Alexander, M., and Anderson, W. (2014) "Civilization arranged in chronological strata": a digital approach to the English semantic space. In: Digital Humanities 2014, 7-12 Jul 2014, Lausanne, Switzerland.

http://eprints.gla.ac.uk/98200

Deposited on: 21 October 2014

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

h⁚umanities

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

"Civilization arranged in chronological strata": A digital approach to the English semantic space

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

Anderson, Wendy
wendy.anderson@glasgow.ac.uk
University of Glasgow

1. Introduction

This paper focuses on the history of the English lexicon, and on displaying a new approach to this history through the database of the *Historical Thesaurus of English* [1] (hereafter abbreviated to HT). It does so by reference to the semantic space of English, following Lehrer's statement that 'the words of a language can be classified into sets which […] divide up the semantic space or the semantic domain in certain ways'. This space is described in this paper as the total accumulation of the various individual semantic fields which make up the language, as represented in the HT database. The paper therefore computationally analyses the size of the English lexicon in these semantic clusters over time, including the metaphorical links which weave between these fields, and so aims to demonstrate the use of the HT in digital humanities by giving a digital analysis of the history of English in ways which were previously not possible.

As part of two wider projects,[2][3] the present paper focuses on describing the general empirical outlines and development of the English semantic space, accompanied by a case study of three contrasting semantic fields and their metaphorical relationships. These are outlined below, following a description of the methodology and theoretical basis of the paper.

2. The *Historical Thesaurus* and Lexical History

The data used in this paper is drawn from the database of the HT, which arranges into hierarchical semantic categories all the recorded words expressed in English from Anglo-Saxon times to the present day, with 793,742 entries within 225,131 categories, each category representing a distinct concept. These concepts are arranged hierarchically and semantically, so that each concept is placed near or within other, similar concepts.

In so doing, the HT unlocks the linguistic and historical data which is currently inaccessible in any usefully-structured way inside historical dictionaries such as the *Oxford English Dictionary* (OED)[4]. As Charlotte Brewer says, with reference to a review in the Times of the OED:

"...even the intensively habitual user [of the OED] could not hope to construct, from an overwhelming multiplicity of individual items, the complete picture, 'the various forms of [...] civilization arranged in chronological strata'..."[5]

Alphabetical arrangement, absent any alternative structure, makes this construction incredibly difficult, if not impossible. But the HT, which structures itself based on meaning and not the alphabet, does give researchers access to this 'complete picture'. This was one of the intentions of the HT from the beginning: Professor Michael Samuels, founder of the project in 1964, saw it as a way of revealing the information about social and cultural change inside and throughout the lexicon which was not easily available for researchers to access.[6] The HT is therefore a massive digital resource for the study of this phenomenon.

3. Semantic Space

Key for the first part of this paper is that the HT, when analysed in database form, gives an indication of rates of lexicalisation in the history of English. This relates to the phenomenon of synonymy, a situation in a language where a number of words are created (or lexicalized) for a single concept (for more on the following discussion, see, amongst others, Lyons 1995[7], Verhagen 2007[8], Hughes 1989[9], and Taylor 2003[10]). While synonymy is a common occurrence in English, as in many other languages, the linguistic insight that synonymy is a form of recategorization, where speakers create a new synonymous term because they wish to reflect a shift in their understanding of, or attitudes towards, a particular concept, allows the use of data on lexicalization rates as an indicator of particular speaker attention to a given concept. Therefore, a situation where there are multiple words for a given concept reflects the evolution of speakers' reactions, attitudes, perceptions and awarenesses of that concept, as human language is too efficient a system to permit there to exist large sets of undifferentiated terms which mean precisely the same thing. The present paper therefore uses this measure as a rough proxy for importance of a concept (just as frequency is used as a similar measure of importance in corpus linguistics, with all the associated issues that varying corpus construction techniques brings with it).

Therefore, for the first time, the database of the HT can give us an empirically based view of English by viewing the changes in the internal structure of the language from an entirely semantic viewpoint. By separating the story of English into the multiple stories of interacting and interrelated semantic fields, this approach can describe the history of English as one generally characterized by overall growth accompanied by occasional trauma which results in sudden expansions or contractions of the English lexicon. The rate of change of each semantic field is therefore a statistic which demonstrates the incidence of such instances of trauma, growing or declining in response to external and internal factors either particular to a semantic field or general to the language as a whole. Such general factors in English include the well-known sudden growth in the mid-1400s which occurs at the start of the English Renaissance, and the Elizabethan and Jacobean spurt which begins a little after 1550, which can be seen in Figure 1:
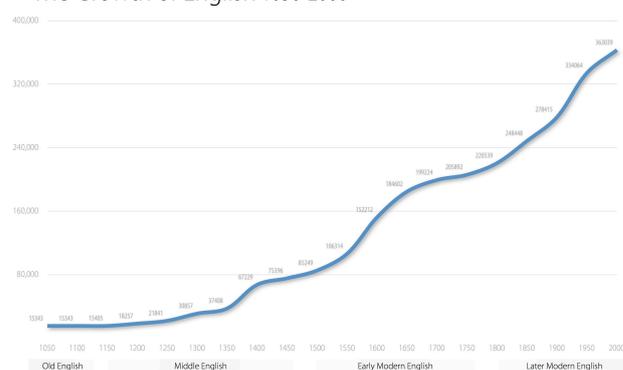

The Growth of English 1050-2000

Fig. 1: The growth of the English language across time, as recorded in the HT.

In addition to presenting an overview of the growth of the semantic space of English between the years 1100 and 2000, the paper will also give short case studies of three aggregate semantic fields (figure 2) and their metaphorical relationships (see section 4 below):

– *02.01.15 Attention and Judgement*, a very large and highly variable category, with an increase of 1000 words in the 1575-1600 period, but a fall of 261 in 1875-1900.
– *03.10.13 Trade and Commerce*, a category which is relatively small but has one of the highest rates of growth spurts, punctuated with long plateaus.
– *03.05.05 Moral Evil*, a category which peaks in 1650 and is one of the rare examples of frequent decline across the

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

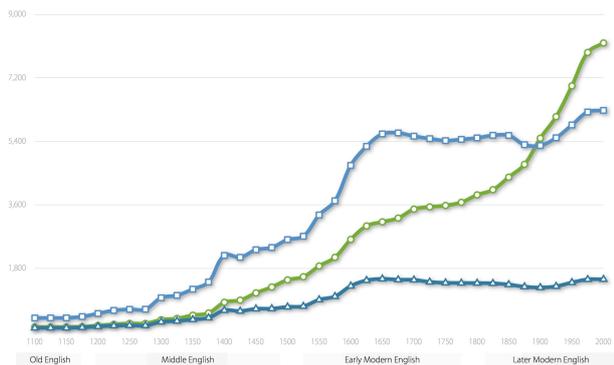## The Growth of Three Semantic Categories 1050-2000

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