DOCUMENT RESUME

ED 443 131 CS 217 209

AUTHOR	Hogenraad, Robert
TITLE	Once Is Not Enough: Statistical Simulation of Textual Data.
PUB DATE	2000-08-00
NOTE	9p.; "Acknowledgment is made to the National Fund for Scientific Research, Belgium." Paper presented at the Seventh Congress of the International Society for the Empirical Study of Literature (IGEL) (Toronto, Ontario, Canada, July 31-August 4, 2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Content Analysis; *Literary Criticism; *Scholarship
IDENTIFIERS	*Empirical Research; Resampling Techniques; *Text Processing (Reading); Textual Analysis; Ulysses (Joyce)

ABSTRACT

Reproduction is part of the regulations of daily speech. Everyday experience shows there are different ways to convey the same information: starting sentences that are quickly rephrased using other words, manuscripts passing through draft states before the writer decides upon the final version. In more ways than one, any text is always but the sample of another one and the words read are but the visible part of an iceberg. James Joyce's "Ulysses" was content analyzed using the Regressive Imagery Dictionary. Among other things, this dictionary allows a person to assess the degree of presence, in texts, of primordial thought content. A negative linear profile is observed, i.e., primordial content decreases as the one-day journey goes by. Pondering this negative linear profile generates wonder about the enterprise of statistical analysis, how little it says, how brittle is a quantitative analysis of a literary work in the face of the uncertainty that surrounds such data. New ways to look at literary data consist of resampling thousands of time data that by nature exist in only one exemplar, bringing empirical study of literature down to what literature is made of--inking out, excisions, alterations, disappearances. The rates of primordial thought content were repeated 2000 times, using a bootstrap algorithm. The best contribution a researcher can make to empirical studies of literature is to "hammer" scholars to see how well resampling allows researchers to assess the degree of uncertainty associated with any measurement. (Contains 3 figures, 5 notes, and 16 references.) (Author/NKA)

Once is not enough: Statistical simulation of textual data[1]

Robert Hogenraad

Université catholique de Louvain,

Louvain-la-Neuve, Belgium

Seventh Congress of the International Society for the Empirical Study of Literature (IGEL), Toronto, July 31-August 4, 2000.

[1] Acknowledgment is made to the National Fund for Scientific Research, Belgium.

Abstract

Reproduction is part of the regulations of daily speech. Everyday experience shows that there are different ways to convey the same information: starting sentences that we quickly rephrase using other words, manuscripts passing through draft states before the writer decides upon the final version. In more ways than one thus, any text is always but the sample of another one and the words we read are but the visible part of an iceberg.

James Joyce's *Ulysses* was content analyzed using the Regressive Imagery Dictionary. Among other things, this dictionary allows one to assess the degree of presence, in texts, of primordial thought content (*love*, *sex*, *food*, *chaos*, *dream*, *flying*, for example). A negative linear profile is observed, i.e., primordial content decreases as the one-day journey goes by.

Pondering this negative linear profile makes one wonder about the enterprise of statistical analysis, how little it tells us, how brittle is a quantitative analysis of a literary work in the face of the uncertainty that surrounds such data. New ways to look at literary data consist of resampling thousands of time data that by nature exist in only one exemplar, bringing empirical study of literature down to what literature is made of, inking out, excisions, alterations, disappearances. The rates of primordial thought content were repeated 2,000 times, using a bootstrap algorithm.

The best contribution one can make to empirical studies of literature is not to collect more p< something that doesn't prove a thing, but to hammer scholars in order to see how well resampling allows one to assess the degree of uncertainty associated with any measurement. Until now, there is no indication that empirical studies of literature so much as noticed the difference that has taken place in processing textual data by a systematic exploration of uncertainty.

Once is not enough: Statistical simulation of textual data

"Rien n'est jamais dit puisqu'on peut le dire autrement."
(Robert Pinget, "*Quelqu'un*". Paris, Minuit, 1965)

In a recent essay, Richard Poirier (1999) has a chapter called "Erasing America" the first part of which is devoted to Baudrillard's (1977) notion of "hyperrealism", according to which "signs no longer designate anything at all, only other signs" (Poirier, 1999, p. 118). Not quite. More elaborately, Poirier takes Baudrillard to task, not for developing a hypothesis like any other, but for failing to see that reproduction has always and forever been the only way in which human beings have been able to acknowledge reality and to talk about it. Reproduction, adds Poirier, "is part of the very inflections of daily speech" (1999, p. 119). Reproducing speech is an idea with some bite in it and I want to use the idea to show that the essential vicariousness of language can indeed be exploited to strengthen the empirical study of literature.

1. The word you read could have been another

Everyday experience shows that there are always different ways to convey the same information. In conversation, we often start sentences that we quickly rephrase using other words[2]. Manuscripts pass often through various draft states before the writer decides upon the final version. Says Auden (1962, p. 96), "The same person could be the author of two unrecognizably different autobiographies; in one, the writer could appear passive, lacking in a capacity for affection, easily bored and smaller than life-size, in the other active, a passionate Knight forever serenading Faith or Beauty, humorless and over-life-size".

According to Benson (1984), the final version of Ian Fleming's (1954) "*Live and Let Die*" is much less violent than the original one. The hero Felix Leiter, who was initially killed by a shark in the original version, survives in the final one (with only one arm and half of his leg gone). There is even a literary genre that exploits the possibility of offering alternative endings to the same story, as in John

[2] Indeed, each word of a text could be seen as a false start.

Fowles' (1969) "*The French Lieutenant's Woman*". In more ways than one thus, any text is always but the sample of another one and the words we read are but the visible part of an iceberg.

2. Report from the front

The 18 chapters of James Joyce's *Ulysses* (1922) were downloaded from the web site http://www.bibliomania.com/Fiction/Joyce/ulysses/ and content analyzed using the PROTAN system of computer-aided content analysis (Hogenraad, Daubies, & Bestgen, 1995). The novel is composed of 307,179 words and (an amazing) 29,184 different words. For reasons that need not occupy us now, we compared each word of the text with each word of Martindale's (1975) Regressive Imagery Dictionary. In very short, this dictionary allows one to assess the degree of presence, in texts, of primordial versus secondary or conceptual thought content. Primordial content[3] (1,827 words) are "found in the world" (*love, sex, food, chaos, dream, flying*, for example) and conceptual or logical content (713 words) are "built into the world" (*money, work, discipline, police, time, justice, law*, to name a few); both have been recognized to be ingredients of the cognitive functioning of the individual and of the social equilibrium of societies.

Each word of *Ulysses* was compared to all the words of the dictionary. Each time a match was found between a word in the text and an entry in the dictionary, that match was recorded and records were added up per chapter (and averaged by the total number of words in each chapter). In the case of *Ulysses*, the decision was made to weight only the rate (relative presence) of primordial thought content in the text and follow this rate across the chapters.

Figure 1 shows the unfolding of primordial thought content [$R^2 = .37$, $F(1, 16) = 9.45$, $p < .01$] over the 18 chapters of *Ulysses*.

[3] Regression is best described as a test of the limits of the possible (Eluard's sentence "*The earth is blue like an orange*"), a challenge to the bounds of social and biological reality ("*She will never make a good husband*"), an inversion of value based on a blurring of differences in general. It is driven by fantasy more than by reality, and by the ignorance of what constitutes institutions and societies.

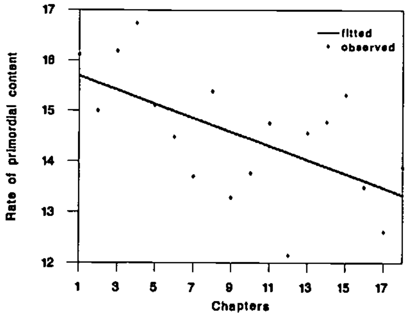

Figure 1. Primordial thought content of the 18 chapters of *Ulysses*, observed and fitted.

Pondering this negative linear profile makes you wonder about the whole enterprise of statistical analysis itself, how little it tells us, how brittle is a quantitative analysis of a literary work in the face of the uncertainty that surrounds such data and of what we can imagine for ourselves just by seeing a work of literature in the making[4]. We have heard enough about null hypothesis statistical testing to do us all for a lifetime. New ways to look at literary data consist of resampling thousands of time data that by nature exist in only one exemplar, bringing empirical study of literature down to what it is made of, inkings out, excisions, alterations, disappearances (Hogenraad, McKenzie, & van Peer, 1997; Hogenraad & McKenzie, 1999). The rates of primordial thought content were repeated 2,000 times, using the Simstat bootstrap algorithm (Péladeau, 1996)[5]. Figure 2 casts into sharp relief the distribution of the 2,000 R^2: 66 R^2 values are around a low .04 (extreme left part of the histogram), but this is irrelevant. The best contribution one can make to empirical studies of literature, therefore, is not to collect more *p*< something that doesn't prove a thing, but to

[4] Just think that Marcel Proust had excised 250 pages from the original manuscript of *Albertine Disparue* –which was eventually published in full, after his death, by his brother Robert Proust under the name of *La Fugitive* in *A la Recherche du Temps Perdu* (Proust, 1999).

[5] Resampling 2,000 times the values of primordial thought content in *Ulysses* is done before you can blink, i.e., in about 34 seconds on a Pentium III/ 500 MHz.

hammer scholars in order to see how well resampling allows one to assess the degree of uncertainty associated with any measurement. Figure 3 shows the fitted profiles resulting from 10 such replications: All 10 R^2 values are consistent –they range from .28 to .62–. It might have gone another way.

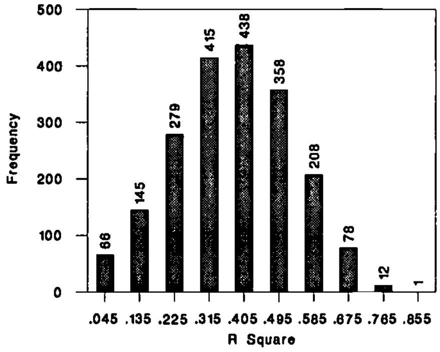

Figure 2. Sampling distribution of the R Square values obtained by 2,000 bootstrap repetitions (primordial thought content of *Ulysses*).

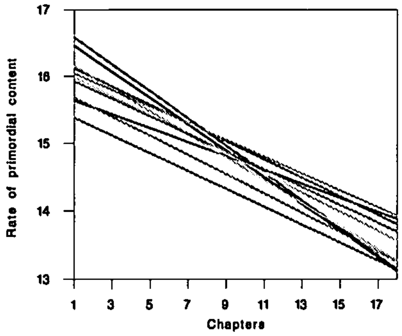

Figure 3. Linear profiles of the rate of primordial thought content of the 18 chapters of *Ulysses* obtained after 10 bootstrap repetitions, fitted values only.

3. Final note

Our ability to repeat a scientific result and to make it more credible is more important than the result per se (Shapin, 1996). Until now, despite much endeavor and writing (Cohen, 1994; Efron & Tibshirani, 1991), there is no indication that empirical studies of literature so much as noticed the difference that has taken place in processing textual data by a systematic exploration of uncertainty.

References

W. H. Auden, W. H. (1962). *The Dyer's Hand*. London: Faber and Faber.

Baudrillard, J. (1977). *Le miroir de la production ou L'illusion critique du matérialisme historique*. Paris: Casterman.

Benson, R. (1984). *The James Bond bedside companion*. New York: Dodd, Mead and Company.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*(12), 997-1003.

Efron, B., & Tibshirani, R. (1991, 26 July). Statistical data in the computer age. *Science, 253*, 390-395.

Fleming, I. (1954). *Live and let die*. London: Jonathan Cape.

Fowles, J. (1969). *The French lieutenant's woman*. London: Cape.

Hogenraad, R., Daubies, C., & Bestgen, Y. (1995). Une théorie et une méthode générale d'analyse textuelle assistée par ordinateur: Le système PROTAN (PROTocol ANalyzer), Version du 2 mars 1995 [A general theory and method of computer-aided text analysis: The PROTAN system (PROTocol ANalyzer), Version of March 2, 1995]. Unpublished document, Psychology Department, Université catholique de Louvain, Louvain-la-Neuve, Belgium, ii + 265 pages. (http://www.psp.ucl.ac.be/~upso/membres/aRH.html)

Hogenraad, R., & McKenzie, D. P. (1999). Replicating text: The cumulation of knowledge in social science. *Quality & Quantity, 33*(2), 97-116.

Hogenraad, R., McKenzie, D. P., & van Peer, W. (1997). Literature has no competitor. *SPIEL, 16*(1-2), 169-174.

Joyce, J. (1922/1990). *Ulysses*. London: Vintage.

Martindale, C. (1975). *Romantic progression: The psychology of literary history*. Washington, DC: Hemisphere.

Péladeau, N. (1996). *Simstat for Windows. User's guide* (Version 1.21d, November 1997). Montréal: Provalis Research. (http://www.simstat.com)

Poirier, R. (1999). *Trying it out in America: Literary and other performances*. New York: Farrar, Straus and Giroux.

Proust, R. and the Nouvelle Revue Française (1999). *Les années perdues de la* *recherche 1922-1931*. Correspondance pour l'édition des volumes posthumes d'*A la Recherche du temps perdu* (N. M. Dyer with A. Rivière and P.-E. Robert, eds.).Paris: Gallimard.

Shapin, S. (1996). *The scientific revolution*. Chicago: The University of Chicago Press.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

CS 217 209

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Once is not enough: Statistical simulation of textual data	
Author(s): Robert Hogenraad	
Corporate Source: 7th Congress, Int'l Society for the Empirical Study of Literature (IGEL), Toronto	Publication Date: July 31-Agust 4, 2000

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	**Level 2A**	**Level 2B**
↑ ☑	↑ □	↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.		
Signature:	Printed Name/Position/Title: Robert Hogenraad, Senior Research Associate, Nat'l Foundation for Scientific Research, Belgium	
Organization/Address: Psych. Dep., Université catholique de Louvain, 10 place du Card. Mercier, B-1348 Louvain-la-Neuve, Belgium	Telephone: +32-(0)10-474411	Fax: +32-(0)10-473774
	E-mail Address: hogenraad@upso.ucl.ac.be	Date: August 9, 2000

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC/REC Clearinghouse
2805 E 10th St Suite 140
Bloomington, IN 47408-2698
Telephone: 812-855-5847
Toll Free: 800-759-4723
FAX: 812-856-5512
e-mail: ericcs@indiana.edu
WWW: http://www.indiana.edu/~eric_rec/

EFF-088 (Rev. 9/97)