [Attack]; 03.11.06.03.06-25 [0008.2906762171] [testing]; 01.15.18-08 [0007.2472005814] [circumstance/occurrence];	BA.05 [Attack]; BI.06.c.01 [Working with metal]; AO.19 [Adversity];
86	,	PUNC		
87	The	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
88	cruel	E3-	02.03.06.13 [0.94117647] [Savage, cruel]; 01.15.21.05.05.01 [0.94444444] [Cruel]; 01.15.21.05.04 [1.00000000] [Fierce];	AT.21 [Harshness, cruelty]; AO.22.d.05 [Ill-treatment]; AO.22.d.04 [Fierceness];
89	wrong	A5.3- A5.1- G2.2-	03.10.03.01-04 [0.90476190] [wrong]; 02.03.05 [0.92857143] [Doing wrong]; 02.03.05.13 [0.93750000] [Unfair, unjust];	BH.12.a [Route/way]; AT.15 [Wrongdoing]; AT.15.i [Unjustness];
90	,	PUNC		
91	the	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
92	scornful	S7.2-	02.02.10-02 [0001.7968987871] [contemptuous]; 02.02.10.02 [0000.6496076121] [Derisive/mocking]; 02.02.10.01-02 [0000.1698401986] [contemptible];	AS.14 [Contempt]; AS.14.a [Derision/ridicule/mockery]; AS.14 [Contempt];

93	ways	X4.2 M6 M3/H3	01.11.01.07-13 [0.92857143] [that way/condition]; 03.10.03.01 [0.95000000] [Route/way]; 01.15.20 [1.00000000] [Manner of action];	AK.01.g [State/condition]; BH.12.a [Route/way]; AO.21 [Manner of action];
94	,	PUNC		
95	The	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
96	painful	B2- E4.1-	01.03.01.09-02 [0.93750000] [causing pain]; 01.03.01.09-01 [1.00000000] [of parts of body]; 01.15.18-03.02 [1.00000000] [inflicting];	AC.01.f [Pain]; AC.01.f [Pain]; AO.19 [Adversity];
97	patience	E3+ K5.2%	02.04.09.04 [0.92307692] [Patience]; 02.04.09.04-06 [0.94444444] [patience in waiting]; 01.06.13.01.03.46 [1.00000000] [Polygonaceae (dock and allies)];	AU.07.a [Patience]; AU.07.a [Patience]; AF.12 [Particular plants/herbs/shrubs];
98	in	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
99	delays	T4-	01.13.09.02.01 [0.93333333] [Delay/postponement];	AM.09.a.01 [Delay/postponement];
100	,	PUNC		
101	Forget	X2.2-	02.01.11.04-01 [0.93750000] [forget, cease to know]; 02.01.11.04-02 [0.93750000] [forget, fail to remember]; 01.15.09.02-03 [0.96000000] [leave undone/fail to perform/carry out];	AR.35.e [Faulty recollection, forgetting]; AR.35.e [Faulty recollection, forgetting]; AO.10.b [Not doing];
102	not	Z6	04.04 [];	ZD [Negative];
103	yet	T1.1.2	01.13.08.03-06 [0025.3883943543] [as formerly/still/to this day]; 02.02.06.01.03-08 [0001.9354741379] [however, nevertheless, notwithstanding]; 01.13.08.06-04 [0000.1415240514] [at some future time/one day];	AM.08.b [The present (time)]; AS.05.b [Qualification, condition, reservation]; AM.08 [Relative time];
104	!	PUNC		
105	S_END	Z99		
106	S_BEGIN	Z99		

107	Forget	X2.2-	02.01.11.04-01 [0.90909091] [forget, cease to know]; 02.01.11.04-02 [0.90909091] [forget, fail to remember]; 01.15.09.02-03 [0.95000000] [leave undone/fail to perform/carry out];	AR.35.e [Faulty recollection, forgetting]; AR.35.e [Faulty recollection, forgetting]; AO.10.b [Not doing];
108	not	Z6	04.04 [];	ZD [Negative];
109	!	PUNC		
110	S_END	Z99		
111	S_BEGIN	Z99		
112	O	Z5	04.03 [Grammatical];	ZC [Grammatical Item];
113	,	PUNC		
114	forget	X2.2-	02.01.11.04-01 [0.91666667] [forget, cease to know]; 02.01.11.04-02 [0.91666667] [forget, fail to remember]; 01.15.09.02-03 [0.95238095] [leave undone/fail to perform/carry out];	AR.35.e [Faulty recollection, forgetting]; AR.35.e [Faulty recollection, forgetting]; AO.10.b [Not doing];
115	not	Z6	04.04 [];	ZD [Negative];
116	this	Z8	04.03 [];	ZC [Grammatical Item];
117	!	PUNC		
118	S_END	Z99		
119	S_BEGIN	Z99		
120	How	Z5 A13.3	01.16.06.01-03 [to what extent]; 04.03 [];	AP.06 [Quantity]; ZC [Grammatical Item];
121	long	T1.1.1	01.13.02-04 [0.90909091] [for a long time];	AM.02 [Duration];
122	ago	T1.1.1	4.1	ZZ [Unrecognised]
123	hath	Z5	04.03 [Grammatical];	ZC [Grammatical Item];
124	been	A3+ Z5	01.11.01.07 [Be/remain in specific state/condition]; 01.16.01.04 [Be the same as]; 04.03 [Grammatical]	AK.01.g [State/condition]; AP.01.d [Identity]; ZC [Grammatical Item];
125	,	PUNC		

126	and	Z5	04.03 [];	ZC [Grammatical Item];
127	is	A3+ Z5	01.11.01.07 [Be/remain in specific state/condition]; 01.16.01.04 [Be the same as]; 04.03 [Grammatical]	AK.01.g [State/condition]; AP.01.d [Identity]; ZC [Grammatical Item];
128	,	PUNC		
129	The	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
130	mind	X1	02.01.02 [0.92857143] [Intellect]; 02.01.13.03 [0.95000000] [Expressed belief, opinion];	AR.03 [Intellect]; AR.45.e [Expressed belief, opinion];
131	that	Z8	04.03 [];	ZC [Grammatical Item];
132	never	T1/Z6	01.13.10.01-03 [0.89473684] [never]; 01.16.06.07-16 [0.94444444] [not at all];	AM.10.b [Infrequency]; AP.06.d [Smallness of quantity/amount/degree];
133	meant	Q1.1 X7+ S6+	02.01.10.02 [0.76923077] [Mean]; 02.04.11.03.02 [0.91304348] [Lament/express grief]; 02.05.04 [0.92307692] [Intend];	AR.29 [Meaning]; AU.15.a [Expression of grief]; AV.03 [Intention];
134	amiss	A5.3-	02.03.03.03-03 [0.90000000] [in a way that falls short]; 02.01.12.08.06.01-02 [0.92000000] [in a wrong way, amiss]; 02.01.12.08.06.01.01 [0.95833333] [Wrongly, erroneously];	AT.09 [Inferiority/baseness]; AR.41.a [Lack of truth, falsity, error]; AR.41.a [Lack of truth, falsity, error];
135	Forget	X2.2-	02.01.11.04-01 [0.93750000] [forget, cease to know]; 02.01.11.04-02 [0.93750000] [forget, fail to remember]; 01.15.09.02-03 [0.96000000] [leave undone/fail to perform/carry out];	AR.35.e [Faulty recollection, forgetting]; AR.35.e [Faulty recollection, forgetting]; AO.10.b [Not doing];
136	not	Z6	04.04 [];	ZD [Negative];
137	yet	T1.1.2	01.13.08.03-06 [0009.3068120975] [as formerly/still/to this day]; 02.02.06.01.03-08 [0007.9123253042] [however, nevertheless, notwithstanding]; 01.13.08.06-04 [0002.6702202192] [at some future time/one day];	AM.08.b [The present (time)]; AS.05.b [Qualification, condition, reservation]; AM.08 [Relative time];
138	!	PUNC		
139	S_END	Z99		
140	S_BEGIN	Z99		

141	Forget	X2.2-	02.01.11.04-01 [0.92857143] [forget, cease to know]; 02.01.11.04-02 [0.92857143] [forget, fail to remember]; 01.15.09.02-03 [0.95652174] [leave undone/fail to perform/carry out];	AR.35.e [Faulty recollection, forgetting]; AR.35.e [Faulty recollection, forgetting]; AO.10.b [Not doing];
142	not	Z6	04.04 [];	ZD [Negative];
143	then	N4 Z5 T1.2	01.13.08.05.02-02 [0.95238095] [next in order/then]; 01.11.03.06 [1.00000000] [Consequently]; 01.13.08.05.02-02.01 [1.00000000] [in a series];	AM.08.d.01 [Succession/following in time]; AK.03.f [Effect/result/consequence]; AM.08.d.01 [Succession/following in time];
144	thine	Z99	01.11.03.01 [For that reason/therefore]	AK.03.a [Cause/reason]
145	own	A9+ S5-	02.06.01-03 [0.93750000] [own]; 02.06.01-03.04 [0.95454545] [own private]; 02.04.13.04 [1.00000000] [Used to a loved one];	AW.01 [Owning]; AW.01 [Owning]; AU.27.c [Terms of endearment];
146	approved	S7.4+ E2+ G1.1	02.02.09.04.01-02 [0.95454545] [approved/accepted]; 02.01.13.08.11.01.01-01 [1.00000000] [certified, verified]; 02.02.06.03-08 [1.00000000] [proved];	AS.12.d [Approval/sanction]; AR.49 [Absence of doubt, confidence]; AS.05.d [Proof];
147	,	PUNC		
148	The	Z5	04.03 [Grammatical]	ZC [Grammatical Item];
149	which	Z8	01.16.01-16 [.which/what/who]; 04.06 [Pronoun]; 04.03 [Grammatical]	AP.01 [Relation/relationship]; ZF [Pronoun]; ZC [Grammatical Item];
150	so	A13.3	01.16.01.09-01 [0.95000000] [so/in such a manner]; 01.11.03.01 [1.00000000] [For that reason/therefore]; 01.13.08.05-10 [1.00000000] [after/afterwards/later];	AP.01.i [Similarity]; AK.03.a [Cause/reason]; AM.08.d [The future/time to come];
151	long	T1.3+	01.13.02-04 [0.92857143] [for a long time];	AM.02 [Duration];
152	hath	Z5	04.03 [Grammatical];	ZC [Grammatical Item];
153	thee	Z8mf	04.06 [];	ZF [Pronoun];
154	so	Z5 A13.3	04.03 [];	ZC [Grammatical Item];
155	loved	E2+	02.04.13 [0.90000000] [Love]; 02.04.13.10 [0.91666667] [Be in love]; 02.04.13.10 [0.92857143] [Be in love/infatuated with];	AU.27 [Love]; AU.29 [Amorous love]; AU.29 [Amorous love];
156	,	PUNC		
157	Whose	Z8	04.06 [];	ZF [Pronoun];

158	steadfast	A2.1-	02.04.13.05-09 [0.88235294] [steadfast/constant in affection]; 01.14.09-11.04 [0.94117647] [firmly fixed]; 01.17.04.01.02-13 [0.94736842] [steadfast];	AU.27.d [Affection, tenderness]; AN.09 [Absence of movement]; AQ.04.a.02 [Nature/attributes of God];
159	faith	E6+ S9 X2.1	03.08 [0.90000000] [Faith]; 03.08.01.05.01 [0.93333333] [A religion/church]; 02.01.13-02 [1.00000000] [system of belief, creed];	BF [Faith]; BF.03.a [A religion/church]; AR.45 [Belief, opinion];
160	yet	T1.1.2	02.02.06.01.03-08 [0011.4410483019] [however, nevertheless, notwithstanding]; 01.13.08.06-04 [0000.5853062571] [at some future time/one day]; 01.13.08.03-06 [0000.4032792314] [as formerly/still/to this day];	AS.05.b [Qualification, condition, reservation]; AM.08 [Relative time]; AM.08.b [The present (time)];
161	never	T1/Z6	01.16.06.07-16 [0.94444444] [not at all]; 01.13.10.01-03 [0.95000000] [never];	AP.06.d [Smallness of quantity/amount/degree]; AM.10.b [Infrequency];
162	moved	M1 E1	01.14 [0.91666667] [Move/be in motion]; 01.14 [0.92857143] [Cause to move/set in motion]; 01.14.01 [0.93333333] [Move body/members];	AN [Movement]; AN [Movement]; AN.01 [Bodily movement];
163	:	PUNC		
164	Forget	X2.2-	02.01.11.04-01 [0.94117647] [forget, cease to know]; 02.01.11.04-02 [0.94117647] [forget, fail to remember]; 01.15.09.02-03 [0.96153846] [leave undone/fail to perform/carry out];	AR.35.e [Faulty recollection, forgetting]; AR.35.e [Faulty recollection, forgetting]; AO.10.b [Not doing];
165	not	Z6	04.04 [];	ZD [Negative];
166	this	Z8	04.03 [];	ZC [Grammatical Item];
167	!	PUNC		

Table A12 Keats positive LL

Rank	tid	thematicheading	AV	Exp AV	LL
1	18	Waterlogged/wild land	53	7.12	128.22
2	1290	Quality of having sides/being a side	79	17.15	122.98
3	700	Wine	24	1	122.02
4	1221	Distance/farness	87	24.79	97.72
5	1737	Deceptive fancy, illusion	18	0.79	88.59
6	1348	Newness/novelty, recency	120	46.54	83.11
7	1148	Absence of colour	51	11.89	73.4
8	928	Quality of voice	45	9.39	73.04
9	2067	Melancholy	35	5.85	70.41
10	830	Physical comfort	35	5.85	70.41
11	2039	Pleasure	114	48.18	66.84
12	1736	Fancy, fantastic notion	28	4.13	62.82
13	3741	Volume	30	5.24	58
14	906	Inaudibility	104	45.57	56.53
15	1271	Surface	35	8.12	50.63
16	3871	Poem/piece of poetry	22	3.63	44.8
17	1011	Coldness	68	27.47	43.62
18	1440	Quietness/tranquillity	119	61.72	42.93
19	2056	Moaning/groaning	20	3.11	42.92
20	1665	Separation/detachment, loosening/unfastening	21	3.47	42.69
21	1280	Closed/shut condition	34	9.1	41.47
22	1277	Condition of being open/not closed	48	16.4	41.33
23	294	Wasting disease	30	7.3	41.05
24	3874	Rhyme	15	1.86	38.63
25	3600	Hiring/letting out	30	7.76	38.18
26	3762	Singing	106	55.35	37.54
27	163	Procreation/reproduction	18	2.93	37.13
28	1506	Lack of violence/severity/intensity	54	21.46	35.75
29	1475	Preservation from injury/destruction	22	4.67	35.05
30	1052	Gas/air in liquid/effervescence	15	2.11	34.99

Table A13 Keats neg LL

Rank	tid	thematicheading	AV	Exp AV	LL
1	2079	Love	8	93.18	132.87
2	142	Life, process of living	29	118.81	99.4
3	2584	Command/bidding	1	37	65.58
4	1323	Year	1	35.61	62.85
5	1126	Light	14	65.78	61.18
6	1340	The present (time)	21	74.13	54.16
7	1582	End/conclusion	2	31.71	49.01
8	1316	Duration	33	83.65	40.62
9	1267	Relative position	1	23.86	39.88
10	3270	Travel in specific course/direction	9	37.36	31.59
11	1382	Swiftness	3	24.25	30.39
12	1233	Vertical extent	2	21.3	29.54
13	1312	Direction	44	88.33	27.84
14	1193	State/condition	21	54.26	27.12
15	2087	Amorous love	16	45.12	25.49
16	837	Bed related to sleep/rest	7	28.65	23.95
17	2105	Pride, arrogance	1	15.4	23.64
18	1988	Goodness, acceptability	32	66.85	22.97
19	911	Resonance/sonority	1	14.79	22.49
20	255	Breathing	10	32.57	21.89
21	1127	Intensity of light	13	37.09	21.28
22	1209	Cause	90	140.52	21.26
23	2931	Soul	2	16.69	21.2
24	1390	Forward movement	1	13.79	20.61
25	1243	Roundness	11	33.32	20.6
26	900	Appearance/aspect	33	64.71	19.34
27	1703	Character, disposition, mood	15	38.67	19.27
28	1180	Existence	7	25.38	19.04
29	1634	Greatness of quantity/amount/degree	11	31.94	18.74
30	1141	Darkness/absence of light	4	19.03	17.86

Table A14 Raw count

s1	heading	CG1	$CG1_n$	CG2	$CG2_n$	CG3	$CG3_n$	CG4	$CG4_n$
AA	The world	2752	633	2715	714	4790	1027	5911	1073
AB	Life	4760	1095	4418	1161	5746	1232	7880	1431
AC	Health and disease	637	147	443	116	482	103	710	129
AD	People	1695	390	1490	392	2091	448	2273	413
AE	Animals	666	153	516	136	778	167	815	148
AF	Plants	1196	275	869	228	1193	256	2081	378
AG	Food and drink	1552	357	1400	368	1750	375	3239	588
AH	Textiles and clothing	276	64	286	75	313	67	497	90
AI	Physical sensation	8900	2048	6092	1601	9988	2141	10985	1994
AJ	Matter	6065	1395	5355	1408	8268	1772	11175	2029
AK	Existence and causation	7185	1653	6157	1619	6014	1289	7716	1401
AL	Space	6063	1395	7136	1876	9524	2041	12815	2327
AM	Time	11042	2541	8117	2134	9866	2115	13091	2377
AN	Movement	4916	1131	4384	1152	6322	1355	8993	1633
AO	Action/operation	9521	2191	7379	1940	7672	1645	8288	1505
AP	Relative properties	7657	1762	6131	1612	6569	1408	7717	1401
AQ	The supernatural	701	161	746	196	536	115	912	166
AR	The mind	8841	2034	7270	1911	9119	1955	10624	1929
AS	Attention, judgment, curiosity	7555	1738	5031	1323	4902	1051	4380	795

AT	Goodness and badness	2335	537	1762	463	1500	322	1899	345
AU	Emotion	13452	3095	8384	2204	10850	2326	12185	2212
AV	Will	1589	366	1182	311	953	204	1281	233
AW	Possession/ownership	3037	699	2281	600	1742	373	1969	357
AX	Language	1588	365	1799	473	2028	435	2225	404
AY	Society	1740	400	2317	609	1801	386	1865	339
AZ	Inhabiting and dwelling	1012	233	1039	273	1145	245	1759	319
BA	Armed hostility	1514	348	1549	407	1437	308	1745	317
BB	Authority	2214	509	2258	594	1988	426	2334	424
BC	Law	145	33	95	25	101	22	183	33
BD	Morality	1616	372	1151	303	853	183	924	168
BE	Education	181	42	101	27	131	28	318	58
BF	Faith	1486	342	1475	388	1317	282	1399	254
BG	Communication	2940	676	2473	650	2918	625	4420	802
BH	Travel and travelling	2584	595	2413	634	3132	671	4277	777
BI	Occupation and work	1725	397	1425	375	2083	446	3490	634
BJ	Trade and finance	876	202	840	221	1115	239	2166	393
BK	Leisure	4350	1001	3374	887	5735	1229	11884	2158

Table A15 Offset aggregate

s1	Label	CG1	$CG1_n$	CG2	$CG2_n$	CG3	$CG3_n$	CG4	$CG4_n$
AA	The world	87.01	33.80	79.39	36.81	140.91	51.45	147.06	44.28
AB	Life	74.03	28.76	74.61	34.59	100.10	36.55	110.87	33.39
AC	Health and disease	26.24	10.19	9.54	4.42	8.31	3.03	8.79	2.65
AD	People	94.03	36.52	54.62	25.33	99.49	36.33	126.94	38.22
AE	Animals	14.02	5.45	10.87	5.04	18.96	6.92	17.38	5.23
AF	Plants	17.11	6.65	15.44	7.16	22.39	8.17	46.43	13.98
AG	Food and drink	29.36	11.40	24.55	11.38	36.11	13.18	69.24	20.85
AH	Textiles and clothing	3.15	1.22	3.57	1.66	4.04	1.47	5.79	1.74
AI	Physical sensation	186.05	72.26	136.69	63.38	201.46	73.55	222.90	67.12
AJ	Matter	85.61	33.25	75.19	34.86	124.61	45.50	156.31	47.07
AK	Existence and causation	135.83	52.76	121.22	56.20	118.98	43.44	143.31	43.15
AL	Space	44.63	17.33	57.18	26.51	74.45	27.18	105.72	31.84
AM	Time	203.40	79.00	156.08	72.37	201.73	73.65	249.21	75.04
AN	Movement	65.57	25.47	52.63	24.40	62.42	22.79	72.78	21.92
AO	Action/operation	130.04	50.51	101.76	47.18	118.93	43.42	112.26	33.80
AP	Relative properties	101.58	39.46	87.73	40.67	110.32	40.28	123.19	37.10
AQ	The supernatural	24.54	9.53	22.51	10.44	17.27	6.31	15.81	4.76
AR	The mind	281.12	109.19	261.83	121.40	352.49	128.70	394.13	118.68
AS	Attention, judgment, curiosity	123.10	47.81	100.04	46.38	113.75	41.53	107.34	32.32
AT	Goodness and badness	43.86	17.04	30.92	14.33	25.35	9.26	29.38	8.85

AU	Emotion	247.34	96.07	163.00	75.58	230.32	84.09	252.21	75.95
AV	Will	30.81	11.97	19.71	9.14	16.03	5.85	21.90	6.59
AW	Possession/ownership	14.37	5.58	10.41	4.83	11.35	4.14	7.47	2.25
AX	Language	12.61	4.90	16.19	7.51	14.71	5.37	19.74	5.95
AY	Society	29.86	11.60	43.67	20.25	30.82	11.25	31.36	9.44
AZ	Inhabiting and dwelling	30.67	11.91	37.62	17.44	40.45	14.77	43.20	13.01
BA	Armed hostility	54.43	21.14	49.88	23.13	50.14	18.31	42.81	12.89
BB	Authority	22.68	8.81	23.90	11.08	24.03	8.77	20.89	6.29
BC	Law	1.64	0.64	1.04	0.48	1.41	0.51	1.41	0.42
BD	Morality	75.25	29.23	44.57	20.66	22.24	8.12	20.28	6.11
BE	Education	3.92	1.52	4.58	2.12	3.93	1.43	8.56	2.58
BF	Faith	59.33	23.05	59.25	27.47	58.04	21.19	60.55	18.23
BG	Communication	86.03	33.42	83.03	38.49	89.69	32.75	130.30	39.23
BH	Travel and travelling	43.85	17.03	35.26	16.35	50.04	18.27	48.68	14.66
BI	Occupation and work	17.50	6.80	18.89	8.76	49.36	18.02	60.04	18.08
BJ	Trade and finance	10.89	4.23	11.35	5.26	11.30	4.13	23.61	7.11
BK	Leisure	63.17	24.54	58.04	26.91	82.99	30.30	263.09	79.22
Total		2574.63	1000.00	2156.78	1000.00	2738.92	1000.00	3320.97	1000.00

Table A16 Log-likelihood

s1	Label	CG1	CG1 Normalised	CG2	CG2 Normalised	CG3	CG3 Normalised	CG4	CG4 Normalised
AA	The world	1684	18.9307073	1761	26.8994593	4791	60.1944919	3274	30.4334489
AB	Life	3605	40.5256531	3068	46.864021	3314	41.6373505	3731	34.6814899
AC	Health and disease	299	3.36121228	133	2.03158892	92	1.15589506	112	1.04109538
AD	People	566	6.36269616	1086	16.5887636	1715	21.5473917	627	5.82827504
AE	Animals	657	7.38567382	451	6.88907219	739	9.28485275	654	6.07925339
AF	Plants	389	4.37294842	650	9.92881801	818	10.2774148	1236	11.4892312
AG	Food and drink	621	6.98097936	741	11.3188525	1106	13.8958689	2187	20.3292464
AH	Textiles and clothing	70	0.78690589	168	2.56621758	202	2.53794351	335	3.11399065
AI	Physical sensation	4797	53.9255362	3423	52.2866832	6325	79.4677857	7031	65.3566216
AJ	Matter	2240	25.1809884	1901	29.0379739	3275	41.1473515	5508	51.1995836
AK	Existence and causation	7488	84.1764468	5146	78.6056884	5721	71.8790833	7477	69.5024122
AL	Space	2586	29.0705517	2602	39.7458223	5108	64.1773043	7070	65.7191459
AM	Time	10417	117.102837	6830	104.328965	8772	110.212082	11679	108.56208
AN	Movement	2094	23.5397275	1145	17.4899948	2562	32.1891647	4040	37.5537977
AO	Action/operation	6429	72.2716849	5747	87.7860263	5411	67.9842195	4138	38.4647561
AP	Relative properties	5672	63.7618598	4053	61.9099991	2860	35.9332596	6785	63.06993
AQ	The supernatural	865	9.72390845	556	8.49295818	703	8.83254598	974	9.05381162
AR	The mind	4835	54.3527137	3879	59.2521309	3866	48.5727209	7244	67.336562
AS	Attention, judgment, curiosity	6731	75.6666217	3209	49.0178108	3487	43.8109358	4628	43.0195484

AT	Goodness and badness	2731	30.7005711	1171	17.8871475	1527	19.1853453	1944	18.0704413
AU	Emotion	11140	125.230451	4789	73.1524761	6084	76.4398432	8469	78.7235427
AV	Will	1549	17.4131031	359	5.48376256	773	9.71203136	830	7.71526041
AW	Possession/ownership	707	7.94774945	386	5.89619039	413	5.18896371	351	3.26271856
AX	Language	1400	15.7381177	2140	32.6887239	1403	17.6273997	1128	10.4853178
AY	Society	638	7.17208508	1361	20.7894174	759	9.53613428	779	7.2411902
AZ	Inhabiting and dwelling	614	6.90228877	479	7.31677512	839	10.5412604	1026	9.53717733
BA	Armed hostility	370	4.15935968	487	7.43897596	345	4.33460649	296	2.75146636
BB	Authority	1168	13.1300868	1241	18.9564049	777	9.76228767	654	6.07925339
BC	Law	40	0.44966051	12	0.18330126	0	0	32	0.29745582
BD	Morality	1162	13.0626377	441	6.73632114	282	3.54306966	282	2.62132944
BE	Education	21	0.23607177	14	0.21385146	20	0.25128154	228	2.11937274
BF	Faith	606	6.81235667	1050	16.0388599	493	6.19408986	543	5.0474535
BG	Communication	842	9.46535366	1550	23.6764122	625	7.85254799	2176	20.226996
BH	Travel and travelling	1181	13.2762264	1283	19.5979592	1448	18.1927832	799	7.42710008
BI	Occupation and work	271	3.04644993	284	4.33812972	642	8.0661373	1226	11.3962762
BJ	Trade and finance	317	3.56355951	271	4.13955336	229	2.87717359	684	6.35811822
BK	Leisure	2154	24.2142183	1599	24.4248923	2066	25.9573827	7402	68.8052501
Total		88956	1000	65466	1000	79592	1000	10757 9	1000

Table A17 Blake [486] tagged

Word	Position _1	Position _2	Position _3	Position _4	Position _5	Position _6
486						
.						
Reeds	Pipe; [made of straw; n; 1384-2000]; †3786; [»215278]; #57.14/0.00	Pipe; [chanter reed; n; 1727-2000]; †3786; [»215283]; #57.14/0.00	Wind-instrument; [reed instrument; n; 1838-2000]; †3785; [»215346]; #13.64/0.09	Wind-instrument; [single reed; n; 1837-2000]; †3785; [»215347]; #13.64/0.09	Wind-instrument; [double reed; n; 1530-2000]; †3785; [»215348]; #13.64/0.09	
of						
Innocence	Simple-mindedness; [Simplicity, simple-mindedness; n; 1385-2000]; †1764; [»117696]; #16.67/0.00	Innocence; [Innocence; n; 1340-2000]; †2824; [»171168]; #12.50/0.14	Innocence; [person; n; 1400-2000]; †2824; [»171169]; #12.50/0.14	Innocence; [freedom from guilt; n; 1559-2000]; †2824; [»171175]; #12.50/0.14		
PIPING	Playing wind instrument; [playing pipe/whistle; n; 1275-2000]; †3765; [»214731]; #13.33/0.00	Playing wind instrument; [playing pipe; aj; 1638-2000]; †3765; [»214745]; #13.33/0.00				
down						
the						
valleys	Valley; [Valley; n; 1297-2000]; †16; [»810]; #4.00/0.00					

wild	Commotion/disturbance/disorder; [In a state of commotion/disorder; aj; 1597-2000]; †1574; [»106619]; #31.58/0.00	Madness, extreme folly; [Madly foolish; aj; 1515-2000]; †1765; [»117704]; #20.00/0.03	Quality of voice; [loud/resonant; aj; 1549-2000]; †928; [»61051]; #10.26/0.05	State of Sea; [agitated; aj; 1205-2000]; †35; [»2026]; #5.26/0.08	Rebelliousness; [Rebellious; aj; 1300-2000]; †2702; [»164179]; #5.26/0.08	Profligacy/dissoluteness/debauchery; [Profligate/dissolute; aj; 1250-2000]; †2838; [»171899]; #5.26/0.08
,						
Piping	Playing wind instrument; [playing pipe/whistle; n; 1275-2000]; †3765; [»214731]; #13.33/0.00	Playing wind instrument; [playing pipe; aj; 1638-2000]; †3765; [»214745]; #13.33/0.00				
songs	Sound/bird defined by; [song; n; 1000-2000]; †481; [»35004]; #36.96/0.00	Commotion/disturbance/disorder; [instance of; n; 1843-2000]; †1574; [»106611]; #31.58/0.09	Singing; [Singing; n; 1000-2000]; †3762; [»214517]; #17.95/0.18	Singing; [instance of; n; 1000-2000]; †3762; [»214518]; #17.95/0.18		
of						
pleasant	Merriment; [Merry; aj; 1530-1782]; †2048; [»128902]; #23.08/0.00	Cause of laughter, joking; [comical; aj; 1583-1760]; †2050; [»129044]; #5.75/0.17	Cause of laughter, joking; [humorous/jesting; aj; 1530-1782]; †2050; [»129049]; #5.75/0.17	Cause of laughter, joking; [jest/joke; vi; 1845-1845]; †2050; [»129065]; #5.75/0.17		
glee	Merriment; [Merriment; n; 1200-2000]; †2048; [»128895]; #23.08/0.00					
,						
On						
a						

cloud	Dimness/absence of brightness; [Grow dim/lose brightness; vi; 1562-2000]; †1142; [»74392]; #6.90/0.00	Unsubstantiality/abstractness; [something lacking substance; n; 1382-2000]; †1184; [»76038]; #6.45/0.04	Cloud; [Cloud; n; 1340-2000]; †135; [»7737]; #5.19/0.08	Cloud; [a cloud; n; 1300-2000]; †135; [»7744]; #5.19/0.08	Cloud; [Become cloudy/overcast; vi; 1562-2000]; †135; [»7807]; #5.19/0.08	Cloud; [Cloud/overcast; vt; 1593-2000]; †135; [»7812]; #5.19/0.08
I						
saw	A companion/associate; [Accompany/associate with; vt; 1300-2000]; †2304; [»148215]; #11.11/0.00	Action of informing; [become informed; vt; 1426-2000]; †3138; [»183494]; #9.76/0.02	Reading; [Read; vt; 1300-2000]; †3215; [»186544]; #9.41/0.04	Reading; [refer to another text for more information; vt; 1608-2000]; †3215; [»186570]; #9.41/0.04	Attention; [Call for attention; in; 1000-2000]; †1900; [»122368]; #8.08/0.09	Understanding; [Understand; vi; 1300-2000]; †1743; [»117114]; #7.69/0.11
a						
child	Youth/young man; [Youth/young man; n; 1382-1888]; †365; [»39886]; #40.00/0.00	Girl; [Girl; n; 1611-2000]; †369; [»39938]; #14.29/0.06	Baby/infant; [Baby/infant; n; 1000-2000]; †370; [»39952]; #14.29/0.06	Baby/infant; [baby girl; n; 1611-2000]; †370; [»39957]; #14.29/0.06		
,						
And						
he						
laughing	Manner of speaking; [with a sneer/laugh, etc.; vt; 1843-2000]; †2210; [»142205]; #8.46/0.00	Laughter; [Laughter; n; 1690-2000]; †2049; [»128921]; #5.21/0.17	Laughter; [instance of; n; 1713-2000]; †2049; [»128922]; #5.21/0.17	Laughter; [Laugh; vi; 1000-2000]; †2049; [»128938]; #5.21/0.17	Laughter; [of inanimate objects; vi; 1386-2000]; †2049; [»128939]; #5.21/0.17	Laughter; [utter with laughter; vt; 1843-2000]; †2049; [»128947]; #5.21/0.17

said	Recitation; [Recite; vt; 1200-2000]; †2215; [»142446]; #16.00/0.00	Action of informing; [inform (a person); vt; 1200-1896]; †3138; [»183476]; #9.76/0.05	Suggestion, proposal; [by uttering; vt; 1596-2000]; †1853; [»120837]; #7.69/0.10	Supposition, surmise, presumption; [supposing; v; 1596-2000]; †1850; [»120693]; #5.26/0.14	Speech; [that which is/can be spoken; n; 1571-2000]; †2208; [»141886]; #3.40/0.19	Speech; [what one has intended to say; n; 1692-2000]; †2208; [»141889]; #3.40/0.19
to						
me						
:						
'Pipe	Pipe; [Pipe; n; 1000-2000]; †3786; [»215277]; #57.14/0.00	Pipe; [call/boatswain's whistle; n; 1638-2000]; †3786; [»215290]; #57.14/0.00	Sound/bird defined by; [thin/shrill; n; 1721-2000]; †481; [»34999]; #36.96/0.05	Sound/bird defined by; [make shrill sound; vi; 1591-2000]; †481; [»35026]; #36.96/0.05	Weeping; [Weep; vi; 1797-2000]; †2057; [»129421]; #23.61/0.09	Singing; [sing shrilly/harshly; vt; 1567-2000]; †3762; [»214651]; #17.95/0.11
a						
song	Sound/bird defined by; [song; n; 1000-2000]; †481; [»35004]; #36.96/0.00	Commotion/disturbance/disorder; [instance of; n; 1843-2000]; †1574; [»106611]; #31.58/0.09	Singing; [Singing; n; 1000-2000]; †3762; [»214517]; #17.95/0.18	Singing; [instance of; n; 1000-2000]; †3762; [»214518]; #17.95/0.18		
about						
a						
Lamb	Innocence; [person; n; 1000-2000]; †2824; [»171169]; #12.50/0.00	Terms of endearment; [Terms of endearment; n; 1553-1820]; †2082; [»130547]; #11.54/0.11				
!						

'						
So	Manner of action; [in this way; av; 1250-2000]; †1501; [»84451]; #17.91/0.00	Manner of action; [in this very manner; av; 1598-2000]; †1501; [»84452]; #17.91/0.00	Manner of action; [in that way; av; 1000-2000]; †1501; [»84453]; #17.91/0.00			
I						
piped	Sound/bird defined by; [chirp/twitter; vi; 1659-2000]; †481; [»35027]; #36.96/0.00	Spot of colour; [small spot/speckle; n; 1676-2000]; †1175; [»75554]; #10.81/0.11				
with						
merry	Merriment; [Merry; aj; 1320-2000]; †2048; [»128902]; #23.08/0.00	Merriment; [Cause to be merry; vt; 1310-1677]; †2048; [»128918]; #23.08/0.00				
cheer	Merriment; [Merriment; n; 1393-1842]; †2048; [»128895]; #23.08/0.00	Consolation, relief; [consolation/relief; n; 1549-1863]; †2044; [»128656]; #21.05/0.04	Consolation, relief; [console/relieve; vt; 1430-2000]; †2044; [»128678]; #21.05/0.04	Consolation, relief; [as food/drink; vt; 1548-2000]; †2044; [»128679]; #21.05/0.04	Consolation, relief; [Console; vr; 1400-1846]; †2044; [»128685]; #21.05/0.04	
.						
'Piper	Smoking; [pipe-smoker; n; 1632-2000]; †866; [»58213]; #11.97/0.00	Wind player; [piper/bagpiper; n; 1000-2000]; †3779; [»215147]; #8.00/0.14				

,						
pipe	Pipe; [Pipe; n; 1000-2000]; †3786; [»215277]; #57.14/0.00	Pipe; [call/boatswain's whistle; n; 1638-2000]; †3786; [»215290]; #57.14/0.00	Sound/bird defined by; [thin/shrill; n; 1721-2000]; †481; [»34999]; #36.96/0.05	Sound/bird defined by; [make shrill sound; vi; 1591-2000]; †481; [»35026]; #36.96/0.05	Weeping; [Weep; vi; 1797-2000]; †2057; [»129421]; #23.61/0.09	Singing; [sing shrilly/harshly; vt; 1567-2000]; †3762; [»214651]; #17.95/0.11
that						
song	Sound/bird defined by; [song; n; 1000-2000]; †481; [»35004]; #36.96/0.00	Commotion/disturbance/disorder; [instance of; n; 1843-2000]; †1574; [»106611]; #31.58/0.09	Singing; [Singing; n; 1000-2000]; †3762; [»214517]; #17.95/0.18	Singing; [instance of; n; 1000-2000]; †3762; [»214518]; #17.95/0.18		
again	Giving back, restitution; [Back (of giving); av; 1300-1662]; †2183; [»139645]; #9.09/0.00	Return; [Return; av; 1000-2000]; †3274; [»189190]; #7.41/0.08	Reciprocal treatment/return of an action; [In return/back; av; 1000-2000]; †1531; [»86971]; #5.26/0.15			
;						
'						
So	Manner of action; [in this way; av; 1250-2000]; †1501; [»84451]; #17.91/0.00	Manner of action; [in this very manner; av; 1598-2000]; †1501; [»84452]; #17.91/0.00	Manner of action; [in that way; av; 1000-2000]; †1501; [»84453]; #17.91/0.00			
I						

piped	Sound/bird defined by; [chirp/twitter; vi; 1659-2000]; †481; [»35027]; #36.96/0.00	Spot of colour; [small spot/speckle; n; 1676-2000]; †1175; [»75554]; #10.81/0.11				
:						
he						
wept	Weeping; [fit(s) of weeping; n; 1836-2000]; †2057; [»129382]; #23.61/0.00	Weeping; [Weep; vi; 1000-2000]; †2057; [»129421]; #23.61/0.00	Weeping; [shed tears (of the eyes); vi; 1567-2000]; †2057; [»129426]; #23.61/0.00	Weeping; [Weep for; vt; 1000-2000]; †2057; [»129432]; #23.61/0.00	Weeping; [shed (tears); vt; 1000-2000]; †2057; [»129440]; #23.61/0.00	
to						
hear	Action of informing; [be informed; vi; 1320-2000]; †3138; [»183467]; #9.76/0.00	Action of informing; [be informed of; vt; 1000-2000]; †3138; [»183493]; #9.76/0.00	Hearing/noise; [Hear; vi; 1000-2000]; †903; [»59882]; #8.11/0.15	Hearing/noise; [listen; vi; 1000-2000]; †903; [»59884]; #8.11/0.15	Hearing/noise; [Hear; vt; 1000-2000]; †903; [»59898]; #8.11/0.15	Hearing/noise; [listen to; vt; 1000-2000]; †903; [»59904]; #8.11/0.15
.						
'Drop	Spot of colour; [Spot of colour; n; 1420-1674]; †1175; [»75549]; #10.81/0.00	Spot of colour; [Spot; vt; 1548-1820]; †1175; [»75580]; #10.81/0.00	Action/process of dripping/falling in drops; [Action/process of dripping/falling in drops; n; 1637-1832]; †1046; [»68355]; #10.34/0.04	Action/process of dripping/falling in drops; [Drip/fall in drops; vi; 1000-2000]; †1046; [»68373]; #10.34/0.04	Action/process of dripping/falling in drops; [be dripping; vi; 1300-1825]; †1046; [»68377]; #10.34/0.04	Action/process of sprinkling; [with/as with drops; vt; 1430-1667]; †1045; [»68351]; #4.55/0.11
thy						

pipe	Pipe; [Pipe; n; 1000-2000]; †3786; [»215277]; #57.14/0.00	Pipe; [call/boatswain's whistle; n; 1638-2000]; †3786; [»215290]; #57.14/0.00	Sound/bird defined by; [thin/shrill; n; 1721-2000]; †481; [»34999]; #36.96/0.05	Sound/bird defined by; [make shrill sound; vi; 1591-2000]; †481; [»35026]; #36.96/0.05	Weeping; [Weep; vi; 1797-2000]; †2057; [»129421]; #23.61/0.09	Singing; [sing shrilly/harshly; vt; 1567-2000]; †3762; [»214651]; #17.95/0.11
,						
thy						
happy	Holiness; [Holy, sacred; aj; 1526-1700]; †2919; [»174661]; #23.08/0.00	Happiness; [Happy; aj; 1525-2000]; †2043; [»128603]; #18.25/0.14	Happiness; [made supremely happy; aj; 1526-1700]; †2043; [»128627]; #18.25/0.14			
pipe	Pipe; [Pipe; n; 1000-2000]; †3786; [»215277]; #57.14/0.00	Pipe; [call/boatswain's whistle; n; 1638-2000]; †3786; [»215290]; #57.14/0.00	Sound/bird defined by; [thin/shrill; n; 1721-2000]; †481; [»34999]; #36.96/0.05	Sound/bird defined by; [make shrill sound; vi; 1591-2000]; †481; [»35026]; #36.96/0.05	Weeping; [Weep; vi; 1797-2000]; †2057; [»129421]; #23.61/0.09	Singing; [sing shrilly/harshly; vt; 1567-2000]; †3762; [»214651]; #17.95/0.11
;						
Sing	Sound/bird defined by; [sing; vi; 1000-2000]; †481; [»35028]; #36.96/0.00	Singing; [instance of; n; 1850-2000]; †3762; [»214518]; #17.95/0.05	Singing; [gift/power of singing; n; 1850-2000]; †3762; [»214521]; #17.95/0.05	Singing; [Sing; vi; 1000-2000]; †3762; [»214591]; #17.95/0.05	Singing; [Sing; vt; 1000-2000]; †3762; [»214628]; #17.95/0.05	Singing; [spend in singing; vt; 1816-1816]; †3762; [»214631]; #17.95/0.05
thy						

songs	Sound/bird defined by; [song; n; 1000-2000]; †481; [»35004]; #36.96/0.00	Commotion/disturbance/disorder; [instance of; n; 1843-2000]; †1574; [»106611]; #31.58/0.09	Singing; [Singing; n; 1000-2000]; †3762; [»214517]; #17.95/0.18	Singing; [instance of; n; 1000-2000]; †3762; [»214518]; #17.95/0.18		
of						
happy	Holiness; [Holy, sacred; aj; 1526-1700]; †2919; [»174661]; #23.08/0.00	Happiness; [Happy; aj; 1525-2000]; †2043; [»128603]; #18.25/0.14	Happiness; [made supremely happy; aj; 1526-1700]; †2043; [»128627]; #18.25/0.14			
cheer	Merriment; [Merriment; n; 1393-1842]; †2048; [»128895]; #23.08/0.00	Consolation, relief; [consolation/relief; n; 1549-1863]; †2044; [»128656]; #21.05/0.04	Consolation, relief; [console/relieve; vt; 1430-2000]; †2044; [»128678]; #21.05/0.04	Consolation, relief; [as food/drink; vt; 1548-2000]; †2044; [»128679]; #21.05/0.04	Consolation, relief; [Console; vr; 1400-1846]; †2044; [»128685]; #21.05/0.04	
!						
'						
So	Manner of action; [in this way; av; 1250-2000]; †1501; [»84451]; #17.91/0.00	Manner of action; [in this very manner; av; 1598-2000]; †1501; [»84452]; #17.91/0.00	Manner of action; [in that way; av; 1000-2000]; †1501; [»84453]; #17.91/0.00			
I						

sung	Sound/bird defined by; [sing; vi; 1000-2000]; †481; [»35028]; #36.96/0.00	Singing; [instance of; n; 1850-2000]; †3762; [»214518]; #17.95/0.05	Singing; [gift/power of singing; n; 1850-2000]; †3762; [»214521]; #17.95/0.05	Singing; [Sing; vi; 1000-2000]; †3762; [»214591]; #17.95/0.05	Singing; [Sing; vt; 1000-2000]; †3762; [»214628]; #17.95/0.05	Singing; [spend in singing; vt; 1816-1816]; †3762; [»214631]; #17.95/0.05
the						
same						
again	Giving back, restitution; [Back (of giving); av; 1300-1662]; †2183; [»139645]; #9.09/0.00	Return; [Return; av; 1000-2000]; †3274; [»189190]; #7.41/0.08	Reciprocal treatment/return of an action; [In return/back; av; 1000-2000]; †1531; [»86971]; #5.26/0.15			
,						
While						
he						
wept	Weeping; [fit(s) of weeping; n; 1836-2000]; †2057; [»129382]; #23.61/0.00	Weeping; [Weep; vi; 1000-2000]; †2057; [»129421]; #23.61/0.00	Weeping; [shed tears (of the eyes); vi; 1567-2000]; †2057; [»129426]; #23.61/0.00	Weeping; [Weep for; vt; 1000-2000]; †2057; [»129432]; #23.61/0.00	Weeping; [shed (tears); vt; 1000-2000]; †2057; [»129440]; #23.61/0.00	
with						
joy	Congratulation; [an instance/expression of; n; 1656-1656]; †1940; [»126255]; #26.67/0.00	Congratulation; [Congratulate (a person); vt; 1483-1701]; †1940; [»126266]; #26.67/0.00	Rejoicing/exultation; [Rejoicing/exultation; n; 1300-2000]; †2046; [»128771]; #19.64/0.12	Rejoicing/exultation; [Rejoice/exult; vi; 1300-1885]; †2046; [»128806]; #19.64/0.12	Rejoicing/exultation; [Rejoice/exult; vr; 1260-1712]; †2046; [»128823]; #19.64/0.12	
to						

hear	Action of informing; [be informed; vi; 1320-2000]; †3138; [»183467]; #9.76/0.00	Action of informing; [be informed of; vt; 1000-2000]; †3138; [»183493]; #9.76/0.00	Hearing/noise; [Hear; vi; 1000-2000]; †903; [»59882]; #8.11/0.15	Hearing/noise; [listen; vi; 1000-2000]; †903; [»59884]; #8.11/0.15	Hearing/noise; [Hear; vt; 1000-2000]; †903; [»59898]; #8.11/0.15	Hearing/noise; [listen to; vt; 1000-2000]; †903; [»59904]; #8.11/0.15
.						
'Piper	Smoking; [pipe-smoker; n; 1632-2000]; †866; [»58213]; #11.97/0.00	Wind player; [piper/bag piper; n; 1000-2000]; †3779; [»215147]; #8.00/0.14				
,						
sit	Action/fact of sitting; [spell of; n; 1832-2000]; †1307; [»98173]; #14.00/0.00	Action/fact of sitting; [manner of; n; 1820-1894]; †1307; [»98174]; #14.00/0.00	Action/fact of sitting; [Be sitting/seated; vi; 1000-2000]; †1307; [»98193]; #14.00/0.00	Action/fact of sitting; [for specific purpose; vi; 1300-2000]; †1307; [»98199]; #14.00/0.00	Action/fact of sitting; [sit down; vi; 1000-2000]; †1307; [»98201]; #14.00/0.00	Action/fact of sitting; [come to a sitting posture; vi; 1300-2000]; †1307; [»98205]; #14.00/0.00
thee						
down	A song; [refrain; n; 1611-1656]; †3756; [»214190]; #8.33/0.00	Attention; [alert; aj; 1812-1850]; †1900; [»122282]; #8.08/0.03	Hill; [Hill; n; 1000-1653]; †12; [»603]; #4.76/0.07	Hill; [sand-hill; n; 1523-1837]; †12; [»622]; #4.76/0.07	Digestive organs; [down throat/into stomach; av; 1582-2000]; †239; [»25368]; #4.35/0.13	Dejection, depression, melancholy; [Dejected; aj; 1610-2000]; †2066; [»129674]; #4.23/0.17
and						

write	Letter-writing; [Write a letter; vi; 1000-2000]; †3251; [»187713]; #31.58/0.00	Letter-writing; [Write (a letter); vt; 1000-2000]; †3251; [»187715]; #31.58/0.00	Letter-writing; [write (something) in a letter; vt; 1400-2000]; †3251; [»187718]; #31.58/0.00	Writing; [Write; vi; 1000-2000]; †3169; [»184496]; #28.57/0.08	Writing; [Write; vt; 1000-2000]; †3169; [»184503]; #28.57/0.08	Writing; [a letter/symbol; vt; 1000-2000]; †3169; [»184504]; #28.57/0.08
In						
a						
book	Written record; [Written record; n; 1000-1681]; †3153; [»183973]; #14.58/0.00	Register/record-book; [Register/record-book; n; 1498-2000]; †3154; [»184021]; #10.00/0.04	Register/record-book; [Register; vt; 1225-2000]; †3154; [»184036]; #10.00/0.04	Treatise/dissertation; [Treatise/dissertation; n; 1000-2000]; †3882; [»220302]; #8.82/0.12	Opera; [libretto; n; 1768-1882]; †3755; [»214176]; #7.69/0.16	
that						
all						
may						
read	Recitation; [that is read out; aj; 1590-2000]; †2215; [»142439]; #16.00/0.00	Recitation; [merely; vi; 1781-2000]; †2215; [»142445]; #16.00/0.00	Recitation; [read aloud; vt; 1000-2000]; †2215; [»142449]; #16.00/0.00	Character, disposition, mood; [make out character; vt; 1611-2000]; †1703; [»115528]; #12.50/0.08	Interpretation of dreams; [Interpret; vt; 1000-2000]; †1882; [»122002]; #11.11/0.11	Reading; [act/spell of; n; 1825-2000]; †3215; [»186491]; #9.41/0.14
.						
'						
So	Manner of action; [in this way; av; 1250-2000]; †1501; [»84451]; #17.91/0.00	Manner of action; [in this very manner; av; 1598-2000]; †1501; [»84452]; #17.91/0.00	Manner of action; [in that way; av; 1000-2000]; †1501; [»84453]; #17.91/0.00			
he						

vanished	Invisibility; [vanishing/disappearing; n; 1650-1872]; †902; [»59779]; #5.88/0.00	Invisibility; [vanish/disappear; vi; 1375-2000]; †902; [»59820]; #5.88/0.00	Invisibility; [by decay/ceasing to exist; vi; 1350-2000]; †902; [»59821]; #5.88/0.00	Invisibility; [cause to vanish/disappear; vt; 1440-2000]; †902; [»59828]; #5.88/0.00		
from						
my						
sight	Sight/vision; [Sight/vision; n; 1200-2000]; †883; [»58823]; #8.47/0.00	Sight/vision; [range/field of; n; 1200-2000]; †883; [»58824]; #8.47/0.00	Sight/vision; [Sight of something; n; 1000-2000]; †883; [»58868]; #8.47/0.00	Sight/vision; [first sight; n; 1471-2000]; †883; [»58869]; #8.47/0.00	Sight/vision; [glimpse; n; 1205-2000]; †883; [»58871]; #8.47/0.00	A vision; [A vision; n; 1000-1825]; †1734; [»116930]; #7.69/0.13
;						
And						
I						
plucked	Cause; [put out of a state/condition; vt; 1387-2000]; †1209; [»77293]; #9.09/0.00	Pushing and pulling; [a sudden/sharp pull; n; 1435-2000]; †1411; [»103445]; #2.94/0.05	Pushing and pulling; [sharply; vi; 1410-2000]; †1411; [»103486]; #2.94/0.05	Pushing and pulling; [pull; vt; 1377-2000]; †1411; [»103526]; #2.94/0.05	Pushing and pulling; [suddenly/sharply; vt; 1500-2000]; †1411; [»103533]; #2.94/0.05	
a						
hollow	Hollowness; [a cavity/hollow; n; 1592-2000]; †1252; [»93106]; #10.87/0.00	Hollowness; [Hollow; aj; 1250-2000]; †1252; [»93123]; #10.87/0.00	Hollowness; [beneath; aj; 1657-2000]; †1252; [»93124]; #10.87/0.00	Hollowness; [Make hollow; vt; 1450-2000]; †1252; [»93138]; #10.87/0.00	Hollowness; [form by hollowing out; vt; 1687-2000]; †1252; [»93142]; #10.87/0.00	Curved surface; [a concave surface; n; 1850-1850]; †1246; [»92673]; #9.09/0.19

reed	Pipe; [made of straw; n; 1384-2000]; †3786; [»215278]; #57.14/0.00	Pipe; [chanter reed; n; 1727-2000]; †3786; [»215283]; #57.14/0.00	Wind-instrument; [reed instrument; n; 1838-2000]; †3785; [»215346]; #13.64/0.09	Wind-instrument; [single reed; n; 1837-2000]; †3785; [»215347]; #13.64/0.09	Wind-instrument; [double reed; n; 1530-2000]; †3785; [»215348]; #13.64/0.09	
,						
And						
I						
made	Art/occupation of writer/author; [Be the author of/write (a work); vt; 1175-2000]; †3856; [»218441]; #17.91/0.00	Character, disposition, mood; [Disposition/character; n; 1674-2000]; †1703; [»115493]; #12.50/0.01	A companion/associate; [A companion/associate; n; 1000-1886]; †2304; [»148191]; #11.11/0.02	Cause; [Cause/bring about; vt; 1000-2000]; †1209; [»77284]; #9.09/0.04	Cause; [in/to a person; vt; 1000-1725]; †1209; [»77286]; #9.09/0.04	Cause; [give rise to; vt; 1175-1834]; †1209; [»77299]; #9.09/0.04
a						
rural	Country dweller; [Country dweller; n; 1513-1831]; †2377; [»151234]; #14.29/0.00	Country as opposed to town; [Country as opposed to town; aj; 1590-2000]; †2385; [»151479]; #9.09/0.09	Music; [Pertaining to music; aj; 1470-1738]; †3738; [»213236]; #6.67/0.18			
pen	Art/occupation of writer/author; [Art/occupation of writer/author; n; 1447-2000]; †3856; [»218404]; #17.91/0.00	Writing instrument; [pen; n; 1657-2000]; †3189; [»185343]; #7.55/0.09	Writing instrument; [quill-pen; n; 1300-2000]; †3189; [»185346]; #7.55/0.09	Writing instrument; [separate point; n; 1657-2000]; †3189; [»185364]; #7.55/0.09	Writing instrument; [pencil; n; 1644-1818]; †3189; [»185370]; #7.55/0.09	

,						
And						
I						
stained	Dimness/absence of brightness; [obscure the lustre of; vt; 1589-1657]; †1142; [»74398]; #6.90/0.00	Ornamental glass-work; [layer of colour; n; 1832-2000]; †3837; [»217555]; #5.13/0.05	Ornamental glass-work; [colour; vt; 1797-2000]; †3837; [»217567]; #5.13/0.05	Colouring matter; [stain; n; 1758-2000]; †1152; [»74734]; #5.00/0.16	Colouring matter; [colouring for glass; n; 1832-2000]; †1152; [»74749]; #5.00/0.16	
the						
water	Weeping; [tears; n; 1362-1840]; †2057; [»129391]; #23.61/0.00	Weeping; [Weep; vi; 1500-2000]; †2057; [»129421]; #23.61/0.00	Weeping; [shed tears (of the eyes); vi; 1400-2000]; †2057; [»129426]; #23.61/0.00	Water; [Water; n; 1000-2000]; †707; [»46002]; #16.67/0.05	Water; [contrasted with wine; n; 1300-1842]; †707; [»46004]; #16.67/0.05	Water; [used to dilute liquors; n; 1382-2000]; †707; [»46005]; #16.67/0.05
clear	Happiness; [Free from trouble/care/sorrow; aj; 1374-1853]; †2043; [»128641]; #18.25/0.00	Transparency/translucence; [Transparent/translucent; aj; 1300-2000]; †1139; [»74268]; #15.00/0.01	Transparency/translucence; [become transparent; vi; 1590-2000]; †1139; [»74278]; #15.00/0.01	Transparency/translucence; [Make transparent/translucent; vt; 1440-2000]; †1139; [»74279]; #15.00/0.01	Innocence; [free from guilt; aj; 1400-2000]; †2824; [»171188]; #12.50/0.06	Quality of voice; [clear; aj; 1300-2000]; †928; [»61080]; #10.26/0.07
,						
And						
I						
wrote	Letter-writing; [Write a letter; vi; 1000-2000]; †3251; [»187713]; #31.58/0.00	Letter-writing; [Write (a letter); vt; 1000-2000]; †3251; [»187715]; #31.58/0.00	Letter-writing; [write (something) in a letter; vt; 1400-2000]; †3251; [»187718]; #31.58/0.00	Writing; [Write; vi; 1000-2000]; †3169; [»184496]; #28.57/0.08	Writing; [Write; vt; 1000-2000]; †3169; [»184503]; #28.57/0.08	Writing; [a letter/symbol; vt; 1000-2000]; †3169; [»184504]; #28.57/0.08

my						
happy	Holiness; [Holy, sacred; aj; 1526-1700]; †2919; [»174661]; #23.08/0.00	Happiness; [Happy; aj; 1525-2000]; †2043; [»128603]; #18.25/0.14	Happiness; [made supremely happy; aj; 1526-1700]; †2043; [»128627]; #18.25/0.14			
songs	Sound/bird defined by; [song; n; 1000-2000]; †481; [»35004]; #36.96/0.00	Commotion/disturbance/disorder; [instance of; n; 1843-2000]; †1574; [»106611]; #31.58/0.09	Singing; [Singing; n; 1000-2000]; †3762; [»214517]; #17.95/0.18	Singing; [instance of; n; 1000-2000]; †3762; [»214518]; #17.95/0.18		
Every						
child	Youth/young man; [Youth/young man; n; 1382-1888]; †365; [»39886]; #40.00/0.00	Girl; [Girl; n; 1611-2000]; †369; [»39938]; #14.29/0.06	Baby/infant; [Baby/infant; n; 1000-2000]; †370; [»39952]; #14.29/0.06	Baby/infant; [baby girl; n; 1611-2000]; †370; [»39957]; #14.29/0.06		
may						
joy	Congratulation; [an instance/expression of; n; 1656-1656]; †1940; [»126255]; #26.67/0.00	Congratulation; [Congratulate (a person); vt; 1483-1701]; †1940; [»126266]; #26.67/0.00	Rejoicing/exultation; [Rejoicing/exultation; n; 1300-2000]; †2046; [»128771]; #19.64/0.12	Rejoicing/exultation; [Rejoice/exult; vi; 1300-1885]; †2046; [»128806]; #19.64/0.12	Rejoicing/exultation; [Rejoice/exult; vr; 1260-1712]; †2046; [»128823]; #19.64/0.12	
to						

hear	Action of informing; [be informed; vi; 1320-2000]; †3138; [»183467]; #9.76/0.00	Action of informing; [be informed of; vt; 1000-2000]; †3138; [»183493]; #9.76/0.00	Hearing/noise; [Hear; vi; 1000-2000]; †903; [»59882]; #8.11/0.15	Hearing/noise; [listen; vi; 1000-2000]; †903; [»59884]; #8.11/0.15	Hearing/noise; [Hear; vt; 1000-2000]; †903; [»59898]; #8.11/0.15	Hearing/noise; [listen to; vt; 1000-2000]; †903; [»59904]; #8.11/0.15
.						

Table A18 Blake [488] tagged

Word	Position_1	Position_2	Position_3	Position_4	Position_5	Position_6
488						
.						
Hear	Attention; [give ear, hearken; vi; 1340-1702]; †1900; [»122289]; #15.15/0.00	Attention; [listen attentively; vt; 1000-2000]; †1900; [»122297]; #15.15/0.00	Action of informing; [be informed; vi; 1320-2000]; †3138; [»183467]; #9.76/0.15	Action of informing; [be informed of; vt; 1000-2000]; †3138; [»183493]; #9.76/0.15		
the						
Voice	Voice/vocal sound; [Voice/vocal sound; n; 1300-2000]; †926; [»60933]; #27.66/0.00	Voice/vocal sound; [faculty of producing; n; 1300-2000]; †926; [»60934]; #27.66/0.00	Voice/vocal sound; [of an individual; n; 1290-2000]; †926; [»60935]; #27.66/0.00	Voice/vocal sound; [sound of voices; n; 1831-1831]; †926; [»60937]; #27.66/0.00	Voice/vocal sound; [sound of specific type of utterance; n; 1325-2000]; †926; [»60945]; #27.66/0.00	Voice/vocal sound; [endow with voice; vt; 1711-2000]; †926; [»60971]; #27.66/0.00
HEAR	Attention; [give ear, hearken; vi; 1340-1702]; †1900; [»122289]; #15.15/0.00	Attention; [listen attentively; vt; 1000-2000]; †1900; [»122297]; #15.15/0.00	Action of informing; [be informed; vi; 1320-2000]; †3138; [»183467]; #9.76/0.15	Action of informing; [be informed of; vt; 1000-2000]; †3138; [»183493]; #9.76/0.15		
the						
voice	Voice/vocal sound; [Voice/vocal sound; n; 1300-2000]; †926; [»60933]; #27.66/0.00	Voice/vocal sound; [faculty of producing; n; 1300-2000]; †926; [»60934]; #27.66/0.00	Voice/vocal sound; [of an individual; n; 1290-2000]; †926; [»60935]; #27.66/0.00	Voice/vocal sound; [sound of voices; n; 1831-1831]; †926; [»60937]; #27.66/0.00	Voice/vocal sound; [sound of specific type of utterance; n; 1325-2000]; †926; [»60945]; #27.66/0.00	Voice/vocal sound; [endow with voice; vt; 1711-2000]; †926; [»60971]; #27.66/0.00
of						

the						
Bard	Performance arts; [minstrel; n; 1449-1805]; †3898; [»221759]; #3.92/0.00					
,						
Who						
present	Attention; [alert; aj; 1548-1864]; †1900; [»122282]; #15.15/0.00	Immediacy; [Immediate; aj; 1563-1793]; †1339; [»89054]; #14.81/0.02	Immediacy; [acting immediately; aj; 1555-1694]; †1339; [»89055]; #14.81/0.02	Direction; [Direct; vt; 1793-1853]; †1312; [»98620]; #13.81/0.05	Offering; [Offer; vt; 1533-2000]; †2184; [»139667]; #11.43/0.07	Offering; [present formally for acceptance; vt; 1509-2000]; †2184; [»139685]; #11.43/0.07
,						
past	End/conclusion; [come/brought to an end; aj; 1400-2000]; †1582; [»107028]; #12.50/0.00	The past; [Of/pertaining to the past/passed away/gone; aj; 1400-2000]; †1341; [»89161]; #8.54/0.11	The past; [just passed; aj; 1300-2000]; †1341; [»89163]; #8.54/0.11	The past; [of time/order; aj; 1450-2000]; †1341; [»89164]; #8.54/0.11		
,						
and						
future	Expectation; [expected; aj; 1374-2000]; †1865; [»121286]; #1.19/0.00					
,						
sees	Attention; [Call for attention; in; 1000-2000]; †1900; [»122368]; #15.15/0.00	Action of informing; [become informed; vt; 1426-2000]; †3138; [»183494]; #9.76/0.02	Understanding; [Understand; vi; 1300-2000]; †1743; [»117114]; #7.69/0.04	Understanding; [Understand; vt; 1000-2000]; †1743; [»117117]; #7.69/0.04	Visibility; [remaining to be seen; v; 1818-1846]; †901; [»59717]; #7.44/0.09	Particular time; [be marked by an event; vt; 1739-2000]; †1320; [»88201]; #7.23/0.11
;						
Whose						

ears	Attention; [an attentive ear; n; 1000-2000]; †1900; [»122271]; #15.15/0.00	Ploughing; [Ploughing; n; 1460-1693]; †736; [»47630]; #10.67/0.08	Ploughing; [Plough (of person); vi; 1000-2000]; †736; [»47672]; #10.67/0.08	Ploughing; [Plough (land); vt; 1000-2000]; †736; [»47688]; #10.67/0.08		
have						
heard	Attention; [give ear, hearken; vi; 1340-1702]; †1900; [»122289]; #15.15/0.00	Attention; [listen attentively; vt; 1000-2000]; †1900; [»122297]; #15.15/0.00	Action of informing; [be informed; vi; 1320-2000]; †3138; [»183467]; #9.76/0.15	Action of informing; [be informed of; vt; 1000-2000]; †3138; [»183493]; #9.76/0.15		
The						
Holy	Holiness; [instance of; n; 1000-2000]; †2919; [»174659]; #15.38/0.00	Holiness; [Holy, sacred; aj; 1000-2000]; †2919; [»174661]; #15.38/0.00				
Word	Speech; [Speech; n; 1000-1728]; †2208; [»141870]; #9.22/0.00	Speech; [that which is/can be spoken; n; 1000-2000]; †2208; [»141886]; #9.22/0.00	Speech; [short; n; 1000-2000]; †2208; [»141893]; #9.22/0.00	Speech; [Speak; vi; 1205-2000]; †2208; [»141966]; #9.22/0.00	Speech; [Speak/say/utter; vt; 1400-1849]; †2208; [»142005]; #9.22/0.00	Speech; [give expression to; vt; 1613-1831]; †2208; [»142008]; #9.22/0.00
That						
walked	Style/manner of dancing; [Style/manner of dancing; vt; 1810-2000]; †4023; [»224898]; #20.00/0.00	Speech; [Speak; vi; 1550-1673]; †2208; [»141966]; #9.22/0.01	Place of resort; [usual haunt; n; 1386-1702]; †2399; [»151842]; #8.33/0.03	Path/place for walking; [in a garden/pleasure-ground; n; 1533-2000]; †3301; [»190987]; #8.33/0.03	Path/place for walking; [shaded/bordered by trees; n; 1596-2000]; †3301; [»190992]; #8.33/0.03	Path/place for walking; [promenade; n; 1840-2000]; †3301; [»190996]; #8.33/0.03

among						
the						
ancient	The past; [one who lived in ancient times; n; 1541-2000]; †1341; [»89144]; #8.54/0.00	The past; [Of/pertaining to the past/passed away/gone; aj; 1490-1793]; †1341; [»89161]; #8.54/0.00	The past; [long-past/old; aj; 1366-2000]; †1341; [»89167]; #8.54/0.00	Old person; [Old person; n; 1502-1837]; †373; [»40023]; #5.56/0.16	Oldness/ancientness; [Old; aj; 1586-2000]; †1344; [»89377]; #5.56/0.16	Oldness/ancientness; [having lasted in some capacity; aj; 1413-1807]; †1344; [»89379]; #5.56/0.16
trees	Refuge/shelter; [specifically; vi; 1700-1902]; †1476; [»82198]; #3.53/0.00	Hanging; [gallows; n; 1500-1847]; †2721; [»165164]; #3.03/0.14				
;						
Calling	Death; [carry off by death; vt; 1526-2000]; †266; [»15505]; #6.54/0.00	Duty/obligation; [a duty/moral necessity; n; 1674-2000]; †2791; [»170017]; #6.12/0.02	Loud cry/shout; [call to animals/birds; n; 1530-2000]; †930; [»61145]; #5.13/0.04	Loud cry/shout; [call; n; 1300-1822]; †930; [»61153]; #5.13/0.04	Loud cry/shout; [word/name; n; 1801-1801]; †930; [»61154]; #5.13/0.04	Loud cry/shout; [call; vi; 1000-2000]; †930; [»61187]; #5.13/0.04
the						
lapsed	Apostasy; [Apostasy; aj; 1664-2000]; †2891; [»173949]; #10.00/0.00	The past; [Of/pertaining to the past/passed away/gone; aj; 1702-1702]; †1341; [»89161]; #8.54/0.14				

soul	Personification; [personification of some quality; n; 1605-1766]; †3103; [»181853]; #10.00/0.00	High intelligence, genius; [High intelligence, genius; n; 1604-2000]; †1745; [»117146]; #7.69/0.05	Fowls; [cuts/parts of fowl; n; 1530-2000]; †614; [»41906]; #7.14/0.09	Life, process of living; [Source/principle of life; n; 1000-1697]; †142; [»8287]; #5.99/0.14	Essence/intrinsic nature; [Essence/intrinsic nature; n; 1596-2000]; †1188; [»76262]; #5.88/0.18	Essence/intrinsic nature; [of a material thing; n; 1658-2000]; †1188; [»76273]; #5.88/0.18
,						
And						
weeping	Weeping; [fit(s) of weeping; n; 1836-2000]; †2057; [»129382]; #9.72/0.00	Weeping; [Weep; vi; 1000-2000]; †2057; [»129421]; #9.72/0.00	Weeping; [shed tears (of the eyes); vi; 1567-2000]; †2057; [»129426]; #9.72/0.00	Weeping; [Weep for; vt; 1000-2000]; †2057; [»129432]; #9.72/0.00	Weeping; [shed (tears); vt; 1000-2000]; †2057; [»129440]; #9.72/0.00	
in						
the						
evening	Day/day-time; [Evening; n; 1440-2000]; †1330; [»88661]; #13.51/0.00					
dew	Precipitation; [Dew; n; 1000-2000]; †136; [»7877]; #25.58/0.00	Precipitation; [Give/fall as dew; vi; 1382-1726]; †136; [»7889]; #25.58/0.00	Precipitation; [Cover with dew; vt; 1821-1821]; †136; [»7891]; #25.58/0.00			
;						
That						
might						

control	Restraint/restraining; [Restraint/restraining; n; 1594-2000]; †2692; [»163738]; #14.29/0.00	Restraint/restraining; [means of restraint/restraining force; n; 1752-2000]; †2692; [»163743]; #14.29/0.00	Restraint/restraining; [one who restrains; n; 1786-1855]; †2692; [»163746]; #14.29/0.00	Restraint/restraining; [hold in check; vt; 1549-1854]; †2692; [»163762]; #14.29/0.00		
The						
starry	Naturally occurring light; [of/pertaining to starlight/bright as the stars; aj; 1608-2000]; †1132; [»73911]; #7.61/0.00	Naturally occurring light; [illuminated by stars; aj; 1374-2000]; †1132; [»73915]; #7.61/0.00				
pole	Point; [pole; n; 1391-2000]; †1611; [»108693]; #8.33/0.00	Point; [to which other points are referred; n; 1849-2000]; †1611; [»108696]; #8.33/0.00				
,						
And						
fallen	Precipitation; [fall of; n; 1593-2000]; †136; [»7870]; #25.58/0.00	Precipitation; [Fall (of precipitation); vi; 1000-2000]; †136; [»7874]; #25.58/0.00	Change; [pass into state, become; vi; 1340-2000]; †1354; [»90057]; #20.00/0.01	Change; [get into specified condition; vi; 1382-2000]; †1354; [»90060]; #20.00/0.01	Waterfall; [Waterfall; n; 1579-2000]; †23; [»1313]; #18.18/0.03	Defeat/overthrow; [Defeat/overthrow; n; 1205-2000]; †1500; [»84383]; #17.14/0.03
,						

fallen	Precipitation; [fall of; n; 1593-2000]; †136; [»7870]; #25.58/0.00	Precipitation; [Fall (of precipitation); vi; 1000-2000]; †136; [»7874]; #25.58/0.00	Change; [pass into state, become; vi; 1340-2000]; †1354; [»90057]; #20.00/0.01	Change; [get into specified condition; vi; 1382-2000]; †1354; [»90060]; #20.00/0.01	Waterfall; [Waterfall; n; 1579-2000]; †23; [»1313]; #18.18/0.03	Defeat/overthrow; [Defeat/overthrow; n; 1205-2000]; †1500; [»84383]; #17.14/0.03
light	Illumination; [illuminated/lit up; aj; 1300-1704]; †1130; [»73778]; #15.22/0.00	Illumination; [well; aj; 1000-2000]; †1130; [»73781]; #15.22/0.00	Illumination; [Be/become illuminated; vi; 1820-2000]; †1130; [»73787]; #15.22/0.00	Illumination; [illuminate; vi; 1700-1700]; †1130; [»73789]; #15.22/0.00	Illumination; [Illuminate; vt; 1000-2000]; †1130; [»73790]; #15.22/0.00	Illumination; [to enable people to see; vt; 1200-2000]; †1130; [»73798]; #15.22/0.00
renew	Backward movement; [return towards point of departure; vi; 1470-1697]; †1391; [»101127]; #12.93/0.00	Speech; [say again/in resumption; vt; 1687-1853]; †2208; [»142051]; #9.22/0.05	Regeneration; [Regenerate; vt; 1382-2000]; †2933; [»174985]; #7.69/0.10	Restoration; [be/become renewed; vi; 1414-1766]; †1464; [»81154]; #7.02/0.15	Restoration; [to flourishing condition; vt; 1535-1726]; †1464; [»81165]; #7.02/0.15	Restoration; [to activity; vt; 1484-2000]; †1464; [»81166]; #7.02/0.15
!						
'O						
Earth	The world; [The world; n; 1000-2000]; †1; [»1]; #22.22/0.00	Ploughing; [Ploughing; n; 1000-1813]; †736; [»47630]; #10.67/0.05	Ploughing; [soil thrown up by plough; n; 1765-1765]; †736; [»47650]; #10.67/0.05	Clay; [for making pottery; n; 1526-1660]; †3526; [»203876]; #8.70/0.15		
,						
O						

Earth	The world; [The world; n; 1000-2000]; †1; [»1]; #22.22/0.00	Ploughing; [Ploughing; n; 1000-1813]; †736; [»47630]; #10.67/0.05	Ploughing; [soil thrown up by plough; n; 1765-1765]; †736; [»47650]; #10.67/0.05	Clay; [for making pottery; n; 1526-1660]; †3526; [»203876]; #8.70/0.15		
,						
return	Direction; [a turn; n; 1681-1802]; †1312; [»98467]; #13.81/0.00	Direction; [back; n; 1450-2000]; †1312; [»98470]; #13.81/0.00	Direction; [turned back; aj; 1700-2000]; †1312; [»98520]; #13.81/0.00	Direction; [reverse the direction of; vt; 1613-2000]; †1312; [»98622]; #13.81/0.00	Backward movement; [return towards point of departure; n; 1390-2000]; †1391; [»101062]; #12.93/0.05	Backward movement; [bringing back to former position; n; 1638-2000]; †1391; [»101065]; #12.93/0.05
!						
Arise	Day/day-time; [Dawn; vi; 1480-1667]; †1330; [»88629]; #13.51/0.00	Audibility; [Be/become audible; vi; 1300-1859]; †904; [»60023]; #13.04/0.05	Insurrection; [Rise in revolt; vi; 1000-1703]; †2707; [»164283]; #11.54/0.11	Effect/result/consequence; [Result; vi; 1205-2000]; †1215; [»77618]; #8.33/0.16		
from						
out	Direction; [away from some thing/place; av; 1000-2000]; †1312; [»98561]; #13.81/0.00	Audibility; [Audible; av; 1382-2000]; †904; [»60021]; #13.04/0.01	End/conclusion; [to the end; av; 1300-2000]; †1582; [»107032]; #12.50/0.02	End/conclusion; [over/finished/expired; av; 1300-2000]; †1582; [»107034]; #12.50/0.02	Offering; [offer by exposure; vt; 1386-1670]; †2184; [»139675]; #11.43/0.05	Effect/result/consequence; [to a result; av; 1534-2000]; †1215; [»77614]; #8.33/0.06
the						

dewy	Precipitation; [Dewy/like dew; aj; 1000-2000]; †136; [»7883]; #25.58/0.00	Precipitation; [consisting of dew; aj; 1820-1827]; †136; [»7884]; #25.58/0.00	Precipitation; [covered/wet with dew; aj; 1000-2000]; †136; [»7888]; #25.58/0.00			
grass	Fat; [of dead bodies; n; 1793-1793]; †224; [»24678]; #9.09/0.00	Land, ground, land mass; [above a mine; n; 1776-2000]; †5; [»233]; #7.44/0.07	Plants defined by habit; [Herb/herbaceous plant; n; 1350-2000]; †545; [»19292]; #5.77/0.13	Plants defined by habit; [herbage/grass; n; 1000-2000]; †545; [»19293]; #5.77/0.13	Plants defined by habit; [stage of growth; n; 1000-1733]; †545; [»19303]; #5.77/0.13	
!						
Night	Day/day-time; [twilight/dusk/nightfall; n; 1205-1703]; †1330; [»88664]; #13.51/0.00	Night; [Night; n; 1000-2000]; †1331; [»88678]; #10.91/0.10	Night; [as a division/period of time; n; 1200-2000]; †1331; [»88679]; #10.91/0.10	Night; [marking lapse of time; n; 1000-1891]; †1331; [»88680]; #10.91/0.10	Night; [marking an occasion/point in time; n; 1000-2000]; †1331; [»88681]; #10.91/0.10	Night; [the kind of night one has had; n; 1667-2000]; †1331; [»88698]; #10.91/0.10
is						
worn	Change to something else, transformation; [from/into; vi; 1555-1805]; †1356; [»90170]; #9.90/0.00	Deterioration; [in quality/character; vi; 1275-2000]; †1995; [»124576]; #9.09/0.03	Deterioration; [with predicative adjective; vi; 1837-1875]; †1995; [»124577]; #9.09/0.03	Appearance/aspect; [Have (specific) appearance; vt; 1611-2000]; †900; [»59639]; #5.33/0.08	Clothing; [Clothing; n; 1570-2000]; †804; [»53173]; #4.65/0.11	Clothing; [by people generally; vi; 1601-1888]; †804; [»53203]; #4.65/0.11
,						
And						
the						

morn	Day/day-time; [Dawn; n; 1000-2000]; †1330; [»88619]; #13.51/0.00	Day/day-time; [Morning; n; 1000-2000]; †1330; [»88630]; #13.51/0.00				
Rises	Audibility; [Be/become audible; vi; 1400-2000]; †904; [»60023]; #13.04/0.00	Present events; [Occur/happen; vi; 1200-1847]; †1217; [»77791]; #11.93/0.01	Insurrection; [an insurrection; n; 1768-1853]; †2707; [»164265]; #11.54/0.02	Insurrection; [Rise in revolt; vi; 1154-2000]; †2707; [»164283]; #11.54/0.02	Swelling; [Swell; vi; 1388-1697]; †286; [»9194]; #11.11/0.04	Travel in specific course/direction; [reach top of; vt; 1808-2000]; †3270; [»188983]; #8.80/0.05
from						
the						
slumbrous	Inaction; [Inactive; aj; 1809-2000]; †1439; [»79239]; #2.78/0.00					
mass.						
'Turn	Bluntness; [Become blunt; vi; 1815-1815]; †1255; [»93245]; #25.00/0.00	Bluntness; [Make blunt; vt; 1568-2000]; †1255; [»93247]; #25.00/0.00	Change; [Change; n; 1597-2000]; †1354; [»90008]; #20.00/0.02	Change; [Change; vi; 1175-2000]; †1354; [»90053]; #20.00/0.02	Change; [pass into state, become; vi; 1303-2000]; †1354; [»90057]; #20.00/0.02	Change; [Change; vt; 1230-1892]; †1354; [»90063]; #20.00/0.02
away	Immediacy; [Immediately; av; 1535-2000]; †1339; [»89060]; #14.81/0.00	Direction; [away from some thing/place; av; 1000-2000]; †1312; [»98561]; #13.81/0.06	Direction; [away from contact/inclusion; av; 1160-2000]; †1312; [»98563]; #13.81/0.06	End/conclusion; [to the end; av; 1340-2000]; †1582; [»107032]; #12.50/0.17		
no						
more						

;						
Why	Surprise, unexpectedness; [Exclamation of surprise; in; 1519-2000]; †1866; [»121410]; #5.48/0.00					
wilt						
thou						
turn	Bluntness; [Become blunt; vi; 1815-1815]; †1255; [»93245]; #25.00/0.00	Bluntness; [Make blunt; vt; 1568-2000]; †1255; [»93247]; #25.00/0.00	Change; [Change; n; 1597-2000]; †1354; [»90008]; #20.00/0.02	Change; [Change; vi; 1175-2000]; †1354; [»90053]; #20.00/0.02	Change; [pass into state, become; vi; 1303-2000]; †1354; [»90057]; #20.00/0.02	Change; [Change; vt; 1230-1892]; †1354; [»90063]; #20.00/0.02
away	Immediacy; [Immediately; av; 1535-2000]; †1339; [»89060]; #14.81/0.00	Direction; [away from some thing/place; av; 1000-2000]; †1312; [»98561]; #13.81/0.06	Direction; [away from contact/inclusion; av; 1160-2000]; †1312; [»98563]; #13.81/0.06	End/conclusion; [to the end; av; 1340-2000]; †1582; [»107032]; #12.50/0.17		
?						
The						
starry	Naturally occurring light; [of/pertaining to starlight/bright as the stars; aj; 1608-2000]; †1132; [»73911]; #7.61/0.00	Naturally occurring light; [illuminated by stars; aj; 1374-2000]; †1132; [»73915]; #7.61/0.00				

floor	Place for dancing; [floor for dancing; n; 1839-2000]; †4022; [»224889]; #8.33/0.00	Perplexity, bewilderment; [nonplus; vt; 1830-2000]; †1857; [»120978]; #7.69/0.04	Parts of building; [floor/storey; n; 1585-2000]; †2408; [»152078]; #7.55/0.08	Factory; [platform/space for carrying out industry; n; 1000-2000]; †3459; [»201133]; #7.14/0.12	Level land; [level place/plain; n; 1400-2000]; †17; [»837]; #5.26/0.15	Floor; [Floor; n; 1000-2000]; †2412; [»152224]; #5.26/0.15
,						
The						
watery	Precipitation; [holding moisture in vapour form; aj; 1377-2000]; †136; [»7872]; #25.58/0.00	Rainbow; [attribute of rainbow; aj; 1600-1755]; †139; [»8050]; #12.50/0.04	Weeping; [suffused/wet with tears; aj; 1447-2000]; †2057; [»129405]; #9.72/0.08	Land, ground, land mass; [wet; aj; 1000-2000]; †5; [»245]; #7.44/0.12	Water; [Of/pertaining to water; aj; 1586-2000]; †1031; [»67669]; #6.67/0.15	Water; [like/of nature of water; aj; 1477-2000]; †1031; [»67672]; #6.67/0.15
shore	Inclination; [a slope; n; 1546-1681]; †1294; [»97013]; #3.47/0.00	Threat/threatening; [Threat/threatening; n; 1375-1650]; †1496; [»84015]; #1.92/0.09	Shore, coast; [Shore/bank; n; 1400-2000]; †9; [»331]; #1.76/0.18	Shore, coast; [Shore/bank; vt; 1832-2000]; †9; [»338]; #1.76/0.18	Shore, coast; [tract between high and low water marks; n; 1622-2000]; †9; [»356]; #1.76/0.18	
,						
Is						
given	Voice/vocal sound; [utter; vt; 1200-2000]; †926; [»60972]; #27.66/0.00	Offering for inspection/consideration; [Offer for inspection/consideration; vt; 1000-2000]; †3096; [»181631]; #20.00/0.02	Offering; [one's hand to be taken; vt; 1382-2000]; †2184; [»139677]; #11.43/0.05	Action of informing; [Give (information); vt; 1449-2000]; †3138; [»183473]; #9.76/0.07	Giving; [Give; vi; 1000-2000]; †2178; [»139325]; #9.68/0.10	Giving; [Give; vt; 1000-2000]; †2178; [»139335]; #9.68/0.10
thee						

till						
the						
break	Change; [Change; vt; 1839-2000]; †1354; [»90063]; #20.00/0.00	Change of direction of movement; [diverge from course; vi; 1677-2000]; †1388; [»100904]; #15.52/0.01	Direction; [change the direction of; vt; 1616-1753]; †1312; [»98626]; #13.81/0.03	Defeat; [break rank; vi; 1598-2000]; †2442; [»153509]; #11.11/0.04	Ploughing; [Plough (land); vt; 1499-1847]; †736; [»47688]; #10.67/0.05	Interruption; [an interruption; n; 1627-2000]; †2213; [»142343]; #10.00/0.06
of						
day	Day/day-time; [Day/day-time; n; 1000-2000]; †1330; [»88606]; #13.51/0.00	Day/day-time; [Dawn; n; 1300-1793]; †1330; [»88619]; #13.51/0.00	Naturally occurring light; [daylight; n; 1340-2000]; †1132; [»73857]; #7.61/0.14			
.						
'						

Table A19 By author for »130497 (Loved one)

words	author	CG	poemID	totals
love;S_END	Alexander Scott. 1520?–158–	CG1	44	1
love;S_END	Richard Edwardes. 1523–66	CG1	46	5
love;S_END	Anonymous. XVI–XVII Century	CG1	55	12
love;S_END	Anonymous. XVI–XVII Century	CG1	58	1
love;S_END	Anonymous. XVI–XVII Century	CG1	66	1
love;S_END	Edmund Spenser. 1552–1599	CG1	82	2
love;S_END	Joshua Sylvester. 1563–1618	CG1	115	1
treasure;S_END	Michael Drayton. 1563–1631	CG1	118	1
love;S_END	Christopher Marlowe. 1564–93	CG1	121	2
love;S_END	Christopher Marlowe. 1564–93	CG1	122	3
love;S_END	William Shakespeare. 1564–1616	CG1	124	1
love;S_END	William Shakespeare. 1564–1616	CG1	162	1
treasure;S_END	John Donne. 1573–1631	CG1	197	1
love;S_END	John Fletcher. 1579–1625	CG1	216	1
sun;S_END	John Webster. ?–1630?	CG1	220	1
love;S_END	William Drummond, of Hawthornden. 1585–1649	CG1	224	1
sun;S_END	William Drummond, of Hawthornden. 1585–1649	CG1	226	1
love;S_END	George Wither. 1588–1667	CG1	237	1
sun;S_END	William Browne, of Tavistock. 1588–1643	CG1	240	1
love;S_END	Robert Herrick. 1591–1674	CG1	247	1
sun;S_END	Robert Herrick. 1591–1674	CG1	247	1
sun;S_END	Robert Herrick. 1591–1674	CG1	261	1
love;S_END	Robert Herrick. 1591–1674	CG1	269	1
treasure;S_END	George Herbert. 1593–1632	CG1	284	1
love;S_END	John Milton. 1608–1674	CG2	313	1
love;S_END	John Milton. 1608–1674	CG2	317	1
love;S_END	William Cartwright. 1611–1643	CG2	332	1
sweet;S_END	Richard Crashaw. 1613?–1649	CG2	337	1
love;S_END	Richard Crashaw. 1613?–1649	CG2	338	1
love;S_END	Richard Crashaw. 1613?–1649	CG2	340	1
love;S_END	Henry Vaughan. 1621–1695	CG2	365	1
love;S_END	Anonymous: Ballads.	CG2	368	1
love;S_END	Anonymous: Ballads.	CG2	376	1
sweet;S_END	Anonymous: Ballads.	CG2	380	1
love;S_END	Anonymous: Ballads.	CG2	392	1

sweet;S_END	Charles Cotton. 1630–1687	CG2	396	1
love;S_END	John Dryden. 1631–1700	CG2	398	1
love;S_END	John Dryden. 1631–1700	CG2	400	2
love;S_END	Sir George Etherege. 1635–1691	CG2	404	1
love;S_END	Aphra Behn. 1640–1689	CG2	412	1
love;S_END	John Wilmot, Earl of Rochester. 1647–1680	CG2	413	1
love;S_END	John Sheffield, Duke of Buckinghamshire. 1649–1720	CG2	418	1
love;S_END	John Cutts, Lord Cutts. 1661–1707	CG2	421	1
love;S_END	George Lyttelton, Lord Lyttelton. 1709–1773	CG3	449	4
love;S_END	Thomas Gray. 1716–1771	CG3	453	1
love;S_END	Thomas Gray. 1716–1771	CG3	455	2
love;S_END	William Collins. 1721–1759	CG3	460	1
love;S_END	Mark Akenside. 1721–1770	CG3	463	1
love;S_END	Robert Cunninghame-Graham of Gartmore. 1735–1797	CG3	469	1
love;S_END	George Crabbe. 1754–1832	CG3	480	1
love;S_END	George Crabbe. 1754–1832	CG3	482	1
love;S_END	William Wordsworth. 1770–1850	CG3	536	1
love;S_END	Sir Walter Scott. 1771–1832	CG3	546	2
love;S_END	Samuel Taylor Coleridge. 1772–1834	CG3	549	1
sun;S_END	Samuel Taylor Coleridge. 1772–1834	CG3	549	3
sun;S_END	Thomas Campbell. 1774–1844	CG3	581	1
love;S_END	Thomas Moore. 1779–1852	CG3	582	2
love;S_END	Thomas Love Peacock. 1785–1866	CG3	594	1
sun;S_END	Percy Bysshe Shelley. 1792–1822	CG3	606	1
love;S_END	Percy Bysshe Shelley. 1792–1822	CG3	613	1
love;S_END	John Keats. 1795–1821	CG3	625	1
love;S_END	Jeremiah Joseph Callanan. 1795–1839	CG3	638	1
love;S_END	George Darley. 1795–1846	CG3	640	1
sun;S_END	Thomas Hood. 1798–1845	CG3	654	1
sweet;S_END	Gerald Griffin. 1803–1840	CG4	663	1
love;S_END	James Clarence Mangan. 1803–1849	CG4	664	1
beloved;S_END	Ralph Waldo Emerson. 1803–1882	CG4	669	1
sweet;S_END	Elizabeth Barrett Browning. 1806–1861	CG4	679	1
love;S_END	Elizabeth Barrett Browning. 1806–1861	CG4	680	1
love;S_END	Elizabeth Barrett Browning. 1806–1861	CG4	682	1
darling;S_END	Helen Selina, Lady Dufferin. 1807–1867	CG4	691	1
love;S_END	Alfred Tennyson, Lord Tennyson. 1809–1892	CG4	707	1
sun;S_END	Alfred Tennyson, Lord Tennyson. 1809–1892	CG4	708	1

love;S_END	Sir Samuel Ferguson. 1810–1886	CG4	712	1
treasure;S_END	Robert Browning. 1812–1889	CG4	728	1
sweet;S_END	John Ruskin. 1819–1900	CG4	744	1
sun;S_END	Matthew Arnold. 1822–1888	CG4	747	1
passion;S_END	Matthew Arnold. 1822–1888	CG4	752	1
sun;S_END	Dante Gabriel Rossetti. 1828–1882	CG4	771	1
sweet;S_END	George Meredith. 1828–1909	CG4	772	1
love;S_END	Alexander Smith. 1829–1867	CG4	777	1
sun;S_END	William Morris. 1834–1896	CG4	800	1
sun;S_END	Algernon Charles Swinburne. 1837–1909	CG4	811	1
love;S_END	Bret Harte. 1839–1902	CG4	813	1
love;S_END	Wilfrid Scawen Blunt. b. 1840	CG4	817	2
passion;S_END	Wilfrid Scawen Blunt. b. 1840	CG4	817	1
sun;S_END	Robert Bridges. b. 1844	CG4	832	1
treasure;S_END	Robert Bridges. b. 1844	CG4	838	1
sun;S_END	Robert Bridges. b. 1844	CG4	839	1
sun;S_END	John Davidson. 1857–1909	CG4	851	1
love;S_END	Francis Thompson. 1859–1907	CG4	875	1

Table A20 By author for »15487 (Die)

words	author	CG	poemID	totals
part;S_END	Sir Thomas Wyatt. 1503–1542	CG1	36	1
die;S_END	Anonymous. XVI–XVII Century	CG1	59	1
part;S_END	Anonymous. XVI–XVII Century	CG1	66	1
die;S_END	Anonymous. XVI–XVII Century	CG1	70	3
die;S_END	Sir Philip Sidney. 1554–86	CG1	89	2
starve;S_END	Michael Drayton. 1563–1631	CG1	116	1
die;S_END	William Shakespeare. 1564–1616	CG1	143	1
end;S_END	William Shakespeare. 1564–1616	CG1	147	1
end;S_END	William Shakespeare. 1564–1616	CG1	164	1
die;S_END	Sir Henry Wotton. 1568–1639	CG1	180	1
part;S_END	John Donne. 1573–1631	CG1	195	1
die;S_END	John Donne. 1573–1631	CG1	199	1
die;S_END	John Donne. 1573–1631	CG1	200	1
die;S_END	John Fletcher. 1579–1625	CG1	211	1
die;S_END	John Fletcher. 1579–1625	CG1	212	2
die;S_END	Francis Beaumont. 1586–1616	CG1	234	1
die;S_END	George Wither. 1588–1667	CG1	237	1
die;S_END	William Browne, of Tavistock. 1588–1643	CG1	242	1
die;S_END	William Browne, of Tavistock. 1588–1643	CG1	245	1
die;S_END	Robert Herrick. 1591–1674	CG1	248	1
part;S_END	Robert Herrick. 1591–1674	CG1	258	1
die;S_END	Robert Herrick. 1591–1674	CG1	261	1
expire;S_END	Robert Herrick. 1591–1674	CG1	263	1
part;S_END	Henry King, Bishop of Chichester. 1592–1669	CG1	279	1
part;S_END	Henry King, Bishop of Chichester. 1592–1669	CG1	280	1
die;S_END	George Herbert. 1593–1632	CG1	281	3
die;S_END	Thomas Carew. 1595?–1639?	CG1	289	1
die;S_END	Edmund Waller. 1606–1687	CG2	305	1
die;S_END	John Milton. 1608–1674	CG2	317	1
die;S_END	Richard Crashaw. 1613?–1649	CG2	338	3
die;S_END	Abraham Cowley. 1618–1667	CG2	352	1
die;S_END	Abraham Cowley. 1618–1667	CG2	353	1
die;S_END	Anonymous: Ballads.	CG2	369	3
die;S_END	Anonymous: Ballads.	CG2	371	1
die;S_END	Anonymous: Ballads.	CG2	374	1
die;S_END	Anonymous: Ballads.	CG2	375	2

die;S_END	Anonymous: Ballads.	CG2	382	1
die;S_END	Anonymous: Ballads.	CG2	387	1
expire;S_END	John Wilmot, Earl of Rochester. 1647–1680	CG2	413	1
die;S_END	John Wilmot, Earl of Rochester. 1647–1680	CG2	415	1
die;S_END	John Wilmot, Earl of Rochester. 1647–1680	CG2	416	2
die;S_END	John Cutts, Lord Cutts. 1661–1707	CG2	421	1
part;S_END	Matthew Prior. 1664–1721	CG2	422	1
part;S_END	Alexander Pope. 1688–1744	CG2	441	1
die;S_END	Alexander Pope. 1688–1744	CG2	441	1
die;S_END	William Broome. ?–1745	CG2	447	1
die;S_END	Samuel Johnson. 1709–1784	CG3	451	1
die;S_END	Thomas Gray. 1716–1771	CG3	453	1
end;S_END	Christopher Smart. 1722–1770	CG3	465	1
die;S_END	Oliver Goldsmith. 1728–1774	CG3	467	1
depart;S_END	William Blake. 1757–1827	CG3	492	1
die;S_END	Robert Burns. 1759–1796	CG3	493	1
buy;S_END	James Hogg. 1770–1835	CG3	514	1
drop;S_END	William Wordsworth. 1770–1850	CG3	515	1
end;S_END	William Wordsworth. 1770–1850	CG3	531	1
die;S_END	William Wordsworth. 1770–1850	CG3	532	1
die;S_END	Sir Walter Scott. 1771–1832	CG3	544	1
depart;S_END	Sir Walter Scott. 1771–1832	CG3	548	1
die;S_END	Sir Walter Scott. 1771–1832	CG3	548	2
die;S_END	Samuel Taylor Coleridge. 1772–1834	CG3	549	1
ghost;S_END	Samuel Taylor Coleridge. 1772–1834	CG3	549	1
sink;S_END	Samuel Taylor Coleridge. 1772–1834	CG3	549	1
depart;S_END	Walter Savage Landor. 1775–1864	CG3	576	1
depart;S_END	Thomas Moore. 1779–1852	CG3	584	1
die;S_END	Percy Bysshe Shelley. 1792–1822	CG3	607	1
die;S_END	Percy Bysshe Shelley. 1792–1822	CG3	611	1
die;S_END	Felicia Dorothea Hemans. 1793–1835	CG3	622	1
sink;S_END	John Keats. 1795–1821	CG3	635	1
die;S_END	George Darley. 1795–1846	CG3	642	1
die;S_END	Thomas Hood. 1798–1845	CG3	653	1
buy;S_END	Thomas Lovell Beddoes. 1803–1849	CG4	667	3
die;S_END	Thomas Lovell Beddoes. 1803–1849	CG4	667	1
die;S_END	Henry Wadsworth Longfellow. 1807–1882	CG4	689	1
die;S_END	Helen Selina, Lady Dufferin. 1807–1867	CG4	691	1
end;S_END	Edward Fitzgerald. 1809–1883	CG4	698	1

die;S_END	Alfred Tennyson, Lord Tennyson. 1809–1892	CG4	704	3
die;S_END	Alfred Tennyson, Lord Tennyson. 1809–1892	CG4	708	1
depart;S_END	Robert Browning. 1812–1889	CG4	716	1
end;S_END	Robert Browning. 1812–1889	CG4	727	1
end;S_END	Aubrey De Vere. 1814–1902	CG4	733	1
end;S_END	Emily Brontë. 1818–1848	CG4	737	1
die;S_END	William (Johnson) Cory. 1823–1892	CG4	758	1
part;S_END	Coventry Patmore. 1823–1896	CG4	764	1
end;S_END	George Meredith. 1828–1909	CG4	772	1
die;S_END	George Meredith. 1828–1909	CG4	772	1
sink;S_END	George Meredith. 1828–1909	CG4	776	1
end;S_END	Christina Georgina Rossetti. 1830–1894	CG4	783	1
die;S_END	Algernon Charles Swinburne. 1837–1909	CG4	809	1
end;S_END	Algernon Charles Swinburne. 1837–1909	CG4	810	1
depart;S_END	Algernon Charles Swinburne. 1837–1909	CG4	810	1
die;S_END	Edmund Gosse. b. 1849	CG4	845	1
die;S_END	Katharine Tynan Hinkson. b. 1861	CG4	877	1

Table A21 S1 Poem titles

PoemID	Author / Poem Title	Word count
John Donne. 1573–1631		
195	Daybreak	61
196	Song	189
197	That Time and Absence proves	172
198	The Ecstasy	160
199	The Dream	294
200	The Funeral	216
201	A Hymn to God the Father	191
202	Death	163
George Herbert. 1593–1632		
281	Virtue	133
282	Easter	79
283	Discipline	181
284	A Dialogue	256
285	The Pulley	182
286	Love	191
Abraham Cowley. 1618–1667		
349	Anacreontics 1	181

350	Anacreontics 2	218
351	Anacreontics 3	87
352	On the Death of Mr. William Hervey	749
353	The Wish	346
Andrew Marvell. 1621–1678		
355	An Horatian Ode	821
356	A Garden	215
357	To His Coy Mistress	379
358	The Picture of Little T. C. in a Prospect of Flowers	319
359	Thoughts in a Garden	565
360	Bermudas	313
361	An Epitaph	166
Henry Vaughan. 1621–1695		
362	The Retreat	247
363	Peace	141
364	The Timber	213
365	Friends Departed	360

Table A22 S2 Poem titles

PoemID	Author / Poem Title	Word count
Robert Herrick. 1591–1674		
247	Corinna's going a-Maying	680
248	To the Virgins, to make much of Time	136
249	To the Western Wind	68
250	To Electra	60
251	To Violets	83
252	To Daffodils	128
253	To Blossoms	131
254	The Primrose	109
255	The Funeral Rites of the Rose	141
256	Cherry-Ripe	69
257	A Meditation for his Mistress	188
258	Delight in Disorder	95
259	Upon Julia's Clothes	55
260	The Bracelet: To Julia	102
261	To Daisies, not to shut so soon	93
262	The Night-piece: To Julia	146
263	To Music, to becalm his Fever	192
264	To Dianeme	88

Table A23 S1 positive collocates at p < 0.0001

collocate	CT	S1	Not S1	S1_exp	RefS1_exp	S1_ll	both_definitions
APPGE;174916	55	18	37	2.7795	52.2205	41.7541771	possessive pronoun, pre-nominal (e.g. my, your, our);Soul
APPGE;174929	31	13	18	1.5666	29.4334	37.3132046	possessive pronoun, pre-nominal (e.g. my, your, our);with regard to moral aspect
DD1;89550	6	5	1	0.3032	5.6968	24.547537	singular determiner (e.g. this, that, another);thereafter/after that
124078;100539	6	5	1	0.3032	5.6968	24.547537	Well;Swift
VM;174949	23	9	14	1.1623	21.8377	24.3939474	modal auxiliary (can, will, would, etc.);of soul: die
CS;151032	7	5	2	0.3538	6.6462	21.6822141	subordinating conjunction (e.g. if, because, unless, so, for);Furnish with inhabitants
APPGE;76262	177	25	152	8.945	168.055	20.8641919	possessive pronoun, pre-nominal (e.g. my, your, our);Essence/intrinsic nature
VM;181524	12	6	6	0.6064	11.3936	19.8074341	modal auxiliary (can, will, would, etc.);Be manifest
S_BEGIN;79424	12	6	6	0.6064	11.3936	19.8074341	S_BEGIN;Abstain/refrain from (action)
157926;GE	8	5	3	0.4043	7.5957	19.5766933	person in authority;germanic genitive marker - (’ or ’s)
129390;S_END	8	5	3	0.4043	7.5957	19.5766933	a tear;S_END
S_BEGIN;90373	5	4	1	0.2527	4.7473	18.9801364	S_BEGIN;lasting, continuous
73597;189125	5	4	1	0.2527	4.7473	18.9801364	Shine;Departing
77287;107032	5	4	1	0.2527	4.7473	18.9801364	bring (a person/thing) into a state/condition;to the end
DD1;127407	5	4	1	0.2527	4.7473	18.9801364	singular determiner (e.g. this, that, another);person/thing
77290;89060	5	4	1	0.2527	4.7473	18.9801364	with force/haste;Immediately

130547;88131	5	4	1	0.2527	4.7473	18.9801364	Terms of endearment;an appointed/fixed time/day/date
88129;98561	5	4	1	0.2527	4.7473	18.9801364	Particular time;away from some thing/place
77290;189127	5	4	1	0.2527	4.7473	18.9801364	with force/haste;Depart/leave/go away
172427;APPGE	5	4	1	0.2527	4.7473	18.9801364	teach (a thing);possessive pronoun, pre-nominal (e.g. my, your, our)
88129;189125	5	4	1	0.2527	4.7473	18.9801364	Particular time;Departing
77290;98561	5	4	1	0.2527	4.7473	18.9801364	with force/haste;away from some thing/place
77290;189125	5	4	1	0.2527	4.7473	18.9801364	with force/haste;Departing
DD1;89182	5	4	1	0.2527	4.7473	18.9801364	singular determiner (e.g. this, that, another);ago
VM;57695	31	9	22	1.5666	29.4334	18.6614718	modal auxiliary (can, will, would, etc.);have orgasm
100860;59629	13	6	7	0.657	12.343	18.6018557	In a straight course;Have (specific) appearance
APPGE;117146	124	19	105	6.2666	117.7334	18.1128987	possessive pronoun, pre-nominal (e.g. my, your, our);High intelligence, genius
CS;155434	9	5	4	0.4548	8.5452	17.9001136	subordinating conjunction (e.g. if, because, unless, so, for);Common soldier
CS;39769	9	5	4	0.4548	8.5452	17.9001136	subordinating conjunction (e.g. if, because, unless, so, for);Man
158284;APPGE	14	6	8	0.7075	13.2925	17.5289731	ordain/prescribe/appoint;possessive pronoun, pre-nominal (e.g. my, your, our)
130547;VM	50	11	39	2.5268	47.4732	17.0253795	Terms of endearment;modal auxiliary (can, will, would, etc.)
124078;89060	21	7	14	1.0613	19.9387	16.5092168	Well;Immediately
APPGE;111119	21	7	14	1.0613	19.9387	16.5092168	possessive pronoun, pre-nominal (e.g. my, your, our);Abundant
100860;1345	10	5	5	0.5054	9.4946	16.5061951	In a straight course;Fountain
CS;133428	10	5	5	0.5054	9.4946	16.5061951	subordinating conjunction (e.g. if, because, unless, so, for);Encourage/embolden
100860;123749	10	5	5	0.5054	9.4946	16.5061951	In a straight course;consider to be, account as
CS;39774	10	5	5	0.5054	9.4946	16.5061951	subordinating conjunction (e.g. if, because, unless, so, for);men collectively

111043;59629	10	5	5	0.5054	9.4946	16.5061951	utterly;Have (specific) appearance
S_BEGIN;79432	10	5	5	0.5054	9.4946	16.5061951	S_BEGIN;leave to another to deal with
100860;122252	10	5	5	0.5054	9.4946	16.5061951	In a straight course;Be attentive, pay attention to
100860;59036	10	5	5	0.5054	9.4946	16.5061951	In a straight course;See
172420;APPGE	6	4	2	0.3032	5.6968	16.4497074	Teach;possessive pronoun, pre-nominal (e.g. my, your, our)
VM;73623	6	4	2	0.3032	5.6968	16.4497074	modal auxiliary (can, will, would, etc.);bright
VM;181526	6	4	2	0.3032	5.6968	16.4497074	modal auxiliary (can, will, would, etc.);strikingly
130547;88194	6	4	2	0.3032	5.6968	16.4497074	Terms of endearment;time/appoint/set a time for
128371;APPGE	6	4	2	0.3032	5.6968	16.4497074	in forbearing/tolerant manner;possessive pronoun, pre-nominal (e.g. my, your, our)
CC;172427	6	4	2	0.3032	5.6968	16.4497074	coordinating conjunction (e.g. and, or);teach (a thing)
87820;189125	6	4	2	0.3032	5.6968	16.4497074	stretch/period/portion of time;Departing
DD1;89642	6	4	2	0.3032	5.6968	16.4497074	singular determiner (e.g. this, that, another);next in order/then
88129;89060	6	4	2	0.3032	5.6968	16.4497074	Particular time;Immediately

Table A24 S2 positive collocates at p < 0.0001

collocate	S2	Not S2	S2_exp	RefS2_exp	S1_ll	both_definitions
142813;PPIO1	15	5	0.6133	19.3867	82.3585391	for something;1st person sing. objective personal pronoun (me)
158298;PPIO1	15	6	0.6439	20.3561	79.786909	demand;1st person sing. objective personal pronoun (me)
76262;128689	11	1	0.368	11.632	69.8412645	Essence/intrinsic nature;Joy/gladness/delight
115527;57498	10	1	0.3373	10.6627	63.0539665	Shape inclinations of, dispose;Refresh/invigorate
76291;57498	10	1	0.3373	10.6627	63.0539665	deprive of essence/quintessence;Refresh/invigorate
115503;128689	10	1	0.3373	10.6627	63.0539665	state of mind;Joy/gladness/delight
76262;57487	10	1	0.3373	10.6627	63.0539665	Essence/intrinsic nature;that which/one who refreshes/invigorates
115503;57498	10	2	0.368	11.632	59.0047785	state of mind;Refresh/invigorate
76262;57498	10	2	0.368	11.632	59.0047785	Essence/intrinsic nature;Refresh/invigorate
S_BEGIN;158298	10	2	0.368	11.632	59.0047785	S_BEGIN;demand
S_BEGIN;142813	10	2	0.368	11.632	59.0047785	S_BEGIN;for something
128474;PPIO1	10	2	0.368	11.632	59.0047785	Please/give pleasure to;1st person sing. objective personal pronoun (me)
128689;PPIO1	10	2	0.368	11.632	59.0047785	Joy/gladness/delight;1st person sing. objective personal pronoun (me)
115497;57498	10	2	0.368	11.632	59.0047785	individual qualities;Refresh/invigorate
142806;PPIO1	10	2	0.368	11.632	59.0047785	Make a request;1st person sing. objective personal pronoun (me)
142752;PPIO1	10	3	0.3986	12.6014	55.8352266	Request;1st person sing. objective personal pronoun (me)
58664;76262	10	3	0.3986	12.6014	55.8352266	Sweet;Essence/intrinsic nature
122814;PPIO1	10	3	0.3986	12.6014	55.8352266	by asking/enquiring;1st person sing. objective personal pronoun (me)

58677;76262	10	3	0.3986	12.6014	55.8352266	Sweetly;Essence/intrinsic nature
58656;76262	10	3	0.3986	12.6014	55.8352266	sweet thing;Essence/intrinsic nature
58653;76262	10	3	0.3986	12.6014	55.8352266	Sweetness;Essence/intrinsic nature
130569;76262	10	4	0.4293	13.5707	53.1912727	Used to a loved one;Essence/intrinsic nature
130547;76262	10	5	0.46	14.54	50.909683	Terms of endearment;Essence/intrinsic nature
77300;PPIO1	10	6	0.4906	15.5094	48.8973712	elicit/call forth;1st person sing. objective personal pronoun (me)
S_BEGIN;122814	7	1	0.2453	7.7547	42.8195359	S_BEGIN;by asking/enquiring
S_BEGIN;142752	7	1	0.2453	7.7547	42.8195359	S_BEGIN;Request
77320;124078	8	6	0.4293	13.5707	37.0072331	Cause/reason;Well
S_BEGIN;77300	7	3	0.3066	9.6934	36.7551461	S_BEGIN;elicit/call forth
77320;84452	8	7	0.46	14.54	35.4632476	Cause/reason;in this very manner
77320;77329	8	7	0.46	14.54	35.4632476	Cause/reason;For that reason/therefore
77320;84451	8	7	0.46	14.54	35.4632476	Cause/reason;in this way
77320;89540	8	7	0.46	14.54	35.4632476	Cause/reason;after/afterwards/later
77320;84453	8	7	0.46	14.54	35.4632476	Cause/reason;in that way
121410;124078	7	4	0.3373	10.6627	34.61412	Exclamation of surprise;Well
158328;124078	7	4	0.3373	10.6627	34.61412	summons;Well
120108;124078	7	4	0.3373	10.6627	34.61412	Puzzle, enigma, riddle;Well
158328;84451	7	4	0.3373	10.6627	34.61412	summons;in this way
158328;84452	7	4	0.3373	10.6627	34.61412	summons;in this very manner
188558;PPHO1	7	4	0.3373	10.6627	34.61412	Travel/proceed/make one's way;3rd person sing. objective personal pronoun (him, her)
158328;77329	7	4	0.3373	10.6627	34.61412	summons;For that reason/therefore
125671;124078	7	4	0.3373	10.6627	34.61412	emphasizing a following statement;Well

158328;84453	7	4	0.3373	10.6627	34.61412	summons;in that way
122484;124078	7	4	0.3373	10.6627	34.61412	with specific form;Well
158328;89540	7	4	0.3373	10.6627	34.61412	summons;after/afterwards/later
125671;84453	7	5	0.368	11.632	32.7963675	emphasizing a following statement;in that way
122484;84452	7	5	0.368	11.632	32.7963675	with specific form;in this very manner
125671;84452	7	5	0.368	11.632	32.7963675	emphasizing a following statement;in this very manner
121410;84452	7	5	0.368	11.632	32.7963675	Exclamation of surprise;in this very manner
122484;84451	7	5	0.368	11.632	32.7963675	with specific form;in this way
120108;89540	7	5	0.368	11.632	32.7963675	Puzzle, enigma, riddle;after/afterwards/later
125671;89540	7	5	0.368	11.632	32.7963675	emphasizing a following statement;after/afterwards/later

Appendix II MySQL

CG1_1_CorpusGroupTable.sql

```
CREATE TABLE `cg1_1` (
  `poemID` int(11) DEFAULT NULL,
  `corp_id` bigint(20) unsigned NOT NULL DEFAULT '0',
  `corp_lemma` text,
  `corp_pos` text,
  `corp_word` text,
  `usas` varchar(90) DEFAULT NULL,
  `hte_match` text,
  `htst_match` text,
  `combined_id` text,
  `corpus_group` text,
  `position_ID` bigint(20) NOT NULL DEFAULT '0',
  KEY `position_id` (`position_ID`) USING BTREE,
  KEY `poemID` (`poemID`) USING BTREE,
  KEY `corp_lemma` (`corp_lemma`(100)) USING BTREE,
  KEY `corp_id` (`corp_id`) USING BTREE
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4
  COLLATE=utf8mb4_0900_ai_ci;
```

LexemeTable.sql

```
CREATE TABLE `lexeme` (
  `htid` int(7) NOT NULL COMMENT 'Unique number for each
  lexeme',
  `catid` int(6) NOT NULL COMMENT 'Unique number for each
  category',
  `word` text NOT NULL COMMENT 'Lemma/headword, with OE
  and OED forms combined for display',
  `wordoe` text COMMENT 'OE form of headword',
  `wordoed` text COMMENT 'OED form of headword',
  `fulldate` varchar(90) DEFAULT NULL COMMENT 'Textual
  form of date',
  `apps` int(4) NOT NULL COMMENT 'Approximate first
  citation (start) date - use with caution',
  `appe` int(4) NOT NULL COMMENT 'Approximate last
  citation (end) date - use with caution',
  `oe` varchar(2) DEFAULT NULL COMMENT 'Old English',
  `oefircon` varchar(1) DEFAULT NULL COMMENT 'Connector
  from Old English',
  `firstdac` varchar(1) DEFAULT NULL COMMENT 'Ante or
  circa for first date',
  `firstd` int(4) DEFAULT NULL COMMENT 'First date',
```

```
  `firstdb` int(4) DEFAULT NULL COMMENT 'B-section of
  first date',
  `firstdbr` varchar(3) DEFAULT NULL COMMENT 'Brackets
  for first date',
  `firmidcon` varchar(1) DEFAULT NULL COMMENT 'Connector
  from first to middle date',
  `firlastcon` varchar(1) DEFAULT NULL COMMENT 'Connector
  from first to last date',
  `middac` varchar(1) DEFAULT NULL COMMENT 'Ante or circa
  for middle date',
  `midd` int(4) DEFAULT NULL COMMENT 'Middle date',
  `middb` int(4) DEFAULT NULL COMMENT 'B-section of
  middle date',
  `middbr` varchar(3) DEFAULT NULL COMMENT 'Brackets for
  middle date',
  `midlascon` varchar(1) DEFAULT NULL COMMENT 'Connector
  from middle to last date',
  `lastdac` varchar(1) DEFAULT NULL COMMENT 'Ante or
  circa for late date',
  `lastd` int(4) DEFAULT NULL COMMENT 'Last date',
  `lastdb` int(4) DEFAULT NULL COMMENT 'B-section of last
  date',
  `lastdbr` varchar(3) DEFAULT NULL COMMENT 'Brackets for
  last date',
  `current` varchar(1) DEFAULT NULL COMMENT 'Current',
  `label` varchar(70) DEFAULT NULL COMMENT 'Usage label',
  `roget` int(3) DEFAULT NULL COMMENT 'Roget''s Thesaurus
  number',
  `catorder` int(3) NOT NULL COMMENT 'Indicates the
  chronological order of lexemes within a category (0
  first)',
  PRIMARY KEY (`htid`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4;
```

CategoryTable.sql

```
CREATE TABLE `category` (
  `catid` int(6) NOT NULL COMMENT 'Unique number for each
  category',
  `t1` varchar(2) NOT NULL COMMENT 'Main sequence
  category, first tier',
  `t2` varchar(2) DEFAULT NULL COMMENT 'Main sequence
  category, second tier',
  `t3` varchar(2) DEFAULT NULL COMMENT 'Main sequence
  category, third tier',
  `t4` varchar(2) DEFAULT NULL COMMENT 'Main sequence
  category, fourth tier',
  `t5` varchar(2) DEFAULT NULL COMMENT 'Main sequence
  category, fifth tier',
  `t6` varchar(2) DEFAULT NULL COMMENT 'Main sequence
  category, sixth tier',
```

```
  `t7` varchar(3) DEFAULT NULL COMMENT 'Main sequence
  category, seventh tier',
  `subcat` varchar(15) DEFAULT NULL COMMENT 'Subcategory,
  no split tiers',
  `pos`
  enum('aj','av','cj','in','n','p','ph','v','vi','vm','vp
  ','vr','vt') NOT NULL DEFAULT 'n' COMMENT 'Part of
  speech',
  `heading` varchar(70) NOT NULL COMMENT 'Textual
  category name',
  `tiering` varchar(8) NOT NULL COMMENT 'Location in
  hierarchy',
  `v1maincat` varchar(21) NOT NULL COMMENT 'Category
  grouping from pre-v.2 hierarchy, from print HTOED (for
  legacy reasons)',
  `mmcat` varchar(4) NOT NULL COMMENT 'Legacy category
  grouping for the Mapping Metaphor project (not
  supported in future versions - to be ignored)',
  `themid` int(4) NOT NULL COMMENT 'Link to unique
  identifer for thematic dataset',
  PRIMARY KEY (`catid`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4;
```

ThematicHeadingsTable.sql

```
CREATE TABLE `thematic_headings` (
  `tid` bigint(20) NOT NULL COMMENT 'Unique number for
  each thematic category',
  `s1` char(4) NOT NULL COMMENT 'Thematic category, first
  tier',
  `s2` char(4) DEFAULT NULL COMMENT 'Thematic category,
  second tier',
  `s3` char(4) DEFAULT NULL COMMENT 'Thematic category,
  third tier',
  `s4` char(4) DEFAULT NULL COMMENT 'Thematic category,
  fourth tier',
  `s5` char(4) DEFAULT NULL COMMENT 'Thematic category,
  fifth tier',
  `thematicheading` varchar(70) NOT NULL COMMENT 'Textual
  thematic category name',
  PRIMARY KEY (`tid`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4;
```

CG1_2_CrossReference.sql

Creates `CG1_2`

```
SELECT
  `phd`.`cg1_1`.`position_ID` AS `position_ID`,
  `phd`.`cg1_1`.`corp_word` AS `corp_word`,
```

```
   `phd`.`cg1_1`.`corp_lemma` AS `corp_lemma`,
   `phd`.`cg1_1`.`corp_pos` AS `corp_pos`,
   `phd`.`lexeme`.`word` AS `word_mixed`,
   `phd`.`lexeme`.`wordoed` AS `word`,
   `phd`.`lexeme`.`htid` AS `htid`,
   `phd`.`lexeme`.`apps` AS `apps`,
   `phd`.`lexeme`.`appe` AS `appe`,
   `phd`.`category`.`heading` AS `heading`,
   `phd`.`category`.`catid` AS `catid`,
   `phd`.`category`.`pos` AS `pos`,
   `phd`.`category`.`tiering` AS `tiering`,
   `phd`.`thematic_headings`.`tid` AS `tid`,
   `phd`.`thematic_headings`.`thematicheading` AS
   `thematicheading`,
   `phd`.`thematic_headings`.`s1` AS `s1`,
   `phd`.`thematic_headings`.`s2` AS `s2`,
   `phd`.`thematic_headings`.`s3` AS `s3`,
   `phd`.`cg1_1`.`corp_id` AS `corp_id`,
   `phd`.`cg1_1`.`poemID` AS `poemID`,
   `phd`.`cg1_1`.`combined_id` AS `combined_id`
FROM (`phd`.`cg1_1`
   LEFT JOIN ((`phd`.`lexeme`
          JOIN `phd`.`category` ON
   ((`phd`.`lexeme`.`catid` = `phd`.`category`.`catid`)))
     JOIN `phd`.`thematic_headings` ON
   ((`phd`.`thematic_headings`.`tid` =
   `phd`.`category`.`themid`))) ON
   ((`phd`.`lexeme`.`wordoed` =
   `phd`.`cg1_1`.`corp_lemma`)))
ORDER BY
   `phd`.`cg1_1`.`corp_id`
```

CG1_3_filter.sql

Creates `CG1_3_filter`

```
SELECT
   `cg1_2`.`corp_id` AS `corp_id`,
   `cg1_2`.`corp_word` AS `corp_word`,
   `cg1_2`.`corp_lemma` AS `corp_lemma`,
   `cg1_2`.`corp_pos` AS `corp_pos`,
   `cg1_2`.`word_mixed` AS `word_mixed`,
   `cg1_2`.`word` AS `word`,
   `cg1_2`.`htid` AS `htid`,
   `cg1_2`.`apps` AS `apps`,
   `cg1_2`.`appe` AS `appe`,
   `cg1_2`.`heading` AS `heading`,
   `cg1_2`.`catid` AS `catid`,
   `cg1_2`.`tiering` AS `tiering`,
   `cg1_2`.`pos` AS `pos`,
   `cg1_2`.`tid` AS `tid`,
```

```
   `cg1_2`.`thematicheading` AS `thematicheading`,
   `cg1_2`.`s1` AS `s1`,
   `cg1_2`.`s2` AS `s2`,
   `cg1_2`.`s3` AS `s3`,
   `cg1_ids`.`poemID` AS `poemID`,
   `cg1_ids`.`position_ID` AS `position_ID`,
   `cg1_ids`.`corpus_group` AS `corpus_group`,
   `cg1_ids`.`combined_id` AS `combined_id`
FROM (`phd`.`cg1_2`
   JOIN `phd`.`cg1_ids` ON ((`cg1_2`.`corp_id` =
   `cg1_ids`.`corp_id`)))
WHERE ((((`cg1_2`.`apps` <= '1700')
         OR isnull(`cg1_2`.`apps`))
     and((`cg1_2`.`appe` >= '1450')
     OR isnull(`cg1_2`.`appe`))
     and(not(`cg1_2`.`corp_pos` in(
                  SELECT
                       `phd`.`claws_filter`.`Tag` FROM
   `phd`.`claws_filter`))))
   OR isnull(`cg1_2`.`corp_pos`)
   or(`cg1_2`.`corp_pos` = 'YSTP')
   or((`cg1_2`.`corp_pos` = 'MC')
   AND isnull(`cg1_2`.`htid`)))
```

DuplicateValues.sql

```
SELECT
   `phd`.`cg1_3_ref`.`position_ID` AS `position_ID`,
   `phd`.`cg1_3_ref`.`corp_id` AS `corp_id`,
   `phd`.`cg1_3_ref`.`corp_word` AS `corp_word`,
   `phd`.`cg1_3_ref`.`corp_pos` AS `corp_pos`,
   `phd`.`cg1_3_ref`.`pos` AS `pos`,
   `phd`.`cg1_3_ref`.`apps` AS `apps`,
   `phd`.`cg1_3_ref`.`appe` AS `appe`,
   `phd`.`cg1_3_ref`.`corp_lemma` AS `corp_lemma`,
   `phd`.`cg1_3_ref`.`word_mixed` AS `word_mixed`,
   `phd`.`cg1_3_ref`.`heading` AS `heading`,
   `phd`.`cg1_3_ref`.`tiering` AS `tiering`,
   `phd`.`cg1_3_ref`.`thematicheading` AS
   `thematicheading`,
   `phd`.`cg1_3_ref`.`catid` AS `catid`,
   `phd`.`cg1_3_ref`.`htid` AS `htid`,
   `phd`.`cg1_3_ref`.`tid` AS `tid`,
   `phd`.`cg1_3_ref`.`s1` AS `s1`,
   `phd`.`cg1_3_ref`.`s2` AS `s2`,
   `phd`.`cg1_3_ref`.`s3` AS `s3`,
   `phd`.`cg1_3_ref`.`poemID` AS `poemID`,
   `phd`.`cg1_3_ref`.`corpus_group` AS `corpus_group`,
   `phd`.`cg1_3_ref`.`combined_id` AS `combined_id`,
   row_number() OVER (PARTITION BY
   `phd`.`cg1_3_ref`.`corp_id`,
```

```
    `phd`.`cg1_3_ref`.`catid` ORDER BY
  `phd`.`cg1_3_ref`.`catid`) AS `dup`
FROM
  `phd`.`cg1_3_ref`
```

CorpusRank.sql[85]

Creates `cg1_rmbycg`

```
SELECT
  `cg1_3_nodups`.`tid` AS `thematic_id`,
  `cg1_3_nodups`.`thematicheading` AS `thematic_heading`,
  count(`cg1_3_nodups`.`tid`) AS `appearance`,
  round((count(`cg1_3_nodups`.`tid`) /
  `thematic_headings_offset`.`them_value`), 2) AS
  `offset_appearance`
FROM (`phd`.`cg1_3_nodups`
  LEFT JOIN `phd`.`thematic_headings_offset` ON
  ((`cg1_3_nodups`.`tid` =
  `thematic_headings_offset`.`themid`)))
GROUP BY
  `cg1_3_nodups`.`tid`,
  `cg1_3_nodups`.`thematicheading`
ORDER BY
  `offset_appearance` DESC
```

RMVbyPoem.sql

```
SELECT
  `phd`.`cg1_3_ref`.`poemID` AS `poemID`,
  `phd`.`cg1_3_ref`.`tid` AS `tid`,
  count(0) AS `MV`,
  round(((count(`phd`.`cg1_3_ref`.`tid`) /
  `thematic_headings_offset`.`them_value`) * 100), 2) AS
  `RMV`
FROM (`phd`.`cg1_3_ref`
  LEFT JOIN `phd`.`thematic_headings_offset` ON
  ((`phd`.`cg1_3_ref`.`tid` =
  `thematic_headings_offset`.`themid`)))
GROUP BY
  `phd`.`cg1_3_ref`.`poemID`,
  `phd`.`cg1_3_ref`.`tid`
ORDER BY
  `phd`.`cg1_3_ref`.`poemID`,
  `RMV` DESC
```

[85] http://www.mysqltutorial.org/mysql-count/

Rank.sql

```
SELECT
   `phd`.`cg1_3_ref`.`position_ID` AS `position_ID`,
   `cg1_rmvbypoem`.`poemID` AS `poemID`,
   `cg1_rmvbypoem`.`tid` AS `tid`,
   `cg1_rmvbypoem`.`MV` AS `MV`,
   `cg1_rmvbypoem`.`RMV` AS `RMV`,
   `phd`.`cg1_3_ref`.`catid` AS `catid`,
   `phd`.`cg1_3_ref`.`heading` AS `heading`,
   `phd`.`cg1_3_ref`.`thematicheading` AS
   `thematicheading`,
   round(percent_rank() OVER (PARTITION BY
   `phd`.`cg1_3_ref`.`position_ID` ORDER BY
   `cg1_rmvbypoem`.`RMV` DESC), 2) AS `percentile_rank`,
   row_number() OVER (PARTITION BY
   `phd`.`cg1_3_ref`.`position_ID` ORDER BY
   `cg1_rmvbypoem`.`RMV` DESC) AS `rowcount`
FROM (`phd`.`cg1_3_ref`
   LEFT JOIN `phd`.`cg1_rmvbypoem` ON
   (((`phd`.`cg1_3_ref`.`poemID` =
   `cg1_rmvbypoem`.`poemID`)
               and(`phd`.`cg1_3_ref`.`tid` =
   `cg1_rmvbypoem`.`tid`))))
GROUP BY
   `phd`.`cg1_3_ref`.`position_ID`,
   `phd`.`cg1_3_ref`.`heading`,
   `phd`.`cg1_3_ref`.`thematicheading`,
   `cg1_rmvbypoem`.`poemID`,
   `phd`.`cg1_3_ref`.`catid`,
   `cg1_rmvbypoem`.`tid`,
   `cg1_rmvbypoem`.`MV`,
   `cg1_rmvbypoem`.`RMV`
ORDER BY
   `phd`.`cg1_3_ref`.`position_ID`,
   `percentile_rank`
```

CG1_4_OrderedByTH.sql

```
SELECT
   `phd`.`cg_ids_ref`.`position_ID` AS `position_ID`,
   `phd`.`cg_ids_ref`.`poemID` AS `poemID`,
   `phd`.`cg_ids_ref`.`corp_word` AS `corp_word`,
   `phd`.`cg_ids_ref`.`corp_pos` AS `corp_pos`,
   `phd`.`cg_ids_ref`.`corp_lemma` AS `corp_lemma`,
   `cg1_3_nodups`.`apps` AS `apps`,
   `cg1_3_nodups`.`appe` AS `appe`,
   `cg1_3_nodups`.`pos` AS `pos`,
```

```
   `cg1_3_nodups`.`word_mixed` AS `word_mixed`,
   `cg1_3_nodups`.`tid` AS `tid`,
   `cg1_3_nodups`.`thematicheading` AS `thematicheading`,
   `cg1_3_nodups`.`catid` AS `catid`,
   `cg1_3_nodups`.`heading` AS `heading`,
   `cg1_3_nodups`.`tiering` AS `tiering`,
   `cg1_3_nodups`.`htid` AS `htid`,
   `cg1_rmvbycg`.`appearance` AS `appearance`,
   `cg1_rmvbycg`.`offset_appearance` AS
   `offset_appearance`,
   `cg1_rank`.`MV` AS `MV`,
   `cg1_rank`.`RMV` AS `RMV`,
   `cg1_rank`.`percentile_rank` AS `percentile_rank`,
   `cg1_3_nodups`.`s1` AS `s1`,
   `cg1_3_nodups`.`s2` AS `s2`,
   `cg1_3_nodups`.`s3` AS `s3`,
   `cg1_3_nodups`.`corpus_group` AS `corpus_group`,
   `cg1_3_nodups`.`combined_id` AS `combined_id`,
  row_number() OVER (PARTITION BY
   `phd`.`cg_ids_ref`.`position_ID` ORDER BY
   `phd`.`cg_ids_ref`.`position_ID`,
     `cg1_rank`.`percentile_rank`,
     `cg1_3_nodups`.`catid`) AS `count_value`
FROM (((`phd`.`cg_ids_ref`
    LEFT JOIN `phd`.`cg1_3_nodups` ON
  ((`phd`.`cg_ids_ref`.`position_ID` =
   `cg1_3_nodups`.`position_ID`)))
  LEFT JOIN `phd`.`cg1_rank` ON
  (((`phd`.`cg_ids_ref`.`position_ID` =
   `cg1_rank`.`position_ID`)
             and(`cg1_3_nodups`.`tid` =
   `cg1_rank`.`tid`)
             and(`cg1_3_nodups`.`catid` =
   `cg1_rank`.`catid`))))
  LEFT JOIN `phd`.`cg1_rmvbycg` ON
  ((`cg1_rmvbycg`.`thematic_id` = `cg1_3_nodups`.`tid`)))
WHERE ((`cg1_rank`.`percentile_rank` < 0.20)
  OR isnull(`cg1_rank`.`percentile_rank`))
ORDER BY
   `phd`.`cg_ids_ref`.`position_ID`,
   `cg1_rank`.`percentile_rank`,
   `cg1_3_nodups`.`catid`
```

CG1_5_SummaryView.sql

```
SELECT
   `cg1_4`.`position_ID` AS `position_ID`,
   `cg1_4`.`corp_word` AS `corp_word`,
   `cg1_4`.`corp_lemma` AS `corp_lemma`,
   `cg1_4`.`corp_pos` AS `corp_pos`,
   `cg1_4`.`poemID` AS `poemID`,
```

```
min((
  CASE `cg1_4`.`count_value`
  WHEN '1' THEN
       concat(`cg1_4`.`thematicheading`, '; [',
`cg1_4`.`heading`, '; ', `cg1_4`.`pos`, '; ',
`cg1_4`.`apps`, '-', `cg1_4`.`appe`, ']; ', '†',
`cg1_4`.`tid`, '; [ ', '>>', `cg1_4`.`catid`, ']; #',
`cg1_4`.`RMV`, '/', `cg1_4`.`percentile_rank`)
  END)) AS `Position_1`,
min((
  CASE `cg1_4`.`count_value`
  WHEN '2' THEN
       concat(`cg1_4`.`thematicheading`, '; [',
`cg1_4`.`heading`, '; ', `cg1_4`.`pos`, '; ',
`cg1_4`.`apps`, '-', `cg1_4`.`appe`, ']; ', '†',
`cg1_4`.`tid`, '; [ ', '>>', `cg1_4`.`catid`, ']; #',
`cg1_4`.`RMV`, '/', `cg1_4`.`percentile_rank`)
  END)) AS `Position_2`,
min((
  CASE `cg1_4`.`count_value`
  WHEN '3' THEN
       concat(`cg1_4`.`thematicheading`, '; [',
`cg1_4`.`heading`, '; ', `cg1_4`.`pos`, '; ',
`cg1_4`.`apps`, '-', `cg1_4`.`appe`, ']; ', '†',
`cg1_4`.`tid`, '; [ ', '>>', `cg1_4`.`catid`, ']; #',
`cg1_4`.`RMV`, '/', `cg1_4`.`percentile_rank`)
  END)) AS `Position_3`,
min((
  CASE `cg1_4`.`count_value`
  WHEN '4' THEN
       concat(`cg1_4`.`thematicheading`, '; [',
`cg1_4`.`heading`, '; ', `cg1_4`.`pos`, '; ',
`cg1_4`.`apps`, '-', `cg1_4`.`appe`, ']; ', '†',
`cg1_4`.`tid`, '; [ ', '>>', `cg1_4`.`catid`, ']; #',
`cg1_4`.`RMV`, '/', `cg1_4`.`percentile_rank`)
  END)) AS `Position_4`,
min((
  CASE `cg1_4`.`count_value`
  WHEN '5' THEN
       concat(`cg1_4`.`thematicheading`, '; [',
`cg1_4`.`heading`, '; ', `cg1_4`.`pos`, '; ',
`cg1_4`.`apps`, '-', `cg1_4`.`appe`, ']; ', '†',
`cg1_4`.`tid`, '; [ ', '>>', `cg1_4`.`catid`, ']; #',
`cg1_4`.`RMV`, '/', `cg1_4`.`percentile_rank`)
  END)) AS `Position_5`,
min((
  CASE `cg1_4`.`count_value`
  WHEN '6' THEN
       concat(`cg1_4`.`thematicheading`, '; [',
`cg1_4`.`heading`, '; ', `cg1_4`.`pos`, '; ',
`cg1_4`.`apps`, '-', `cg1_4`.`appe`, ']; ', '†',
`cg1_4`.`tid`, '; [ ', '>>', `cg1_4`.`catid`, ']; #',
`cg1_4`.`RMV`, '/', `cg1_4`.`percentile_rank`)
```

```
    END)) AS `Position_6`,
   `cg1_4`.`corpus_group` AS `corpus_group`,
   `cg1_4`.`combined_id` AS `combined_id`
FROM
   `phd`.`cg1_4`
GROUP BY
   `cg1_4`.`corp_word`,
   `cg1_4`.`corp_lemma`,
   `cg1_4`.`corp_pos`,
   `cg1_4`.`position_ID`,
   `cg1_4`.`poemID`,
   `cg1_4`.`corpus_group`,
   `cg1_4`.`combined_id`
```

Log_Likelihood.sql

```
SELECT
   `cg_th`.`tid` AS `tid`,
   `cg_th`.`thematicheading` AS `thematicheading`,
   `cg_th`.`s1` AS `s1`,
   `cg_th`.`s2` AS `s2`,
   `cg_th`.`s3` AS `s3`,
   `cg_th`.`keats_aggregate` AS `keats_aggregate`,
   `cg_th`.`totals` AS `totals`,
   `cg_thref`.`Not_keats` AS `Not_keats`,
  round(((12340 * (`cg_th`.`keats_aggregate` +
   `cg_thref`.`Not_keats`)) / (12340 + 531686)), 2) AS
   `keats_exp`,
  round(((531686 * (`cg_th`.`keats_aggregate` +
   `cg_thref`.`Not_keats`)) / (12340 + 531686)), 2) AS
   `ref_exp`,
  round( if(isnull(`cg_th`.`keats_aggregate`), 0, (2 *
   ((`cg_th`.`keats_aggregate` *
   log((`cg_th`.`keats_aggregate` / ((12340 *
   (`cg_th`.`keats_aggregate` + `cg_thref`.`Not_keats`)) /
   (12340 + 531686))))) + (`cg_thref`.`Not_keats` *
   log((`cg_thref`.`Not_keats` / ((531686 *
   (`cg_th`.`keats_aggregate` + `cg_thref`.`Not_keats`)) /
   (12340 + 531686)))))))), 2) AS `keats_ll`
FROM (`phd`.`cg_th`
   LEFT JOIN `phd`.`cg_thref` ON ((`cg_th`.`tid` =
   `cg_thref`.`tid`)))
GROUP BY
   `cg_th`.`tid`,
   `cg_th`.`thematicheading`,
   `cg_th`.`s1`,
   `cg_th`.`s2`,
   `cg_th`.`s3`,
   `cg_th`.`keats_aggregate`,
   `cg_th`.`totals`,
```

```
   `cg_thref`.`Not_keats`
ORDER BY
   `cg_th`.`tid`
```

Collmatch.sql

```
SELECT
   `phd`.`cg_4_ref`.`position_ID` AS `position_ID`,
   (
     CASE WHEN isnull(`phd`.`cg_4_ref`.`corp_pos`) THEN
          `phd`.`cg_4_ref`.`corp_word`
     WHEN isnull(`phd`.`cg_4_ref`.`pos`) THEN
          `phd`.`cg_4_ref`.`corp_pos`
     ELSE
          `phd`.`cg_4_ref`.`catid`
     END) AS `determiner`,
   `phd`.`cg_4_ref`.`corp_word` AS `corp_word`,
   `phd`.`cg_4_ref`.`corp_pos` AS `corp_pos`,
   `phd`.`cg_4_ref`.`pos` AS `pos`,
   `phd`.`cg_4_ref`.`apps` AS `apps`,
   `phd`.`cg_4_ref`.`appe` AS `appe`,
   `phd`.`cg_4_ref`.`corp_lemma` AS `corp_lemma`,
   `phd`.`cg_4_ref`.`word_mixed` AS `word_mixed`,
   `phd`.`cg_4_ref`.`heading` AS `heading`,
   `phd`.`cg_4_ref`.`tiering` AS `tiering`,
   `phd`.`cg_4_ref`.`thematicheading` AS
   `thematicheading`,
   `phd`.`cg_4_ref`.`catid` AS `catid`,
   `phd`.`cg_4_ref`.`htid` AS `htid`,
   `phd`.`cg_4_ref`.`tid` AS `tid`,
   `phd`.`cg_4_ref`.`appearance` AS `appearance`,
   `phd`.`cg_4_ref`.`offset_appearance` AS
   `offset_appearance`,
   `phd`.`cg_4_ref`.`MV` AS `MV`,
   `phd`.`cg_4_ref`.`RMV` AS `RMV`,
   `phd`.`cg_4_ref`.`percentile_rank` AS
   `percentile_rank`,
   `phd`.`cg_4_ref`.`s1` AS `s1`,
   `phd`.`cg_4_ref`.`s2` AS `s2`,
   `phd`.`cg_4_ref`.`s3` AS `s3`,
   `phd`.`cg_4_ref`.`count_value` AS `count_value`,
   `phd`.`cg_4_ref`.`corpus_group` AS `corpus_group`,
   `phd`.`cg_4_ref`.`poemID` AS `poemID`,
   `phd`.`cg_4_ref`.`combined_id` AS `combined_id`
FROM
   `phd`.`cg_4_ref`
```

Collpairs.sql

```
SELECT
   `e1`.`position_ID` AS `position_ID`,
   `phd`.`corpus_id`.`corpus_group` AS `CG`,
   `phd`.`corpus_id`.`poemID` AS `poemID`,
   min((
     CASE WHEN (`e1`.`count_value` = '1') THEN
          concat((
               CASE WHEN (`e1`.`corp_lemma` IS NOT NULL)
   THEN
                    `e1`.`corp_lemma`
               ELSE
                    `e1`.`corp_word`
               END), ';', (
               CASE WHEN ((
                    SELECT
                         `e2`.`corp_lemma` FROM
   `phd`.`cg_collmatch_ref` `e2`
                    WHERE ((`e2`.`position_ID` =
   (`e1`.`position_ID` + 1))
                    and(`e2`.`count_value` = '1'))
          ORDER BY
               `e2`.`position_ID`
          LIMIT 1) IS NOT NULL) THEN
   (
     SELECT
          `e2`.`corp_lemma` FROM `phd`.`cg_collmatch_ref`
   `e2`
     WHERE ((`e2`.`position_ID` = (`e1`.`position_ID` +
   1))
     and(`e2`.`count_value` = '1'))
ORDER BY
   `e2`.`position_ID`
LIMIT 1)
ELSE
   (
     SELECT
          `e2`.`corp_word` FROM `phd`.`cg_collmatch_ref`
   `e2`
     WHERE ((`e2`.`position_ID` = (`e1`.`position_ID` +
   1))
     and(`e2`.`count_value` = '1'))
ORDER BY
   `e2`.`position_ID`
LIMIT 1)
               END))
     END)) AS `1-1+words`,
   min((
     CASE WHEN (`e1`.`count_value` = '1') THEN
          concat(`e1`.`determiner`, ';', (
```

```
                SELECT
                     `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '1'))
       ORDER BY
            `e2`.`position_ID`
       LIMIT 1))
  END)) AS `1-1+`,
min((
  CASE WHEN (`e1`.`count_value` = '1') THEN
       concat(`e1`.`determiner`, ';', (
                SELECT
                     `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '2'))
       ORDER BY
            `e2`.`position_ID`
       LIMIT 1))
  END)) AS `1-2+`,
min((
  CASE WHEN (`e1`.`count_value` = '1') THEN
       concat(`e1`.`determiner`, ';', (
                SELECT
                     `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '3'))
       ORDER BY
            `e2`.`position_ID`
       LIMIT 1))
  END)) AS `1-3+`,
min((
  CASE WHEN (`e1`.`count_value` = '1') THEN
       concat(`e1`.`determiner`, ';', (
                SELECT
                     `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '4'))
       ORDER BY
            `e2`.`position_ID`
       LIMIT 1))
  END)) AS `1-4+`,
min((
  CASE WHEN (`e1`.`count_value` = '1') THEN
       concat(`e1`.`determiner`, ';', (
                SELECT
```

```
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '5'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `1-5+`,
min((
  CASE WHEN (`e1`.`count_value` = '1') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '6'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `1-6+`,
min((
  CASE WHEN (`e1`.`count_value` = '2') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '1'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `2-1+`,
min((
  CASE WHEN (`e1`.`count_value` = '2') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '2'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `2-2+`,
min((
  CASE WHEN (`e1`.`count_value` = '2') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
```

```
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '3'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `2-3+`,
min((
  CASE WHEN (`e1`.`count_value` = '2') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '4'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `2-4+`,
min((
  CASE WHEN (`e1`.`count_value` = '2') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '5'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `2-5+`,
min((
  CASE WHEN (`e1`.`count_value` = '2') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '6'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `2-6+`,
min((
  CASE WHEN (`e1`.`count_value` = '3') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
```

```
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '1'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `3-1+`,
min((
  CASE WHEN (`e1`.`count_value` = '3') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '2'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `3-2+`,
min((
  CASE WHEN (`e1`.`count_value` = '3') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '3'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `3-3+`,
min((
  CASE WHEN (`e1`.`count_value` = '3') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '4'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `3-4+`,
min((
  CASE WHEN (`e1`.`count_value` = '3') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
```

```
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '5'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `3-5+`,
min((
  CASE WHEN (`e1`.`count_value` = '3') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '6'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `3-6+`,
min((
  CASE WHEN (`e1`.`count_value` = '4') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '1'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `4-1+`,
min((
  CASE WHEN (`e1`.`count_value` = '4') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '2'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `4-2+`,
min((
  CASE WHEN (`e1`.`count_value` = '4') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
```

```
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '3'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `4-3+`,
min((
  CASE WHEN (`e1`.`count_value` = '4') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '4'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `4-4+`,
min((
  CASE WHEN (`e1`.`count_value` = '4') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '5'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `4-5+`,
min((
  CASE WHEN (`e1`.`count_value` = '4') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '6'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `4-6+`,
min((
  CASE WHEN (`e1`.`count_value` = '5') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
```

```
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '1'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `5-1+`,
min((
  CASE WHEN (`e1`.`count_value` = '5') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '2'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `5-2+`,
min((
  CASE WHEN (`e1`.`count_value` = '5') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '3'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `5-3+`,
min((
  CASE WHEN (`e1`.`count_value` = '5') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '4'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `5-4+`,
min((
  CASE WHEN (`e1`.`count_value` = '5') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
```

```
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '5'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `5-5+`,
min((
  CASE WHEN (`e1`.`count_value` = '5') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '6'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `5-6+`,
min((
  CASE WHEN (`e1`.`count_value` = '6') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '1'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `6-1+`,
min((
  CASE WHEN (`e1`.`count_value` = '6') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
                    `e2`.`determiner` FROM
`phd`.`cg_collmatch_ref` `e2`
                WHERE ((`e2`.`position_ID` =
(`e1`.`position_ID` + 1))
                and(`e2`.`count_value` = '2'))
        ORDER BY
             `e2`.`position_ID`
        LIMIT 1))
  END)) AS `6-2+`,
min((
  CASE WHEN (`e1`.`count_value` = '6') THEN
        concat(`e1`.`determiner`, ';', (
                SELECT
```

```
                        `e2`.`determiner` FROM
   `phd`.`cg_collmatch_ref` `e2`
                    WHERE ((`e2`.`position_ID` =
   (`e1`.`position_ID` + 1))
                    and(`e2`.`count_value` = '3'))
          ORDER BY
               `e2`.`position_ID`
          LIMIT 1))
     END)) AS `6-3+`,
   min((
     CASE WHEN (`e1`.`count_value` = '6') THEN
          concat(`e1`.`determiner`, ';', (
                    SELECT
                         `e2`.`determiner` FROM
   `phd`.`cg_collmatch_ref` `e2`
                    WHERE ((`e2`.`position_ID` =
   (`e1`.`position_ID` + 1))
                    and(`e2`.`count_value` = '4'))
          ORDER BY
               `e2`.`position_ID`
          LIMIT 1))
     END)) AS `6-4+`,
   min((
     CASE WHEN (`e1`.`count_value` = '6') THEN
          concat(`e1`.`determiner`, ';', (
                    SELECT
                         `e2`.`determiner` FROM
   `phd`.`cg_collmatch_ref` `e2`
                    WHERE ((`e2`.`position_ID` =
   (`e1`.`position_ID` + 1))
                    and(`e2`.`count_value` = '5'))
          ORDER BY
               `e2`.`position_ID`
          LIMIT 1))
     END)) AS `6-5+`,
   min((
     CASE WHEN (`e1`.`count_value` = '6') THEN
          concat(`e1`.`determiner`, ';', (
                    SELECT
                         `e2`.`determiner` FROM
   `phd`.`cg_collmatch_ref` `e2`
                    WHERE ((`e2`.`position_ID` =
   (`e1`.`position_ID` + 1))
                    and(`e2`.`count_value` = '6'))
          ORDER BY
               `e2`.`position_ID`
          LIMIT 1))
     END)) AS `6-6+`
FROM (`phd`.`cg_collmatch_ref` `e1`
   LEFT JOIN `phd`.`corpus_id` ON ((`e1`.`position_ID` =
   `phd`.`corpus_id`.`position_ID`)))
GROUP BY
   `e1`.`position_ID`
```

Collcount.sql

```
SELECT
   `t`.`p` AS `p`,
   count(0) AS `combcount`,
   left(`t`.`p`, (locate(';', `t`.`p`) - 1)) AS
   `first_tag`,
   right(`t`.`p`, (length(`t`.`p`) - locate(';',
   `t`.`p`))) AS `second_tag`
FROM (
   SELECT
     `phd`.`cg_collpairs_ref`.`1-1+` AS `p`
   FROM
     `phd`.`cg_collpairs_ref`
   UNION ALL
   SELECT
     `phd`.`cg_collpairs_ref`.`1-2+` AS `p`
   FROM
     `phd`.`cg_collpairs_ref`
   UNION ALL
   SELECT
     `phd`.`cg_collpairs_ref`.`1-3+` AS `p`
   FROM
     `phd`.`cg_collpairs_ref`
   UNION ALL
   SELECT
     `phd`.`cg_collpairs_ref`.`1-4+` AS `p`
   FROM
     `phd`.`cg_collpairs_ref`
   UNION ALL
   SELECT
     `phd`.`cg_collpairs_ref`.`1-5+` AS `p`
   FROM
     `phd`.`cg_collpairs_ref`
   UNION ALL
   SELECT
     `phd`.`cg_collpairs_ref`.`1-6+` AS `p`
   FROM
     `phd`.`cg_collpairs_ref`
   UNION ALL
   SELECT
     `phd`.`cg_collpairs_ref`.`2-1+` AS `p`
   FROM
     `phd`.`cg_collpairs_ref`
   UNION ALL
   SELECT
     `phd`.`cg_collpairs_ref`.`2-2+` AS `p`
   FROM
```

```
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`2-3+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`2-4+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`2-5+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`2-6+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`3-1+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`3-2+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`3-3+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`3-4+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`3-5+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`3-6+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
```

```
  `phd`.`cg_collpairs_ref`.`4-1+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`4-2+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`4-3+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`4-4+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`4-5+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`4-6+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`5-1+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`5-2+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`5-3+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`5-4+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
UNION ALL
SELECT
  `phd`.`cg_collpairs_ref`.`5-5+` AS `p`
FROM
  `phd`.`cg_collpairs_ref`
```

```
  UNION ALL
  SELECT
    `phd`.`cg_collpairs_ref`.`5-6+` AS `p`
  FROM
    `phd`.`cg_collpairs_ref`
  UNION ALL
  SELECT
    `phd`.`cg_collpairs_ref`.`6-1+` AS `p`
  FROM
    `phd`.`cg_collpairs_ref`
  UNION ALL
  SELECT
    `phd`.`cg_collpairs_ref`.`6-2+` AS `p`
  FROM
    `phd`.`cg_collpairs_ref`
  UNION ALL
  SELECT
    `phd`.`cg_collpairs_ref`.`6-3+` AS `p`
  FROM
    `phd`.`cg_collpairs_ref`
  UNION ALL
  SELECT
    `phd`.`cg_collpairs_ref`.`6-4+` AS `p`
  FROM
    `phd`.`cg_collpairs_ref`
  UNION ALL
  SELECT
    `phd`.`cg_collpairs_ref`.`6-5+` AS `p`
  FROM
    `phd`.`cg_collpairs_ref`
  UNION ALL
  SELECT
    `phd`.`cg_collpairs_ref`.`6-6+` AS `p`
  FROM
    `phd`.`cg_collpairs_ref`) `t`
GROUP BY
  `t`.`p`
ORDER BY
  `combcount` DESC
```

Colldescs.sql

```
SELECT
  `phd`.`cg_collcount_ref`.`p` AS `p`,
  `phd`.`cg_collcount_ref`.`combcount` AS `combcount`,
  `phd`.`cg_collcount_ref`.`first_tag` AS `l`,
  `phd`.`cg_collcount_ref`.`second_tag` AS `r`,
  (
```

```
  CASE WHEN
regexp_like(`phd`.`cg_collcount_ref`.`first_tag`, '^[0-
9]') THEN
  (
       SELECT
            `phd`.`category`.`heading`
       FROM
            `phd`.`category`
       WHERE (`phd`.`cg_collcount_ref`.`first_tag` =
`phd`.`category`.`catid`))
ELSE
  (
       SELECT
            `phd`.`claws_tags`.`Description`
       FROM
            `phd`.`claws_tags`
       WHERE (`phd`.`cg_collcount_ref`.`first_tag` =
`phd`.`claws_tags`.`Tag`))
  END) AS `l_definition`, (
  CASE WHEN
regexp_like(`phd`.`cg_collcount_ref`.`second_tag`,
'^[0-9]') THEN
  (
       SELECT
            `phd`.`category`.`heading`
       FROM
            `phd`.`category`
       WHERE (`phd`.`cg_collcount_ref`.`second_tag` =
`phd`.`category`.`catid`))
ELSE
  (
       SELECT
            `phd`.`claws_tags`.`Description`
       FROM
            `phd`.`claws_tags`
       WHERE (`phd`.`cg_collcount_ref`.`second_tag` =
`phd`.`claws_tags`.`Tag`))
  END) AS `r_definition`, concat((
  CASE WHEN
regexp_like(`phd`.`cg_collcount_ref`.`first_tag`, '^[0-
9]') THEN
  (
       SELECT
            `phd`.`category`.`heading` FROM
`phd`.`category`
       WHERE (`phd`.`cg_collcount_ref`.`first_tag` =
`phd`.`category`.`catid`))
ELSE
  (
       SELECT
            `phd`.`claws_tags`.`Description` FROM
`phd`.`claws_tags`
```

```
          WHERE (`phd`.`cg_collcount_ref`.`first_tag` =
   `phd`.`claws_tags`.`Tag`))
    END), ';', (
    CASE WHEN
   regexp_like(`phd`.`cg_collcount_ref`.`second_tag`,
   '^[0-9]') THEN
    (
          SELECT
               `phd`.`category`.`heading` FROM
   `phd`.`category`
          WHERE (`phd`.`cg_collcount_ref`.`second_tag` =
   `phd`.`category`.`catid`))
   ELSE
    (
          SELECT
               `phd`.`claws_tags`.`Description` FROM
   `phd`.`claws_tags`
          WHERE (`phd`.`cg_collcount_ref`.`second_tag` =
   `phd`.`claws_tags`.`Tag`))
    END)) AS `both_definitions`
FROM ((`phd`.`cg_collcount_ref`
   LEFT JOIN `phd`.`category` ON
   (((`phd`.`cg_collcount_ref`.`first_tag` =
   `phd`.`category`.`catid`)
               and(`phd`.`cg_collcount_ref`.`second_tag` =
   `phd`.`category`.`catid`))))
   LEFT JOIN `phd`.`claws_tags` ON
   (((`phd`.`cg_collcount_ref`.`first_tag` =
   `phd`.`claws_tags`.`Tag`)
               and(`phd`.`cg_collcount_ref`.`second_tag` =
   `phd`.`claws_tags`.`Tag`))))
ORDER BY
   `phd`.`cg_collcount_ref`.`combcount` DESC
```

Appendix III Digital appendices

DA1 List of poems.xlsx

A numbered list of all poems in the OBEV collection, available as a cross-reference to the corpus group divisions described in Appendix Table A1 (p.198). Referred to in §4.2.2.

DA2 Post-processing output for all corpus groups

The post-processing results for all poems in the four corpus groups is provided under CG1_4.csv, CG2_4.csv, CG3_4.csv, and CG4_4.csv for CG1, CG2, CG3, and CG4, respectively. These tables were produced by the CG1_4_OrderedByTH.sql query, as discussed in §4.6.4.

DA3 Summary view results for all corpus groups

The post-processing results for all poems in the four corpus groups is provided under CG1_5.csv, CG2_5.csv, CG3_5.csv, and CG4_5.csv for CG1, CG2, CG3, and CG4, respectively. These tables were produced by the CG1_5_SummaryView.sql query, as discussed in §4.6.5.

DA4 CG_collpairs.csv

All tag pairs for up to six distinct semantic tags or identifiers for each node, with a maximum of 36 possible variations for each node. See §7.3.2 (p.145) for example of the first two pair sets. Referred to in §7.3.2.

DA5 CG_collpairs_nopunc.csv

A modified version of DA4, showing tag pairs after removing punctuation from the corpus. Section §7.3.3.a (p.152) explains the justification and impact of this modification.

DA6 S1andS2positivecolls.xlsx

All positive collocates for the S1 (Metaphysical poets, §7.5.1) and S2 (Cavalier poets, §7.5.2) samples. See Table A23 and Table A24 for the first 50 results from each sample, as discussed in §7.6.