To appear in: **IJGLSA**, ***Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis*** (Berkeley), 18.1 (Spring 2013)

"Culturomics" and the representation of the language of the Third Reich in digitized books

Klaas Willems

Faculty of Arts and Philosophy
University of Ghent, Blandijnberg 2, B-9000 Ghent (Belgium)
Klaas.Willems@ugent.be

Abstract

This article reviews the major findings of a case study exploring the language of the Third Reich by means of the recently introduced computational tool *Google Books Ngram Viewer* (http://books.google.com/ngrams). This tool has been designed to investigate cultural trends and salient semiotic developments in history on the basis of the digital corpus of *Google books* on the World Wide Web. The aim of the article is to examine the reliability and overall usefulness of the new instrument for conducting fine-grained "culturomic" investigations on the basis of very large monolingual corpora.

1. Introduction

In their article "Quantitative analysis of culture using millions of digitized books", J.-B. Michel et al. (2011) make an interesting case for a new subdiscipline of cultural-semiotic studies called "culturomics". The term refers to the investigation of cultural trends and salient semiotic developments in the history of mankind which are examined on the basis of the digitized books provided by *Google books* on the World Wide Web: "Culturomics is the application of high-throughput data collection and

analysis to the study of human culture" (Michel et al. 2011, p. 181). To date, the *Google books* corpus is the largest machine-readable corpus of printed data available to the scientific community. To demonstrate the potential resources this corpus offers for culturomic studies, the authors present some dazzling numbers. Meanwhile 15 million books have been digitized, which is about four percent of all books ever printed since the invention of the printing press in the 15th century. From this material, Google constructed a corpus of over five million books in 2009, the equivalent of 500 billion words, which can be accessed through *Google Books Ngram Viewer* (http://books.google.com/ngrams).

The present article reviews the findings of a focused case study performed by means of *Ngram Viewer*. Given the ongoing debate about the reliability of the language data retrieved from the World Wide Web for frequency-based linguistic analysis (see, e.g., Keller and Lapata 2003 and Kilgarriff 2007), the aim of the article is to examine the reliability and usefulness of the new instrument for conducting sufficiently fine-grained culturomic investigations on the basis of extremely large monolingual corpora. The case study reported on draws on the currently available German corpus which contains 37 billion words (see Michel et al. 2011, p. 176). The object of the case study are the first attestations and subsequent variations in usage frequency of 50 randomly chosen German expressions that are commonly regarded as typical of the language of the Third Reich.

2. The language of the Third Reich

2.1. The language of Nazi Germany lends itself particularly well to an evaluation of computational tools such as *Ngram Viewer* (cf. Willems 2012). According to Schmitz-Berning (2000, p. vii), the Nazi period can be divided into a first part from 1918 to 1933, in which the National Socialists rose to power (*Kampfzeit* 'battle time, time of struggle'), and a second part which lasted from 1933 to 1945 (*das Dritte Reich* 'the Third Reich'). However, for the sake of convenience, I will use the name "Third Reich" to refer to the entire period in this article.

It is broadly accepted that only a small number of the non-technical expressions of "Nazi-German" were actually coined during the Nazi period (1918–1945) (see Klemperer 2000, Sternberger, Storz and Süskind 1968, Schmitz-Berning 2000, Michael and Doerr 2002). Unlike technical jargon such as *Blutschutzgesetz* '(Nuremberg) Blood Protection Act',[1] *K-Schein* (*Kriegsausbildungsschein*) 'wartime training certificate (…) issued upon completion of a Hitler Youth wartime training program', *Reichskulturkammer* 'Reich Chamber of Culture' etc. and terminology such as *Atlantikwall* 'Atlantic wall', *Hitlerjugend* 'Hitler Youth', *NSDAP* (*Nationalsozialistische Deutsche Arbeiterpartei*, the official name of the Nazi Party), most Nazi-German expressions – mainly word formations – already existed in the German lexicon prior to 1918 but they became much more frequent with the advent of the National-Socialist state.

[1] Throughout this article, the English translations provided alongside the German expressions are taken from Michael and Doerr, *Nazi-Deutsch/Nazi German* (2002). Michael and Doerr's lexicon contains some 6,000 entries.

The 50 expressions that for the purpose of the present study were entered into *Ngram Viewer* are nouns, verbs, adjectives, and adverbs, all of which were extracted from Michael and Doerr (2002), together with their translations. (Note that the vast majority of expressions assembled in Michael and Doerr's lexicon are nouns). The case study largely confirms the findings of earlier studies of the language of the Third Reich, with however a few notable exceptions which call for an explanation.

2.2. The largest group of expressions – 40 in number, presented in alphabetical order below – are all attested in German books published prior to 1918, but they show a significant rise in frequency in the ensuing decades. This group includes the following expressions:

(1) *arisch* ('Aryan'), *artverwandt* ('racially related'), *aufnorden* ('to Nordicize'), *ausrotten* ('to tear out root and branch, to eradicate'), *Bestleistung* ('best performance'), *blutlich* ('blood related'), *Dienststelle* ('government department/office'), *dritte*(*s*) *Reich* ('Third Reich'), *Ehrengericht* ('Honor Court'), *Einbruch* ('break through'), *Entvolkung* ('degermanization'), *erbkrank* ('hereditary ill'), (*der*) *Führer* ('the Leader'), *Gau* ('district, province'), *Gefolgschaft* ('entourage, followers (loyal to Hitler)'), *gemeinschaftsunfähig* ('community unsuitable'), *Generalgouvernement* ('General Government (in eastern Poland)'), *gigantisch* ('enormous'), *Großdeutschland* ('Greater Germany'), *Herzland* ('heartland'), *judenfrei* ('free of Jews'), *Kadavergehorsam*

(‘corpse-like obedience’), *Kameraden* (‘comrades’), *Kulturboden* (‘cultural soil’), *Landjahr* (‘year in the country (on a farm)’), *Meintat* (‘archaic for crime’), *Menschentum* (‘humanity, German race’), *Musterbetrieb* (‘model company’), *organisch* (‘organic’), *planmäßig* (‘according to plan’), *rassisch* (‘racial’), *raumfremd* (‘alien to an area’), *Reichsbürger* (‘Reich citizen’), *Sippe* (‘kinship, family, clan’), *Strafexpedition* (‘punishment expedition’), *verpolt* (‘having become Polish’), *völkisch* (‘ethnic, racial, national’), (*gesundes*) *Volksempfinden* (‘[healthy] national feeling’), *Volkskörper* (‘people’s body’), *Zusammenballung* (‘crowding together’).

After reaching a peak between 1918 and 1945, the usage frequency of these expressions again decreases towards the pre-1918 level by the end of World War II or in the immediately following years. This is illustrated in Fig. 1 for the expression *Kulturboden* ‘cultural soil’.

Fig. 1. *Kulturboden* (‘cultural soil’) in German books from 1880 to 1980 (frequencies 1890: 2.0e-5, 1941: 1.1e-4, 1980: 1.5e-5)

Like all ensuing graphs in this article, the graph in Fig. 1 is displayed without statistical smoothing, which means that no averaging between subsequent years has been applied and only "raw data" is presented. The x-axis represents the years of publication (e.g., 1880–1980 in Fig. 1), the y-axis the frequency of the expression, that is, the percentage of the German word *Kulturboden* among the total set of unigrams (one word lexemes) in the 37 billion words corpus of digitized German books.[2]

The list in (1) is revealing. Even expressions whose negative connotation is now regarded very strong because of their idiomatically charged connection with the Third Reich can already be found in publications that appeared prior to the National-Socialist era. There are, for instance, two early records of the query term *gemeinschaftsunfähig* ('community unsuitable') in the corpus, one from 1894, the other from 1908, whereas a similar pseudo-formal word formation such as *blutbedingt* ('conditioned by blood') (Michael and Doerr 2002, p. 102) appears to be of more recent origin, the earliest attestations dating from the 1920s. On the other hand, *blutlich* ('blood related') is also among the expressions that can be traced back to the first half of the 19th century. Note that none of these adjectives are listed in a

[2] The early decades of the 16th century are represented by only a few books per year, but by 1800 the corpus grows to almost 100 million words per year and by 2000 this number increases to 11 billion words per year (Michel et al. 2011, p. 176). *Ngram Viewer* is limited to clusters of five lexemes, i.e., 5-grams (see Michel et al. 2011, p. 176). For the purpose of the present article, the query terms were restricted to unigrams. Note that a unigram is considered "common" if its frequency is greater than one per billion, i.e., 1.0e-9 (Michel et al. 2011, p. 176).

major dictionary of current German such as Duden's 10-volume edition *Das große Wörterbuch der deutschen Sprache* (Duden 1999).

Like the adjective *blutlich*, the noun *Meintat* ('archaic for crime') is among the German expressions that had long fallen in disuse but regained currency during the Third Reich, only to be dropped again in the first years after its collapse. This historical development accounts for the graph in Fig. 2:

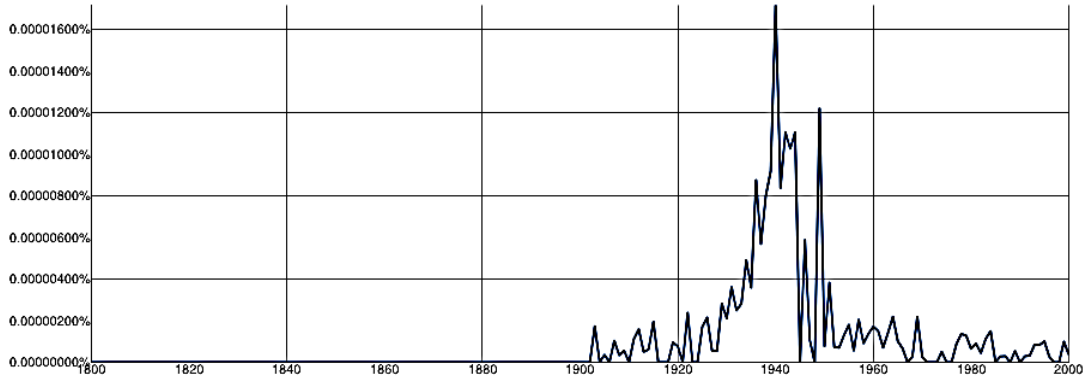

Fig. 2. *Meintat* ('archaic for crime) in German books from 1800 to 2000 (frequencies 1907: 1.0e-6, 1940: 1.6e-5, 1952: 1.0e-6)

The graph in Fig. 2 is strikingly different from the one in Fig. 3 which shows the usage frequency of the proper name *Einstein* in the German corpus between 1900 and 1970. As pointed out by Michel et al. (2011), *Ngram Viewer* is a particularly useful tool to detect (e.g. Nazi) censorship: "Suppression of a person or an idea leaves quantifiable fingerprints" (Michel et al. 2011, p. 181; see also Bohannon 2011).

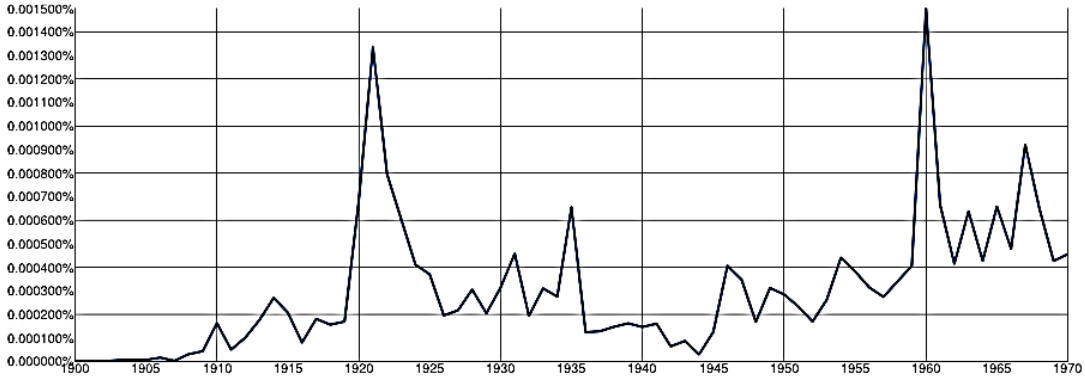

Fig. 3. *Einstein* in German books from 1900 to 1970
(frequencies 1921: 1.3e-3, 1942: 1.0e-4, 1946: 4.0e-4, 1960: 1.5e-3)

The French-based loanword *fanatisch* ('fanatical') (Fig. 4) shows an evolution in the German corpus that is very similar to Fig. 1, with a steady increase in frequency since the end of World War I and an equally steady decline in frequency between 1946 and 1955. The culturomic significance of this finding can be measured when it is compared to the entirely different evolution of the original adjective *fanatique* in the corpus of French books since the 19th century. In the digitized French corpus (which contains 45 billion words), the use of the expression *fanatique* has been falling steadily since the mid-19th century (Fig. 5).

Fig. 4. *fanatisch* (‘fanatical’) in German books from 1880 to 1980
(frequencies 1890: 4.0e-5, 1945: 1.6e-4, 1980: 8.5e-5)

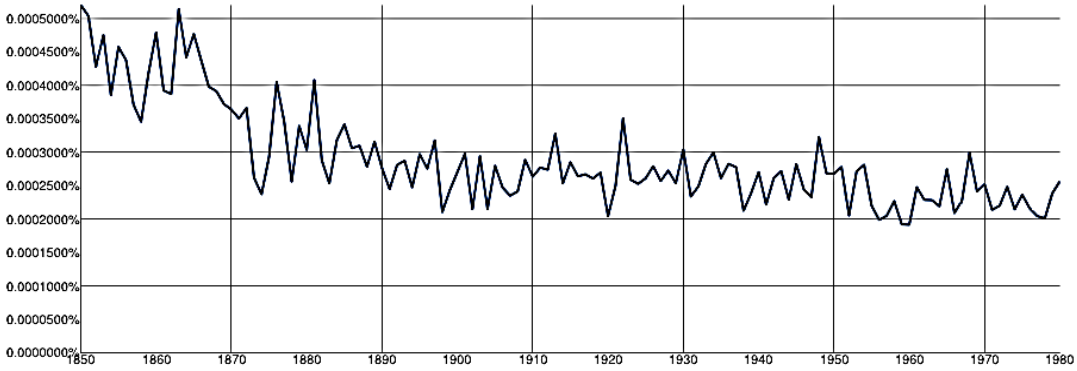

Fig. 5. *fanatique* (‘fanatical’) in French books from 1850 to 1980
(frequencies 1850: 5.0e-4, 1945: 2.8e-4, 1980: 2.5e-4)

Not all expressions listed in (1) have retained a National-Socialist connotation in modern German. This applies, for instance, to currently “neutral” words such as *Bestleistung* (‘best performance’), *Einbruch* (‘break through’), *gigantisch* (‘enormous’), and *planmäßig* (‘according to plan’). However, the frequency of these words, too, has been falling steadily since 1945 or thereabout.

2.3. Other expressions insofar deviate from the pattern in Fig. 1 that their rise in frequency continues after the Third Reich without major interruptions. Our case study of 50 items unearthed 4 such expressions, which are listed in (2):

(2) *brutal* ('brutal, cruel'), *Eintopf* ('one-pot meal'), *Großoffensive* ('great offensive'), *schlagartig* ('all of a sudden').

Fig. 6 displays the graph for *Großoffensive*, of which Michael and Doerr (2002, p. 197) write: "Near the end of World War II, Goebbels' term meant to inspire hope for a successful German counterattack."

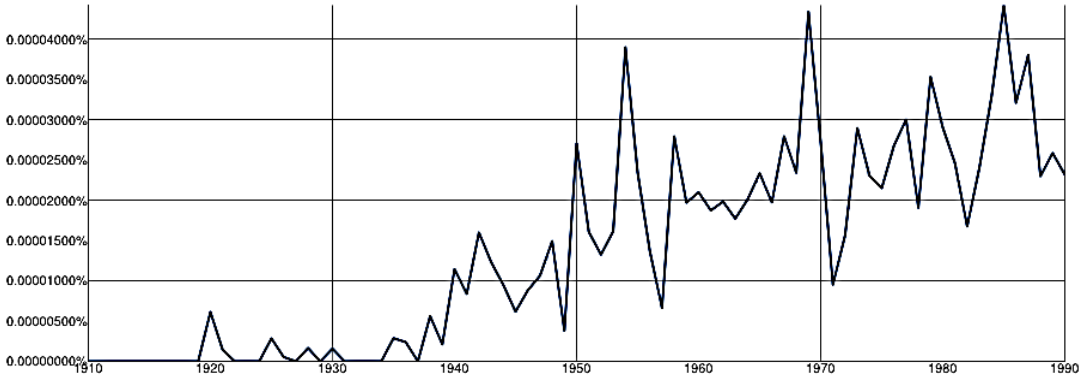

Fig. 6. *Großoffensive* ('great offensive') in German books from 1910 to 1990 (frequencies 1920: 5.0e-6, 1944: 1.0e-5, 1977: 3.0e-5)

The graph points to the conclusion that, contrary to the words listed in (1), *Großoffensive* is among the expressions that have remained common in the German language after World War II, losing its National-Socialist connotation in the language at an early stage. Note that the pattern in Fig. 6 could in principle be caused

by the frequent use of *Großoffensive* in books dealing with the Third Reich from 1945 onwards. In order to see whether this conclusion is warranted, it is necessary to have a closer look at the actual usage of the query term in the publications of the post-war period. *Ngram Viewer* provides this possibility by giving below each graph year ranges that lead the user directly to the query term in the published books in the digital corpus. With respect to *Großoffensive*, such an additional analysis reveals that the continued increase in the expression's usage frequency is not primarily due to metalinguistic usage in scholarly volumes. The expression became a neutral designation for any major offensive in modern German. However, only a qualitative corpus analysis can determine when exactly the change of connotation took place.

The graph of an expression such as *brutal* ('brutal, cruel') shows a similar development (Fig. 7). When exactly *brutal* lost the clearly positive connotation it had for the National Socialists (see Michael and Doerr 2002, p. 108: "brutal, cruel. Fanatical and ruthless, decided without compromise, positive connotation for Nazis"), remains to be established.

Fig. 7. *brutal* ('brutal') in German books from 1890 to 1990
(frequencies 1890: 7.5e-5, 1940: 3.0e-4, 1990: 3.7e-4)

2.4. Still different is the development of an expression such as *Volkskanzler* 'People's Chancellor'. The graph in Fig. 8 shows how this newly created epithet, defining Hitler as "the guide of the German people" (Michael and Doerr 2002, p. 424), was enormously successful between 1931 and 1936.

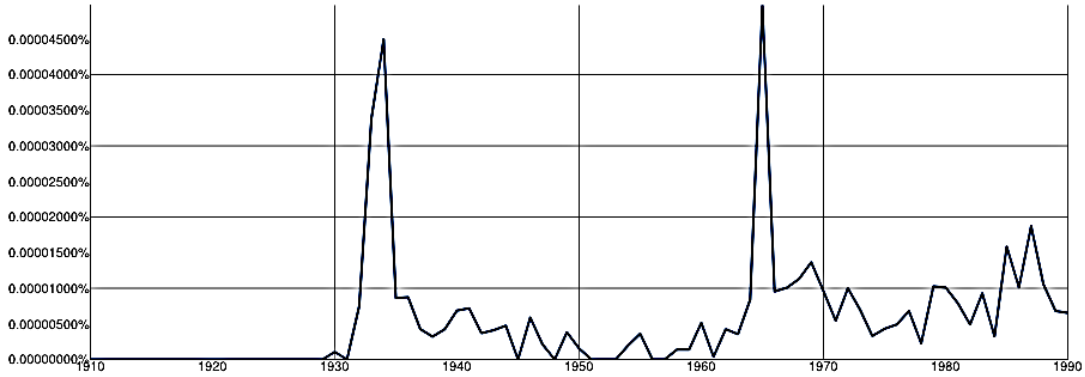

Fig. 8. *Volkskanzler* ('People's Chancellor') in German books from 1910 to 1990 (frequencies 1934: 4.0e-5, 1947: 2.5e-7, 1965: 5.0e-5)

Remarkably, after being almost away from the screen for more than twenty years, the expression again became very popular in the mid-1960s. This is not due to a suddenly increased scholarly interest in the Third Reich but because German politician Ludwig Erhard (1897–1977), who was elected Chancellor of the Federal Republic of Germany in 1963, was also widely called a *Volkskanzler* in Germany – apparently without any reminiscence of the connotation the expression had three decades earlier.

2.5. *Ngram Viewer* succeeds in dispelling the belief that the expression *charakterlich* (‘according to character’) is an “adjective coined by the Nazis meaning the German character” (Michael and Doerr 2002, p. 112).[3] The word can be traced back to the 1890s, e.g., in publications dealing with psychology and literary studies (Fig. 9).

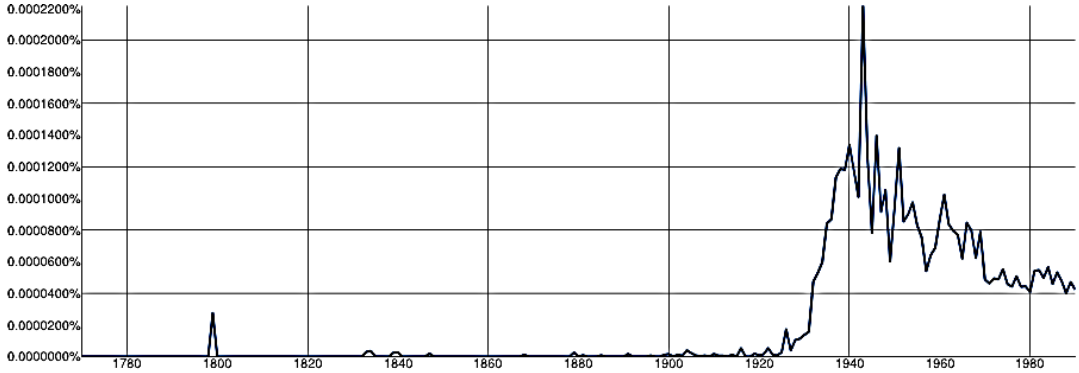

Fig. 9. *charakterlich* (‘according to character’) in German books from 1770 to 1990 (frequencies 1926: 2.0e-5, 1943: 2.2e-4, 1990: 4.0e-5)

However, the indication in Fig. 9 that the earliest attestations of *charakterlich* date back to the end of the 18th century does not prove reliable. It appears that several expressions which are formally similar to *charakterlich* – for instance *charakterisch* (now obsolete in the German language) and the currently common neutral word *charakteristisch* (‘characteristic’) – are occasionally erroneously scanned and digitized as *charakterlich* in 18th and 19th German sources. This is proof that the “Optical Character Recognition” program (‘OCR’, Michel et al. 2011, p. 176) used by *Ngram Viewer* is not flawless when applied to older German printed texts, giving

[3] One finds the assumption that *charakterlich* was a new coinage of the National Socialists also expressed in Klemperer (2000, p. 180), Sternberger, Storz and Süskind (1968, 13 and 37–44), Schmitz-Berning (2000, p. 132–133), among others.

rise to lemmatization errors which may distort the quantitative results. The fact that many older German books are printed in Gothic script (*Fraktur*) may well be responsible for most of these errors. For example, the German translation of Latin *denotare* in Friedrich Karl Kraft's *Deutsch-Lateinisches Lexikon* (1843, p. 445) reads *deutlich erklären*. The first word of the translation is however lemmatized as *blutlich* by *Ngram Viewer*, mistaking the initial cluster [de-] for [bl-] and thus wrongly providing for one of the few early attestations of the expression *blutlich* in German books.[4] The extent to which such lemmatization errors actually occur and influence the quantitative results is unclear, but it is arguably a matter of considerable importance if *Ngram Viewer* is to assume the significance it is designed to have for the culturomic investigation of older documents.

3. Conclusion

The majority of the German expressions analyzed in this study, viz. 80 percent of the sample, conforms to the type expected on the basis of the literature that deals with the language of the Third Reich. These expressions were already present in the German lexicon but started to rise in frequency around 1920 and peaked in the 1930s or early 1940s. After World War II, their frequency for the most part dropped drastically, yet they continued to be used in post-war publications, either with or

[4] Although Michel et al. (2011, p. 176) point out that the dates and places of publication were provided by the libraries and publishers, the publication years may be prone to similar errors. For instance, Carl Caesar Leonhard's *Taschenbuch für die gesammte Mineralogie* published in *1813* (but not printed in Gothic script) is presented as a book published in *1815*.

without the National-Socialist connotation (see (1)). The number of expressions whose frequency continued to rise without major interruptions after 1945 turns out to be significantly smaller (see (2)).

The effect of occasional lemmatization errors such as those pointed out above on the outcome of the findings has yet to be determined, in particular regarding pre-20th century German publications which are often in Gothic script. Nevertheless, it is safe to say that *Google Books Ngram Viewer* constitutes a valuable computational tool for cultural and semiotic analyses which can be used to good effect in the representation of quantitative findings based on large corpora of publications. It bears pointing out, however, that *Ngram Viewer* is merely an instrument to retrieve potentially useful data from large datasets, not an end in itself, and careful interpretation of the findings remains essential. Once periodicals, newspapers, etc. which for the time being are excluded from the corpus (Michel et al. 2011, p. 176, 181), will also be systematically covered, the effectiveness of the new tool will undoubtedly further increase.

References

Bohannon, John. 2011. “Google Books, Wikipedia, and the future of culturomics”. *Science* 331, 14 January, p. 135.

Duden. 1999. *Das große Wörterbuch der deutschen Sprache*. 10 vol.s. Mannheim, etc.: Dudenverlag.

Keller, Frank and Mirella Lapata. 2003. “Using the Web to obtain frequencies for unseen bigrams”. *Computational Linguistics* 29: 3, pp. 459–484.

Kilgarriff, Adam. 2007. “Googleology is bad science”. *Computational Linguistics* 33: 1, pp. 147–151.

Klemperer, Victor. 2000. *The Language of the Third Reich*. London & New York: Continuum. (German original: Klemperer, Victor. 1947. *LTI. Notizbuch eines Philologen*. Berlin: Aufbau Verlag.)

Michael, Robert and Karin Doerr. 2002. *Nazi-Deutsch/Nazi German: An English Lexicon of the Language of the Third Reich.* Newport, CT: Greenwood.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden et al. 2011. “Quantitative analysis of culture using millions of digitized books”. *Science* 331, 14 January, pp. 176–182.

Schmitz-Berning, Cornelia. 2000. *Vokabular des Nationalsozialismus*. Berlin & New York: W. de Gruyter.

Sternberger, Dolf, Gerhard Storz and Wilhelm E. Süskind. 1968. *Aus dem Wörterbuch des Unmenschen. Neue erweiterte Ausgabe mit Zeugnissen des Streites über die Sprachkritik* (3rd edition). Hamburg & Düsseldorf: Claassen (1st edition 1957).

Willems, Klaas. 2012. “‘Google Books Ngram Viewer’ und historische Computerlexikologie: Der Sprachgebrauch der NS-Zeit”. *Muttersprache* 122: 2, pp. 81–101.