Computers and the Humanities **35:** 237–253, 2001.

Computing Historical Consciousness. A Quantitative Inquiry into the Presence of the Past in Newspaper Texts

THIJS POLLMANN
Utrecht Institute of Linguistics OTS, Drift 8, 3512 BS Utrecht, The Netherlands
E-mail: Thijs.Pollmann@let.uu.nl

R. HARALD BAAYEN
Interfaculty Research Unit for Language and Speech (IWTS), University of Nijmegen & Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
E-mail: baayen@mpi.nl

Abstract. In this paper, some electronically gathered data are presented and analyzed about the presence of the past in newspaper texts. In ten large text corpora of six different languages, all dates in the form of years between 1930 and 1990 were counted. For six of these corpora this was done for all the years between 1200 and 1993. Depicting these frequencies on the timeline, we find an underlying regularly declining curve, deviations at regular places and culturally determined peaks at irregular points. These three phenomena are analyzed.

Mathematically spoken, all the underlying curves have the same form. Whether a newspaper gives much or little attention to the past, the distribution of this attention over time turns out to be inversely proportional to the distance between past and present. It is shown that this distribution is largely independent of the total number of years in a corpus, the culture in which it is published, the language and the date of origin of the corpus. The phenomenon is explained as a kind of forgetting: the larger the distance between past and present, the more difficult it is to connect something of the past to an item in the present day. A more detailed analysis of the data shows a breakpoint in the frequency vs. distance from the publication date of the texts. References to events older than approximately 50 years are the result of a forgetting process that is distinctively different from the forgetting speed of more recent events.

Pandel's classification of the dimensions of historical consciousness is used to answer the question how these investigations elucidate the historical consciousness of the cultures in which the newspapers are written and read.

1. Introduction

The aim of this paper is to investigate some aspects of the use of historical knowledge with electronic means. To be more precise, we want to present and analyze some quantitative data about the presence of the past in newspaper texts. The data consist of frequencies of dates in the form of years. The languages of the text corpora, their ages and the cultures in which they had a function, are independent variables.

Years are used as pegs for many well-known historical events: 1066, 1517, 1813, 1917 etc., but also as points on an imaginary time-line, to measure distances in time. In this paper years are to be taken in this second sense: a means of giving some structure to the past and of supporting our understanding of 'earlier' and 'later' in history.

Of course, years are not historical knowledge by themselves. But we take it that years in normal language use are signs that a writer is referring to something in the past. The data are supposed to give some insight into the frequency with which a literate public of (mostly) non-historians is asked to pay attention to aspects of human history, and how this attention is distributed over the past.

Years in texts are a kind of words. This quality allows for some investigations which relate to the field of word frequency studies. An accepted methodological principle in this field says that the frequency of aspects of linguistic usage can be used as an indicator of the underlying mental organization that makes this usage possible. Years turn out to be rather suited as a means to study some fundamental aspects of processing knowledge of the past. This is due to a series of properties. Years are discrete entities, they are easy to detect, they form a clearly defined linguistic category (although few grammars describe their peculiarities) and they are mostly unambiguous (although they may refer to an endless range of events, etc. in the past). To most years, one cannot refer by other simple linguistic expressions. They are frequent in normal written language use, are easy to process statistically, and – what is more – to process statistically with 'time' as independent variable. Years form a time-series. We shall make use of these properties in the analysis of the distribution of year frequencies over the time-line. By doing this we obtain a quantified picture of the decay of attention to what once happened and is now receding into the past, a picture of the passage of time.

This paper is structured as follows. In Section 2 we shall explain how the empirical data have been collected. In Section 3, there will be a first analysis of these data. Section 4 brings a more sophisticated statistical analysis of the data, which will sharpen some of the characteristics that were signaled already in previous sections. In Section 5 we shall try to answer the question whether these analyses might be said to reflect historical consciousness, using a classification of aspects of historical consciousness by Pandel.

2. The Data

In the framework of the research on which we report here, the first author collected all occurrences of years in a set of four large corpora of written language. To make comparisons between the sources possible, the collections have been kept apart. All four corpora consist of newspaper texts. The material dates from 1994 (plus in one case the first four months of 1995). Not all electronically readable text corpora recognise numbers as words. In making concordances of years, we eventually made use of Microconcord.

The first collection consists of the years that emanated from the CD-rom edition of the German daily *Frankfurter Allgemeine Zeitung*. The total number of words in this corpus can be estimated as 26 million. The size of the FAZ-corpus was computed by a count of the total number of articles multiplied by the mean-length of a representative part of the articles. We refer to this collection as **FAZ**.

The second corpus we used, is the so-called 27mln corpus of the *Instituut voor Nederlandse Lexicologie* (Institute for Dutch lexicology). It consists of 27 million word forms originating from the editorial columns of the 1994 editions and those of the first four months of 1995 of *NRC/Handelsblad*, a Dutch quality newspaper.[1] The word forms in this corpus have been linguistically coded to be used for all kinds of linguistic research. We refer to this year-collection as **NRC**. The third corpus is the 1994 edition on CD-rom of the *International Herald Tribune*, an American newspaper for an international readership. The size of the corpus can be estimated as to 18 million words. The collection of years is henceforth called **IHT**. The fourth collection originated from the 1994 CD-rom edition of *de Volkskrant*, a leading Dutch daily with a national distribution. The CD-rom edition contains 19 million words. This size we computed by means of the given frequencies of some high-frequency words of functional categories (prepositions, adverbs etc.) in the 27mln corpus and the given size of this corpus. In estimating the size of the IHT-corpus we used in a comparable way the frequency-lists of American English in Kuçera and Francis (1965). This collection is referred to as to **VK**.[2]

From these corpora we collected the years which refer to the past between 1993 and 1200 and those of the future between 1996 and 2100.

Below we will refer to three other collections of years of the period between 1990 and 1930. Their sources were the CD-rom editions of *The (London) Times and Sunday Times (1994)*, of the French newspaper *Le Monde (1994)* and of the Spanish *El Mundo (first semester of 1994)*. The special ways in which these corpora has been filed made it impossible to find an easy way to collect all the years between 1200 and 2100.

Collecting the years from the corpora involved two steps. In the first stage, we collected electronically all numbers in the ranges we mentioned. Subsequently, we sifted from those sets the years on the basis of the sense of the sentences in which the numbers happened to occur. In this way, the years were separated from numbers which refer to minimum wages, numbers of employees in a firm, points in a sports competition etc. For numbers with a relatively high frequency (>50), this sifting was done by extrapolation from a representative sample. Most of the time, texts with listed numerical information like sports results, weather reports etc., were lacking in the corpora we used; sometimes, however, one finds complete surveys of election results. Happily, at no point one comes across Stock Exchange reports.

Of course, years do occur that refer to dates before 1200, but their number is small. This is why they were kept outside the collections. Abbreviations of years, like ‘18’ in ‘1914–1918’ or ‘68’ were also excluded. It was not easy to locate these

Table I. Numbers of years and 'year-densities' in four corpora of newspaper texts and two word frequency-lists

	1	2	3	4	5	6
	Size of the corpus	Total number of years	Past	Future	3 in % of 2	Year-density 2:1
FAZ	26,200,000	80571	72259	8312	89.7	1:325
NRC	27,000,000	69440	63466	5974	91.4	1:388
IHT	19,000,000	44764	40854	3910	91.2	1:424
VK	18,000,000	26833	22656	4177	84.4	1:670
WFEng60	1,000,000	1572	1492	80	94.9	1:645
WFDut69	720,000	709	650	59	91.7	1:1015

forms systematically. The same applies for expressions like 'the 1860s' which are quite frequent in English texts. These too were kept outside the collections.

Additionally, the number of years in two word frequency lists were counted. The first, henceforth **WFEng60**, is based on English texts which were published in the United States in 1960 (Kuçera and Francis, 1965). This corpus has a size of roughly 1 million words. The other, which we will call **WFDut69**, has as its base a collection of Dutch oral and written language dating from 1969–1970 (Uit den Boogaart, 1975). This collection consists of 720,000 words. Because of the fact that only frequency data were available, and not the texts, so that we could not separate the years from other numbers, these two collections have to be used with some circumspection.

The numbers of years in the different corpora can be presented schematically as follows. Cf. Table I.

From this table, it will be immediately clear that the collections are rather large. **FAZ**, with its 80,000 items, is by far the largest. This might be a consequence of the size of the corpus (26,200,000 words), but it cannot be totally explained by this fact. Of the newspaper collections, **VK** has the smallest number of years. This is also partly a consequence of the size of the corpus from which the data stem, but there must be other factors involved.

Accepting the estimates of the sizes of the corpora presented above, one can compute the 'year-densities' of the different corpora, which we define as the number of years divided by the number of words in the corpus (column 6 in Table I). The year-density of the *Frankfurter Allgemeine Zeitung* is the largest of all the newspaper corpora: one in each 325 word forms is a year. The vast majority refers to the past. The year density of *NRC/Handelsblad* turns out to be larger than that of the *International Herald Tribune*, but smaller than that of the *Frankfurter Allgemeine Zeitung*. Clearly, *de Volkskrant* has the smallest year density of all the newspaper corpora. It seems probable that the differences in year density are

linked to the general character of the newspaper. We will turn to this in Section 4 below.[3]

Table I, column 5, gives the proportion of the numbers of years in each of the collections that refer to the past. In all collections this turns out to be about 90%, with an exception for *de Volkskrant* (84.4%) and the corpus of texts on which the frequency-list of the American English is based (94.9%). According to these figures, newspapers are rather homogenous in the distribution of their attention to the past and the future. We shall not try to explain this.

3. A Further Analysis

In this section, we will analyze the data somewhat further. Especially the frequencies with which these data in the corpora occur will be of interest.

Of course, we would not expect all years to occur equally frequently in the collections. We will expect 1492 to occur more often than 1491, and 1945 to be more frequent than 1946 etc. More generally, we will expect frequencies to diminish as the distance to the present-day increases. It is a rather obvious, but striking feature of our historical consciousness that we pay less attention to parts of the past, as these parts become farther off. Pandel (1991) reports on an inquiry, in which students who were asked to mention historical events, most often referred to historical events of the twentieth century, and less often to events in the Middle Ages. Sometimes one finds the intuition that a poor 'historical consciousness' or hodiecentrism, as it is sometimes called, will most of all neglect the past farther off (cf. Van Berkel, 1985). According to this, we might have to expect that a corpus that on average contains few years would pay proportionally less attention to the distant past than a corpus that reveals more attention to the past in general.

These conjectures have to do with the distribution of the year frequencies over the time-line. In the analysis, we shall concentrate on these distributional phenomena.

To make a first acquaintance with the kinds of phenomena the data will confront us with, we present the data from **FAZ**, **NRC** and **IHT** concerning the years 1990–1930. Cf. Figure 1.

The graph exhibits three notched, but otherwise regularly declining curves. There are small peaks for the years 1980, 1970, 1960 etc., and some peaks elsewhere, among which the striking one at 1945 and 1944. These are the three things we always find when we plot the distribution of the frequencies over the time line: there is an underlying regularly declining curve, we find deviations at regular places, and we have peaks at irregular points.

We shall discuss these features in the three subsections below.

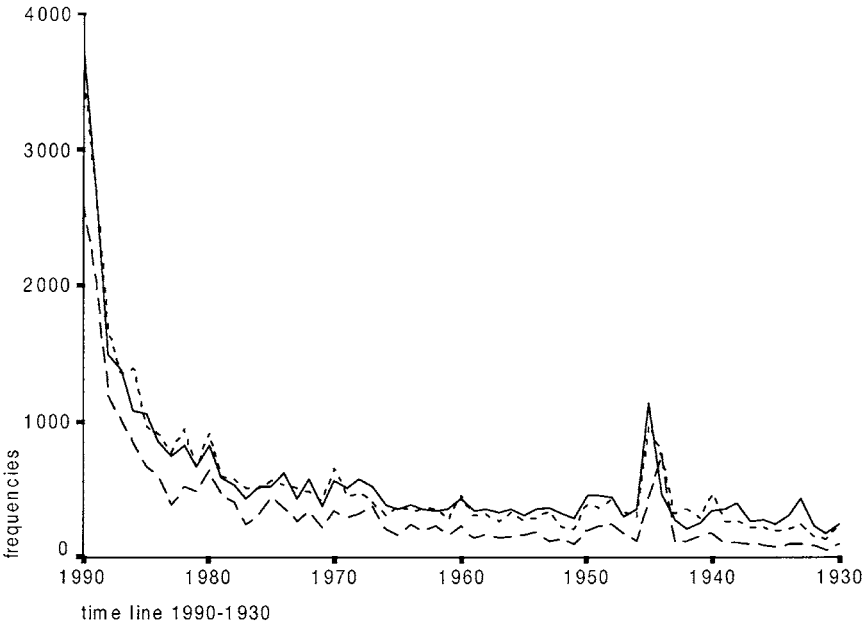

Figure 1. Frequencies of the years 1990–1930 occurring in three newspaper corpora FAZ (line), IHT (strokes), NRC+ (small strokes).

3.1. REGULAR DEVIATIONS

Regular deviations of the curve have something to do with the roundness of the number values of the years. All 'round years' occur more frequently than one would expect on the basis of the declining curve. We find this phenomenon not only for the period 1990–1930, but in all periods. The year 1400 is more frequent than 1410 or 1390, and 1650 occurs more often than 1640 etc. Round numbers are more suitable for indicating estimates. Round years clearly indicate estimated points on the time-line.

Roundness is not an absolute, but a relative property of numbers: numbers are more or less round. In general, numbers which score high on the roundness scale are more frequent in normal language use than less round numbers. This is a property of all uses of numbers in natural language and turns out to be true for years also (cf. Jansen and Pollmann, in preparation). We come across another kind of regular deviation in the data. These are the years that were anniversaries in 1994. For example, in the collections **FAZ**, **NRC** and **IHT**, the years 1894 and 1844 are represented more often than the neighboring years of 1893 or 1845.

Of course, this is a consequence of the fact that in 1994 all kinds of events have been commemorated that took place one hundred or one hundred and fifty years earlier. The *International Herald Tribune* even has a daily column in which events of 50, 75 and 100 years ago are commemorated. Clearly, the years 1944, 1919 and 1894 are greatly overrepresented in **IHT**. Evidently, 'commemoration years'

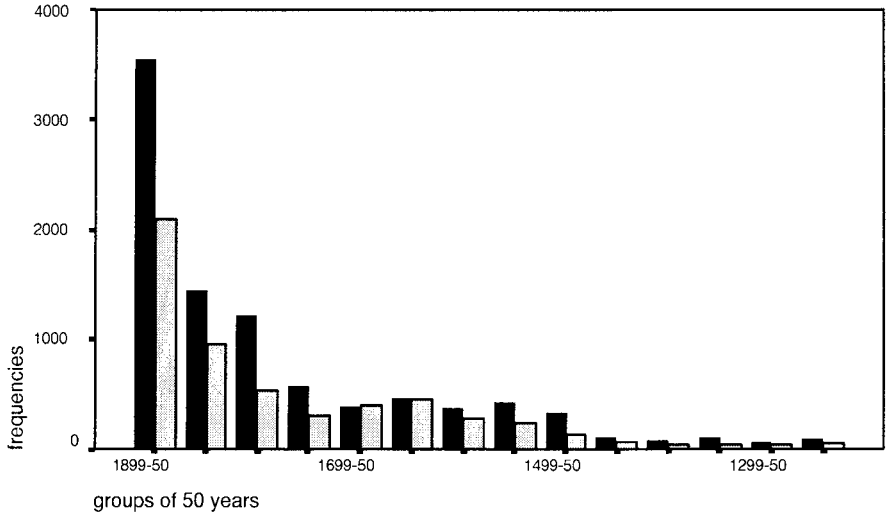

Figure 2. Number of years over the period 1899–1250 in sets of 50-year periods in **FAZ** (dark) and **NRC** (light).

form peaks. The figures for the years 1920, 1919 and 1918 in **IHT** are 88, 365 and 90 respectively; those of the years 1895, 1894 and 1893 are 18, 361 and 21 respectively. For this reason one might conclude that about 300 of the high number of occurrences of '1944' are an effect of the commemorational character of this year. The figures for the years 1945, 1944 and 1943 in **IHT** are 444, 772 and 118.

3.2. IRREGULAR DEVIATIONS

Figure 1 also shows irregularities that we cannot relate to a property of the years as numbers. Apparently they reflect the special role some historical episodes have in the present day. This is of course true for the very striking presence of the years 1944 and 1945 in the data, but also for other years: 1985 in **NRC**, 1948 in all collections; 1982 and 1968 in **FAZ** and **NRC**; 1949, 1938 and 1933 in **FAZ**.

It is obvious to look for an explanation first in the special position these years have in the historical consciousness of the people for whom the newspaper is intended.

In the present, some historical figures, events and developments are more important than others. Obviously, they are better suited to illustrate, elucidate or explain aspects of the present day. That the use of years reflects the special importance of some years or periods for the contemporary culture becomes visible in Figure 2.

This figure covers the period 1899 to 1250, and represents numbers of years in sets of fifty years. The years originate from **FAZ** and **NRC**. It is easy to see that the German newspaper has more years from the second part of the nineteenth

and of the eighteenth century and from the first part of the sixteenth century, whereas the Dutch daily contains many years referring to the seventeenth (and the second part of the sixteenth) century. Here, too, one may say that these 'peaks' mirror the significance of these periods in the German and Dutch cultures. The importance of the Golden Age for present-day Dutch culture is confirmed by the *Cultureel Woordenboek*, a Dutch dictionary of cultural literacy, which counts far more facts from this period as part of the general education of the contemporary Dutch than events of the eighteenth or nineteenth century. For the five fifty-year periods between 1750 and 1500 the totals in the *Cultureel Woordenboek* are 25, 42, 69, 76 and 19. As far as the German data are concerned, it is beyond doubt that the periods we mentioned, the periods of Romanticism and Goethe and Schiller and of the German unification and Bismarck, play important roles in today's German cultural identity. However, we could not find independent quantitative evidence to explain the high numbers of years which cause the peaks in **FAZ**.

3.3. THE CURVE

Setting aside the incidental, regular or irregular deviations of the curve, we still have the general curve. And from the point of view of our research project it is this general curve that is most interesting. Although it is easy to predict that we shall come across fewer years the further we go back in time, this does not a priori mean that the general form of the curve is itself predictable. Nevertheless, Figure 1 gives rise to the suspicion that the curve itself exhibits some regularities too. After all, the slope of the curve looks the same for each of the three sets of data. All three curves demonstrate a rather steep slope on the left which steepness diminishes the more we come to the right.

Approximately the same curve we get, when we plot the data over a larger range of time. Figure 3 presents the plots for the period 1990–1690 for **FAZ**, **NRC** and **IHT**. The distribution of the frequencies over the time-line is in percentages of the 10-year averages of years in the respective corpora. Here we see also an inverse-like curve: the frequencies seem to be inversely proportional to the distance in time. These considerations suggest the following claims.

- The distribution of the attention to the past is to a large extent independent of the language and the cultures in which the newspapers have been written and find their audiences. It is as if a universal force in the human mind by and large regulates how our attention to the past will be distributed. The distribution of the attention to the past is independent of the total amount of attention given to phenomena of the past, as measured in the total amount of years. Whether a given corpus has a high or a low year density the distribution of the attention over the past remains largely the same. It is as if "historical interest generates historical interest" no matter to which part of the past this interest is directed. To put it otherwise, hodiecentrism (the absence of interest in the parts of the past of a long time ago) is always accompanied by a lack of interest in

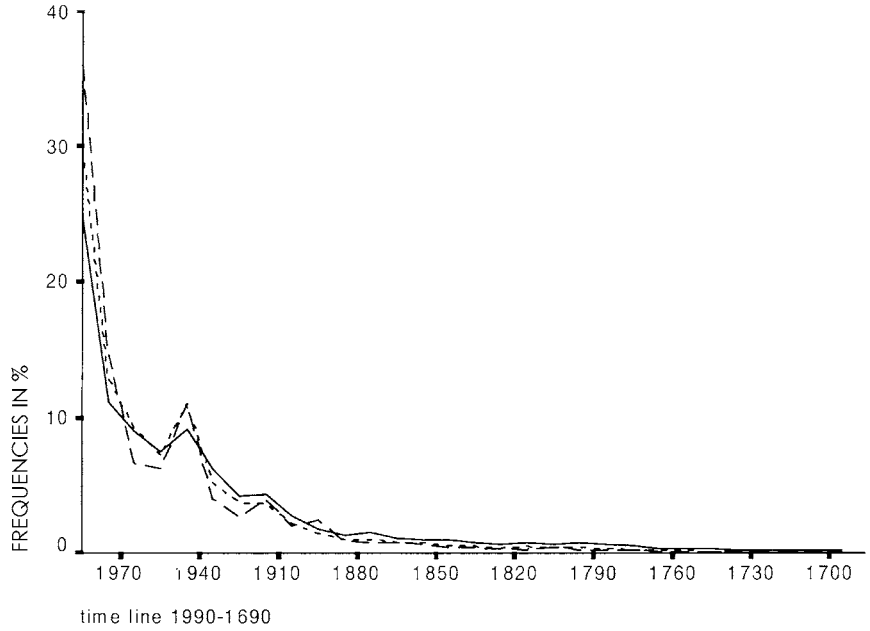

Figure 3. The frequencies of the ten year averages of years between 1990 and 1690 in three corpora FAZ (line), IHT (strokes), NRC (small strokes) (in percentages of the total number of years in this range in each corpus).

history in general. However, a deeper statistical analysis as will be presented in Section 4 brings to light a marked difference between the IHT on the one side and FAZ and NRC on the other. The IHT has less year-types. It presents significantly fewer years more often.

- The corpora on which the frequency data of WFEng60 and WFNed69 are based, dating from 1960 and 1969/70 respectively, show a distribution of the years over the time line which resembles those of Figure 3. We take this as an indication that the distribution of the attention to the past is independent of the date of origin.

These generalizations give rise to some questions about the nature of the phenomena under scrutiny. In the next two sections, we will broaden the scope in two directions. In Section 4, we will support these conclusions with the help of statistical techniques developed for the analysis of word frequency distributions. In that section, we will also relate the findings to the study of word frequency phenomena in general. In Section 5, we will investigate what we might conclude on the basis of these analyses about the broad concept of "historical consciousness".

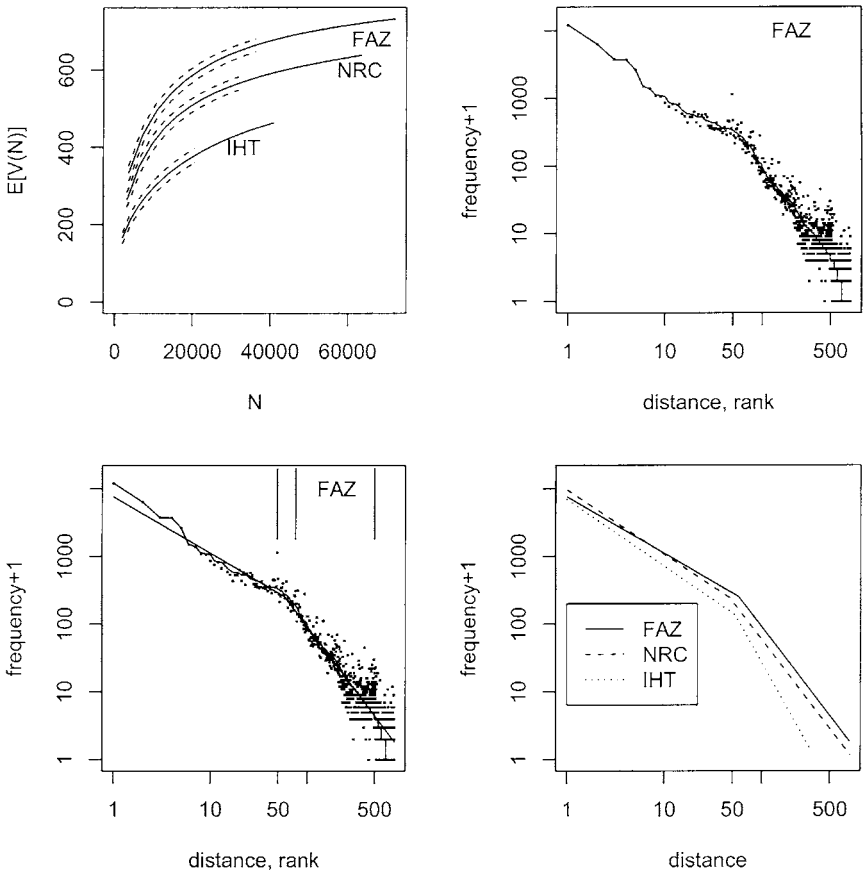

Figure 4. The expected growth curve of the number of year types E[V(N)] as a function of the number of year tokens N for FAZ, NRC, and IHT (upper left panel), the rank-frequency distribution (solid line) and the distance-frequency distribution of the FAZ (upper right panel), the rank-frequency and distance-frequency distributions with a bipartite linear fit (lower left panel; the vertical line segments highlight the years 1945, 1918, and 1500), and the bipartite linear fits for all three newspapers (bottom right).

4. Statistical Analysis

This section has a two-fold aim. We will first show that the year frequency distributions of the FAZ, NRC and IHT differ significantly with respect to their richness of historical referencing. We will then proceed to show that they also have a surprising property in common, namely, a discontinuity in the distance-frequency distribution suggesting that within a time span of 50 years individual experience allows for the recall of a greater spectrum of events.

For the question of possible newspaper-specific richness with respect to historical referencing, consider the upper left panel of Figure 4, which plots by means of solid lines the way in which the number of different year types mentioned increases

Table II. Lexical statistics for the IHT, the FAZ, and the NRC

	FAZ	NRC	IHT
N	72259	63466	40854
$V(N)$	731	638	463
$V(1,N)$	81	104	118
$\check{S}$	781.83	720.55	680.98
K	461.75	427.93	576.77
Z	8.6792	8.2446	8.2815
b	0.0318	0.0248	0.0146
γ	−0.5752	−0.5156	−0.4164
X^2	21.63	74.94[a]	26.71
p	0.0613	0.0000	0.0136

The high X^2 value for the fit for the NRC is due to severe irregularities in the head of the frequency spectrum of the NRC and not to a systematic qualitative lack of goodness of fit.

with the number of year tokens using binomial interpolation (Good and Toulmin, 1956; Muller, 1977). The dashed lines represent 95% confidence intervals around each vocabulary growth curve for the intervals for which the confidence intervals can be calculated without further parametric assumptions (Chitashvili and Baayen, 1993). The non-overlapping confidence intervals show that the three newspapers are quite dissimilar with respect to the extent to which they refer to years in the past, not only in terms of tokens, but also in terms of the types expected for equal numbers of tokens. The FAZ displays the greatest richness with respect to historical referencing, while the IHT is relatively poor in this respect.

Table I provides some further statistics illustrating the ranking that is apparent in the plot of growth curves. This ranking is found not only for the numbers of tokens N and the numbers of types $V(N)$, but also for the number of years $\check{S}$ that a newspaper might have referenced for an infinitely large corpus of newspaper issues from 1994 and for the parameters b and γ of the generalized inverse Gauss-Poisson model (Sichel, 1986) on which these estimates are based. Complementary to these measures of type richness, Yule's K is a measure of repetitiveness. Not surprisingly, the newspaper with the lowest referential richness displays the highest value for K.

The present year-frequency distributions differ markedly from standard word frequency distributions. For the latter, the number of hapax legomena $V(1,N)$, the types occurring with token frequency 1, tends to comprise at least half of the total number of types. This reflects the fact that large numbers of word types typically do not appear even in very large corpora (Baayen, 2000). By contrast, the scarcity of hapax legomena in the year-frequency distribution of the FAZ correlates with the fact that this newspaper references 731 out of the 782 types it might have mentioned in the limit of $N \rightarrow \infty$. The observed number of types comes close even to the

logically possible maximum number of year references, 800, the number of years spanning the years for which references were collected (1993–1194).

The solid line in the upper rigt panel of Figure 4 represents the Zipfian rank-frequency plot in the double logarithmic plane, with the highest-frequency year being assigned rank 1, the next highest frequency year rank 2, and the unseen years, which have frequency 0, the highest ranks. For word frequency distributions, bi-logarithmic rank-frequency plots generally tend to reveal a straight line (Zipf, 1949) or, more often, a slightly convex curve (Mandelbrot, 1953). Note that for the present data, the rank-frequency curve reveals a non-Zipfian convex curvature at the very right-hand side of the plot that ties in with the scarcity of unseen year types. Note, furthermore, that the present rank-frequency relation appears to consist of two roughly linear segments with different slopes that meet at a breakpoint located approximately around rank 50, instead of displaying a gradual downward trend of the Zipf-Mandelbrot type.

To understand what is at issue here, consider the distance-frequency distribution also plotted in the upper right panel of Figure 4 by means of dots, the distance being the number of years a given year type is removed in history from 1994. Thus, 1993 has distance 1, 1992 distance 2, etc. It is a remarkable property of the distance-frequency distribution that, in spite of the scatter of year frequencies due to some years hosting more important events than others, it still closely follows the rank-frequency distribution. In fact, the rank-frequency curve emerges as a kind of expected value of the distance-frequency curve, which reveals exactly the same discontinuity at around distance 50 as the rank-frequency curve at around rank 50.

This discontinuity shows that the relation between year frequency f and year distance d cannot be simply modeled along Zipfian lines as an exponential relation

$$f = a/d^b \tag{1}$$

which transforms into a linear relationship in the double logarithmic plane,

$$\log(f) = \log(a) + b\log(d), \tag{2}$$

nor as a Zipf-Mandelbrot relation of the form

$$f = a/(d + c)^b \tag{3}$$

Instead, we need a more complex linear model of the form

$$f = a_0 + a_1(d_i - d_n) + a_2(d_i - d_n)\mathbf{I}_{[i>n]} \tag{4}$$

where we select the breakpoint such that the deviance of the model is minimized. For the FAZ, the optimal model has a significant breakpoint for $n = 59$ ($F(1,798) = 6240.72$, $p < 2.2\text{e-}16$ for a_1 and $F(1,797) = 168.04$, $p < 2.2\text{e-}16$ for a_2), indicating a breakpoint in 1935. The bottom left panel of Figure 4 adds this fit to the data of the upper right panel, and also highlights the years 1945, 1918, and 1500 by means of vertical line segments. For the NRC, a significant breakpoint is found at $n = 47$,

i.e., in 1947 ($F(1,798) = 5604.16$, $p < 2.2e\text{-}16$ for a_1 and $F(1,797) = 144.33$, $p < 2.2e\text{-}16$ for a_2). Finally, the breakpoint for the IHT is located at $n = 54$, i.e., in 1940 ($F(1,303) = 2669.81$, $p < 2.2e\text{-}16$ for a_1 and $F(1,303) = 167.03$, $p < 2.2e\text{-}16$ for a_2).[4]

Although the exact values of the breakpoints are approximate, given the slightly undulating curve for the nearest distances and the increasing scatter for larger distances, it is clear that for all three newspapers we have a real change in the way years are referencing the past of the last 50 years, or the more remote parts of the time line. The bottom right panel of Figure 4 illustrates this similarity for the three newspapers jointly. In spite of the three newspapers giving rise to year-frequency distributions that differ substantially with respect to year-richness, they are all reveal the same kind of linear relation in the double-logarithmic plane and are subject to the same discontinuity in the distance-frequency relation. Clearly, the distribution of attention to the past is indeed to a large extent independent of the language and the cultures in which our newspapers originate.

Finally consider the interpretation of the slopes of the two line segments in these fits, $E_1 = a_1$ and $E_2 = a_1 + a_2$ in (4), in the light of the number of different events that are referenced in a particular year. For years at a small distance, the number of different events is likely to be large, while for distant years it is more likely that the same event is referenced by all or nearly all of the year tokens. If this is indeed the case, we may interpret the gradients E_1 and E_2 as measures of referential concentration. For small distances, the relevant gradient, E_1, is small compared to the gradient for large distances, E_2. For small distances, therefore, the referential concentration is small, indicating a wide variety of different events being referenced. By contrast, the larger values for E_2 indicate greater lexical concentration, with a smaller number of events being referenced more intensively. This interpretation is analogous to the well-known relation between polylexy and frequency, with higher frequency words having more meanings and shades of meaning than lower frequency words (Koehler, 1986). In other words, the breakpoint analysis suggests that 'history' begins around distance 50, with specific events that are generally accepted as being important to be commemorated in the collective mind. For shorter distances, memory of individual experiences allows for the recall of a greater spectrum of events.

5. Historical Consciousness

Do newspapers reflect the historical consciousness of the culture in which they are written and read? In a theoretical introduction to one of the few existing empirical studies of historical consciousness, Pandel (1991) has presented a useful survey of seven dimensions that might be found in this complex concept. Notably the dimensions 'time consciousness' and 'consciousness of identity' seem to be of some relevance in the framework of the present inquiry.[5] In Pandel's view, 'time consciousness' is the cognitive faculty of contrasting the past with the present or

the future. The 'consciousness of identity' makes it possible for the individual to discern 'the own group' from 'the group to which others belong', insofar as this is connected to a time-perspective, i.e. if the actions of people in the past are characterized as the actions of someone of 'the own group'. Although these concepts are not very clearly defined, we can use them in tentatively answering the question of what our investigations have brought to light about the attention of newspapers to the past.

According to Pandel, people have ideas about the time that makes up the present. Theoretically, the present might be considered as a point in time. This does not alter the fact, however, that people experience the present as something with duration. This intuition seems to be correct. People can talk about 'the present time', 'this time', 'now' referring to a certain stretch of time. In accordance with these expressions one can say that consciousness of the past does not start at precisely the moment before this present moment. 'The past', 'formerly', 'in bygone days', 'later', 'soon', it all starts at some distance from this very moment. When people are asked to estimate when the present began or when "nowadays" started or how old things maximally might be to call them contemporary, they go back some years or so, relating the 'end' of the past mostly to a change in their lives (getting a new job, a new house, the death of a partner, entering a new phase in education etc.). In this view 'someone's own present' should take up a period of ten years at the most. For this period of time the newspaper data do not show anything in particular. Of course, one can think of a type of research that might be able to teach us something about the 'length' of the present, measured by the use of the expressions we mentioned, but the frequencies of the years do not bring to light anything interesting in this respect. There is simply no indication in the data that the most recent years have an existence in our minds that differs from the other parts of the past. However, as the detailed statistical analysis revealed, there is something in the data that cannot but interpreted as a breakpoint in the way we use the available knowledge of the past. This breakpoint seems to be at a distance of about 50 years, varying between 59 to 47. The process of forgetting things older than 50 years seems to go quicker than the forgetting of more recent dates. It seems reasonable to think that this has something to do with the way we relate to the past farther away. Memories of living people will be more varied, more individually colored then the written historical accounts, which present the past in a more or less standardized and canonized form. The 50 years distance might be the point where first hand knowledge of the past changes in the knowledge which has been passed down to us by stories told by others.

'Time-consciousness' also contains, says Pandel, an idea about the 'Dichtigkeit der Ereignisse' (litt. density of events). The individual knows varying numbers of events of different periods of the past. 'Events' can be extended to 'people', 'states of affairs' and 'objects'. The historical consciousness of different historical periods contains a different number of things. We might assert this without claiming that these periods in reality saw a different number of important events, etc. This

component of 'time consciousness' can easily be found in the data, sc. in the general form of the data-lines. As argued elsewhere, there is some reason to identify this general change in 'event density' with a sort of forgetting, a 'forgetting' which is the effect of a diminishing attention to phenomena which claimed this attention for a certain period in the past (cf. Pollmann, 1998a). The curves have the form of 'forgetting-curves' (cf. Friedman, 1990: 33). A functional explanation of this is obvious: the greater the distance between past and present, the more difficult it is to connect something of the past to an item in the present day. This is probably true not only for journalists and newspapers, but for their readers as well. We might be dealing with a universal property of the human mind.

A third component of 'time consciousness', says Pandel, is the human inclination to attribute a special meaning for the present day to parts of time, the past or the future. Nazism had a special interest in German antiquity and in the future of a thousand-year reign. In the Renaissance this interest concerned Greek-Roman antiquity, and in the period of Romanticism it was the Middle Ages. Clearly, these periods of special interest one finds in the data we presented. The Dutch Golden Age, the period of Romanticism and of the 'Reichsgründung' and the period of the Second World War and the 'Third Reich' have a place in the 'time-consciousness' of the Germans, c.q. the Dutch. As we argued above, in the light of the generally diminishing attention to the past that happened to be expressed in the curves, these episodes form parts of the 'consciousness of identity' of the cultures in which these newspapers are written and read.[6]

The data enable us also to make a few remarks on the presence of historical knowledge in newspapers.

In a recent study on the function and use of historical knowledge, the German historian Schörken has argued that history has just a marginal position in newspapers. Newspapers live by the topicalities of the day, and history is –, as common sense will have it – not topical at all. 'The marginal position of history in the daily press is caused by the pressure of topical subjects, the event character and the pragmatical structure of most of the news-items' (1995: 124). Schörken's opinion does not get any empirical underpinning, but he does not stand alone (Bieber, 1986). There is, however, some reason to doubt the correctness of Schörken's position, which can be found in the argumentative and narrative structure of newspaper texts. Journalists are expected to provide interpretation and context in addition to the facts of the news. It is even stated, for example, in the American ethical code for journalists, that news items be presented 'in a context that give them meaning' (MacManus, 1992). It might be expected that interpretation and context would be found among other things in information about the history of the news, i.e. about developments that preceded the events of the day. In orientating themselves in the world, non-historians do use some knowledge of the past, their primary aim not being to understand the past as such, but rather the world of today. This is the case for individuals as well as for public institutions like newspapers, which are in our

society the medium by which we learn about our world. In performing this function newspapers will use historical knowledge.

The figures presented so far make it clear, in our opinion, that the past has an obvious place in newspaper columns. Roughly two per thousand word forms in the newspapers are years related to the past. We can conclude on this basis that Schörken's contention concerning the marginal position of the past in the newspapers is not in accordance with the facts. Schörken supposes that the media present history not as knowledge, but as information, recollection, discussion and emotion, i.e. as '*Vergegenwärtigung*' (representation) of the past. He might be right if one is looking for autonomous writings and pieces of historical knowledge. But reading a newspaper looking for the past in this way, one can easily overlook the fact that attention to the past arises from the argumentative and narrative structure of these texts. Journalists do not deny their core task in writing about the past. Their attention to the past finds its inducement in the topics of the day; – history in the newspaper is applied history. Schörken's position is comparable to the complaint that there is no economics or political science in the newspapers, by someone who is overlooking the fact that journalists write about economical and political issues all the time.

Notes

[1] We thank the Institute of Dutch Lexicology for its permission to use the 27mln-corpus.

[2] In this paper, we refer to parts of the time-line by two years, of which the first indicates the year which is nearest to the present. In the figures the reader will find the present at the left-hand side. By presenting the data in such an a-historical manner, we want to stress that in this paper the reader is encouraged to look to the past from a contemporary point of view.

[3] In the British National Corpus, the largest text corpus in existence (100,000,000 words), we found a year density of approximately 400. The texts in the BNC date from the period 1975–1993 and are not all taken from newspapers. Cf. Thijs Pollmann (1998b).

[4] In the case of the IHT, application of (4) results in a breakpoint at distance 306, resulting in large deviance for the smaller distances. The breakpoint at $n = 54$ was obtained by restricting i to the range [1,306].

[5] In addition to this type of historical consciousness Pandel postulates a 'reality consciousness' concerning the difference between reality and fiction; concerning a 'historicity consciousness' concerning the difference between the changeable and the static; a 'political consciousness' concerning divisions of power; a 'social-economic consciousness' concerning the rich-poor contrast; and a 'moral consciousness' concerning the difference between good and evil. On these five dimensions the data do not bring special things to light.

[6] Pandel mentions two other components of 'time consciousness': the need for subdivisions of the past into periods; and the tendency to render a story-like cohesion to historical events. Neither aspect can be found in the data.

References

Baayen, R.H. *Word Frequency Distributions*. Kluwer Academic Publishers (to appear).

Berkel, K. van. "Inleiding". In *Geschiedenis: een hoofdvak*. Eds. A.Th. van Deursen et al. Leiden, 1985, pp. 1–4.

Bieber, Horst. "Geschichte als Hintergrund in Kommentar und Leitartikel". *Geschichtsdidaktik*, 11 (1986), 357–363.

Chitashvili, R.J. and R.H. Baayen. "Word Frequency Distributions". In *Quantitative Text Analysis*. Eds. G. Altmann and L. Hrebícek. Trier: Wissenschaftlicher Verlag Trier, 1993, pp. 54–135.

Friedman, W.J. *About Time*. Cambridge, 1990.

Good, I.J. and G.H. Toulmin. "The Number of New Species and the Increase in Population Coverage, When a Sample is Increased". *Biometrika*, 43 (1956), 45–63.

Jansen, C.J.M. and M.M.W. Pollmann. "On Round Numbers" (in preparation).

Koehler, R. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer, 1986.

Kuçera, Henry and W. Nelson Francis. *Computational Analysis of Present-Day American English*. Providence Rhode Island, 1965.

Mandelbrot, B. "An Information Theory of the Statistical Structure of Language". In *Communication Theory*. Ed. W.E. Jackson. New York: Academic Press, 1953, pp. 503–512,

McManus, John H. (1992) "What Kind of Commodity is News?" *Communication Research*, 19, 787–805.

Muller, Ch. *Principes et méthodes de statistique lexicale*. Hachette, Paris, 1977.

Pandel, H.-J. "Dimensionen des Geschichtsbewußtsein. Ein Versuch, seine Struktur für Empirie und Pragmatik diskutierbar zu machen". *Geschichtsdidaktik*, 12(2), 130–142.

Pandel, H.-J. "Geschichtlichkeit und Gesellschaftlichkeit im Geschichtsbewußtsein. Zusammenfassendes Resümee empirischer Untersuchungen". In *Geschichtsbewußtsein empirisch*. Eds. Bodo von Borries, Hans-Jürgen Pandel and Jörn Rüsen. Pfaffenweiler, 1991, pp. 1–23.

Pollmann, Thijs. "On Forgetting the Historical Past". *Memory and Cognition*, 26(2) (1998a), 320–329.

Pollmann, Thijs. "The Process of Cognitive Distance: A Quantitative Analysis of Some Aspects of Historical Culture". *Historical Social Research*, 23(4) (1998b), 79–93.

Pollmann, Thijs. "Forgetting and the Ageing of Scientific Publications". *Scientometrics*, 47(1) (2000), 43–54.

Schörken, Rolf. *Begegnungen mit Geschichte. Vom ausserwissenschaftlichen Umgang mit der Historie in Literatur und Medien*. Stuttgart, 1995.

Sigurd, Bengt. "Round Numbers". *Language and Society*, 17 (1988), 243–252.

Yule, G.U. *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press, 1994.

Zipf, George Kingsley. *Human Behavior and the Principle of the Least Effort. An Introduction to Human Ecology*. New York: Hafner, 1949.

Zipf, George Kingsley. *The Psycho-Biology of Language*. Cambridge MA: M.I.T. Press, 1968.