Culturomics and Genre: *Wrath* and *Anger* in the 17th Century

Hans-Jürgen Diller
Ruhr-Universität Bochum

1. The tools

1.1. Culturomics and the Google Ngram Viewer

My first introduction to "Culturomics" was at the Helsinki Corpus Festival in late September 2011, by Andreas Jucker, although at the time I wasn't aware of it. In the course of his opening plenary lecture, which was about *courtesy* and *politeness*, Andreas showed a diagram which, I believe, was generated by the Google Books Ngram Viewer <books.google.com/ngrams>. Given my obsession with word pairs, I immediately thought: what can be done with *courtesy* and *politeness* can also be done with *passion* and *emotion* or with *wrath* and *anger*. I soon discovered that there is a whole new science (or what claims to be a new science) behind Google Books, and the Google Books Ngram Viewer in particular. The name of that science is *Culturomics*, of course a portmanteau formation from *Culture* and *Genomics*.

In January 2011 a group of scientists associated with the Harvard Cultural Observatory (among them Steven Pinker) published an article in *Science* (Michel *et al.* 2011a), announcing that Culturomics "extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities". The article was based on a corpus of over 5 million books "(~4% of all books ever published)" which had "emerged from Google's effort to digitize books" (p. 176, col. 1). The English sub-corpus amounts to about 361 billion words (col. 2). The English-language entries from the year 2000 alone would take 80 years to read – "without interruptions for food or sleep" (*ib.*).

In their pioneering studies they analyse the frequency of "slavery", and (somewhat unsurprisingly) they discover that "[t]he use of 'slavery' peaked during the Civil War […] and then again during the civil rights movement". Other studies concern the waning frequency of irregular verbs, the changing fame of various personalities and censorship during the Third Reich (where the need for careful interpretation, which is emphasized by the authors themselves, is particularly great).[1]

The corpora on which the Ngram Viewer is based were originally created in July 2009 <http://books.google.com/ngrams/info>.[2] *Ngram* (or *n-gram*) is simply a superordinate term for *1-gram* ("unigram"), *2-gram* ("bigram"), *3-gram* etc. For practical reasons, Michel *et al.* stop at 5-grams. A 1-gram is either a word or a punctuation mark. 2-, 3-, 4-grams etc. are combinations of a corresponding

[1] <http://www.sciencemag.org/content/331/6014/176/suppl/DC1> contains a link "SOM Data" which opens a file "6D) Manual annotations of the most suppressed and most enhanced individuals (German list, 33-45)", where Konrad Adenauer is given a "suppression score" of 127.981, whereas Kurt Schumacher scores only 91.7682. Adenauer, Lord Mayor of Cologne and President of the Prussian State Council until 1933 and Federal Chancellor 1949-63, was imprisoned by the Nazis for several short periods. Schumacher, Social Democratic member of the *Reichstag* 1930–33 and party leader 1945–52, spent more than 10 years in concentration camps <http://en.wikipedia.org>. Since the suppression score reflects the difference between pre- and post-Nazi prominence and Nazi-time obscurity (p. 181, col. 3), the difference between the two scores is probably due to Adenauer's greater visibility before 1933 and after 1945 rather than to either man's obscurity between those years. And we must, of course, remember that censorship or "suppression" are not the same as persecution and physical suffering.

[2] Thanks to progress in OCR, improved versions of all sub-corpora were created in 2012, with the exception of the "Google Million".

 Selected Proceedings of the 2012 Symposium on New Approaches in English Historical Lexis (HEL-LEX 3), ed. R. W. McConchie et al., 54-65. Somerville, MA: Cascadilla Proceedings Project.

number of 1-grams. The frequency of a n-gram "is computed by dividing the number of instances of the n-gram in a given year by the total number of words in the corpus in that year" (col. 2).

Use of the Ngram Viewer is quite simple. To trace the relative frequency of any n-gram through any time span you merely visit <http://books.google.com/ngrams>, where you used to be shown the frequency of "Atlantis" (a 1-gram) and "El Dorado" (a 2-gram) in "English" from 1800 to 2000.[3] You then enter the time span and the n-grams that interest you in the appropriate boxes, and the result is the desired graph. The present study, which tries to combine the Ngram Viewer and the Google Book Search to analyse the use and frequency of *anger* and *wrath* in the 17th century after Shakespeare, is in many ways a continuation of Diller (Forthcoming b), which studies the use of those two words in Shakespeare's time. Before we can analyse the word-use we must give some account of the potential of the Ngram Viewer in combination with the Google Book Search. The Ngram Viewer has "filled many students of culture with enthusiasm" because they think it enables them to "quantify and empirically investigate cultural trends, political currents, collective mentalities and even the complexity of cultures" (Krischke 2012: p. N5, col. 4).

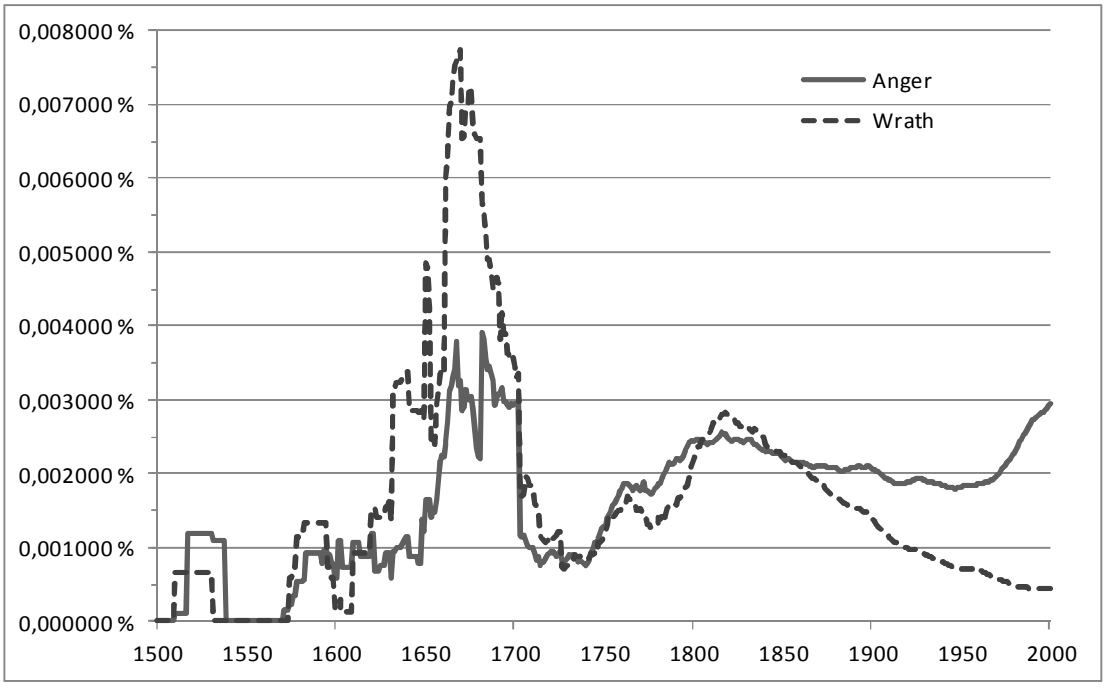

Figure 1: Relative frequency of *anger* and *wrath*, 1500-2000
(adapted from http://books.google.com/ngrams/)

[3] Michel *et al.* restrict their study of English to 1800–2000, "the most carefully curated of the datasets" (2011b: 16). Between June and October 2012 "Atlantis" and "El Dorado" have been replaced by "Albert Einstein", "Sherlock Holmes", and "Frankenstein".

1.2. What the Google Ngram Viewer can show

Figure 1 shows the relative frequencies of *wrath* and *anger* from 1500 to 2000 in the "English One Million" corpus with "smoothing of 10". "Smoothing of 10" means that the diagram points above any particular year (say, 1600) do not represent the frequencies of just that year but of 1600 plus the ten preceding and following years (i.e. 1590–1610), divided by 21 (=10+1+10). The default smoothing of the Ngram Viewer is set at 3. Smoothings help us discover long-term tendencies: the higher the smoothing, the more long-term the observable tendency. Smoothing of 0 would yield the exact values for every single year (=0+1+0). It would result in a very spiky line which, so to speak, would represent a century as 100 trees but not as a wood. Fig. 1 shows abrupt rises and drops at or near years 1630, 1650, and 1700.

"English One Million", also called "the Google Million" or "Eng-1M", is one of five English corpora, the others being "English", "American English", "British English", and "English Fiction".[4] The "Google Million" was chosen because it seems to be the most carefully curated of the English corpora. Its description is also the most detailed one and deserves to be quoted in full:

> All [books in Eng-1M] are in English with dates ranging from 1500 to 2008. No more than about 6000 books were chosen from any one year, which means that all of the scanned books from early years are present, and books from later years are randomly sampled. The random samplings reflect the subject distributions for the year (so there are more computer books in 2000 than 1980).
>
> (from <http://books.google.com/ngrams/info>)

The Ngram Viewer gives only percentages, but absolute figures can be obtained with a trick. Setting the smoothing at 0 and clicking the "here" link at the bottom (as in Fig. 1) gives access to the number of n-grams (1- to 5-grams) in all corpora for all years included in them. For 1-grams even "total counts" are given, i.e. the full number of 1-grams for each year. To obtain absolute figures, we have to read the percentage for any year off our diagram and multiply it by the number of 1-grams for that particular year. Since 1-grams are not exactly the same as words and since reading a quantity off a diagram converts analogue into digital information, the resulting figures will not be absolutely accurate, but they will give us an idea of the order of magnitude: in our period the absolute frequency of words like *anger* and *wrath* will not usually rise above two-digit figures for any one year.

Michel *et al.* (2011b: 13) claim that "the Eng-1M corpus more closely resembles a traditional 'balanced' corpus" than the other corpora. At the same time, they freely admit that the balancing algorithm employed is "very crude" and "might in fact skew the resulting corpus in ways that made it less reflective of texts actually published at the time." The authors also recognize the need for "the development of high-throughput methods for balancing large corpora consistent with a particular corpus-builder's desiderata".

Even as a 'balanced' corpus, the "Google Million" shares an important weakness with the rest of the Google corpora: it is not open to inspection, there is, to my knowledge, no publicly available list of the books that compose the corpus. In short, the data represented in the graphs cannot be checked by the user. Although the authors are aware that the scanned books often contain extraneous editorial matter, which should be eliminated before a word count, Michel *et al.* (2011b: 5) give only a very general picture of their elimination principles and practices, which they call "Structure Extraction". Personal experience suggests that structure extraction can be a time-consuming and error-prone process; so any frequency calculations should be taken with a pinch of salt.

---

[4] There are also corpora of Chinese (simplified), French, German, Hebrew, Spanish, and Russian. For more detailed descriptions see <http://books.google.com/ngrams/info> and Michel *et al.* (2011b: 13). "English Fiction", incidentally, presents a picture that is quite different from that of the "Google Million": from 1600 to the present, *anger* is in a majority over *wrath* that is interrupted only from 1628 to 1644. The reader can obtain this piece of information by generating a diagram from the "English Fiction" corpus for the years 1625-1645 with smoothing of 3. Smoothing of 0 will show that *wrath* tops *anger* only in 1631, 1634, 1639, and 1641.

1.3. From the Ngram Viewer to the Google Book Search

The lack of verifiability or falsifiability which this involves is addressed as a "frequently asked question" <http://www.culturomics.org/Resources/faq>: "All these n-grams are out of context. I can't tell how they are being used." To find out about the contexts, the authors recommend a "trick": "look at the hits" from any particular year "by eye". The "hits" can be obtained from below the graphs (such as Fig. 1), where you find "'interesting' year ranges for your query terms. Clicking on those will submit your query directly to Google Books […], but the results are returned from the full Google Books corpus", not only from the corpus you selected <http://books.google.com/ngrams/info>.

That last qualifier means, in plain English, that the problem of falsifiability is not really solved, perhaps not even seen. The "trick" does give us contexts, but they are not exclusively derived from the textual universe that is represented in the graphs. Incidentally, the study of the use of more than one word affords an illustration of the consequences of that missing link, as will be shown in Table 2 and early in Section 3. Still, the books which contain the "hits" from any year range can serve, in many ways, as a corpus, even though we have no way of knowing the textual universe which that corpus represents.

Looking at contextualized data "by eye" offers many insights, especially when you want to explore fine differences in the usage of two or more words that may be said to be in competition with each other – fight in a semantic battlefield, as Kiricsi (2003–04; 2005: 163–86) puts it. Diller (Forthcoming b) looks at such contexts in the year range of 1575–1620. Together with data derived from Gevaert (2007) and the British National Corpus <http://corpus.byu.edu/bnc/>, this enables us to form an over-all, long-term picture, as found in Table 1.

Table 1: Frequency of *wrath* and *anger* in the history of English

Year	Source	*anger*	*wrath*	Sum	*wrath* / *anger*
c1400	Gevaert, p. 152	105	798	903	7.60
c1500	Gevaert, p. 174	301	386	687	1.28
c1600	Google Book Search[5]	624	599	1223	0.96
c1990	BNC	3671	343	4014	0.09

The data are highly selective, but they do give us some idea of the change in the ratio of *wrath* to *anger*. At first sight this picture, however incomplete, looks quite plausible. It reflects the well-known fact that *wrath* is in decline while *anger* is on the rise. Even modern popular translations of the Bible, like the "Contemporary English Version", the "Easy-to-Read Version", and the "New Century Version", return no match for *wrath* <http://www.biblegateway.com/>. If, as Diller (2012; Forthcoming a) suggests, *wrath* describes primarily an active reaction to status violation while *anger* denotes an active response to the blocking of personal goals,[6] then the tendency is also plausible in that it shows the *longue durée* of a growing importance of the individual. A similar observation has been made by Wierzbicka & Harkins (2001: 17), who postulate a "shift from the Shakespearean *wrath* to modern *anger*" which "both reflects, and constitutes an aspect of, the democratisation of society and the passing of the feudal order".

But the picture is too neat: the selection of century points may be plausible, but it is nevertheless arbitrary and may hide more short-term developments which deserve our attention, too. A return to the longitudinal diagram offered by the Ngram Viewer tells us that this is indeed the case. Fig. 1 shows an enormous hump for *wrath* in the 17th century, marked by two almost vertical rises around 1630 and 1660, to be followed by a similar drop shortly after 1700.

Clearly, the long-term tendency shown in Table 1 may be interrupted by short- or mid-term influences. To identify these influences we can avail ourselves of the "interesting" years which link us

[5] These figures were generated by clicking an "'interesting' year range" at the bottom of Fig. 1 which was subsequently set at 1575–1620.

[6] An earlier version of this claim may be found in Diller (1994). See Gevaert (2007) for a detailed critique, Diller (Forthcoming a) for an anti-critique.

to the Google Book Search and thus to specific, searchable texts. Following range 1500–1640 for "wrath" we will obtain about 3,430 hits; the 1500–1641 range for "anger" returns 3,340. The hits can be sorted by "relevance" or "date". And we can do still better: once we are in the Google Book Search we can tailor the year ranges to our personal needs, even cut them down to single years.

In other words, we can obtain almost any data, and we can arrange them in a multitude of different ways. But we know very little about the database and the textual universe of which it might be a sample. It is certainly not a sample of all English texts ever printed or transmitted to us. It is a sample of the googled books of some libraries, mostly US American. As such, it will reflect the preferences of 19th-century librarians as much as those of 17th-century English writers, publishers and readers. To that extent it is not reflective of 17th-century English culture, but of the reception of that culture in the 19th century. Nothing in Michel *et al.* (2011a, b) suggests that the authors have consulted such traditional philological tools as the *New Cambridge Bibliography of English Literature* (*NCBEL*) or the *English Short Title Catalogue* (*ESTC*). But even if their database fails the criterion of representativeness it will satisfy the criterion of diversity (Biber et al. 1998: 246–53) – enough to overcome the weakness of the Google Ngram Viewer which was pointed out by Nunberg (2010): in spite of its huge database it "can't sort books by genre or topic".

Nunberg's objection is a valid one, but there is also a valid rebuttal to it: even if the Ngram Viewer "can't sort by genre or topic", we can. With the "trick" mentioned above we can access texts with high frequencies of *anger* and/or *wrath* and sort them by genre or topic – or at least a fair number of them. Topic and/or genre are often announced in the title. There will of course be titles which are inconclusive, but they are surprisingly few and can be simply omitted from the sorting. More serious are the misleading titles: a novel may, of course, masquerade as a travelogue or a diary, or as an exchange of letters: if we don't happen to know the books behind the titles, we will be misled. But the damage is not as bad as we may think at first: a novel meant to pass for a travelogue, for instance, will also adopt the linguistic features, including the vocabulary, of a travelogue. Still, if we make a claim about, say, the use of *wrath* in private letters, we would not want to be caught presenting fictitious letters as evidence. Even if such errors cannot be altogether avoided we want to keep their number as low as possible. To obtain this goal we shall combine two approaches. In the first we begin with the texts which Google offers us as frequent users of *wrath* and/or *anger* and try to categorize them. In the second we begin with categorized texts, try to find them on the Internet and see how frequently they use *wrath* and/or *anger*. As far as possible, the categories in both approaches will be broadly the same. From the first approach we expect a plausible explanation of the "hump" observed in Fig.1. The second approach, it is hoped, will provide a corpus of texts which is more "reflective of texts actually published at the time", and less reflective of the acquisition preferences of 19th-century libraries.

2. The Google Book Search: selecting a period, distinguishing genres and/or topics

Sorting by genre and (to a lesser extent) by topic was attempted by Diller (Forthcoming b), with a genre list originally inspired by the Helsinki Corpus (for details see Diller *et al.* 2010). As Diller (Forthcoming b) was concerned with *anger* and *wrath* in the time of Shakespeare it seemed best to select data from 1575 to 1620. The data were thus taken from around the turn of a century, which had the added advantage of making them comparable to those of Gevaert and the BNC. But if we want to understand the "hump" in Fig. 1, we should extend the year range at least to the decades surrounding 1630, the year of the first abrupt rise.

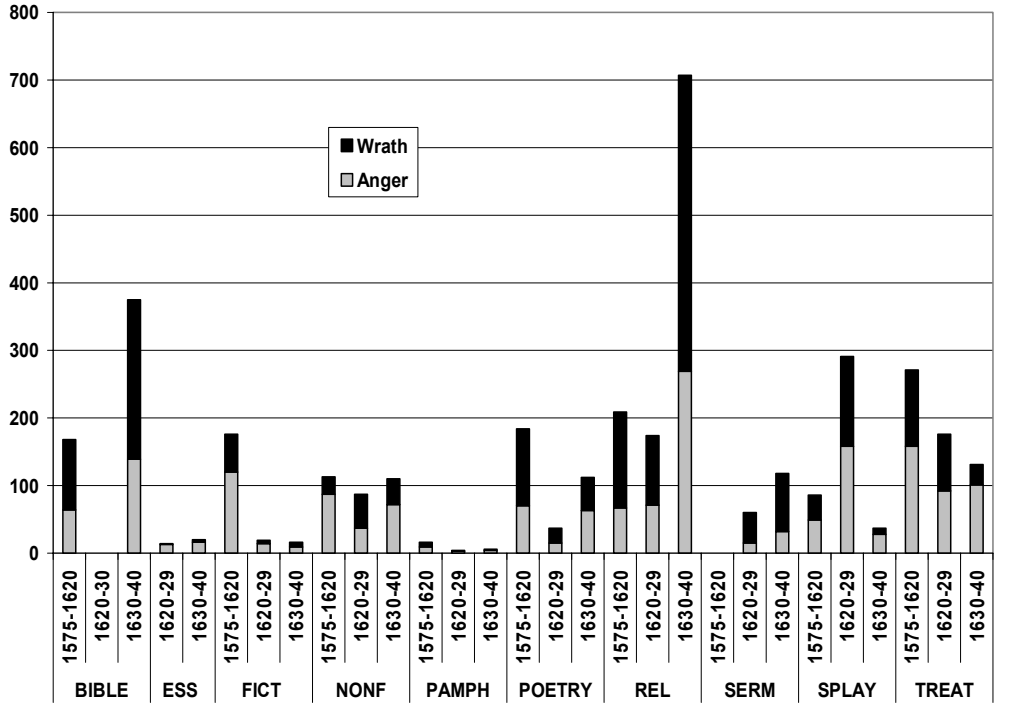

Figure 2: Frequencies of *wrath* and *anger* in selected genres, Google Book Search 1575–1640[7]

The categorization applied in Fig. 2 is largely derived from the *NCBEL* (Watson 1974, 1971) and has the advantage of being familiar to the scholarly community. But it does certainly not qualify as a strict taxonomy: BIBLE and SERM(on) could clearly be subsumed under REL(igion); there is also an overlap between REL and TREAT(ise), as well as between REL and POETRY. And almost all stage plays (as well as a fair amount of fiction) are written in verse and would thus qualify as POETRY. Religion is clearly not a genre; on the other hand it is too important a factor to be ignored in our categorization. To minimize the ambiguity which these inconsistencies entail, the following principles were observed: all stage plays are categorized as S(tage)PLAY; all poetry, whether religious or not, fictional or not, was categorized as POETRY. REL subsumes religious treatises and tracts as well as a few minor religious genres such as prayer or meditation, thus reserving TREAT for non-religious treatises. With these precautions the division is heuristically useful, giving us reasonably homogeneous groups of texts for which we can calculate a "*wrath/anger* ratio" and thus distinguish between *wrath*-preferring and *anger*-preferring genres. As Fig. 2 shows, BIBLE, REL and SERM clearly belong to the first, while FICT(ion) and NONF(iction) belong, equally clearly, to the second group. SPLAY and TREAT take an intermediate position. POETRY, which for 1575–1620 shows the same *wrath/anger* ratio as BIBLE, drops slightly in 1620–29 and significantly in 1630–40.[8]

3. An alternative look at the Internet

The differences between genres, topics, and decades are certainly suggestive but as we said earlier, they are not necessarily reflective of English culture in the 17th century. There is also a considerable discrepancy between "Eng-1M" and the "Google corpus as a whole": the abrupt rise in the *wrath/anger* ratio which Fig. 1 shows for around 1630 is not reflected in Fig. 2. While Fig. 1 indicates, for the

[7] The tables from which Figs. 2 and 3 are generated are omitted for reasons of space. They can be obtained from the author.

[8] Against these findings, Chadwyck-Healey's English Poetry Full-Text Database shows poetry as almost consistently *wrath*-preferring.

1630s, a ratio of about 3, that of the totals underlying Fig. 2 is much more stable, rising only from 0.96 in 1575–1620 to 1.23 in 1630–40 (Table 2).

Table 2: *Anger* and *wrath* 1575–1640
(for details cf. Fig: 2)

Period	Books[9]	*anger*	*wrath*	*wrath/anger* ratio
1575–1620	140	624	599	0.96
1620–30	73	419	444	1.06
1630–40	111	733	899	1.23

The only thing we can say with any confidence is that religious texts show a strong and exceptional preference for *wrath* as opposed to *anger*. To obtain a more accurate picture of the use of both words in the 17th century we must look elsewhere, and we must take our steps in a very different order. Instead of proceeding from a graph to a loosely related body of texts and then categorizing these according to genre and topic, we should start with a list of categorized texts and then try to find as many of them as possible in computer-readable, on-line form. Thanks to the Google Digitizing Initiative and the Internet Archive, finding computer-readable texts is no longer a problem, although text readability (OCR quality) often is. Still, it seems reasonable to expect that poor OCR (=Optical Character Recognition) will affect the study of *anger* no more than the study of *wrath*. Any skewing of the results is therefore unlikely or at least acceptable in an exploratory study like this. When a text or book is available in more than one version, the best OCR can be found by concordancing all versions for the same keyword and sorting contexts.

A good list of categorized texts is provided by the *NCBEL*, which has been used as the starting-point for a "European Database of Descriptors of English Electronic Texts" (EuDDEET, cf. Diller *et al.* 2010). As the inspiration for EuDDEET came from De Smet's CLMET (Corpus of Late Modern English Texts, cf. De Smet 2005), its compilation has proceeded in reverse chronological order, from the 19th and early 20th century to the 18th and 17th; it is now approaching the early modern period. As is well known, the volumes of the *NCBEL* cover the consecutive periods (600–1660, 1660–1800, 1800–1900, 1900–1950), but within these volumes the main division is into genres and topics: "Religion", "History" and "Philosophy" appear side by side with "Poetry", "Drama", "Fiction", etc. Since EuDDEET follows this principle, the data derived from it will never be strictly comparable to those derived from Google. Google pretends that a text's place on the time axis is unproblematic – with the result, e.g., that an edition of Shakespeare's *Richard III*, "Printed by I. Norton, 1634" appears under that year. A genre-driven categorization must face the fact that the problem of dating varies from genre to genre: diaries and collections of letters are often published after the writers' death, which makes the year of publication unsatisfactory. But to assign every letter or diary entry to the year or even day of writing is impracticable. By contrast, the dating of stage plays can often be accurate to the year.

For our purposes the diversity of genres is more important than the consistency of dating principles. Since our previous findings suggest the importance of religion for the choice of *wrath* vs. *anger* it was necessary to study religious texts. The highly public character of religion in the 17th century made it advisable to study a genre with more private concerns as well; the diary was chosen as such a genre. To balance these non-literary genres, fiction and drama were included. Religious texts and diaries were arranged chronologically according to author's birth, the chronology of dramatic and fictional texts was determined by their publication or performance.[10]

The text files selected were analysed with the concordancing software Antconc (Anthony 2011). As far as possible, editions of collected works were chosen as a text base. The results are shown in Table 3.

[9] Number of books found by Google Book Search to contain *anger* and/or *wrath.*

[10] Religion: from Richard Sibbes (1577?–1635) to Peter King (1669–1734); diaries: from James Melville (1556–1614) to Sir John Reresby (1634–89); prose fiction: from *Friar Rush* (1568) to Congreve's *Incognita* (1692); stage plays: from Lyly's *Sapho and Phao* (1584) to Shirley's *Contention of Ajax and Ulysses* (1658).

Table 3: *Wrath* and *anger* in selected genres (categorization from *NCBEL*)

Genre	No. of writers	*anger*	*wrath*	*wrath*/*anger* ratio
Religion	36	2869	6543	2.20
Diaries	18	133	140	1.04
Fiction	12[11]	61	7	0.11
Drama	19	910	291	0.32

A few preliminary observations are necessary before we can comment on Table 3. To make the figures truly comparable, we would have to compute the size of the text files from which they are taken. In the framework of this short exploratory article that is unfeasible, and since we are interested in the ratio between frequencies rather than in the (absolute or relative) frequency of individual words, the omission seems tolerable. But serious statistics is of course impossible on such a basis. Still, the difference between genres as regards the *wrath*/*anger* ratio is considerable and seems to confirm our distinction between *anger*-preferring and *wrath*-preferring genres. The ratio for fiction is much lower than in the first approach, but that may be due to the surprisingly small number of authors and texts. That small number in turn may be due to the fact that fiction, as defined by *NCBEL*, is only beginning in our period, but also to the acquisition policies of 19th-century American librarians. However cautiously we interpret our figures, the strong position of *wrath* in religious language is confirmed.

The importance of religion is also underlined when we take a closer look at the diaries. Behind the almost total balance of 1.04 there are individuals like George Fox (1624–91), the founder of the Society of Friends, whose ratio is 7.25:1 (29/4), and Samuel Pepys (1633–1703) who uses *anger* 29 times and *wrath* only once. Then there is Alexander Brodie of Brodie (1617–1680), who uses both words no less than 54 times. These few figures are enough to indicate that with other diarists the frequency of both words is quite low. In the case of diaries, individual interest and style are clearly more important than any conceivable conventions of the genre. It would be interesting to scan diaries not just for *anger* and *wrath*, but for a longer list of emotion words. Such an undertaking could provide a measure of the extent to which diarists are concerned with psychological problems. The religious associations of *wrath* can also be shown by a very simple, even crude, indicator: the frequency with which *God* collocates with *wrath* as opposed to *anger*. The term *collocation* is here used to contrast with *colligation*: *colligation* means the co-occurrence of two or more lexemes within a syntactic construction; *collocation* is defined only quantitatively: the co-occurrence of two or more lexemes within a given distance in a syntactically undefined co-text. Antconc enables the researcher to determine a "window size", i.e. the number of characters preceding and following the search string. Setting the window at 50 yields a co-text of between 11 and 14 words. The frequency of *God* in the co-text of *wrath* and *anger*, respectively, varies significantly from genre to genre, as Table 4 shows.

Table 4: Frequency of *God* in co-text of *wrath* and *anger* in various genres

Genre	N(*wrath*)	N(*God*)	N(*God*/ *wrath*)	N(*anger*)	N(*God*)	N(*God*/ *anger)*
Religion	6543	3221	0.49	2969	913	0.31
Diaries	140	56	0.40	133	20	0.15
Fiction	7	0	0.00	61	1	0.02
Drama	301	19	0.06	904	24	0.03

In religious texts both *wrath* and *anger* collocate strongly with *God*: every other token of *wrath* and every third token of *anger* has *God* in its co-text, and this does not count the many pronouns and metaphors doing duty for *God*, like *the wrath of the Lord*, *the wrath of the Lamb*, and many others. In the diaries the collocation of *wrath* with *God* is only slightly below that of the religious texts (40%),

[11] Anonymous works were treated as works by different authors.

while that of *anger* with *God* is considerably lower (only 15%). In fiction the association between *God* and either lexeme seems negligible; in drama it is also low, though collocation of *wrath* with *God* is twice that of *anger*. In both genres *anger* is much more frequent than *wrath*. Fiction, defined here as strictly a prose genre, uses *wrath* hardly at all.

The richest harvest is offered by drama – for various reasons. First of all, drama is a highly productive and popular genre during our period. Second, and hardly less important, an impressive amount of research has produced a broad consensus which provides non-specialists like the present writer with the necessary information on such matters as dates, authorship, and dramatic sub-genres (see, above all, Harbage, Schoenbaum & Wagonheim 1989). And finally, the genre of the period is remarkably well represented in the Google corpus. There are many editions which are admittedly out of copyright and out of date, but which are still useful for our present purposes.

Given the uncertainty of authorship and the prevalence of collaborative authorship at the time, it would be unwise to pay too much attention to the apparent or real preferences of individual authors. But one or two observations are in order even at this stage. With the exception of Heywood, all authors living into the 17th century use *anger* more than *wrath*; writers dying before 1600 prefer *wrath*. Heywood's exceptional behaviour is easily accounted for when we look at our figures in terms of dramatic (sub-)genre: it is entirely due to his dramatizations of classical mythology which Harbage assigns to an otherwise rare genre of "Classical Legend".[12] Without these plays Heywood's *wrath/anger* ratio would be 0.71 instead of 1.07.

Fig. 3 shows the frequency of *anger* and *wrath* in the four most important dramatic sub-genres. At the same time it shows the number of plays in the corpus selected for analysis and the number of plays listed in Harbage 1989. The two lines run largely in parallel, which suggests that the corpus is a fair sample. From 1580 to 1658, comedy always shows more *anger* than *wrath* – with the exception of 1581–90 which is due to a single play: Greene's *Scottish History of James IV* (1590) which, with 2 occurrences of *wrath* and 0 of *anger*, figures in Harbage 1964 as "History" but has been re-categorized as a "Romantic Comedy" in Harbage 1989. Without *James IV* the comedies of the 1580s would show a *wrath/anger* ratio of 0.5 instead of 1.5. Clearly, these figures are too low to tell us anything except that neither *wrath* nor *anger* were at all frequent in the comedies of the period. After 1600 only one genre (tragicomedy 1601–10) shows more *wrath* than *anger*, but again the figures are too low to have much significance (10 times *wrath* and 7 times *anger* in 5 plays).

[12] These are the five "Age" plays (*Golden* (pr. 1611), *Silver* (pr. 1612), *Brazen* (pr. 1613), and the two-part *Iron Age* (pr. 1632) as well as *Love's Mistress* (performed 1634, pr. 1636), which is based on Apuleius' *Golden Ass*.

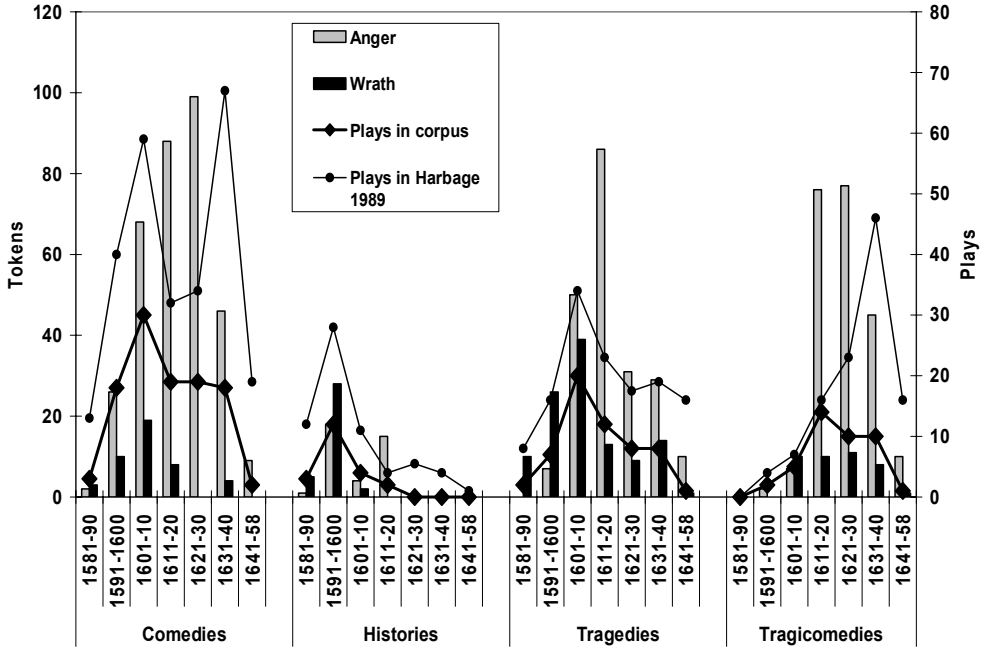

Figure 3: *Wrath* and *anger* in dramatic sub-genres 1581–1658 (acc. to Harbage 1989)

4. Conclusions

We have given abundant though perhaps unsurprising evidence that the frequency of a word is closely connected with the genre in which it is used. Looking at the Internet Archive through the spectacles of traditional bibliography we can do what the Google Ngram Viewer can't do: sort by genre and topic. But the really important question has hardly been tackled: what accounts for the affinity between lexeme and genre? The answer may well differ from word to word and from genre to genre. In the case of *wrath* vs. *anger* a first answer might be rank. But rank in which hierarchy? The answer is threefold: rank of the values associated with the word, rank of the topic which the genre (or text) is about, and rank of the text genre. The last hierarchy applies above all, to literary genres. Wierzbicka & Harkins (2001: 17) and Diller (2012; Forthcoming a) seem agreed that *wrath* differs from *anger* in its association with high rank. In religious writing the association of *wrath* with hierarchy is particularly pronounced: wrath is usually experienced by God; when experienced by humans it is one of the deadly sins, i.e. a revolt against God and closely related to pride. In diaries the association between *wrath* and matters religious is again close, even though the diary is, of course, not a religious text category. In drama the demise of *wrath* after 1600 in favour of *anger* coincides with the rise of tragicomedy and the comparative decline of tragedy. This fact agrees with the observation, only touched on in this paper, that *wrath* continues to be more frequent than *anger* in poetry (fn. 6). In spite of the rise of prose in narrative and also drama, poetry retains its high prestige. In the context of the Renascence it is worth noting that Spenser's epic, the *Faerie Queene*, uses *wrath* 84 times and *anger* only 12 times (Osgood 1915: 987, 27), whereas Sidney's prose romance, the *Arcadia*, shows 45 instances of *anger* against only 8 cases of *wrath* (Sidney 2003; Diller Forthcoming b). The distinction between *anger* and *wrath* had been occasionally discussed since the late Middle Ages (Pecock 1924: 110; Hilton 2000: I 70; Diller Forthcoming a). Hilton even goes so far as to deny that "fleischly angers" are deadly sins: they are merely obstacles to the contemplative life. The distinction lives on as peculiarly Christian and is defended against the anti-emotionalism of the Stoics (Bright 1586; Downame 1609; cf. Diller Forthcoming b).

This quantitative analysis calls for a qualification of Wierzbicka & Harkins' claim: *wrath* is "Shakespearean" only in the sense that Shakespeare uses it more frequently than we do, but except for his history plays (14 : 21), he prefers *anger* to *wrath* (63 : 57). As the comparison between the *Faerie Queene* and the *Arcadia* has shown, the rise of *anger* is earlier than most of Shakespeare's plays. In the long run it may be due to the process of "democratisation", but in the shorter run it shows the progress of civilization and individualization.

References

Anthony, Lawrence. 2011. AntConc (Version antconc3.2.4w.exe) [Computer Software]. Tokyo, Japan: Waseda University. Available from <http://www.antlab.sci.waseda.ac.jp/>.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

BibleGateway. A searchable online Bible in over 100 versions and 50 languages. Available at <http://www.biblegateway.com/> (Last visited 24/2/12).

Bright, Timothy. 1586. *A Treatise of Melancholy: Containing the Causes thereof*. London: Vautrollier. *Google Book Search*, accessed via link "1500–1641 – anger" in Fig. 1 (Last visited 7/9/12).

British National Corpus (BNC). Available at <http://corpus.byu.edu/bnc/> (Last visited 22/2/12).

Chadwyck-Healey's English Poetry Full-Text Database: http://cdroms.digibib.net. Available (on subscription) via http://www.ub.rub.de/DigiBib/Datenbank/Anglistik-db.htm (last visited 1/9/12).

De Smet, Hendrik. 2005. A corpus of Late Modern English texts. *ICAME Journal* 29: 69–82.

Diller, Hans-Jürgen. 1994. Emotions in the English lexicon: A historical study of a lexical field. In F. Fernández, M. Fuster & J.J. Calvo (eds.), *English Historical Linguistics 1992*, 219–234. Amsterdam/Philadelphia: John Benjamins.

Diller, Hans-Jürgen. 2012. *ANGER* and *TĒNE* in Middle English. In Manfred Markus, Yoko Iyeiri, Reinhard Heuberger & Emil Chamson (eds.), *Middle and Modern English Corpus Linguistics*. Amsterdam: John Benjamins, 109–124.

Diller, Hans-Jürgen: Forthcoming a. *Ssoong* on Ifaluk, *ANGER* and *WRATH* in Middle English: Historical Semantics as bridge-builder. In Laura Wright & Richard Dance (eds.), *Fourteen Essays on Middle English*. Frankfurt: Peter Lang.

Diller Hans-Jürgen. Forthcoming b. *Wrath* and *Anger* in the time of Shakespeare. In Agnieszka Pokojska & Agnieszka Romanowska (eds.), *Eyes to Wonder, Tongues to Praise. Volume in Honour of Professor Marta Gibińska*. Cracow: Jagiellonian University Press.

Diller, Hans-Jürgen, Hendrik De Smet & Jukka Tyrkkö. 2010. A European Database of Descriptors of English Electronic Texts. *The European English Messenger* 19,2: 29–35.

Downame [also Downham], John. 1609. *A Treatise of Anger*. London: Welby. Available at Early English Books Online <http://eebo.chadwyck.com/home> (Last visited 24/2/12).

Gevaert, Caroline. 2007. *The history of ANGER. The lexical field of ANGER from Old to Early Modern English*. Ph. D. diss. Katholieke Universiteit Leuven. Available at <https://repository.libis.kuleuven.be/dspace/bitstream/1979/893/2/thesisgedrukt.pdf> (Last visited 24/2/12).

Harbage, Alfred. 1989. *Annals of English Drama 975–1700. An Analytical Record of All Plays, Extant or Lost, Chronologically Arranged and Indexed by Authors, Titles, Dramatic companies, &c.* 3rd ed. rev. by Sylvia Stoler Wagonheim. London: Routledge. 2nd ed. rev. by S. Schoenbaum. London: Methuen, 1964. 1st ed. 1940.

Hilton, Walter. 2000. *The Scale of Perfection*, ed. by Thomas Bestul. Kalamazoo MI: Medieval Institute Publications. Available at <http://www.lib.rochester.edu/camelot/teams/bestul.htm> (Last visited 1/9/12).

Kiricsi, Ágnes. 2003–2004. From *Mod* to *Minde* – Report from a Semantic Battlefield. In Kathleen E. Dubs (ed.), *What does it Mean*, 217–239. Piliscsaba: Pázmány Péter Catholic University.

Kiricsi, Ágnes. 2005. *Semantic Rivalry of* Mod/Mood *and* Gemynd/Minde *in Old and Middle English Literature*. Ph. D. diss. Budapest Eötvös Loránd University of Sciences, Faculty of Humanities.

Krischke, Wolfgang. 2012. "Nach der Genomik nun die Kulturomik". *Frankfurter Allgemeine Zeitung* #201 (29 August 2012), p. N5.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011a. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331: 176–182 (Published online ahead of print: 12/16/2010). Available at: <http://www.sciencemag.org/content/331/6014/176.full.pdf> (last visited 3/6/12).

Michel, Jean-Baptiste*, et al.* 2011b. Supporting Online Material to previous article. Corrected 11 March 2011. Available at:

<http://www.sciencemag.org/content/suppl/2010/12/16/science.1199644.DC1/Michel.SOM.revision.2.pdf**>** **(**last visited 3/6/12**).**

Nunberg, Geoff. 2010. Humanities research with the Google Books corpus. Available at <http://languagelog.ldc.upenn.edu/nll/?p=2847> (December 16, 2010). (last visited 30/5/12).

Osgood, Charles Grosvenor. 1915. *A Concordance to the Poems of Edmund Spenser*. Washington, D.C.: The Carnegie Institution. Available at <http://www.archive.org/> (Last visited 6/9/12).

Pecock, Reginald. 1924. *The Folewer to the Donet*, ed. by Elsi Vaughan Hitchcock. EETS OS 164. London: Oxford University Press.

Sidney, Philip. 2003. *The Countess of Pembroke's Arcadia (1590)*. Transcribed by Risa Bear. Renascence Editions. Available at <https://scholarsbank.uoregon.edu/xmlui/bitstream/handle/1794/812/countess.pdf?sequence=1> (Last visited 6/9/12).

Spenser, Edmund. 1912. *The Poetical Works*, ed. and intr. by J. C. Smith and E. de Selincourt. London: Oxford University Press.

Watson, George (ed.). 1974, 1971. *New Cambridge Bibliography of English Literature*. Vol. 1 (600–1660): 1974, Vol. 2 (1660–1800): 1971. Cambridge: Cambridge University Press.

Wierzbicka, Anna, & Jean Harkins. 2001. Introduction. In Jean Harkins & Anna Wierzbicka (eds.), *Emotions in Crosslinguistic Perspective*. Berlin: Mouton de Gruyter, 1–34.

Selected Proceedings of the 2012 Symposium on New Approaches in English Historical Lexis (HEL-LEX 3)

edited by R. W. McConchie,
Teo Juvonen, Mark Kaunisto,
Minna Nevala, and Jukka Tyrkkö

Cascadilla Proceedings Project Somerville, MA 2013

ISBN 978-1-57473-455-3 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Diller, Hans-Jürgen. 2013. Culturomics and Genre: *Wrath* and *Anger* in the 17th Century. In *Selected Proceedings of the 2012 Symposium on New Approaches in English Historical Lexis (HEL-LEX 3)*, ed. R. W. McConchie et al., 54-65. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2836.