Digital Scholarship in the Humanities Advance Access published October 1, 2015

Metaphor, Popular Science, and Semantic Tagging: Distant reading with the *Historical Thesaurus of English*

Marc Alexander and Fraser Dallachy
University of Glasgow, UK

Scott Piao, Alistair Baron and Paul Rayson
Lancaster University, UK

Abstract

The use of metaphor in popular science is widespread to aid readers' conceptions of the scientific concepts under discussion. Almost all research in this area has been done by careful close reading of the text(s) in question, but this article describes—for the first time—a digital 'distant reading' analysis of popular science, using a system created by a team from Glasgow and Lancaster. This team, as part of the SAMUELS project, has developed semantic tagging software which is based upon the UCREL Semantic Analysis System developed by Lancaster University's University Centre for Computer Corpus Research on Language, but using the uniquely comprehensive *Historical Thesaurus of English* (published in 2009 as *The Historical Thesaurus of the Oxford English Dictionary*) as its knowledge base, in order to provide fine-grained meaning distinctions for use in word-sense disambiguation. In addition to analyzing metaphors in highly abstract book-length popular science texts from physics and mathematics, this article describes the technical underpinning to the system and the methods employed to hone the word-sense disambiguation procedure.

Correspondence:
Marc Alexander,
12 University Gardens,
University of Glasgow
G12 8QQ, UK.
E-mail:
marc.alexander@glasgow.ac.uk

1 Introduction

The SAMUELS project addresses a very real and growing problem: the need to search increasingly large corpora of textual data in an effective and focused way. Truly effective searching of text is currently hindered by a need to search using word forms, while in actual use almost all searches are aimed at the 'meaning' behind that word form. This would be acceptable should each word form have only one meaning, but this is far from the case—for example, the *Historical Thesaurus of English* (Kay *et al.*, 2009; henceforth HT) recognizes 104 noun meanings of the word-form 'set', and so current search technology means that a user must often sift through results which include all of these possibilities, so that mathematical number sets (HT cat. 01.16.04.04.01-02) are entangled with sets of dancers (HT 03.13.05.07.06-02), and even potentially badger sets (HT 01.05.19.05.06-10.05).[1] While these possibilities can be filtered based on other words in the search context, there is still an extensive need for manual checking and filtering, and the searcher is further hindered by their

doi:10.1093/llc/fqv045

frequent inability to specify the word's grammatical part of speech—meaning that all 408 meanings of 'set' across all parts of speech may be returned in a search. Overall, in the HT data set, 62% of English word forms refer to more than one meaning (67 word forms in English have more than 100 possible meanings, 2,580 have more than 20 possible meanings, and 111,127 have more than one possible meaning).

As searching is an increasingly essential part of effectively using the information contained in rapidly growing textual data sets such as digitized books, Internet news, and social media content, this problem of word forms and polysemy harms business and the general public just as much as academia; analysis of textual data allows companies to finesse their business strategies, general Internet users to find the information they require more rapidly, and researchers to identify patterns in data sets too large to be 'read' by a human researcher. In place of using these word forms, therefore, a key development in tackling the issue of search difficulty is the development of a capability to run semantic searches, in which the 'meanings' of words are primary rather than their word form proxy.

To this end, the SAMUELS project aims to exploit the enormous potential of the HT as a comprehensive lexical and semantic database of the language. Members of University Centre for Computer Corpus Research on Language (UCREL) at Lancaster University already possess a highly successful semantic tagger in the form of the UCREL Semantic Analysis System (USAS). This system utilizes a thesaurus based on the Longman Lexicon of Contemporary English (McArthur, 1981; cf. Piao *et al.*, 2005). The SAMUELS project aims to be a significant step forward from the USAS tagger by honing the level of detail it can achieve. This improvement is based on the implementation of the HT database as the key knowledge base of a disambiguating semantic tagger. Containing as it does more than 700,000 word senses arranged into 225,000 categories, the HT is a powerful tool for correctly labeling the meanings of words in a text.

This article focuses on a particular use of such a semantic tagger: the ability to undertake 'distant reading', to use Franco Moretti's (2013) term, or 'macroanalysis' from the perspective of Matt Jockers (2013), to achieve a large-scale 'specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models.' (Moretti, 2005, p. 2). The article therefore first gives an overview of the Historical Thesaurus Semantic Tagger (HTST), its relationship to the UCREL-developed tagging software that has preceded it, and the methods by which it identifies the HT meaning code with which each word should be labeled. In the second section, it details the linguistic methods dependent upon the HT data set which have been employed to refine the disambiguation of word senses. In the third and final section, it describes a test-case application of the HTST: automatic identification of semantic domains used analogically in popular science texts.

2 The Tagging System

The system we call the HTST extends a suite of existing corpus annotation tools developed in the UCREL[2] research center at Lancaster University. The particular tools incorporated in the HTST include VARiant Detector (VARD)[3] (Baron & Rayson, 2008), Constituent Likelihood Automatic Word-tagging System (CLAWS)[4] (Garside and Smith, 1997), and USAS (Rayson *et al.*, 2004). The first tool, VARD, is used to identify spelling variants, particularly those in historical texts, and link them to modern standard spellings by employing a number of dictionaries, phonetic matching, an edit distance metric, letter replacement heuristics, and statistical models. CLAWS is a part-of-speech (POS) tagger, which has been used to annotate a wide range of corpora and has been proven to be one of the most accurate English POS taggers (Leech and Smith, 2000). The final tool, USAS (UCREL Semantic Annotation System), is semantic annotation software that employs a coarse-grained semantic taxonomic scheme containing 21 major categories and 232 subcategories aimed primarily at modern English texts.

This combination of tools provides a range of functionalities of text annotation which are either

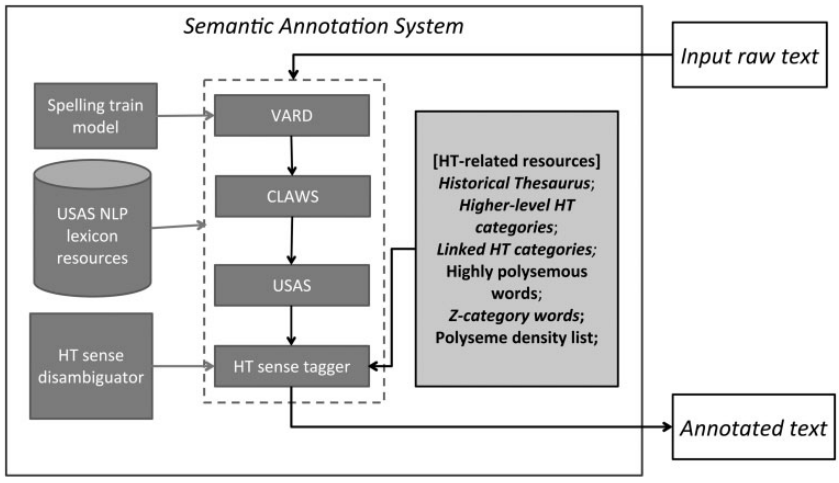

Fig. 1 Architecture of the HTST system

necessary or helpful in assigning the correct HT code for each of the words in a text. For example, the VARD tool enables us to search the HT database using standard spellings of words, CLAWS helps us to retrieve relevant semantic information from the HT by constraining the search with POS information, and USAS provides broader semantic information for words, which is helpful for the disambiguation of HT semantic concepts. Based on these tools, a new tagger component has been developed by incorporating the *Historical Thesaurus*, which provides a large-scale semantic lexical resource that employs a highly fine-grained semantic classification scheme. The system also employs a set of related sub-lexicons, such as one providing default senses for highly polysemous words, which are used to assist the semantic disambiguation of words. With the set of annotation tools included in the HTST system, it is capable of producing a multi-layered annotation of texts. The individual tools form a pipeline system and each of the tools adds its own layer of annotation to the input text. Figure 1 illustrates the architecture of the HTST system.

In addition to single words, the HTST system is also capable of identifying and annotating multi-word expressions (MWEs) as single semantic units. Figure 2 shows a sample annotation output that contains MWE annotation. The MWEs 'bear in mind', 'cost of living', and 'New York' are annotated with semantic concepts of 'Memory, keeping in mind', 'Expenditure', and 'Geographical name' respectively. We estimate that at least 16% of running text (tokens) consists of MWEs, based on the number marked by USAS. By coincidence around 16% of the entries (types) in the HT are MWEs. HTST's capability to annotate MWEs significantly improves the quality of semantic analysis of text since it permits phrasal verbs, compound nouns, named entities such as people, places, and organizations as well as non-compositional idiomatic expressions, to be treated as single units.

Currently, for a given text, the system produces six layers of annotation, including lemmas for the input words, POS information for words, USAS semantic codes, MWE flags (indicating whether or not a word is part of a MWE), HT full sense codes, and HT thematic sense codes. For example, for the input word 'children', the system assigns the following tags/codes:

- Lemma: '*child*'
- POS: '*NN2*' (plural noun)
- USAS semantic tag: '*S2mf/T3-*' (people/new and young)
- MWE flag: '*0*' (not part of MWE)
- HT sense code: '*01.04.04.04*' (HT code for 'child' category)
- Thematic level sense code: '*AD.03.d*' (thematic code for 'child' category)

Annotation result:

TOKEN	LEMMA	POSTAG	SEMTAG1	MWE	SEMTAG2	SEMTAG3
S_BEGIN	NULL	NULL	Z99	0	NULL	NULL
You	you	PPY	Z8mf	0	04.06 [];	ZF [Pronoun];
must	must	VM	S6+ A7+	0	02.05.02-04.01.01 [0.88888889] [at the time (in virtual oblique narration)]; 02.05.02-04 [0.91666667] [be absolutely compelled/obliged];	AV.01.b [Necessity];
bear	bear	VVI	X2.2+	1:3:1	[MWE] 02.01.11.01 [Retain in the memory Retain in the memory]	AR.35 [Memory, keeping in mind]
in	in	II	X2.2+	1:3:2	[MWE] 02.01.11.01 [Retain in the memory Retain in the memory]	AR.35 [Memory, keeping in mind]
mind	mind	NN1	X2.2+	1:3:3	[MWE] 02.01.11.01 [Retain in the memory Retain in the memory]	AR.35 [Memory, keeping in mind]
the	the	AT	Z5	0	04.03 [Grammatical]	ZC [Grammatical Item];
cost	cost	NN1	I1.3	2:3:1	[MWE] 03.12.20.02-07.10 [Spend cost of living]	BJ.01.y.02 [Expenditure]
of	of	IO	I1.3	2:3:2	[MWE] 03.12.20.02-07.10 [Spend cost of living]	BJ.01.y.02 [Expenditure]
living	living	NN1	I1.3	2:3:3	[MWE] 03.12.20.02-07.10 [Spend cost of living]	BJ.01.y.02 [Expenditure]
is	be	VBZ	A3+ Z5	0	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];
higher	high	JJR	N3.7++ N5++ A11.1++	0	01.12.05.07 [0.92307692] [High in position]; 02.04.10.10 [0.92857143] [Merry]; 01.16.06.03.01 [0.93750000] [Great in degree];	AL.05.g [High position]; AU.12.a [Merriment]; AP.06.a.01 [High/intense degree];
in	in	II	Z5	0	04.03 [Grammatical]	ZC [Grammatical Item];
New	new	NP1	Z2	3:2:1	04.01.02 [Geographical Name];	ZA02 [Geographical Name];
York	york	NP1	Z2	3:2:2	04.01.02 [Geographical Name];	ZA02 [Geographical Name];
.	PUNC	YSTP	PUNC	0	NULL	NULL
S_END	NULL	NULL	Z99	0	NULL	NULL

Fig. 2 A sample HTST annotation output

In the SAMUELS project, we focus on assigning correct HT sense codes to the words. With respect to the HT semantic annotation, the system is designed to produce two layers of HT sense codes. The first layer consists of the original HT semantic codes which are derived from the HT semantic classification scheme consisting of over 225,000 highly fine-grained semantic categories. This layer of annotation provides highly specific semantic categories such as '01.12.05.09.01-08.05 peg/nail' and '01.16.07.05.08.03.01 Loosen/unfasten/untie'. In order to make the sense classification and semantic codes more manageable for human researchers, these categories are grouped into 4,033 thematic semantic categories, such as 'AB.17.e.05 Skin' and 'AP.05.a Measurement of length', which still provide a fairly fine classification scheme for a deep semantic analysis of text.

Word sense disambiguation has been a difficult challenge for computational linguistics and Natural Language Processing. This is particularly the case when attempting to assign codes from the highly fine-grained HT semantic classification scheme. We are exploring and testing a range of methods

and techniques to achieve a highly accurate semantic disambiguation (see Section 3 below). Currently, the main generic algorithm we have implemented in our prototype system is a context distance-based method. Thus, where a word/MWE has more than one candidate HT category, these are ranked by their relative distances to the surrounding context. Here the distance can be defined in various ways, and different techniques can be used to calculate such a distance. We started with a word-based algorithm, drawing upon the high-quality brief definitions (named headings in HT, similar to semantic primes) of each semantic concept provided by the HT. Most HT semantic codes consist of multiple layers, in which each of the layers is defined by a heading. For example, in the HT code '03.12.20.02-07.10' for 'cost of living', there are six nested headings 'Society', 'Trade and finance', 'Management of money', 'Expenditure', 'expenses', and 'cost of living' that correspond to the sub-layers '03', '03.12', '03.12.20', '03.12.20.02', '03.12.20.02-07', '03.12.20.02-07.10', respectively. This allows us to obtain a set of words which describe and define the given HT semantic category, and which can be compared against the words in the surrounding context for measuring distance.

Below is an outline of the disambiguation process. Given a word with multiple candidate HT sense categories:

- For each of the candidate categories, extract all possible parent categories and collect headings (simple definitions) of them, in addition to that of current heading. The words in the headings form a feature set $HW_i = \{h_1, h_2, \ldots, h_m\}$.
- Collect up to five content words from each side of the key word/MWE. Together with the target word/MWE w_t, they form a context feature set $CW = \{w_t, w_1, w_2, \ldots, w_n\}$.
- Measure the Jaccard Distance (Choi *et al.*, 2010) between CW and each HW_i, and select the candidate categories (up to three) that have the closest distances to the context.
- If the previous steps fail,
  - Check core HT categories of the key word/MWE from a manually compiled list.
  - If not found, check for default HT categories from a polyseme density list.
- The resulting codes are then mapped into the thematic sense codes.

In our first test on 10 manually annotated sample texts from various genres and domains, this approach obtained precisions ranging from 71.74% (ENRON email corpus[5]) to 80.33% (Hansard Corpus[6]). Generally, the first iteration of the tagger performed better on formal texts. The lower performance on noisy data such as email is partially due to the fact that the annotation tools that are used as preprocessors were trained on standard English texts, and hence performed less well on noisy text. The exception to this rule is VARD, which is trained by default to deal with historical texts; adaptation and training is underway to apply VARD to Computer-Mediated Communication, such as email (Tagg *et al.*, 2014). The second half of the present SAMUELS project focuses on adding further resources (see Section 3 below) to substantially improve these accuracy figures.

Because HTST is resource-intensive software which depends on large lexical data sets and employs complex computational algorithms in each of the stages described above, it is built as a web service to achieve fast speed and scalability. Such a design brings benefits of allowing flexible remote access to the system. Currently the HTST system can be accessed in three ways:

(1) Access via a demo Web site,[7] which provides limited access for a quick trial.
(2) Access via a graphical user interface client tool, which users can run on their computer to process larger data.
(3) Access via server client software, which can be used to connect programmatically from users' software system for efficiently processing data on a large scale.

As the project progresses, the HTST system will be integrated into the corpus processing and retrieval site Wmatrix API,[8] which will support wider and more convenient access to the system along with corpus indexing and retrieval functions for different research communities.

3 Disambiguation

In addition to the distance algorithm described above, the HTST uses a range of new techniques to accurately determine the meaning of a word form. A number of these are yet to be implemented (and form the second half of the present development project), but initial tests and manual analyses show they have the potential to drastically improve word sense disambiguation.

Firstly, the HT data set contains information about the dates of usage for any word form in any particular meaning—for example, the word form 'wine' has been used to mean the alcoholic drink produced from grapes since Old English, but has only been used to mean a deep crimson color since around 1895, and is only used as an intransitive verb (meaning to drink wine, as in 'to wine and dine') after 1829. Of the 18 possible senses of the word 'wine', only 4 were in use in the 1400s, and only 10 are in use today. The system therefore allows date filtering based on the date of an input text in order to narrow down potential word sense matches.

Secondly, we have a new technique to use as a proxy of the importance of a word sense. Word sense importance—that is to say, how prominent or core a particular word sense is compared to another—is generally calculated as a frequency measure, with the rate of occurrence of a word in an appropriate corpus or text collection being used as a measure of how relatively important that word sense is compared to other senses. We cannot accurately do this for data in the past—the corpus of historical English is skewed by our knowledge of only those texts which happen to survive. Instead, we use polyseme density as a measure of importance—those word forms which bud off additional meanings and parts of speech of the same word form 'in very similar meanings' are the more prominent of those word forms, generally speaking, than the more isolated word forms. Returning to 'wine', 6 of the 18 meanings of this word form are to do with the alcoholic drink, all of which are derived from the first chronological sense in this area, meaning the drink itself. These word senses include the verbs 'wine', meaning to stock wine (as in 'to wine the King's Cellar') or to drink wine, or nouns meaning the glass from which one drinks wine or non-grape wines (such as parsley wine). The density of the word form 'wine' in the semantic area of drinking means that it is likely that the core or most prominent sense of 'wine' is that of the drink itself. This changes throughout time—in the Old English period, the semantic density of the form 'wine' points to the core sense of being that of a friend or protector, as found in *Beowulf* (for example, line 457 *For gewyrhtum Þu, wine min Beowulf*, 'So it is as to fight in our defense, our friend Beowulf' in the verse translation by Alexander, 2013). We can therefore use this measure of polyseme density to weight more heavily meanings of higher recorded prominence.

A related idea is that of human scale distance. The HT data set is highly precise and moves from the most general concepts to the most precise, with precise concepts being deeper in the HT hierarchy than more general ones. This level of relative precision shifts significantly between categories, and so we have developed for the HT data set a moving cutoff line for each semantic category, which we term a 'human scale' depth. We follow the cognitive linguists Fauconnier and Turner (2002, p. 312) in characterizing human scale as a situation with 'direct perception and action in familiar frames that are easily apprehended by human beings: An object falls, someone lifts an object, two people converse, one person goes somewhere. They typically have very few participants, direct intentionality, and immediate bodily effect and are immediately apprehended as coherent'. In short, the human scale 'is the level at which it is natural for us to have the impression that we have direct, reliable, and comprehensive understanding' (Fauconnier and Turner, 2002, p. 323). For the HTST, then, the closer a particular candidate sense is to a human scale concept, the more likely it may be that this is a sense used in a text. By using this weighting, we can use this measure to compensate for the tendency in other semantic taggers to not know the difference between a highly obscure and arcane sense and one which is more often encountered by human beings—while still having the ability to tag a word with the arcane sense if other factors in the HTST algorithms support it.

There are further lexical resources and knowledge bases which we intend to evaluate the usefulness of in order to develop HTST further. For some highly polysemous words, such as 'run' (302 possible meanings), 'strike' (256), 'fall' (206), 'cast' (187), 'round' (179), 'turn' (174), 'point' (169), 'slip' (165), 'pass' (160), 'shoot' (159), 'take' (158), and so forth, we have determined some default senses to help when the tagger has no other way of determining a word sense—so that, for example, the verb 'take', in the absence of any other disambiguating information, is most likely to be used in HT sense 02.06.13, roughly meaning to move a thing from a place into one's possession. We are also working on using other indicators from the surrounding context and document topics as weighting factors (a document known to be about politics, for example, can have political senses more heavily weighted). We plan to use the USAS semantic information of the target word's collocations in the disambiguation process, to use semantic collocation data. And finally, in the current stage of the project, we will improve the HTST system using information extracted from word sense definitions and example sentences supplied by our partners at Oxford University Press and the *Oxford English Dictionary*. Using these example sentences as a training set (with a large number of high-quality sample sentences and collocations provided for each possible word sense, adding up to a data set of hundreds of millions of words), we aim to improve our accuracy figures drastically.

Many of these new disambiguation techniques are new to this project, and arise from the unique combination of the existing UCREL technologies and the unprecedented scope of the HT data set. There are yet more to be developed, and we believe that in the future we can achieve some remarkable accuracy figures using these novel techniques.

4 Metaphor and Distant Reading

As stated above, one of the intended uses of the HTST system is to facilitate the ability to perform 'distant reading' of texts—the automatic analysis of contents and extraction of data from those texts, as opposed to the 'close reading' traditionally practised, in which a human researcher would be required to read the texts in full him/herself. One of the aims of 'distant reading' is, therefore, to construct and analyze metadata about large-scale collections of information in a way that does not require detailed and time-consuming research on individual texts (Moretti, 2013). As a proof of concept for this type of application, the tagger was employed on two popular science texts. It was hypothesized that automatic semantic tagging of a text could be used to reveal the use of analogical language in that text, on the basis of the semantic domains which are represented throughout it. Extensive use of domains which are not directly relevant to the subject matter of the text would, in theory, be the result of their use as source domains from which to draw analogical imagery, as the author describes an abstract concept which is directly relevant to their subject matter using imagery derived from a more concrete concept. We therefore aimed at an automatic identification of analogical usage in these texts using the tagger to find these areas.

This is in keeping with modern research on metaphor and the key role which metaphorical thinking plays in cognition. While we know that 'figurative meaning is part of the basic fabric of linguistic structure' (Dancygier and Sweetser, 2014, p. 1), it is also the case—for the purposes of our current article—that 'as soon as one gets away from concrete physical experience and starts talking about abstractions or emotions, metaphorical understanding is the norm' (Lakoff, 1993, p. 205). The need for a more empirical approach, as we describe here, in the field of metaphor studies and cognitive linguistics is clearly set out in Gibbs (2006).

4.1 Methodology

The texts chosen were Brian Greene's (2004) *The Fabric of the Cosmos* (FC), which describes the ways in which humankind has conceived of space and time throughout history, and Marcus du Sautoy's (2003) *The Music of the Primes* (MP), which discusses number theory and particularly the Riemann Hypothesis, a mathematical speculation about the distribution of prime numbers. These books were deemed suitable on the basis that quantum physics

and mathematics are concerned primarily with abstract concepts. A non-specialist audience (such as would be expected to constitute a large proportion of the readership of popular science books) is likely to require some explication of these concepts, which increases the probability that the author will employ analogical strategies.

Alongside their abstract subject matter, the texts were evaluated for likely quality through the criteria that they should be written by respected and trustworthy sources, and have received generally positive reader reviews, suggesting that they are reliable and largely accurate representations of the subjects they discuss. Both Brian Greene and Marcus du Sautoy are practising academics at world-leading research institutes—Greene is Professor of Mathematics and Physics at Columbia University, du Sautoy is Professor of Mathematics at the University of Oxford and Fellow of New College—and are thus in a position to write these books in question from a strong professional background.

The two texts were purchased as eBooks and converted to a plain text format which was then annotated using the HTST software, producing a version of the text marked-up with the HT codes appropriate to each word.[9] Using a modified version of the Wmatrix tool developed at Lancaster's UCREL (Rayson, 2008), a list of the most prevalent semantic domains was created, and this list was then contrasted against a test corpus in order to identify domains which were represented more than would be expected in text without a specific focus or subject matter. The domains were ranked for statistical significance on the basis of log-likelihood scores.[10] The test corpus used consisted of a million sentences culled at random from the English version of Wikipedia.[11]

The basis for the choice of this corpus was twofold. The first consideration was that this material should be broadly similar in style to the books under analysis; that is to say that it should be factual, and without any marked stylistic flourishes such as particularly imagery-laden or poetic language which would affect the weighting of its semantic content. The second consideration was that the content of the corpus should also not itself be weighted toward any particular semantic domain; this should be the case owing to the random selection of the sentences. It was hoped, therefore, that when the semantic contents of the popular science books were compared with that of the Wikipedia corpus, semantic domains which stood out as being used more heavily in the books than in the corpus would indicate a significant use of these concepts over and above that which would have been expected in other writing.

Heavily employed semantic domains should be either directly relevant to the subject matter of the books (i.e. the concepts of 'space' and 'time' would be expected to occur frequently in FC, while 'mathematics' and 'number' might be expected in MP) or, where they were not relevant, in theory represent the use of analogical language to express concepts which 'were' directly relevant to the subject of the books (for example, the domain of shape might be predicted within MP, as mathematics often describes the results of plotting variables on a graph as if they have a 'shape' or 'trajectory', such as a 'curve' or 'incline'/'decline'. Some few domains were not automatically assignable to the categories of relevant or analogical, and so a set of intermediate categories were manually assigned to the relevant or analogical categories by the analyst.

Once the domains returned by the tagger software were evaluated and the 'relevant' domains discounted from consideration, the remaining domains were investigated for evidence of metaphorical or analogical material. There was again a need for researcher intervention here, especially for a degree of close reading. Although to some extent it was possible to predict why a metaphorical link might exist through reading the publisher's blurb on the online bookshop pages from which the books were purchased, it was still necessary to have a reliable idea of the content of the text so that strong analogical links could be more reliably identified.

More accurate identification of analogy and metaphor was aided by the division of the text into units of roughly equal length (approximately 500 words per unit). Within these the use of words which most commonly instantiated an analogical semantic domain were counted, and the result visualized as graphs. This allowed rapid visual identification of the portions of a text which made extensive use of particular domains.

4.2 Results

The analysis of both texts did result in identification of analogical material, although with different distribution patterns; whereas FC had the largest cluster of its analogical language present in the final five chapters of the book, MP appeared to have consistent metaphors of distance and direction which ran throughout its length.

Upon analysis using the log-likelihood measure, FC was found to contain eight key semantic domains (listed with their log-likelihood scores in parentheses):

01.12 Space (LL 13655.8)
01.12.01 Distance (LL 6344.8)
01.10.07.05.04.08 Photon (LL 4912.5)
01.13.07 Reckoning of time (LL 3603.5)
01.08.01.15 Textile manufacture (LL 3193.5)
03.13.03.02.08.02 Stringed instruments (LL 2277.7)
03.13.03.03.09.14 Pattern/design (LL 1949.8)
01.08.01.14.01.03 Woven fabric (LL 1922.2)

The first four of these can be considered relevant to the subject material of the book, as they are very specifically and literally concerned with the concepts of time, space, and photons. The second group of four (highlighted in bold) are, however, not directly relevant to the subject matter of the book, and are therefore candidates for analogical language use. Figure 3 displays representative units from the text of FC; the seven most frequent non-grammatical words from non-relevant domains ('fabric', 'feel', 'figure', 'new', 'region', 'sense', and 'string') are displayed, with the two words ('string' and 'fabric') which are linked closely to the four bold domains above shown in blue. From this graph it is clear that the word 'string' in particular is used heavily in the final third of the text, as is 'fabric'. (Not all of the remaining words are analogical; 'sense' and 'feel' are metaphorical as they refer to getting a mental understanding of something, not physically sensing or feeling something, but the remainder of this list of most common words are not particularly metaphorical.) A reading of FC—or, indeed, consultation of its contents list—makes apparent the reason for the prevalence of such analogical concepts, as well as their distribution in the final section of the book.

FC describes and evaluates conceptions of space and time throughout history in a chronological sequence. From chapter 12 to the end of the text (at chapter 16), the author discusses aspects of quantum physics which suggests that the universe as being composed of 'strings'—one-dimensional quantum objects which are hypothesized to form quantum particles through 'vibrating' in different manners. The term 'string' is, in itself, metaphorical, coined by theoretical physicists seeking an appropriate word and image for the hypothetical concept. It is convenient and beneficial for authors discussing string theory to extend this metaphor in their writing, as Greene does, seizing upon an everyday concept with which most readers would therefore be familiar in order to explicate the highly abstract concept of 'strings' in quantum physics. Greene takes this a stage further to conceptualize space and time as a 'fabric' which is woven from these 'strings', which accounts for the presence of the 'woven' category as well as for 'pattern/design' where these apply to the imagined space-time textile (cf. Greene, 2004, pp. 486–7).

MP contains a less diverse range of analogical categories although it does so throughout the length of the text rather than in clusters within it. Two main semantic domains emerge as employed analogically: 01.12.01.01 Distance/farness and 01.12.06 Direction. The use of these domains is instantiated in words such as 'way', 'far', 'line', 'level', and 'point'.

A reading of the text of MP establishes the reason for the use of these categories. Du Sautoy, in handling the very abstract concepts of number patterns, employs an analogy throughout the text by which he conceives of the axes of a graph as delimiting a physical space or landscape in which the numbers are distributed as if they were physical objects or features of the landscape (cf. du Sautoy, 2003, p. 85).

The combined evidence of FC and MP is therefore encouraging proof that a semantic tagging tool such as the HTST can be fruitfully used for this type of research. It successfully identified key semantic domains using a 'distant reading' approach, allowing a researcher to focus efforts on the identification of relevant and analogical domains from this list. Filtering of relevant domains on the basis of

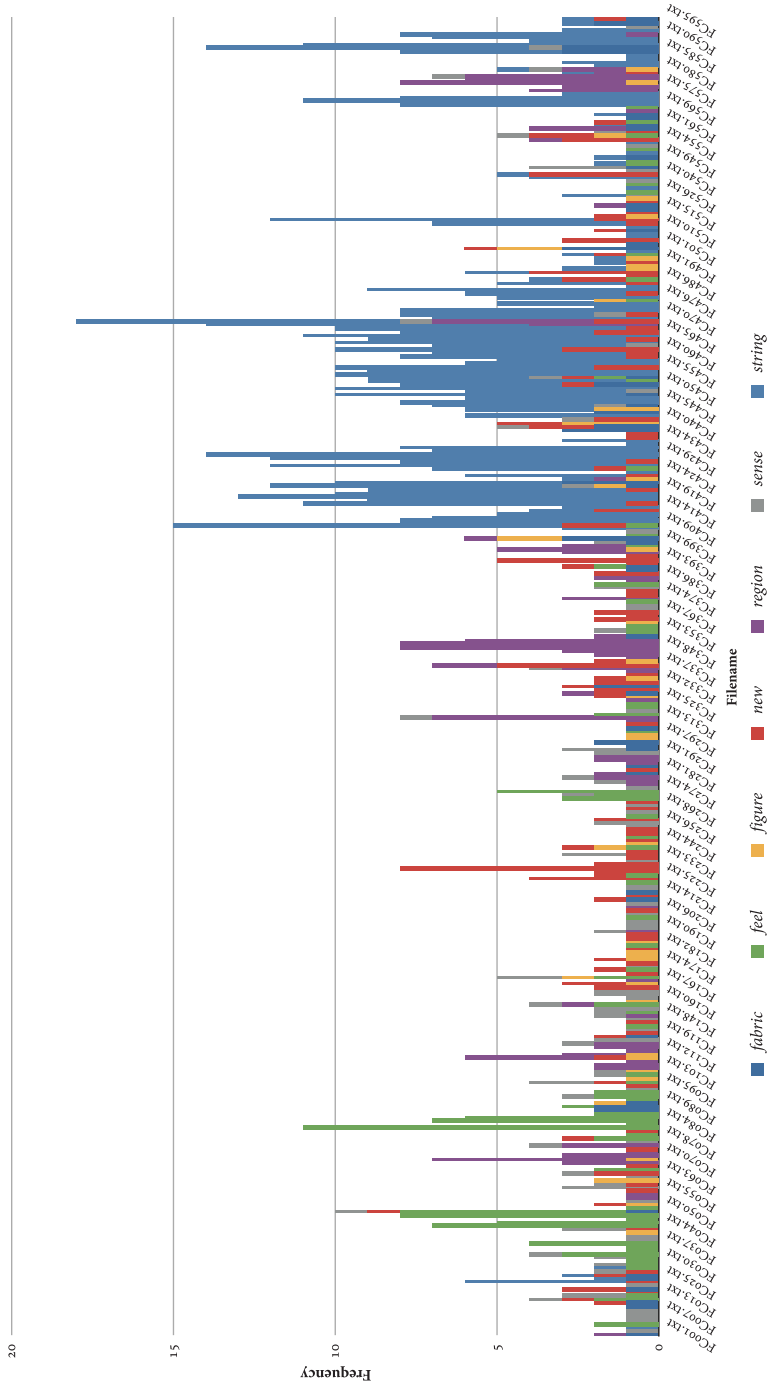

Fig. 3 Distribution of most frequent non-grammatical words from non-relevant domains in FC

contextual information was then possible, which left the researcher's main task as the evaluation of the analogical domains to establish the reasons for their presence. The tagging software, therefore, successfully enabled research on the use of analogy in popular science by correctly extracting the semantic domains which were being employed for analogical purposes.

5 Conclusion

Overall, the application of precise semantic annotation, based on high-quality humanities data and new computational techniques, produces compelling results for the analysis of textual data sets. Chief among these, separate from the major advances in corpus linguistics which a semantic tagger at this level of granularity can provide, is a new way of applying meaning-based 'distant reading' to book-length texts. We have shown it is possible to achieve new results and ways of viewing texts, particularly in the area of metaphorical clusters, and the continued advancement of the HTST system, using the new techniques we outline above, will open up further new ways of looking at English text in ways we are only now beginning to discover.

Funding

This work was supported by the Arts and Humanities Research Council in conjunction with the Economic and Social Research Council [grant AH/L010062/1]. This grant is jointly held by M.A., A.B., P.R., Jean Anderson of the University of Glasgow, Professor Dawn Archer of the University of Central Lancashire, Professor Jonathan Hope of Strathclyde University, Professor Lesley Jeffries of the University of Huddersfield, Professor Christian Kay of the University of Glasgow, and Dr Brian Walker of the University of Huddersfield.

References

Alexander, M. (2013). *Beowulf*. London: Penguin.

Baron, A. and Rayson, P. (2008). *VARD 2: A Tool for Dealing with Spelling Variation in Historical Corpora, Proceedings of the Postgraduate Conference in Corpus Linguistics*. Birmingham, UK: Aston University.

Choi, S., Cha, S., and Tappert, C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1): 43–8.

Dancygier, B. and Sweetser, E. (2014). *Figurative Language*. Cambridge: Cambridge University Press.

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

Fauconnier, G. and Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York, NY: Basic Books.

Garside, R. and Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In Garside, R., Leech, G. and McEnery, A. (eds), *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman, pp. 102–21.

Gibbs, R. W. (2006). Introspection and cognitive linguistics: should we trust our own intuitions? *Annual Review of Cognitive Linguistics*, 4(1): 135–51.

Greene, B. (2004). *The Fabric of the Cosmos: Space, Time and the Texture of Reality*. New York, NY: Alfred A. Knopf.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.

Kay, C., Roberts, J., Samuels, M., and Wotherspoon, I. (eds) (2009). *Historical Thesaurus of the Oxford English Dictionary*. Oxford: Oxford University Press. www.glasgow.ac.uk/thesaurus

Lakoff, G. (1993). The contemporary theory of metaphor. In Ortony, A. (ed.), *Metaphor and Thought*. 2nd edn. Cambridge: Cambridge University Press, pp. 202–51.

Leech, G. and Smith, N. (2000). *Manual to Accompany the British National Corpus (Version 2) with Improved Word-class Tagging*. Lancaster, UK: UCREL, Lancaster University. http://ucrel.lancs.ac.uk/bnc2/bnc2postag_manual.htm

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Piao, S. S. L., Archer, D., Mudraya, O., *et al.* (2005). A large semantic lexicon for corpus annotation. *Corpus Linguistics*, July: 14–17.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4): 519–49.

Rayson, P., Archer, D., Piao, S., and McEnery, T. (2004). 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, pp. 7–12.

du Sautoy, M. (2003). *The Music of the Primes: Why an Unsolved Problem in Mathematics Matters*. London: Harper Perennial.

Tagg, C., Baron, A., and Rayson, P. (2014). ''i didn't spel that wrong did i. Oops'': analysis and normalisation of SMS spelling variation. In Cougan, L. A. and Cédrick, F. (eds), *SMS Communication: A Linguistic Approach*. Amsterdam: John Benjamins, pp. 217–37.

Notes

1 Given HT categories are appropriate for the current version at the time of writing (4.2). A guide to the versions of the HT is available at http://historicalthesaurus.arts.gla.ac.uk/versions-and-changes

2 For further information about UCREL, see http://ucrel.lancs.ac.uk/

3 For further information about VARD, see http://ucrel.lancs.ac.uk/vard/

4 For further information about CLAWS, see http://ucrel.lancs.ac.uk/claws/

5 Sampled from the ENRON email corpus: https://www.cs.cmu.edu/~./enron/

6 This sample was taken from the corpus tagged in the JISC-funded Parliamentary Discourse project http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/parliamentarydiscourse/

7 HTST test Web site is at http://phlox.lancs.ac.uk/ucrel/semtagger/english

8 For further details of Wmatrix, see http://ucrel.lancs.ac.uk/wmatrix/

9 This was carried out under fair dealing defences for research in UK copyright law.

10 http://ucrel.lancs.ac.uk/llwizard.html. The original source of the log-likelihood formula is Dunning (1993).

11 Created by David MacIver and available at http://goo.gl/gBDI3. The English Wikipedia is at http://en.wikipedia.org/wiki.