‘The Various Forms of Civilization Arranged in Chronological Strata’: Manipulating the *HTOED*

Marc Alexander – University of Glasgow
m.alexander@englang.arts.gla.ac.uk

1. Introduction

The complete *Historical Thesaurus of the Oxford English Dictionary* (*HTOED*) is an unparalleled companion to the *Oxford English Dictionary* itself for the study of words. Its database, containing a little under 800,000 sense entries, is a resource which can be used in many different ways beyond the preparation of the printed volumes.[1] This article uses the *HTOED* database to take some steps towards generating data-rich displays of some of the data stored within the thesaurus itself. With access to the database and means of organizing what they find there, researchers can unlock data contained but not accessible within the *OED* (the principal ‘parent’ of the *HTOED* (Kay *et al* 2009: xiii)) due to its alphabetical arrangement. Charlotte Brewer, identifying the problem of scope, refers to the statement in *The Times* from which this article’s title is taken:

> ...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’… (Brewer 2007: 232)

[1] For information about the structure of the original database, see Kay and Chase (1987), Wotherspoon (1992) and Wotherspoon (this volume). The database used here was based on that described in those articles, with a large number of structural modifications by the author.

By taking some steps towards visually displaying *HTOED* data, we can glimpse ways in which this thesaurus moves toward giving us a 'complete picture'.

2. The Data

As outlined elsewhere in this volume (in the chapters by Kay and Wotherspoon), the data stored within the *HTOED* is a fine-grained conceptual hierarchy containing almost all of the recorded words in English, arranged semantically. Each category of words is nested within other, wider categories, so that, for example, the verb category *Live dissolutely* is within *Licentiousness*, itself adjacent to *Guilt* and *Rascalry* and within the wider category *Morality*. This hierarchical structure differs from the organization of many other thesauri, such as that of Peter Mark Roget. While Roget's categories exist in a single linear sequence, *HTOED* categories can relate to others either horizontally (on the same hierarchical level) or vertically (on a higher or lower level, either containing or being contained by another category). In addition, each concept is able to contain a series of subcategories within itself, separate from the main sequence. It is this complex hierarchical structure which helps make the *HTOED* database so useful for visualization: each individual point in the hierarchy can contain both word entries for the concept represented by that point, and also all the conceptual descendants which follow it, each surrounded by siblings of similar meaning.

The size of the *HTOED* also makes it amenable to computational analysis. The current version of the database (as of early 2010) contains 793,747 entries, compared to *OED2*'s 616,500 (Algeo 1990: 137), all within 236,346 categories, each representing a distinct concept. Taking into account each field stored within it, the database itself contains approximately 22.7 million pieces of data.

3. Hierarchy and Visualisation

To best display any data visually, an analyst attempts to increase informational density while simultaneously maximising what de

Beaugrande and Dressler call *informativity* (1981: 17ff). This is a careful act because, beyond an ideal 'peak', as information density increases within a fixed space, information transfer rapidly approaches zero (Tufte, 2006). *HTOED* data presents a challenge in this area due to its hierarchical nature.

The normal metaphor for visualising hierarchy is a tree-like system, like that often used in organisation charts. *HTOED* data is, however, far too large to be used in such a way – even a spider-like tree or hypertree could not represent the thesaurus, whose largest category alone (the adverb *Immediately*) contains over 250 synonyms. Alternative and emerging tree-like representations were investigated (see Robertson *et al* 1991, 2002 and Furnas and Zacks 1994), but none could appropriately represent the scale of the *HTOED*.

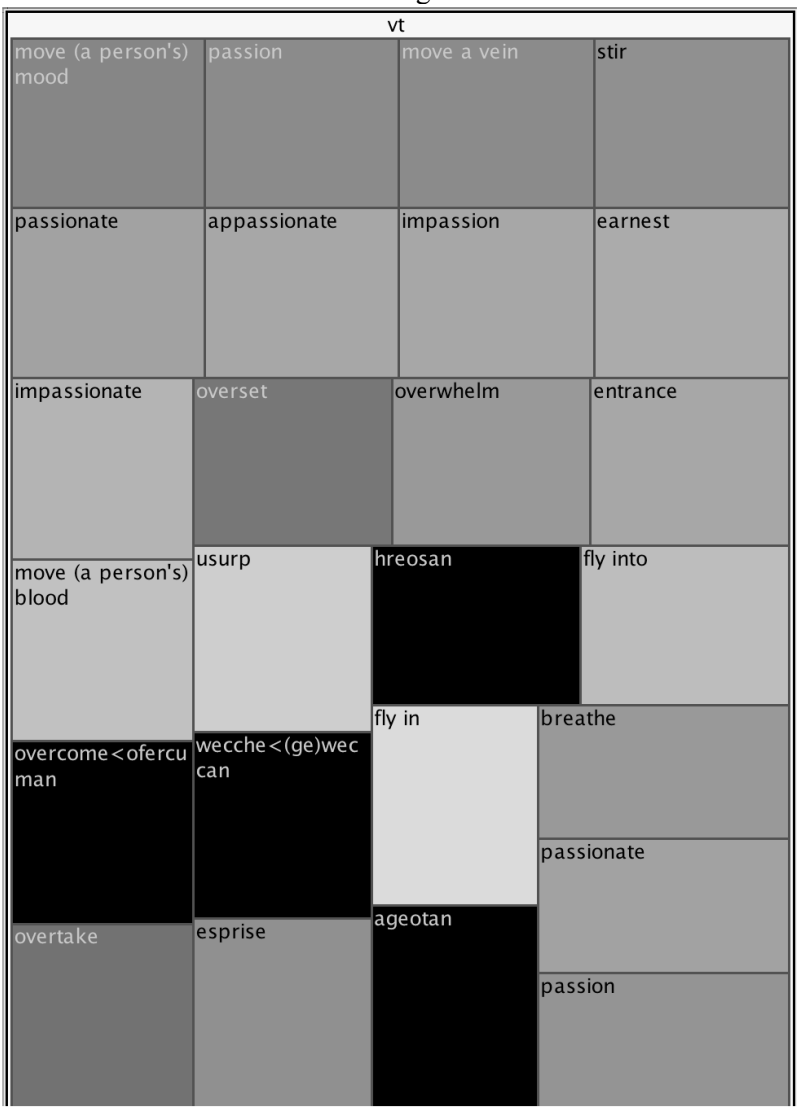
An alternative way of displaying hierarchy two-dimensionally is by representing each category as a nested object on a plane, such as a rectangle. This technique produces a 'treemap' (see Shneiderman 2009), wherein each entry in a hierarchy is represented by a rectangle which is large enough to contain smaller rectangles representing its descendants while simultaneously being itself small enough to nest within further rectangles representing its parent categories, rather like a Russian matryoshka doll. In short, a treemap structure takes the organisational chart metaphor of SENIOR IS UP and replaces it with SENIOR IS BIG.

4. *Passion, Love* and *Hate*

Figure 1 shows the *HTOED* data for the transitive verbs for *Affect with passion/strong emotion* (Kay *et al* 2009: 1053) as a treemap. Each rectangle is equal in area, if not identical in dimension, and the algorithm varies these dimensions while keeping area constant in order to ensure it tessellates precisely.[2] The visualisation in Figure 1 shows an extra dimension of data by adding shading to each rectangle, with the shade representing the first cited date of

[2] The algorithm used throughout this paper is Treemap-Squarified, written by Ben Bederson and Martin Wattinberg and available as an open source Java implementation from the University of Maryland Human-Computer Interaction Lab at http://www.cs.umd.edu/hcil/treemap/.

each word, mapped onto a linear scale between 1000CE (black) and 2000CE (white). This necessitates converting the *Old English* value in *HTOED* to an integer value of 1000, and the *Current* value to 2000. These values were chosen to give a reasonable linear scale.



02.02.15 (vt.): *Affect with passion/strong emotion* by first cited date

This visualization, although large and thus relatively data-poor, displays features of interest. It maintains the principle that word entries should be the most salient part of the image while it attempts to add as much information as feasible, for instance, adding a display of the first cited date and an overlay of the word form, as here. This extra dimension can vary, however: Figure 2 shows all of the *Strong feeling/passion* category with the parts of speech differently coloured (white to black in alphabetical order of adjective, adverb, interjection, noun, verb intransitive, verb reflexive and verb transitive). From the visualization, one quickly perceives that, unlike some categories, *Strong feeling/passion* is dominated by both adjectives and nouns (compare 01.02.08.02 *Beverage*, a category dominated by nouns, or 01.05.08.04.01 *Swift movement*, dominated by verbs).

Figure 2: 02.02.15 (all): *Strong feeling/passion* by part of speech

HTOED data can also be filtered to contextually represent a subset of the data. Figure 3 shows a subset of the *Ardour/fervour* category, displaying only those words which have citation evidence after 1870CE. This date was deliberately chosen to approximate

usage in modern times (due to the age of parts of the *OED*). This allows an analyst to compare the overall category above with the recent category, and to visually inspect areas of interest, such as the small number of words from Old English (eight), and the relatively large number from Middle English.

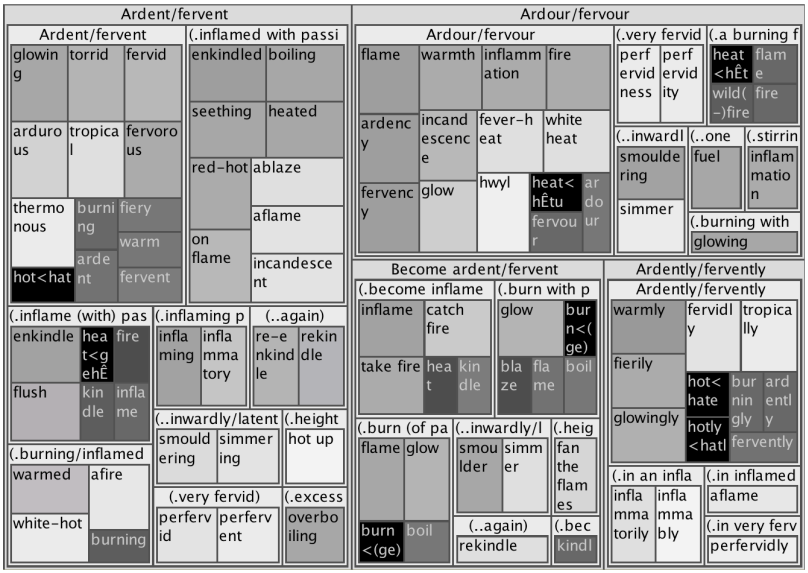


Figure 3: 02.02.15.01 (all): *Ardour/fervour,* current citation evidence only

The other option to add information, alongside colour/shading, is to vary the size of each rectangle. While this is technically possible, it is not ideal for linguistic data. The general principle for visualisation is to ensure that the number of variable dimensions should not exceed the number of dimensions in the data itself; citation dates, parts of speech and other information contained in the *HTOED* are generally linear or categorical, and so unidimensional. As Edward Tufte says:

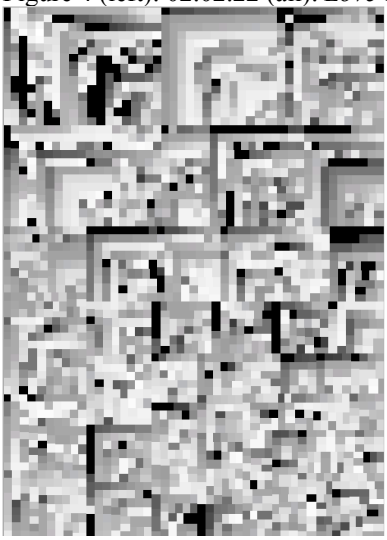
> There are considerable ambiguities in how people perceive a two-dimensional surface and convert that perception into a one-dimensional number. Changes in physical area on the surface of a graphic do not reliably produce proportional changes in perceived areas. […] These designs cause so

> many problems that they should be avoided. (Tufte 2001: 71; see also Macdonald-Ross 1977)

Experiments in this area have shown that Tufte's advice is best followed with the *HTOED* visualizations.

Lastly, the final axis which can be varied with respect to these treemaps is scale; that is, the size of each word-rectangle can be reduced, increasing the range of data which each can display. This loses one dimension of data, as small rectangles do not have enough space available for text to be superimposed upon them (such as the word form the rectangle represents), although this can also be an advantage in focusing the visual display on wider, more data-dense patterns. Figure 4 shows the category *Love*, with the first-cited date scale introduced above, and can be compared to Figure 5, *Hate*. It is immediately apparent that *Hate* is a 'darker' category than *Love*; in the monochrome scale used here, a category which appears 'dark' is one which is relatively early, with most lexical innovation occurring in *Hate* in the Old and Middle English periods (note that Old English is always represented as black, as it is a binary category in the *HTOED* database).

Figure 4 (left): 02.02.22 (all): *Love* by first cited date



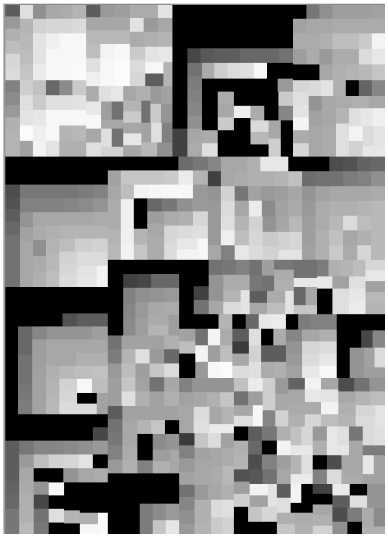


Figure 5 (right): 02.02.23 (all): *Hate* by first cited date

5. Mapping English

The sections above explored the visualization of the *HTOED* at the level of lexical and categorical exploration; what might be termed the 'mid-level' of the thesaurus. Moving back further in the hierarchy lets us view the semantic structure of the English language as a whole.

Present-Day English

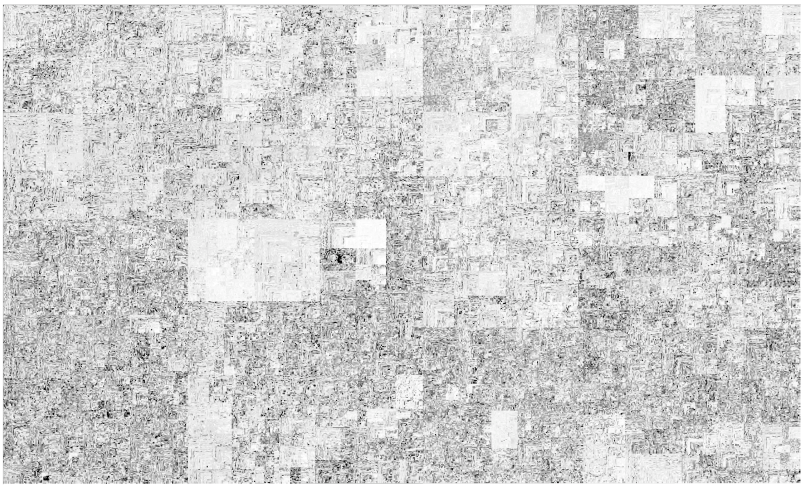


Figure 6: Present-day English as shown by the *HTOED*, shaded by first cited date

Figure 6 is a treemap showing all of present-day English in the *HTOED*, with every word represented by a small dot of ink. Those black dots are present-day words which originated in Old English, and white dots represent those which entered the language much more recently. Again, the map is arranged by semantic field, so words in close semantic proximity are also physically close to one another on the diagram. Such a visualisation is not very useful without a key, which Figure 7 provides.

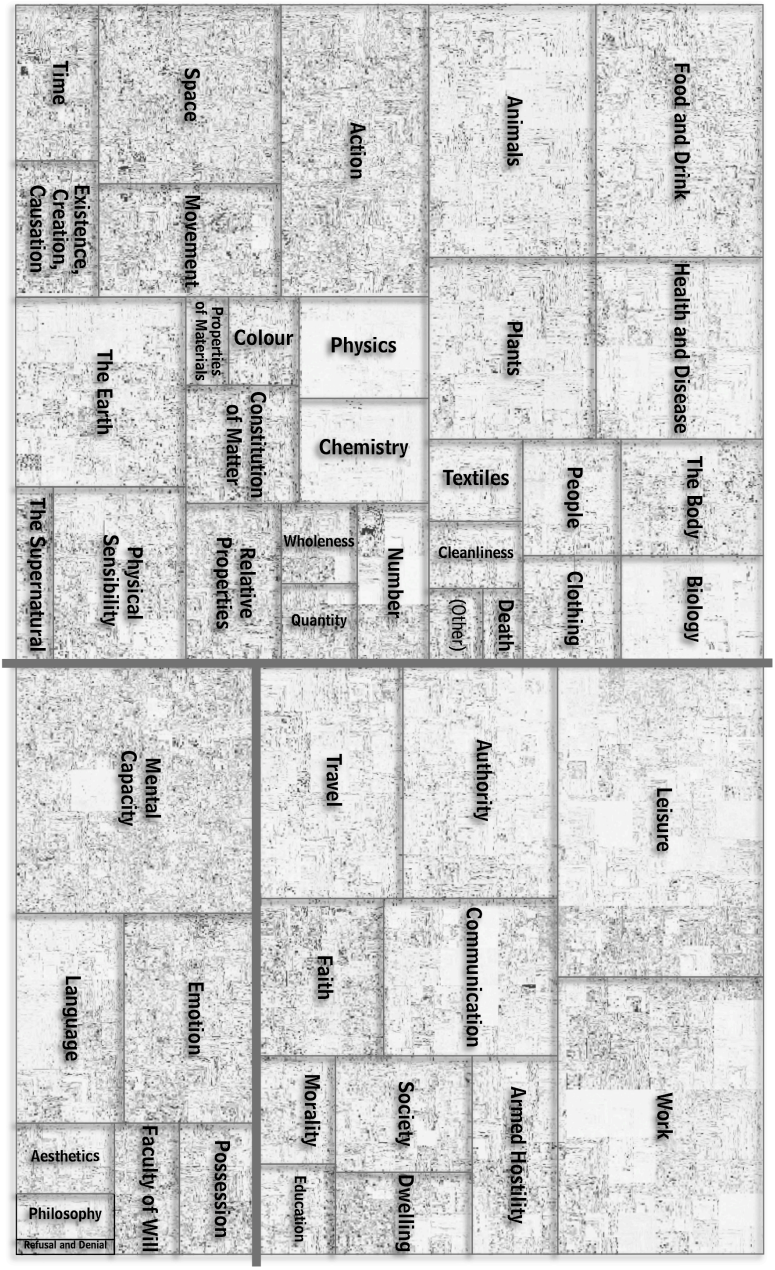


Figure 7: A key for the present-day English treemap

One unintended side-effect of this visualization is that it produces what might be called a 'patchwork' effect. Areas such as *Physics* and *Chemistry* are quite light, as are parts of *Number* (which includes *Mathematics*) and *Language*. Although unexpected, this is natural – such light patches are areas of recent lexical innovation, made up of clusters of words first cited in the *OED* in recent years (or, rather, from the late nineteenth century onwards – recent from the perspective of much of *OED2*, an issue *OED3* will address). Therefore, we can expect a 'patchwork' effect in areas affected by rapid social, technological or academic growth, such as *Computing* (inside *Number*, adjacent to *Mathematics*), *Physics, Chemistry, Linguistics, Communication, Travel*, and so on. Conversely, darker and therefore older patches cover existence in *Time and Space, Creation, Causation, Faith, Emotion*, and the parts of *Number* which refer to *Arithmetic* or *Enumeration*.

Diachronic 'Slices'

This effect is pronounced in present-day English, but if other selections of the data are taken, it reduces somewhat. If the present-day data is thought of as a 'slice' of the *HTOED*, then other such slices can be taken between the Old English period and the present day.

Figure 8 shows three such 'slices' of the data, centred on major literary-historical figures of the Late Modern, Early Modern and Middle English periods. Proceeding backwards from the present day, the first treemap is that of Samuel Johnson, and shows those entries first cited before his death in 1784 and last cited after his birth in 1709, a total of 247,933 senses. The second covers those senses we have evidence of being in use during the life of Shakespeare (1564-1616, 207,930 senses) and the third illustrates the same for Geoffrey Chaucer (an approximation of 1340-1400, with 73,432 senses).

Looking at these treemaps, there is a visible reduction between present-day English and that of Johnson in the patchwork effect observed in Figure 6, although it is still observable in places (most notably in the area of *Leisure*). A further reduction is visible in the semantic space of Shakespeare where these shades are much more

evenly distributed, and by the time of Chaucer almost no delineated rectangular patches of innovation can be found.

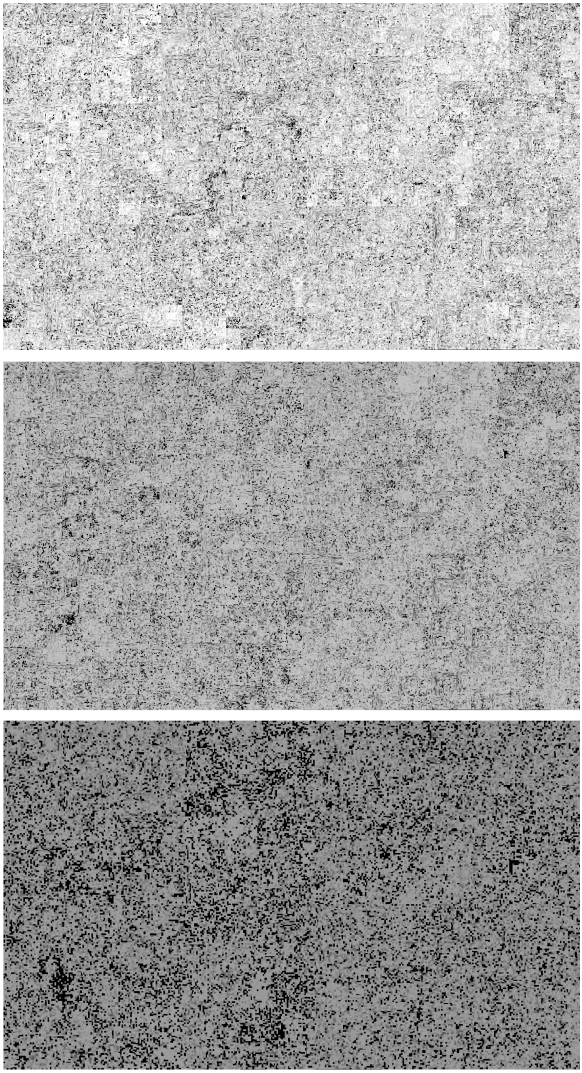


Figure 8: From top to bottom: English during the life of Samuel Johnson, William Shakespeare and Geoffrey Chaucer respectively, shaded by first cited date

One point to note here is that although the images in Figure 8 have been displayed at identical size for clarity's sake, when the size of each word is fixed then the relative size of these images varies greatly (as is apparent from the difference in granularity of the word 'blocks' between Johnson and Chaucer). To disambiguate, Figure 9 shows the relative sizes of the semantic spaces of all four high-density treemaps in this section.

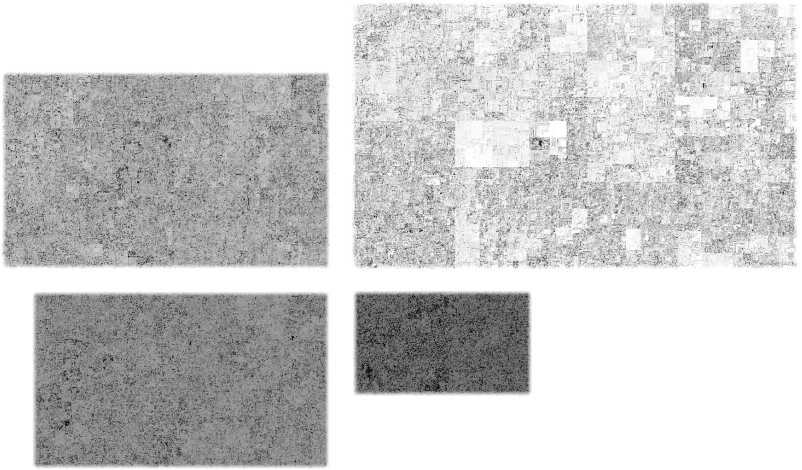


Figure 9: Relative sizes of *HTOED* visualization 'slices'; anti-clockwise from top right, present-day English, English in the age of Johnson, English in the age of Shakespeare and English in the age of Chaucer

Overlays

Finally in this section, it is possible, using two colour scales, to overlay one set of data results on an existing chart. As a demonstration, Figure 10 overlays on the present-day English treemap all those words which have been first cited in the *OED/HTOED* since 1983 (that is, in the lifetime of the current author). Some patterns show up which bear further investigation – a small cluster in *Computing* highlighting the large number of neologisms in that area in the 1980s and 1990s, and a surprisingly large cluster in *Leisure* which shows a jump in music-related neologisms (especially in 03.11.03.01.03.03 (n.) *Pop music*, with entries such as *gangsta rap* 1990–, *techno-house* 1991– and *trip*

hop 1994–; cf p.1705 of Kay *et al* 2009). Such overlays can show the location of selected senses and link the study of neologisms, word forms or other information with their semantic distribution. Beyond this, they should also be of interest to researchers into the practice of lexicography, giving (as here) information about recent emphases in the collection of neologisms pre-*OED3*, as it is unlikely that the high number of musical terms in the overlay shown in Figure 10 is proportionally correlated with developments in the English language itself, but is rather related to the nature of the *OED Additions* series within the data as a whole.

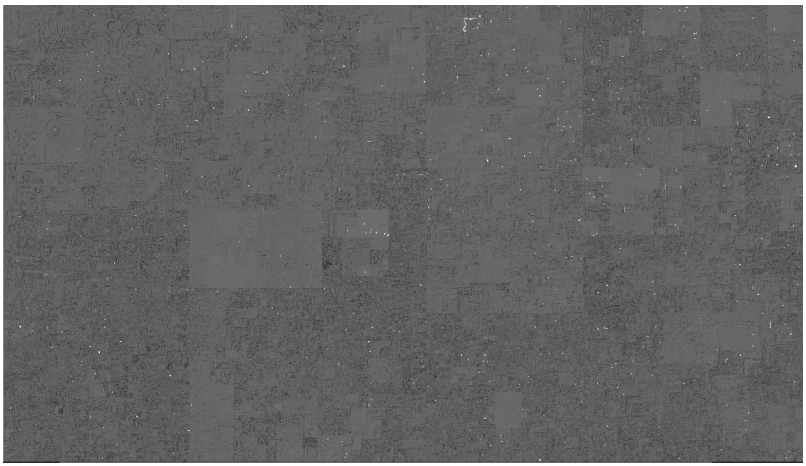


Figure 10: Neologisms since 1983 as recorded in *HTOED*

Similarly, Figure 11 similarly overlays onto present-day English the famously polysemous word-form *set* in *HTOED*, which appears in 313 distinct categories. Its distribution is not as clustered as the neologisms above, with remarkably few clusters – demonstrating the extensive spread of the word form across the semantic space of English.

Figure 11: The senses of the word-form *set*, as recorded in *HTOED*

6. Summary

In the ways outlined above, such visual displays of *HTOED* data can provide useful entry points to a large, complex lexicographical and lexicological dataset. Firstly, in a pedagogical sense, displays could give students and others a new way of looking at lexicological data, and of exploring them as an application of semantic field theory.[3] Secondly, as computer displays and online dictionary interfaces become more polished, new ways of encouraging exploration of lexical data online are needed to replace the lost experience of browsing a printed dictionary, rather than only providing users with a blank search interface. And finally, such visualizations can point analysts towards areas of possible semantic, lexical or cultural interest, whether areas of trauma in the history of the language, areas of rapid growth, or areas of relative stability. All of these could be made easily visible through some

[3] Adrienne Lehrer (1974: 15) describes semantic fields as the theory that 'the words of a language can be classified into sets which are related to conceptual fields and divide up the semantic space or the semantic domain in certain ways'. If Figure 7 above shows the semantic space of English, then Figure 6 can be said to show the 'word space' of the language, a plane containing lexical items and parallel to the semantic space, but without explicit semantic divisions.

enhancement and development of the first steps given here towards displaying and exploring the data stored within the *HTOED* itself.

References

Algeo, J. (1990). 'The Emperor's New Clothes: The second edition of the society's dictionary'. *Transactions of the Philological Society*, 88(2), pp. 131-150.

de Beaugrande, R.-A., and W. Dressler (1981). *Introduction to Text Lingustics*. Longman, London.

Brewer, C. (2007). *Treasure-House of the Language: The Living* OED. Yale University Press, New Haven, CT.

Furnas, G.W., and J. Zacks (1994) 'Multitrees: Enriching and reusing hierarchical structures'. *Human Factors in Computing Systems: Proceedings of the CHI '94 Conference 1994*, pp 330-336.

Kay, C., and T. Chase (1987) 'Constructing a Thesaurus Database'. *Literary and Linguistic Computing* 2(3), pp. 161–163.

Kay, C., J. Roberts, M. Samuels, and I. Wotherspoon (2009). *Historical Thesaurus of the OED*. Oxford University Press, Oxford.

Lehrer, A. (1974). *Semantic Fields and Lexical Structure*. North-Holland Publishing Co., Amsterdam.

Macdonald-Ross, M. (1977) 'How Numbers Are Shown: A Review of Research on the Presentation of Quantitative Data in Texts'. *Educational Technology Research and Development*, 25(4), pp. 359-409.

Robertson, G.G., J.D. Mackinlay, and S.K. Card (1991). 'Cone Trees: Animated 3D Visualizations of Hierarchical Information'. *Proceedings of the SIGCHI conference on human factors in computing systems: Reaching through technology*, pp.189-194.

Robertson, G.G., K. Cameron, M. Czerwinski, and D. Robbins (2002) 'Animated Visualization of Multiple Intersecting Hierarchies'. *Journal of Information Visualization*, 1(1), pp.50-65.

Shneiderman, B. (2009) 'Treemaps for space-constrained visualization of hierarchies',

http://www.cs.umd.edu/hcil/treemap-history/index.shtml, accessed 26 March 2010.
Simpson, J. & E. Weiner (eds) (1989) *The Oxford English Dictionary*, 2nd. ed., Oxford University Press, Oxford.
Tufte, E. (2001). *The Visual Display of Quantitative Information*, 2nd edition. Graphics Press, Cheshire, CT.
Tufte, E. (2006). *Beautiful Evidence*. Graphics Press, Cheshire, CT.
Wotherspoon, I. (1992) 'Historical Thesaurus Database Using Ingres', *Literary and Linguistic Computing*, 7, 4, pp. 218-225.