The 25th Joint International Conference of the Association for Literary and Linguistic Computing and Association for Computers and the Humanities
and
The 6th Joint International Conference of the Alliance of Digital Humanities Organizations

digital humanities
Lausanne – Switzerland '14

Unil
UNIL | Université de Lausanne

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

International Program Committee

Melissa Terras, chair
Deb Verhoeven, vice-chair

John Bradley
Jieh Hsiang
Jane Hunter
Aimée Morrison
Bethany Nowviskie
Dan O’Donnell
Sarah Potvin
James Smithies
Takafumi Suzuki
Tomoji Tabata
Toru Tomabechi
Glen Worthey
Vika Zafrin

Local Organizing Committee

Claire Clivaz, co-local organizer
Frédéric Kaplan, co-local organizer

Karl Aberer
Jeannette Frey
Benoît Garbinato
Philippe Kaenel
Isabelle Kratz
Enrico Natale
Lukas Rosenthaler
Süsstrunk Sabine
Michael Stolz
François Vallotton
Boris Vejdovsky
Dominique Vinck

Silver Level Sponsors

Yandex Europe AG

Bronze Level Sponsors

CLARIN
Common Language Resources and Technology Infrastructure

Partners

infoclio.ch
Swiss National Science Foundation

Organizers

Alliance of Digital Humanities Organizations (ADHO)
École Polytechnique Fédérale de Lausanne (EPFL)
Université de Lausanne (UNIL)

Conference Volunteers

Cyril Bornet
Diane Brousse
Vincent Buntinx
Giovanni Colavizza
Olivier Dalang
Isabella di Lenardo
Heidi Dowding
Sebastien Dupont
Anthony Durity
Mikal Eckstrom
Maud Ehrmann
Slimane Fouad
Alicia Foucart
Hannah Jacobs
Penny Johnston
Andrea Mazzei
Elisa Nury
Paul O'Shea
Anna Pytlowany
Yannick Rochat
Jörg Röder
Dario Rodighiero
Elifsu Sabuncu
Jillian Saucier
Qiaoyu Shi
Sree Ganesh Thottempudi

history of English, with a loss of 246 words between 1650 and 1900.

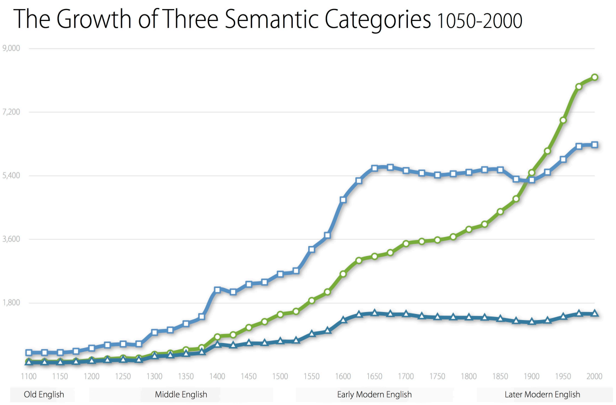

Fig. 2: The growth of three semantic fields. Square: 02.01.15 Attention and Judgement; Circle: 03.05.05 Moral Evil; Triangle: 03.10.13 Trade and Commerce.

Each of these reflect both global trends in the history of English (such as those above, in addition to relative plateaus in the 1700s) while also containing their own internal factors, such as shifts in religious emphasis and in broader economic and industrial patterns.

Not all of these factors are expected; there is no mention in the literature of the rise and fall of lexicalization in the semantic field of *Moral Evil*, nor in many of the other unusual patterns in the rate-of-change data described in this paper. The new data described here gives rise, in the tradition of digital humanities, to the necessity for further explanations from a range of humanities disciplines, such as linguistics, history and literary studies (see Alexander and Struan 2013[11] for an interdisciplinary study in a further semantic field).

4. Metaphoricity

Beyond these rates of change, each semantic field above has metaphorical links to other areas of the language, which the HT can reveal to us. Far from being a solely literary technique, much of all language is figurative – recent research has shown somewhere between 8% and 18% of English discourse is metaphorical, with an average of every seventh word being a metaphor.[12]

This is problematic, as while advances are being made in the semantics of digital texts, alongside emerging concepts of a semantically-aware Web, we are at a very early stage in comprehensively and systematically understanding English metaphor, and therefore at an early stage of being able to accurately deal digitally with the meanings encoded in those texts. By mapping the HT's semantic categories onto one another in order to analyse the degree of lexical overlap in different conceptual fields, we can provide results which will comprehensively demonstrate the widespread, systematic and far-reaching impact of metaphor on English. This is the aim of the *Mapping Metaphor* project at Glasgow,[13] which provides some of our data in this paper, demonstrating empirically the systematic lexical connections between our case study fields (such as that between attention and vision, or evil and darkness).

5. Conclusion

Overall, as well as giving an overview of the history of the English semantic space and its metaphorical interrelationships, the paper also argues for a semantically-informed history of English which operates from a top-down approach, picking out broad patterns and the connections between various semantic categories in order to highlight for analysis those noteworthy elements in a large sea of data. As ever, such large-scale analyses are only possible through a combination of database techniques, statistical analysis, visual displays of complex datasets, and humanities scholarship.

References

1. **Kay, C., J. Roberts, M. Samuels, and I. Wotherspoon** (eds). (2009). *Historical Thesaurus of the Oxford English Dictionary*. Oxford: Oxford University Press.
2. **Alexander, M.** (2012). *Patchworks and Field-Boundaries: Visualising the history of English.* Conference paper at Digital Humanities 2012. Hamburg: University of Hamburg.
3. **Anderson, W., M. Alexander, E. Bramwell, C. Kay, and C. Hough.** (2013)–. *Mapping Metaphor with the Historical Thesaurus*. Glasgow: University of Glasgow. www.glasgow.ac.uk/metaphor
4. **Simpson, J. and E. Weiner** (eds). 1989. *The Oxford English Dictionary*, 2nd edition. Oxford: Oxford University Press.
5. **Brewer, C.** (2007). *Treasure-House of the Language: The Living OED*. New Haven, CT: Yale University Press. Page 232.
6. **Samuels, M.L.** (1972). *Linguistic Evolution: With Special Reference to English*. Cambridge: Cambridge University Press. Page 180.
7. **Lyons, J.** (1995). *Linguistic Semantics*. Cambridge: Cambridge University Press.
8. **Verhagen, A.** (2007). *Construal and Perspectivization.* In The Oxford Handbook of Cognitive Lingustics, eds D. Geeraerts and H. Cuyckens. Oxford: Oxford University Press. 48-81.
9. **Hughes, G.** (1989). *Words in Time: A social history of the English vocabulary*. Oxford: Basil Blackwell.
10. **Taylor, J.R.** (2003). *Linguistic Categorisation*, 3rd edition. Oxford: Oxford University Press.
11. **Alexander, M and A. Struan**. (2013). *'In countries so unciviliz'd as those?': Notions of Civility in the British Experience of the World*. In Experiencing Imperialism, eds M. Farr and X. Guégan. London: Palgrave Macmillan.
12. www.glasgow.ac.uk/metaphor

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

Alexander, Marc
marc.alexander@glasgow.ac.uk
University of Glasgow

Anderson, Jean
jean.anderson@glasgow.ac.uk
University of Glasgow

Baron, Alistair
a.baron@lancaster.ac.uk
Lancaster University

Dallachy, Fraser
fraser.dallachy@glasgow.ac.uk
University of Glasgow

Kay, Christian
christian.kay@glasgow.ac.uk
University of Glasgow

Piao, Scott
s.piao@lancaster.ac.uk
Lancaster University

Rayson, Paul
p.rayson@lancaster.ac.uk
Lancaster University

1. Introduction

This paper describes and implements a computational procedure for semantically analysing analogy in large bodies of text using a semantic annotation system based on the database of the *Historical Thesaurus of English*.[1] In so doing, it demonstrates the value of a comprehensive and fine-grained semantic annotation system for English within corpus linguistics. Using log-likelihood measures on its semantically-annotated corpus of abstract popular science, the paper therefore demonstrates the existence, the extent, and the location of significant metaphorical content in this corpus. In so doing, it applies a version of Franco Moretti's 'distant reading' programme in the analysis of literary history to non-narrative texts, as well as continuing work on integrating meaning into the methodologies of corpus linguistics.[2]

1.1. Analogy and Popular Science

Following the 1980 publication of George Lakoff and Mark Johnson's *Metaphors We Live By*,[3] it has been frequently stated that human beings, as embodied minds perceiving the mental, social and physical worlds around them, understand abstractions in terms of concrete entities. While this is a well-explicated concept in cognitive linguistics and psychology, few studies have yet aimed to establish both the extent and operation of this in a large corpus of discourse. The standard methodology in cognitive linguistics tends to rely on introspection and the intuitions of native speakers, at the expense of empirical data.[4] This lack of rigour has resulted in results which, though "intuitively appealing", are criticized "for lacking a clear set of methodological decision principles".[5] Following earlier work we have undertaken on the investigation of analogy and metaphor in English from empirical groundings,[6] [7] in this paper we discuss a methodology for identifying these textual phenomena automatically, and in so doing aim to open up cognitive linguistics to more digital humanities techniques, in addition to demonstrating the use of automated semantic annotation and disambiguation techniques at an unprecedented level of granularity.

1.2. The Corpus

We take as our initial data two book-length popular science texts which focus on explaining abstract concepts to a non-specialist audience, and therefore provide the greatest potential for the analysis of non-literary analogy - metaphor theory tells us that these should therefore be rich in non-abstract analogies. The corpus is therefore made up of Brian Greene's 2004 *The Fabric of the Cosmos* and Marcus du Sautoy's 2003 *The Music of the Primes*, although we have subsequently tested the methodology on other popular science texts.

Through the procedure we describe in 3.1 below to analyse metaphor and analogy in these texts, we identify a range of domains which are unusually frequent in these texts and which are not pertinent to their subject matter (that is, not in the areas of physics, mathematics or general science). We then demonstrate in the remainder of section 3 that these domains are those analogies used systematically and consistently across the texts to elucidate and explicate the abstract concepts the books are focused on discussing. In order to do this, we identify all the semantic domains mentioned in these texts at very high levels of precision, using an annotation system built around the unprecedented detail found in the database of the *Historical Thesaurus*.

2. Semantic Annotation

Semantic tagging and annotation is, we argue, the best solution we have to address the problem of searching and aggregating large collections of textual data: at present, historians, literary scholars and other researchers must search texts and summarize their contents based on word forms. These forms are highly problematic, given that most of them in English refer to multiple senses – for example, the word form "strike" has 181 *Historical Thesaurus* meaning entries in English, effectively inhibiting any large-scale automated research into the language of industrial action; "show" has 99 meanings, prohibiting effective searches on, say, theatrical metaphors or those of emotional displays. In such cases, much time and effort is expended in manually disambiguating and filtering search results and word statistics.

To resolve this problem, we use in this paper an early version of the Glasgow-Lancaster Semantic Annotation System, which we are currently developing at both of those universities. GL-SAS is a tool for annotating large corpora with meaning codes from the *Historical Thesaurus*, enabling us to search and aggregate data using the 236,000 precise meaning codes in that dataset, rather than imprecise word forms. These *Thesaurus* category codes are over one thousand times more precise than USAS, the current leader in semantic annotation in English corpus linguistics.[8] The system automatically disambiguates these word meanings using existing computational disambiguation techniques alongside new context-dependent methods enabled by the *Historical Thesaurus'* dating codes and its fine-grained hierarchical structure. With our data showing that 60% of word forms in English refer to more than one meaning, and with some word forms referring to close to two hundred meanings, effective disambiguation is essential to GL-SAS.

3. Results

3.1. Methodology

The 600,000 word corpus we outline above were lemmatised and then processed through our annotation system, resulting in texts with each word being annotated with a Historical Thesaurus meaning code. We then aggregated those codes into a dataset which summarised the frequency of each meaning code in the text, and took that frequency list and compared it to a reference corpus made up of a 14m word corpus of random selections from Wikipedia, to provide a comparison against standard expository text. Our comparison was based on a log-likelihood significance measure,[9] which identifies, to an acceptable degree, those semantic domains which are mentioned unusually frequently in our popular science texts by comparison to the reference corpus, and therefore indicates a text's "key" domains (where the log-likelihood values are greater than around 20)[10] - those domains which reflect what a text is "about".[11]

3.2. *The Fabric of the Cosmos*

Brian Greene's 2004 *The Fabric of the Cosmos* discusses theoretical physics and its relation to the concepts of space and time. Its key semantic domains are given in Table 1:

HT Category	Category Name	Log-Likelihood Value
01.05.07	Space	13655.8
01.05.07.01	Distance	6344.8
01.04.07.05.04.08	Photon	4912.5
01.05.06.07	Computation of time	3603.5
01.02.09.15	**Spinning textiles**	**3193.5**
03.11.03.01.08.02	**Stringed instruments**	**2277.7**
03.11.03.02.09.14	**Pattern/design**	**1949.8**
01.02.09.14.01.03	**Woven fabric**	**1922.2**

While the first four domains are within the *Thesaurus* categories which refer to the text's topic, and therefore expected, the next four (in bold) are not immediately relevant to the book's topic. Looking for these domains in the text itself, chunked into 591 smaller files of 320 words each, we get a distribution like this:

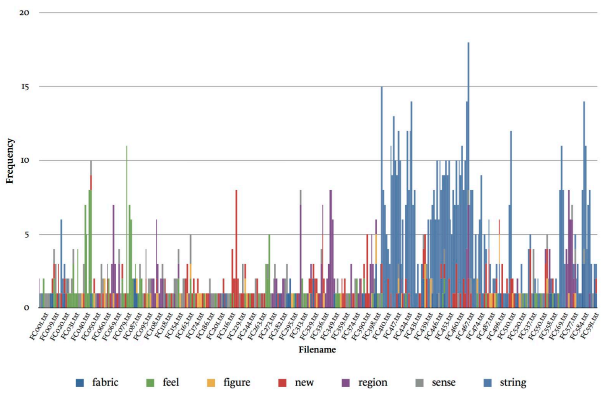

Fig. 1: Analogical textual clusters in The Fabric of the Cosmos, shown by frequency of key semantic domains

(Here, the Thesaurus codes have been replaced by words representing those categories, for ease of reading.)

The peak three-quarters of the way through the text indicates an area rich in mentions of textiles, and looking at this point in the text we find passages such as:

Since we speak of the ‘fabric’ of spacetime, the suggestion goes, maybe spacetime is stitched out of strings much as a shirt is stitched out of thread. That is, much as joining numerous threads together in an appropriate pattern produces a shirt’s fabric, maybe joining numerous strings together in an appropriate pattern produces what we commonly call spacetime’s fabric. Matter, like you and me, would then amount to additional agglomerations of vibrating strings.[12]

The areas we have identified through the log-likelihood analysis are therefore those areas rich in metaphors of fabric and strings (as other examples show) which are used by the author to discuss physics. We can therefore use this technique to pinpoint areas of significant use of metaphor or analogy in a text.

3.3. *The Music of the Primes*

As a check of the methodology, the same technique shows that in this particular book, which discusses prime number theory, there are highly key domains of *travel* and *landscape* in use alongside mathematical terms. Going to sections particularly rich in these domains gives analogical content over a long stretch, introduced by the following extract:

Gauss’s two-dimensional map of imaginary numbers charts the numbers that we shall feed into the zeta function. The north-south axis keeps track of how many steps we take in the imaginary direction, whilst the east west axis charts the real numbers. We can lay this map out flat on a table. What we want to do is to create a physical landscape situated in the space above this map. The shadow of the zeta function will then turn into a physical object whose peaks and valleys we can explore.[13]

4. Conclusion

We therefore demonstrate in this paper the use of a very fine-grained semantic annotation system, and establish the utility of such detailed annotations by describing a digital technique for discovering not only the existence of systematic metaphorical content but also its location and where it clusters. We believe that this result is significant in its own right, particularly for scholars of metaphor or cognitive linguistics, but we will also show that this represents only one of the uses to which highly-granular semantically annotated data can be put.

References

1. **Kay, C., J. Roberts, M. Samuels, and I. Wotherspoon** (eds). (2009). *Historical Thesaurus of the Oxford English Dictionary*. Oxford: Oxford University Press. See also historicalthesaurus.arts.gla.ac.uk .

2. **Rayson, Paul**. (2008). *From Key Words to Key Semantic Domains*. International Journal of Corpus Linguistics 13.4. 519-549.

3. **Lakoff, George & Mark Johnson**. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

4. **Gibbs, Raymond W.** (2006a). *Introspection and Cognitive Linguistics: Should We Trust Our Own Intuitions?* Annual Review of Cognitive Linguistics 4(1). 135-151.

5. **Evans, Vyvyan & Melanie Green.** (2006). *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press. Page 780.

6. **Alexander, Marc & Christian Kay**. (2011) [2010]. *Mapping Metaphors Across Time with the Historical Thesaurus*. Conference paper at Helsinki Corpus Festival: The Past, Present, and Future of English Historical Corpora, University of Helsinki, Finland. Based on an earlier paper at The 3rd UK Cognitive Linguistics Conference, University of Hertfordshire.

7. **Alexander, Marc**. (2011). *Meaning Construction in Popular Science An Investigation into Cognitive, Digital, and Empirical Approaches to Discourse Reification*. University of Glasgow: Ph.D. thesis.

8. ucrel.lancs.ac.uk/usas

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

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

11. **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. London: Routledge. 516-530.

12. **Greene, Brian R**. (2004). *The Fabric of the Cosmos: Space, Time and the Texture of Reality*. Alfred A Knopf: New York. Page 486-7.

13. **du Sautoy, Marcus**. (2003). *The Music of the Primes: Why an Unsolved Problem in Mathematics Matters*. London: Harper Perennial. Page 85.

The Cryptic Novel: A Computational Taxonomy of the Eighteenth-Century Literary Field

Algee-Hewitt, Mark
mark.algee-hewitt@stanford.edu
Stanford University

Eidem, Laura
lmeidem@stanford.edu
Stanford University

Heuser, Ryan
heuser@stanford.edu
Stanford University

Law, Anita
anital@stanford.edu
Stanford University

Llewellyn, Tanya
tanya.llewellyn@gmail.com
Stanford University

Overview