Spheres of “public” in eighteenth-century Britain

Mark J. Hill (presenter), Antti Kanner, Jani Marjanen, Ville Vaara, Leo Lahti, Eetu Mäkelä, Mikko Tolonen

Revised Abstract for Nordic Digital Humanities Conference in Helsinki, March 2018

The eighteenth-century saw a transformation in the practices of public discourse. With the emergence of clubs, associations, and, in particular, coffee houses, civic exchange intensified from the late seventeenth century. At the same time print media was transformed: book printing proliferated; new genres emerged (especially novels and small histories); works printed in smaller formats made reading more convenient (including in public); and periodicals - generally printed onto single folio half-sheets - emerged as a separate category of printed work which was written specifically for public consumption, and with the intention of influencing public discourse (such periodicals were intended to be both ephemeral and shared, often read, and then discussed, publically each day). This paper studies how these changes may be recognized in language by quantitatively studying the word “public” and its semantic context in the Eighteenth-Century Collections Online (ECCO).

While there are many descriptions of the transformation of public discourse (both contemporary and historical), there has been limited research into the language revolving (and evolving) around “public” in the eighteenth-century. Jürgen Habermas (2003: 2-3) famously argues that the emergence of words such as “*Öffentlichkeit*” in German and “publicity” in English are indicative of a change in the public sphere more generally. The conceptual history of “*Öffentlichkeit*” has been further studied in depth by Lucian Hölscher (1978), but a systematic study of the semantic context of “public” in British eighteenth-century material is missing. Studies that have covered this topic, such as Gunn (1989), base their findings on a very limited set of source material. In contrast, this study, by using a large-scale digitized corpus, aims to supplement earlier studies that focus on individual speech acts or particular collections of sources, and provide a more comprehensive account of how the language of “public” changed in the eighteenth century.

The historical subject matter means that the study is based on the ECCO corpus. While ECCO is in many ways an invaluable resource, a key goal of this study is to be methodologically sound from the perspective of corpus-linguistics and intellectual history, while developing insights which are relevant more generally to sociologists and historians. In this regard, ECCO does come with its own particular problems: both in terms of content and size.

With regard to content: OCR mistakes remain problematic; its heterogeneity in genres can skew investigations; and the unpredictable nature of duplicate texts introduced by numerous reprints of certain volumes must be taken into account. However, many of these problems can be mitigated in different ways. For example, in specific cases we compare findings with the, much smaller, ECCO TCP (an OCR corrected subset of ECCO). We have further used the English Short Title Catalogue (ESTC) to connect textual findings with relevant metadata information contained in the catalogue. By merging ESTC metadata with ECCO, one can more easily use existing historical knowledge (for example, issues around reprints and multiple editions) to engage with the corpus.

With regard to size: the corpus itself is too big to run with automatic parsers (in terms of computing time and resources). We have therefore extracted a separate, and smaller, corpus (with the help of ESTC metadata) to do more complex and demanding analyses. The subcorpus is roughly 0.2% the size of the original corpus (this number remains relatively stable for each decade of the century), yet is still made of 19,945,971 tokens. Additionally, results of analyses on the sub-corpora were replicated (in a much simpler and cruder form) on the whole dataset, offering results which corroborate initial observations.

The size constraints provide their own advantages, however. The smaller subsections were chosen to represent pamphlets and other similar short documents by extracting all documents with less than 10406 characters in them. Compared to other specific genres or text types, this proved to be a successful method when attempting to define a meaningful subcorpus. Advantages include: limiting effects of reprints; including a relatively large number of individual writers in the analysis; subjects covered by pamphlets tend to be historically topical, and thus cleaner evidence for claims in terms of diachronic meaning; and as shorter texts, inspecting single occurrences in their original context is much more efficient as things such as theme, context, and writer’s intentions reveal themselves comparatively quickly compared to larger works. Thus, issues around distant and close reading are more easily overcome. In addition, we are able to compare semantic change between the larger corpus and the more rapidly shifting topical and political debates found in pamphlets, which offers its own historical insights.

In terms of specific linguistic approaches, analysis started with examinations of contextual distributions of “public” by year. Then, by changing the parameters of this analysis (for example, by defining the context as a set of syntactic dependencies engaged by public, or as collocation structures of a wider lexical environment) different aspects of the use of “public” can be brought to the foreground.

As syntactic constraints govern possibilities of combinations of words in shorter ranges of context, the narrower context windows contain a lot of syntactic information in addition to collocational information. Because of this syntactic restrictedness of close range combinations, the semantic relatedness of words with similar short range context distributions is one of degree of mutual interchangeability and, as such, of metaphorical relatedness (Heylen, Peirsman, Geeraerts, Speelman 2008). Wider context windows, such as paragraphs, are free from syntactic constraints, and so semantic relatedness between two words with similar wide range context distributions carries information from frequent contiguity in context and can be described as more metonymical than metaphorical by nature, as is visible from applications based on term-document-matrices, such as topic modelling or Latent Semantic Analysis (cf. Blei, Ng and Jordan (2003) and Dumais (2005))

The syntactic dependencies were counted by analysing the pamphlet subcorpus using Stanford Lexical Parser (Cheng and Manning 2014). Results show changes in the tendency to use "public" as an adjective attribute and in compound positions. Since in English the overwhelmingly most frequent position for both adjective attributes and compounding attributes is preceding head words, this analysis could be adequately replicated using bigrams in the whole dataset. Lexical environments have been analysed by clustering second order collocations (cf. Bertels and Speelman (2014)) and replicated by using a random sampling from the whole dataset to produce the second order vectors.

The study of all bigrams relating to "public" (such as "public opinion", "public finances", "public religion") in ECCO provides for a broader analysis of the use of "public" in eighteenth-century discourse that not only focuses on particular compounds, but provides a better idea of which domains "public" was used in. It points towards a declining trend in relative frequency of religious bigrams during the course of the eighteenth century and rise in the relative frequency of secular bigrams - both political and economic. This allows us to present three arguments: First, it is argued that this is indicative of an overall shift in the language around "public" as the concept's focus changed and it began to be used in new domains. This expansion of discourses or domains in which "public" was used is confirmed in the analyses of a wider lexical environment. Second, we also notice that some collocates to public, such as "public opinion" and "public good", gained a stronger rhetorical appeal. They became tropes in their own right and gained a future orientation in political discourse in the latter half of the eighteenth century (Koselleck 1972). Third, by combining the results of the distributional semantics of "public" in ECCO with information extracted from ESTC, one can recognize how different groups used the language relating to "public" in different ways. For example, authors writing on religious topics tended to use "public" differently from authors associated with the enlightenment in Scotland or France.

There are two important upshots to this study: the methodological and the historical. With regard to the former, the paper works as a convincing case study which could be used as an example, or workflow, for studying other words that are pivotal to large structural change. With regard to the latter, the work is of particular historical relevance to recent discussions in eighteenth century intellectual history. In particular, the study contributes to the critical discussion of Habermas that has been taking place in the English-speaking world since the translation of his *Structural Transformation of the Public Sphere* in 1989, while also informing more traditional historical analyses which have not been able to draw tools from the digital humanities (Hill 2017).

References

Bertels, Ann and Dirk Speelman (2014). “Clustering for semantic purposes. Exploration of semantic similarity in a technical corpus.” *Terminology* 20:2, pp. 279–303. John Benjamins Publishing Company.

Blei, David, Andrew Y. Ng and Michael I. Jordan (2003). “Latent Dirichlecht Allocation.” *Journal of Machine Learning Research* 3 (4–5). Pp. 993–1022.

Chen, Danqi and Christopher D Manning (2014). “A Fast and Accurate Dependency Parser using Neural Networks.” *Proceedings of EMNLP 2014*.

Dumais, Susan T. (2005). Latent Semantic Analysis. *Annual Review of Information Science and Technology*. 38: 188–230.

Gunn, J.A.W. (1989). “Public opinion.’ *Political Innovation and Conceptual Change* (Edited by Terence Ball, James Farr & Rusell L. Hanson). Cambridge: Cambridge University Press.

Habermas, Jürgen (2003 [1962]). *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Cambridge: Polity.

Heylen, Christopher, Yves Peirsman, Dirk Geeraerts and Dirk Speelman (2008). “Modelling Word Similarity: An Evaluation of Automatic Synonymy Extraction Algorithms.” *Proceedings of LREC 2008*.

Hill, Mark J. (2017), “Invisible interpretations: reflections on the digital humanities and intellectual history.” *Global Intellectual History* 1.2, pp. 130-150.

Hölscher, Lucian (1978), “‘Öffentlichkeit.’” Otto Brunner et al. (Hrsg.) *Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland*. Band 4, Stuttgart, Klett-Cotta, pp. 413–467.

Koselleck, Reinhart (1972), “‘Einleitung.’” Otto Brunner, Werner Conze & Reinhart Koselleck (hrsg.), *Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland*. Band I, Stuttgart, Klett-Cotta, pp. XIII–XXVII.