

Aging in the Treasure: Some Methods for Evaluating Content Words in Large Data Bases

Paul A. Fortier, Kevin J. Keen, Marc Fortier

Paul A. Fortier (Author for Contact), Centre on Aging, University of Manitoba, Winnipeg, Man., R3T 2N2, CANADA.

Kevin J. Keen, Department of Statistics, University of Manitoba, Winnipeg, Man., R3T 2N2, CANADA.

Marc Fortier, St. Paul's High School, 2200 Grant Ave., Winnipeg, Man., R3P 0P8, CANADA.

KEYWORDS: database, content analysis, time-series analysis

AFFILIATION: University of Manitoba, University of Manitoba, St. Paul's High School

E-MAIL: Fortier@ccm.UManitoba.ca
(Contact author)
KJKeen@cc.UManitoba.ca
Fortier@ccu.UManitoba.ca

FAX NUMBER: 204-275-5781

PHONE NUMBER: 204-474-9841 (office)
204-474-9313 (department)
204-475-5853 (home)

1. Background

Because of the demographic changes brought about by control both of fertility and infectious diseases in humans, people are living longer. In order to come to terms with these facts it is useful to examine how society has viewed aging and old people at various stages of history. The Trésor de la Langue Française database contains roughly a thousand primarily literary texts published between 1789 and 1964, that is to say the period during which France developed from a feudal, agrarian state to a modern industrial society. It is of interest to examine the use of terms relating to aging in this society, and, given the volume of data, a statistical approach would seem warranted. Methods of time series analysis are used to study the theme of aging through frequency counts from the Trésor. By simply plotting the frequency count of a collection of the most frequently occurring words evoking the thematic construct of age, it is possible to see clearly the evolution of allusions to aging in a substantial portion of modern French literature and thus, by inference, the evolution of attitudes regarding age in one western European culture.

2. The Data

All numbers analyzed in this paper are drawn from the *Dictionnaire des Fréquences*, published in 1971 by the Trésor de la Langue Française (Imbs). The preface to this work explains that it is based on 416 primarily literary texts published from the Revolution (1789) to the late 19th century, and 586 late 19th and early 20th century texts. The total number of words recorded is 70 million after the exclusion of 2.2 million proper nouns and foreign words.

Volume III of the *Dictionnaire* contains the frequencies of the most frequently occurring words in the database divided into 15 time segments of roughly ten years each, but varying on occasion between 27 years (1789-1815), and 5 years (1933-37). The 4,000 most frequent words in each time segment were retained so that a total of 6947 words are covered. The number of words and the number of texts included in each time segment are not reported but it seems reasonable to estimate that the number of texts per time segment varies between 50 and 75, and the number of words sampled between 4 and 6 million. Frequencies in each time segment are normalized to a base of 10 million words to facilitate comparison between segments.

The vocabulary of aging which will be examined is made up of six words, *âgé*, *vieillard*, *vieillesse*, *vieilli*, *vieillir* and *vieux*. The frequencies for *grand-père* are furnished but those for *grand-mère* are not, so it was decided not to include this word in order not to introduce gender bias into the data.

3. Frequencies in the Data

In order to facilitate understanding of the data, the relative frequencies found in the *Dictionnaire* have been adjusted to a base of 100,000 words, roughly the size of a substantial (300 page) novel. The sum of the frequencies of the six words evoking aging in each time segment varies between 21.00 and 58.31. Ignoring any temporal relationship in the data, calculations show the mean of sums is 38.29 and the standard deviation is 9.38. The observed total frequencies thus fall between -1.84 and +2.14 standard deviations from the mean. Given 14 degrees of freedom, none of these totals generate a t-statistic significant even at the 0.05 level. However, examination of a time series plot of the total frequency reveals a striking pattern of temporal relationship. There is thus a clear need for more sophisticated statistical techniques, other than those based on the assumption of independently distributed data, because that assumption leads to an inflated estimate of the sampling variability of the total frequencies.

It is known that when a large number of measurements are made in a relatively homogeneous population, these measurements tend to have the

well-known Gaussian or normal distribution. When a frequently occurring phenomenon is studied in a relatively small corpus, for example the period in Zola's novels, this tendency is also observed (Brunet).

With regard to the underlying distribution of the data in this study, after removing temporal dependencies and structures, it is anticipated that the distribution of resulting residuals from the data will, by analogy, also tend to the Gaussian distribution.

4. Methodology

Even an unsophisticated technique such as plotting word frequencies as a function of the start of time periods for passages assayed in the Trésor can reveal much about the evolution of attitudes with time. Details on the graphical analysis of time series to suspect – if not detect – features such as a trend, oscillations, or a random component can be found in Kendall, Stuart and Ord (1983).

A more recent and powerful analytical tool is available in the form of change-point analysis is to determine the date at which a time series changes in a manner stipulated by a model thus allowing the researcher to speculate about possible reasons for the date of change. Descriptions of these statistical tools can be found in Kotz and Johnson (1989).

5. Results

Given below is a time series plot for the frequency of words per 100,000 for all words selected that evoke the theme of aging.

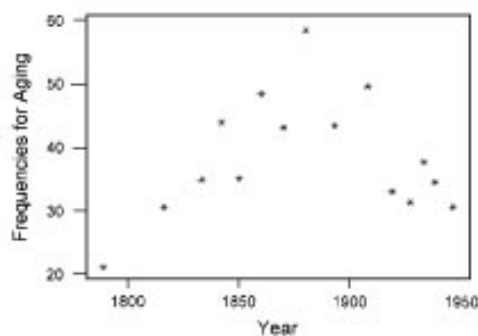


Figure 1: Frequencies for the Theme of Aging

The total vocabulary of aging tends to be used with increasing frequency from 1789 to about 1880, after which its use steadily decreases. It can be noted that the period from the beginning of the Revolution to the abdication of President MacMahon in 1879 was characterized by social unrest,

culminating in four revolutions, the industrialization of the country without a concomitant improvement in standards of living, and constant contest between a reactionary and a forward looking view of governmental structures (Wright).

The data indicate that these tensions correspond to an increasing use of the vocabulary of aging until a crisis point somewhere around 1880. After that turning point, a liberal ideology generally dominates French society and outlook, something which the data suggest brings about fewer allusions to aging.

The proportion of the use of *vieillard* (the only clearly pejorative term) to the other vocabulary of aging, has a different pattern as evidenced from the time series plot (below) of the relative frequency of the word *vieillard*.

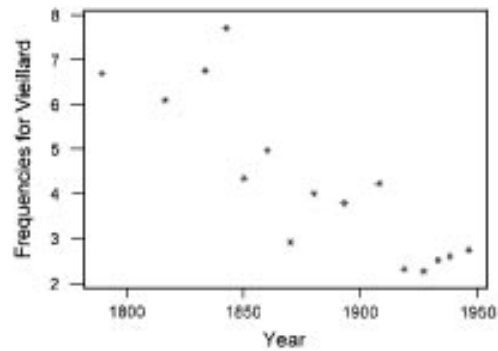


Figure 2: Frequencies for the Word Vieillard

The frequency of *vieillard* decreases sharply from 1789 to the 1860s, after which time the use of this word shows much smaller fluctuations in a continuing downward trend. The significance of this pattern is perhaps prophetic, since 1860 cannot be identified as a turning point in French history or society. It may be that the decreasingly pejorative vocabulary of aging is predicting the triumph of liberalism which will not take place in the political realm until late 1879. But clearly this phenomenon requires more study.

6. Conclusion

The study of the vocabulary of aging in a Trésor de la Langue Française database does lead to results of interest to the historian of French society. This small preliminary study also demonstrates the usefulness of the techniques of time series analysis in general and the suggestion of success for the application of both intervention model analysis and change-point analysis in studying content words in large databases.

References

- Brunet, Étienne, 1985. "La phrase de Zola." *La Critique Littéraire et l'Ordinateur*. Montréal: Derval and Lenoble, pp. 111–57.
- Imbs, Paul, 1971. *Dictionnaire des Fréquences*. 4 vols. Nancy: C.N.R.S.-T.L.F.
- Kendall, Sir Maurice, Alan Stuart and J. Keith Ord, 1983. *The Advanced Theory of Statistics, Volume 3, Design and Analysis, and Time Series, Fourth Edition*. New York: Macmillan.
- Kotz, Samuel and Norman L. Johnson, 1989. *Encyclopedia of Statistical Science*. New York: John Wiley and Son.
- Wright, Gordon, 1987. *France in Modern Times*. 4th ed. New York: Norton.