The Seven Words of the Virgin: Identifying change in the discourse context of the concept of virginity in Early Modern English

Susan Fitzmaurice
s.fitzmaurice@sheffield.ac.uk
University of Sheffield

Justyna Robinson
justyna.robinson@sussex.ac.uk
University of Sussex

Iona Hine
i.hine@sheffield.ac.uk
University of Sheffield

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

Kathryn Rogers
k.m.rogers@sheffield.ac.uk
University of Sheffield

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

Michael Pidd
m.pidd@sheffield.ac.uk
University of Sheffield

Seth Mehl
seth.mehl.10@ucl.ac.uk
University of Sheffield

Matthew Groves
m.i.groves@sheffield.ac.uk
University of Sheffield

Brian Aitken
brianaAitken@glasgow.ac.uk
University of Glasgow

Introduction

The Linguistic DNA project (LDNA) is an AHRC-funded collaborative project (AHRC grant AH/M00614X/1) between the universities of Sheffield, Glasgow, and Sussex which is designing automatic processes to investigate the emergence and development of concepts in pre-1800 CE print. Employing Early English Books Online, manually-transcribed through the Text Creation Partnership (EEBO-TCP) as its primary dataset, supplemented by Eighteenth Century Collections Online (ECCO-TCP) and other high-quality 18th-century text collections, the project is developing and refining a processing pipeline which assembles groupings of words bound together by their contextual use in printed discourse. The project is charting development of these discourse-embedded word groups across time, investigating how they are shaped by historical and literary contexts, the boundaries and overlap between the groupings, and the interaction of 'encyclopedic' groupings with more traditional 'thesaurus'-style semantic fields.

This paper discusses results from a branch of the project which is investigating incidences of rapid change in the size of semantic categories as represented in *The Historical Thesaurus of English* (Kay et al., 2016). Development of concepts through size of *Thesaurus* categories has been investigated previously (cf. Alexander and Struan, 2013; Jürgen-Diller, 2014), although the extra dimension provided by the outputs of the LDNA processor allows a dramatic leap forward in such research by enabling identification of instances in which change in discourse-embedded word groupings acts as catalyst for corresponding rapid change in the semantic fields of English.

The present case study investigates words relating to the concept of 'virginity', utilising processed time-slice subsets of EEBO-TCP as snapshots of the discourse context for these words in Early Modern English print. By building sample word-groupings (the term 'cluster' is here avoided to avoid confusion with cluster or network analysis word clusters) for each of the subsets, it establishes the discourse context of 'virginity' words at different points in the timespan covered by EEBO-TCP. Comparison of these groupings suggests change in focus of language users, through which a largely religious context of use opens out to a secular and then a poetic literary context, suggesting that society's consciousness of this concept and the scale on which it was discussed enlarged dramatically in the period covered by the sub-corpora.

Methodology

In order to select a *Historical Thesaurus* category for analysis, an average pattern of change over time was established. The *Thesaurus* arranges the words of the English language into a semantic hierarchy that is seven category levels deep with the potential

for up to four further sub-category levels within any given category. Owing to the incredibly fine-grained nature of the sense categorisation in the *Thesaurus*, it was necessary to 'cut' the hierarchy at human scale using a thematic category set, developed during the AHRC- and ESRC-funded SAMUELS project (Grant AH/L010062/1), which is intended to allow *Thesaurus* users to find information at a level that is salient to human beings – i.e. neither too general nor too detailed.

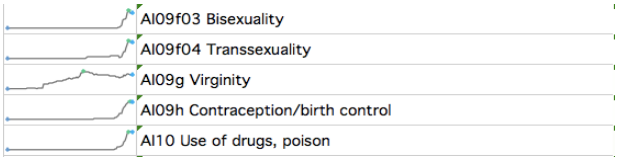

Figure 1: Sparkline showing growth of 'Virginity' category from 1000 CE to 2000 CE in context of surrounding thematic categories (which are themselves unusual as they are lexicalised only in later periods of English)

The number of lexemes within each category level was counted, and lexemes were filtered to include only those active within the approximate time range of the EEBO-TCP collection, i.e. 1475-1700 CE. This data was aggregated so that the change in the mean contents of a category could be viewed across time, and decade-to-decade percentage changes calculated. Individual categories were then compared to this average category change, and a deviation of more than 5% from the average change considered to be significant. Out of the categories which were marked as statistically unusual from this process, category 'AI09g Virginity' was selected as promising because the items in its lexis had a relatively low number of homographs that could skew the results towards irrelevant information.

Testing of the LDNA processor outputs is being conducted on select subsets extracted to provide snapshots across the EEBO time-period. The subsets used for this paper cover the periods 1520-39, 1550-59, 1610-11, and 1649. They are designed to contain a similar number of tokens; the progressively contracting timespans reflect the concomitant growth of printed material throughout the 15th to 17th centuries. Each token in the text is regularised, lemmatised, and tagged with a NUPOS part of speech tag via the MorphAdorner pipeline developed by Martin Müller and Philip Burns (Burns, 2013). Data is then gathered by the LDNA processor for the token's co-occurrences within 100- and 200-word bi-directional windows which are intended to simulate paragraph-like sections of the proximate discourse (cf. Fitzmaurice *et al.* forthcoming). Pointwise Mutual Information (PMI) is used to provide a statistic for likelihood of word co-occurrences; a minimum PMI value of 0.5 was arrived at experimentally for identifying node-collocate pairs to be considered interesting in initial stages of investigation.

Seven items – 'maid,' 'maiden,' 'maidenhead,' 'undefiled,' 'vestal,' 'virgin,' and 'virginity' – in the 'Virginity' category were found to be present consistently across the subsets (although an eighth – 'virginal' – was present in the 1520-39 and 1610-11 subsets). The co-occurrences were then processed to identify those which occurred with multiple items in this list. Words which co-occurred with four or more items were investigated further.

Results

Comparison of the co-occurrence results across the five text subsets shows a consistent shift in the patterns of word association with 'Virginity' category items. The words 'woman' and 'widow' remain strongly associated with the terms across all the subsets, demonstrating societal preoccupation with female rather than male virginity. The most evident change in the grouping is movement from a predominately religious discourse context into the secular world. In the 1520-39 subset, the Virgin Mary is intimately related to discussion of virginity. In the shared collocates listing, *mother* collocates with all seven of the 'Virginity' lexemes, 'mary' with six, 'angel,' 'bless,' 'hymn,' 'nativity,' and 'nazareth' with five each. Of these, only 'mother' maintains a strong association with 'Virginity' words throughout the EEBO period, appearing with four items in the 1610-11 text set and five in 1649.

The secularisation of the term is suggested by the prevalence in later subsets of words relating to marriage, reflecting what appears to be a growing focus on wedlock being preceded by virginity. 'Marry' gradually increases its association with the node items, collocating with four, then five, then seven from 1520 to 1649. 'Marriage' and 'wife' both enter the shared collocate group in 1550, and remain there through to 1649, whilst 'matrimony' is present in 1550, drops out in 1610, and returns in 1649.

The extensive list of shared collocates in the 1649 sub-corpus strongly reflects the greater prevalence of literary fiction and poetry in printers' output and reinforces that virginity is a topic for which the discourse context is expanding; where it was easy to intuitively group 'marry,' 'marriage,' and 'wife' together, the 1649 collocates do not form easily identifiable groupings.

Discussion

The consistency of the core items found in the subsets is interesting in its disparity with the *Thesaurus* data, where the increase in the number of terms present in the 'Virginity' category suggests

that there should be an expanding number of items found throughout these subsets. The most likely explanation for this is loss of low frequency information through a combination of cut-off values intended to reduce noise for later clustering experiments, and difficulty in normalising/lemmatising low frequency items. A clear outcome of the analysis is the confirmation that the category of 'Virginity' contains core vocabulary which remains almost unchanged in over a century (i.e. 1520-1649), primarily a consistent group of seven items which co-occur with 'Virginity' category words.

This study demonstrates that understanding of semantic development can be enriched by such cross-analysis of discursive-concept word groups with *Thesaurus* semantic fields and the word groups which travel through time with multiple items of *Thesaurus* categories.

Bibliography

Alexander, M. and Struan, A. (2013). "'In Countries so Unciviliz'd as Those?': The language of incivility and the British experience of the world." In Farr, M. and Guégan, X. (eds), *Experiencing Imperialism: The British abroad since the Eighteenth Century, vol. 2*. London: Palgrave Macmillan. pp. 232-249.

Burns, P. R. (2013). *MorphAdorner v2: A Java library for the morphological adornment of English language texts*. Evanston, IL. Northwestern University. http://morphadorner.northwestern.edu/morphadorner/download/morphadorner.pdf. (last accessed on 31st March, 2017).

Diller, H-J. (2014). *Words for Feelings*. Anglistische Forschungen vol. 446. Heidelberg: Universitätsverlag Winter

Fitzmaurice, S., Robinson, J., Alexander, M., Hine, I., Mehl, S. and Dallachy, F. (forthcoming). "Linguistic DNA: Investigating conceptual change in early Modern English discourse." *Studia Neophilologica*.

Kay, C., Roberts, J., Samuels, M., Wotherspoon, I. and Alexander, M. (eds) (2016). *The Historical Thesaurus of English*, version 4.2. Glasgow: University of Glasgow. http://historicalthesaurus.arts.gla.ac.uk/.