Word association measure

Bernard Desgraupes and Sylvain Loiseau
<bernard.desgraupes@u-paris10.fr>, <sylvain.loiseau@univ-paris13.fr>

April 16, 2013

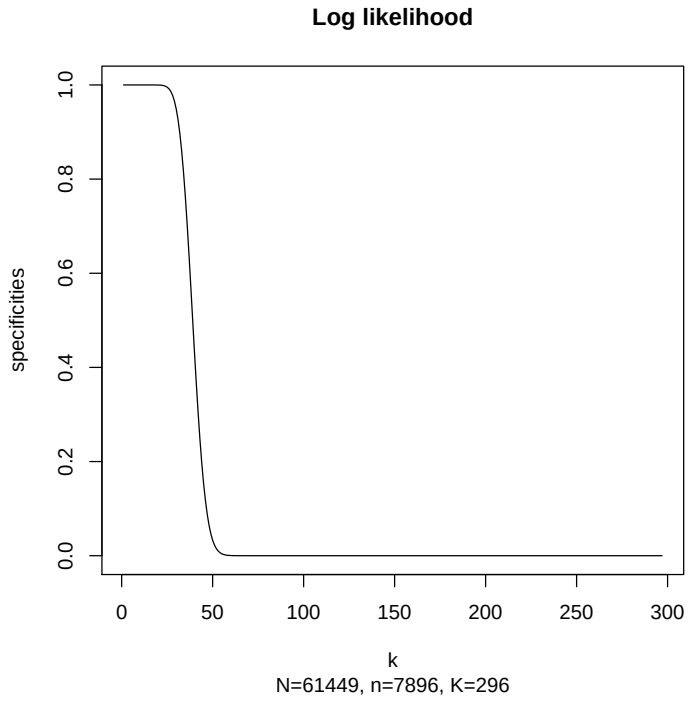
Abstract

Contents

1 Introduction

2 Indicator of word association

2.1 Log-likelihood



2.2 Specificities

This indicator has been proposed by Lafon in "Sur la variabilité de la fréquence des formes dans un corpus", *Mots*, 1, 1980, 127–165 (`http://www.persee.fr/web/revues/home/prescript/article/mots_0243-6450_1980_num_1_1_1008`).

It takes four arguments:

- N the total size of the corpus (in number of occurrences)
- n the size of the sub corpus (in number of occurrences)
- K the frequency of the form under scrutiny in the corpus
- k the frequency of the form under scrutiny in the subcorpus

Consider these parameters for the lexical form *peuple* in three public discourses by Robespierre in a corpus of 10 discourses containing $N = 61449$ occurrences in total (Lafon 1980) :

Discours	N	n	K	k
4	61449	6903	296	14
5	61449	7896	296	53
8	61449	2063	296	16

For each line we can compute the expected frequency of the form $(K \times n/N)$ and mark + if the form is more frequent than expected or − otherwise.

Discours	N	n	K	k	expected	$k > expected$
4	61449	6903	296	14	32.80	−
5	61449	7896	296	53	37.52	+
8	61449	2063	296	16	9.80	+

The form *peuple* is less frequent in the fourth discourse than expected. On the contrary, *peuple* is more frequent than expected in the fifth and eigth discourses.

If the observed frequency is less than the expected frequency, we compute the sum of the probability for a frequency lesser or equal to the observed frequency ($Prob(X \leq k)$). If the observed frequency is greater than the expected frequency, we compute the sum of the probability for a frequency greater to the observed frequency ($Prob(X > k)$) (Lafon 1980 : 152).

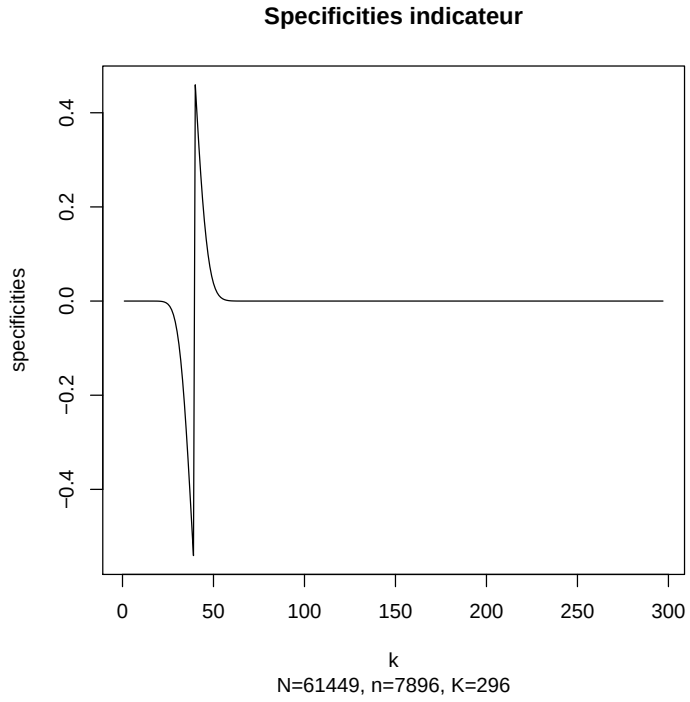
In both cases, the more unexpected is the frequency, the smaller is the indicator.

Discours	N	n	K	k	expected	$k > expected$	cumulative extreme probability
4	61449	6903	296	14	32.80	−	0.0000669371
5	61449	7896	296	53	37.52	+	0.0077234888
8	61449	2063	296	16	9.80	+	0.0433282491

According to this indicator, the second case is more "surprising" than the third or, in other terms, *peuple* is more attracted by, or specific to the second discourse than to the third.

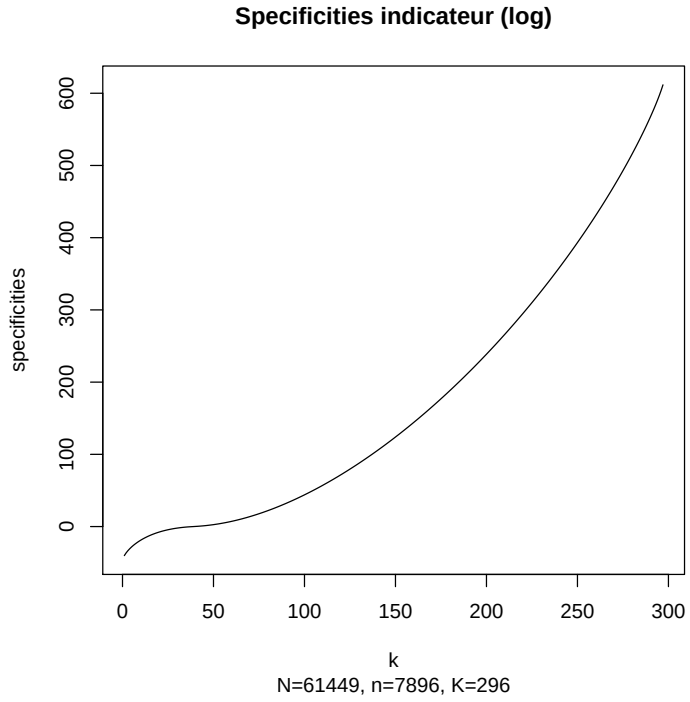
According to relative frequency, one could conclude the other way around: $53/7896 = 0.0067 < 16/2063 = 0.0078$ (cf. Lafon 1980 : 152).

For the fifth discourse above ($N = 61449$, $n = 7896$), the possible frequencies of *peuple* range from 0 to 296 (if all occurrences of *peuple* where in this discours). Here is the corresponding values for the specificities indicator:



2.3 Log of specificities

The log of the probability; with ”-” sign if specificity is negative and ”+” if it is positive.



3 Comparison of indicators

4 Distribution of the specificities of a form across sub-corpus

5 Bibliographie

KILGARRIFF, Adam, 1996. ■ Which words are particularly characteristic of a text? A survey of statistical approaches ■. Dans : Proc. AISB Workshop on Language Engineering for Document Analysis and Recognition. . pp. 33-40

Dunning, T. 1993. "Accurate methods for the statistics of surprise and coincidence." Computational Linguistics. 19(1). Pp 61–74.

Hofland, K. and Johanssen, S. 1989. Frequency analysis of English vocabulary= and grammar, based on the LOB corpus. Oxford: Clarendon.

Kilgarriff, A. 1996. "Which words are particularly characteristic of a text? A survey of statistical approaches." Proceedings, ALLC-ACH '96. Bergen, Norway.
`http://www.cse.iitb.ac.in/~shwetaghonge/prec_recall.pdf`

DUNNING, Ted, 1993. ■ Accurate methods for the statistics of surprise and coincidence ■. Dans : Computational linguistics. 19/1. MIT Press, pp. 61-74
`http://acl.ldc.upenn.edu/J/J93/J93-1003.pdf`

CHAUDHARI, Dipak L, DAMANI, Om P & LAXMAN, Srivatsan 2011 ■ Lexical co-occurrence, statistical significance, and word association ■ Dans : Proceedings of the Conference on Empirical Methods in Natural Language Processing (Edinburgh, Scotland, UK, July 27-31). pp. 1058-68

http://www.aclweb.org/anthology-new/D/D11/D11-1098.pdf