# wam package: computing Word association measure

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

October 26, 2015

**Abstract**

---

## Contents



---

# 1 Introduction: Indicators of word association

```
> library(wam);
```

## 1.1 Introduction

This package contains various functions for computing word association measure as well as a high level function for conveniently applying these functions on all the lexical types of a corpora.

## 1.2 Value

All association measure functions are prefixed with "wam." and return a numeric vector indicating the association strengh between the word(s) under scrutiny and subcorpus/ora.

These association measures, unless otherwise stated in the help pages of the functions, are positive when the word is over-represented ("attracted"), and negative when the word is under-represented.

In absolute value, the more the word is over-representend or under-represented, the more the association measure givien is high.

## 1.3 Arguments

All association measure functions have the following signature: ($N$, $n$, $K$, $k$), where:

1. $N$ is the total size of the corpora
2. $n$ is the size of the subcorpora
3. $K$ is the total frequency of the form under scrutiny in the corpora
4. $k$ is the sub-frequency of the form under scrutinty in the subcorpora

This can be easily turn into the "contingency table" representation used in some presentation (according to Stefan Evert UCS documentation) :

| | word | $\neg word$ | T |
|---|---|---|---|
| subcorpus | O11<br>E11 | O12<br>E12 | R1 |
| $\neg subcorpus$ | O21<br>E21 | O22<br>E22 | R2 |
| Totals | C1 | C2 | N |

where :

- N = total words in corpus (or subcorpus or restriction, but they are not implemented yet)
- C1 = frequency of the collocate in the whole corpus

- C2 = frequency of words that aren't the collocate in the corpus
- R1 = total words in window
- R2 = total words outside of window
- O11 = how many of collocate there are in the window
- O12 = how many words other than the collocate there are in the window (calculated from row total)
- O21 = how many of collocate there are outside the window
- O22 = how many words other than the collocate there are outside the window
- E11 = expected values (proportion of collocate that would belong in window if collocate were spread evenly)
- E12 = " " (proportion of collocate that would belong outside window if collocate were spread evenly)
- E21 = " " (proportion of other words that would belong in window if collocate were spread evenly)
- E22 = " " (proportion of other words that would belong outside window if collocate were spread evenly)

Conversion from $N$, $n$, $K$, $k$ notation :

| | word | $\neg word$ | $T$ |
|---|---|---|---|
| subcorpus | $k$ | $n-k$ | $n$ |
| $\neg subcorpus$ | $K-k$ | $N-K-(n-k)$ | $N-n$ |
| Totals | $K$ | $N-K$ | $N$ |

Conversion to $N$, $n$, $K$, $k$ notation :

- $N = N$
- $n = O11 + O12$
- $K = O11 + O21$
- $k = O11$

## 1.4 Recycling arguments

For all functions arguments are recycled.

## 2 The indicators

### 2.1 Introduction

All word association functions will be illustrated with data from the robespierre dataset. We will consider the subfrequency of the lexical types *peuple* in a subcorpora containing the fourth discourse by Robespierre. Is this type over- or under-represented in this discouse, according to its frequency in the corpus of all the discourses?

```
> data(robespierre, package="wam")
> head(robespierre)

  types parts   k     N    K    n
1    de    D1 464 61449 3173 8395
2    la    D1 365 61449 2788 8395
3   les    D1 281 61449 2123 8395
4    et    D1 227 61449 1708 8395
5    le    D1 200 61449 1351 8395
6     l    D1 188 61449 1287 8395

> peuple_D4 <- robespierre[robespierre$types=="peuple" & robespierre$parts == "D4",]
> peuple_D4

     types parts  k     N   K    n
495 peuple    D4 14 61449 296 6903

> N <- peuple_D4$N
> n <- peuple_D4$n
> K <- peuple_D4$K
> k <- peuple_D4$k
```

A graph of the function is provided for all the possible values of $k$. The interval of the possible value of $K$ is $[0, min(k, n)]$. (it is not possible to have more occurrences of the lexical type than the size of the subcorpus or more than the total frequency of the type in the whole corpus).

```
> maxk <- min(K,n)
> maxk

[1] 296

> allk <- 0:maxk
```

The expected subfrequency of "peuple" if the subcorpus D4 (the fourth discourse by Robespierre) is: $K \times n/N$. The mode is the closest positive integers to the expected frequency.

```
> expected = round(K * n / N, 2)
> expected
```

```
[1] 33.25
```

```
> mode <- floor((n+1)*(K+1)/(N+2));
> mode
```

```
[1] 33
```

A form is over-used (attracted) if the subfrequency $k$ in the subcorpus is greater than the expected expected frequency and under-used otherwise. Here, the form peuple is under-represented.

## 2.2 Log-likelihood

See Dunning 1993.

```
> wam.loglikelihood(N, n, K, k);
```

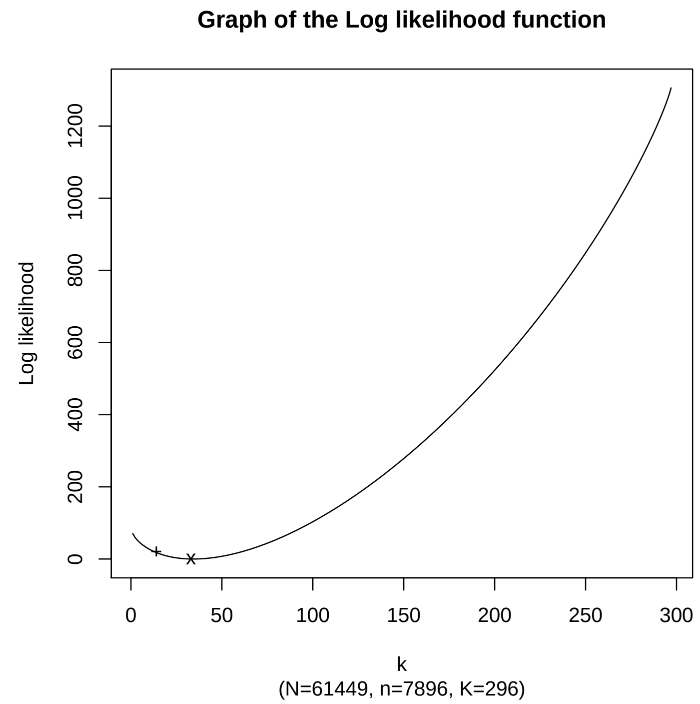
```
[1] 15.7202
```

```
> wam.loglikelihood(N, n, K, mode);
```

```
[1] 0.002162702
```

Graph of the function :

```
> plot(wam.loglikelihood(N, n, K, allk),
+      type="l", xlab="k", ylab="Log likelihood",
+           main="Graph of the Log likelihood function",
+           sub="(N=61449, n=7896, K=296)")
> points(k, wam.loglikelihood(N, n, K, k), pch="+")
> points(mode, wam.loglikelihood(N, n, K, mode), pch="x")
```




## 2.3 Specificities

See Lafon 1980.

```
> wam.specificities(N, n, K, k, method="base");
```

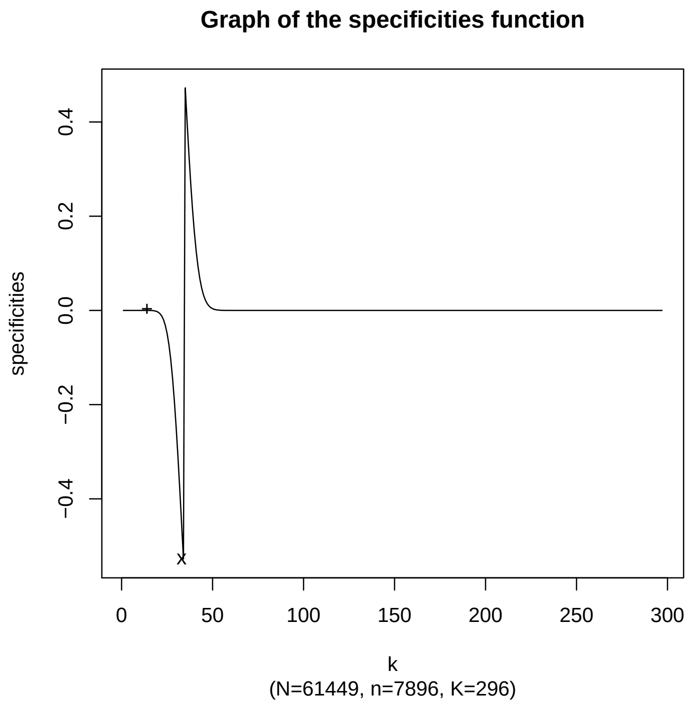
```
[1] -6.693709e-05
```

```
> wam.specificities(N, n, K, mode, method="base");
```

```
[1] -0.5276713
```

Graph of the function:

```
> plot(wam.specificities(N, n, K, allk, method="base"),
+      type="l", xlab="k", ylab="specificities",
+           main="Graph of the specificities function",
+           sub="(N=61449, n=7896, K=296)")
> points(k, wam.specificities(N, n, K, k, method="base"), pch="+")
> points(mode, wam.specificities(N, n, K, mode, method="base"), pch="x")
```

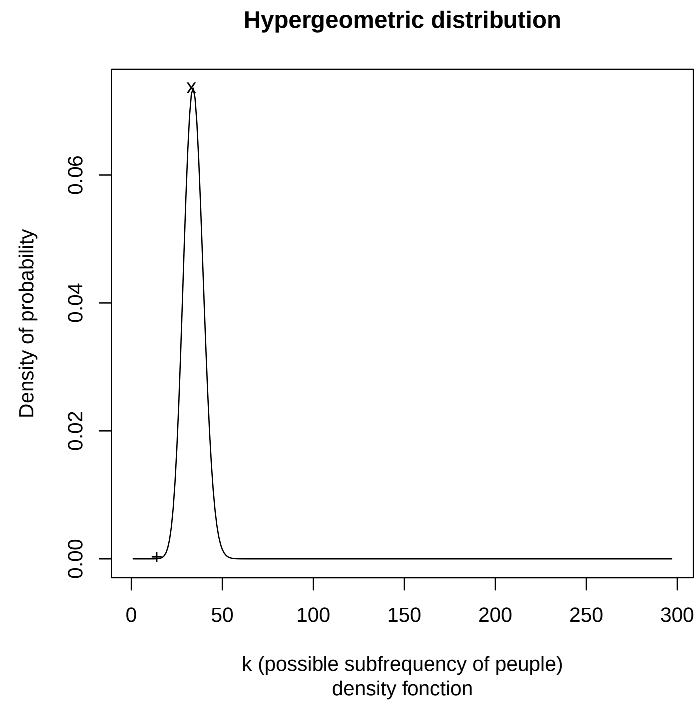



#### 2.3.1 Analysis of the specificities indicator : Standard indicator (method="base")

The presentation below follows (Lafon, 1980).

The hypergeometric distribution give the probability associated with a drawing without replacement. For all the possible subfrequencys of *peuple* in the fourth discourses we can compute the density of probability in the hypergeometric distribution. The graph contains also the observed frequency as well as the mode.

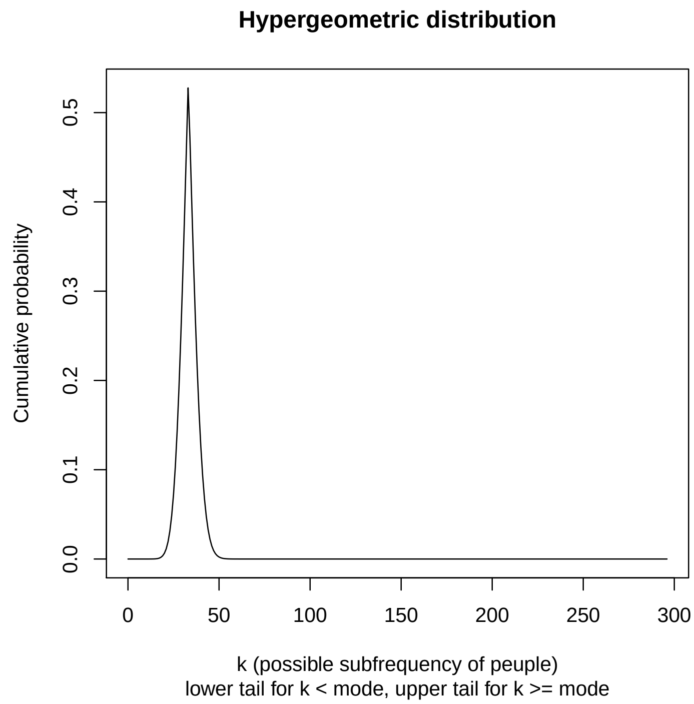
```
> plot(dhyper(allk, K, N-K, n),
+ type="l", xlab="k (possible subfrequency of peuple)", ylab="Density of probability",
+ main="Hypergeometric distribution", sub="density fonction")
> points(k, dhyper(k, K, N-K, n), pch="+")
> points(mode, dhyper(mode, K, N-K, n), pch="x")
```




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 ($Prob(X \leq k)$) – that is, the cumulative probability.

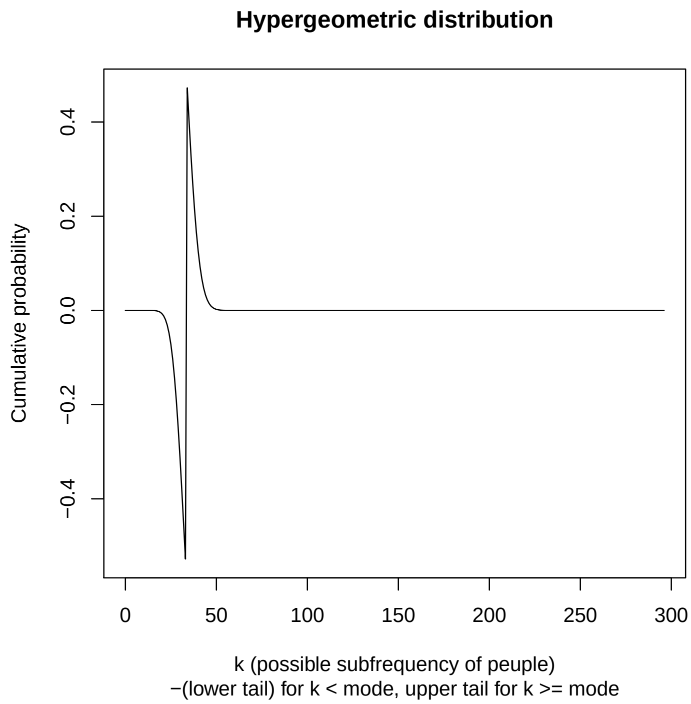
If the observed frequency is greater than the expected frequency, we compute the sum of the probability for a frequency greater to the observed frequency ($Prob(X > k)$) (Lafon 1980 : 141) – that is, the cumulative probability for the upper tail of the distribution.

```
> y <- ifelse(allk <= mode, phyper(allk, K, N-K, n),
+                           phyper(allk-1, K, N-K, n, lower.tail=FALSE))
> plot(allk, y,
+         type="l", xlab="k (possible subfrequency of peuple)",
+         ylab="Cumulative probability",
+         main="Hypergeometric distribution",
+         sub="lower tail for k < mode, upper tail for k >= mode")
```



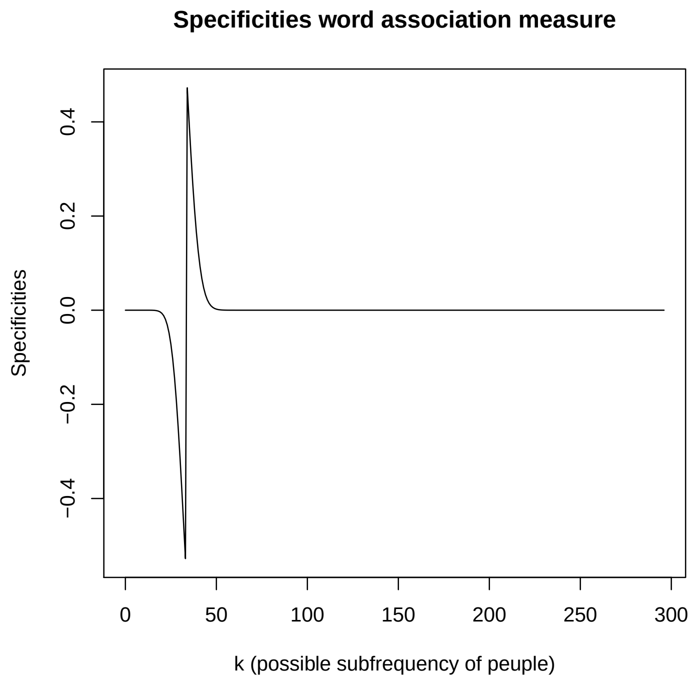

We add a sign: negative if the frequency is lower than expected, positive if it is greater.

```
> y <- ifelse(allk <= mode, phyper(allk, K, N-K, n),
+                           phyper(allk-1, K, N-K, n, lower.tail=FALSE))
> y <- ifelse(allk <= mode, -y, y);
> plot(allk, y,
+        type="l", xlab="k (possible subfrequency of peuple)",
+        ylab="Cumulative probability",
+        main="Hypergeometric distribution",
+        sub="-(lower tail) for k < mode, upper tail for k >= mode")
```




It is the standard Specificities function (Lafon 1980), as implemented in the function wam.specificities with method="base" :

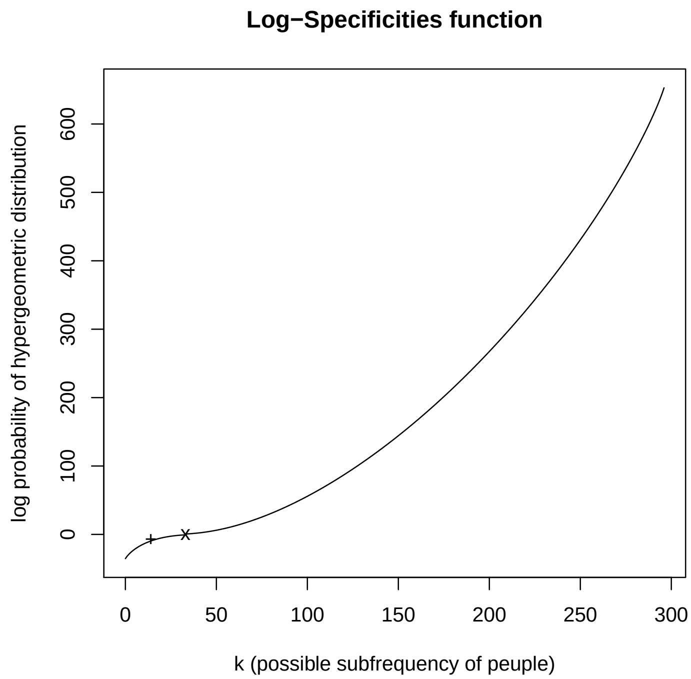
```
> plot(allk, wam.specificities(N, n, K, allk, method="base"),
+         type="l", xlab="k (possible subfrequency of peuple)",
+         ylab="Specificities",
+         main="Specificities word association measure")
```




#### 2.3.2 Analysis of the specificities indicator : log (method="log")

In order to ease the reading, log are used:

```
> y <- ifelse(allk <= mode, phyper(allk, K, N-K, n, log.p=TRUE),
+                          phyper(allk-1, K, N-K, n, lower.tail=FALSE, log.p=TRUE))
> y <- ifelse(allk <= mode, -abs(y), abs(y));
> plot(allk, y,
+         type="l", xlab="k (possible subfrequency of peuple)",
+         ylab="log probability of hypergeometric distribution",
+         main="Log-Specificities function");
> points(k, phyper(k, K, N-K, n, log.p=TRUE), , pch="+")
> points(mode, phyper(mode, K, N-K, n, log.p=TRUE), pch="x")
```



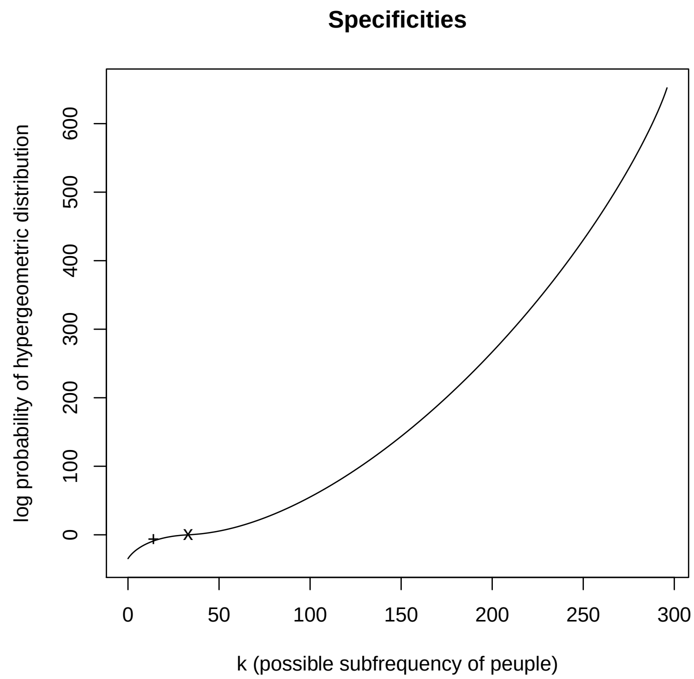

Another issue is that the mode is different from 0:

```
> phyper(mode, K, N-K, n, log.p=TRUE)
```

```
[1] -0.6392818
```

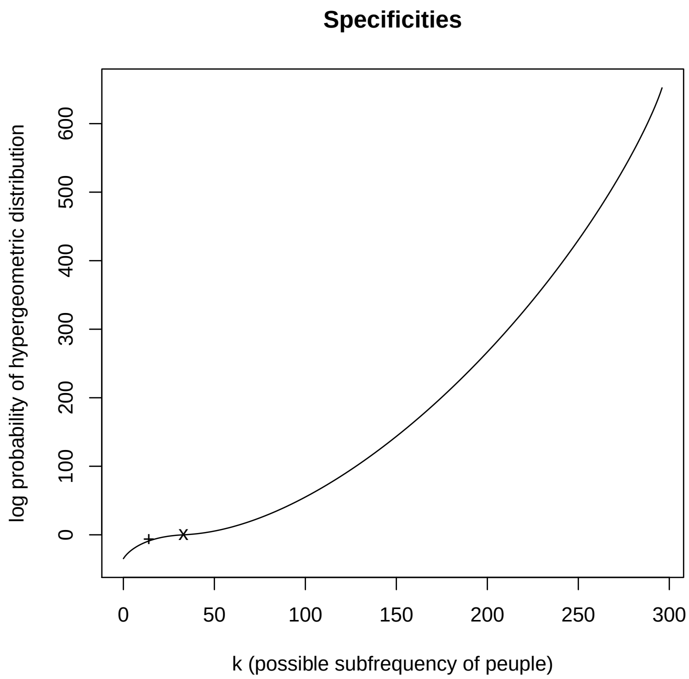
in order to have $mode = 0$, the value of the mode is substracted from all values:

```
> y <- ifelse(allk <= mode, phyper(allk, K, N-K, n, log.p=TRUE),
+                          phyper(allk-1, K, N-K, n, lower.tail=FALSE, log.p=TRUE))
> cdmo <- phyper(mode, K, N-K, n, log.p=TRUE);
> y <- ifelse(allk <= mode, -abs(cdmo-y), abs(cdmo-y));
> plot(allk, y,
+         type="l", xlab="k (possible subfrequency of peuple)",
+         ylab="log probability of hypergeometric distribution",
+         main="Specificities")
> points(k, wam.specificities(N, n, K, k, method="log"), pch="+")
> points(mode, wam.specificities(N, n, K, mode, method="log"), pch="x")
```




That is the wam.specificities function with method="log":

```
> plot(allk, wam.specificities(N, n, K, allk, method="log"),
+        type="l", xlab="k (possible subfrequency of peuple)",
+        ylab="log probability of hypergeometric distribution",
+        main="Specificities")
> points(k, wam.specificities(N, n, K, k, method="log"), pch="+")
> points(mode, wam.specificities(N, n, K, mode, method="log"), pch="x")
```




where the mode is 0 :

```
> wam.specificities(N, n, K, mode, method="log");
```

```
[1] 0
```

## 2.4 z

```
> wam.z(N, n, K, k);
```

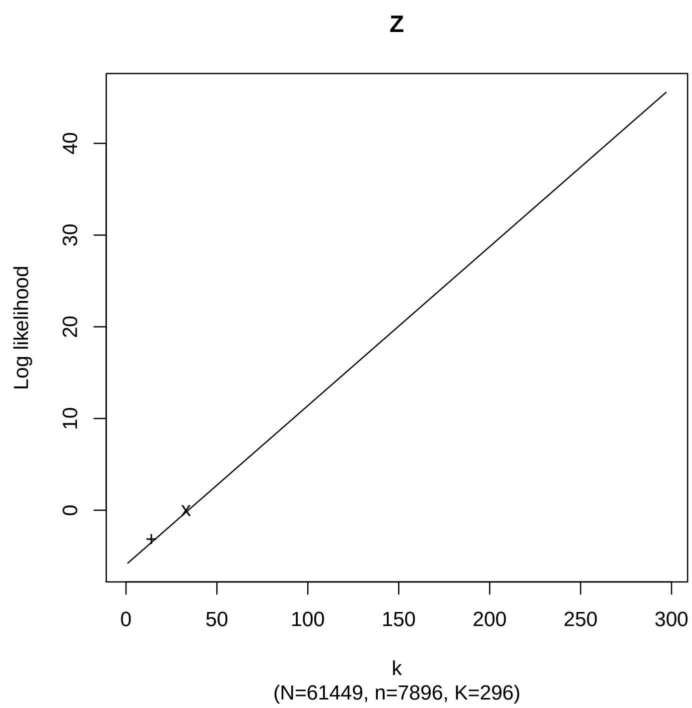
```
[1] -3.338591
```

```
> wam.z(N, n, K, mode);
```

```
[1] -0.04366125
```

Graph of the function :

```
> plot(wam.z(N, n, K, allk),
+       type="l", xlab="k", ylab="Log likelihood",
+            main="Z",
+            sub="(N=61449, n=7896, K=296)")
> points(k, wam.z(N, n, K, k), pch="+")
> points(mode, wam.z(N, n, K, mode), pch="x")
```




## 2.5 t

```
> wam.t(N, n, K, k);
```

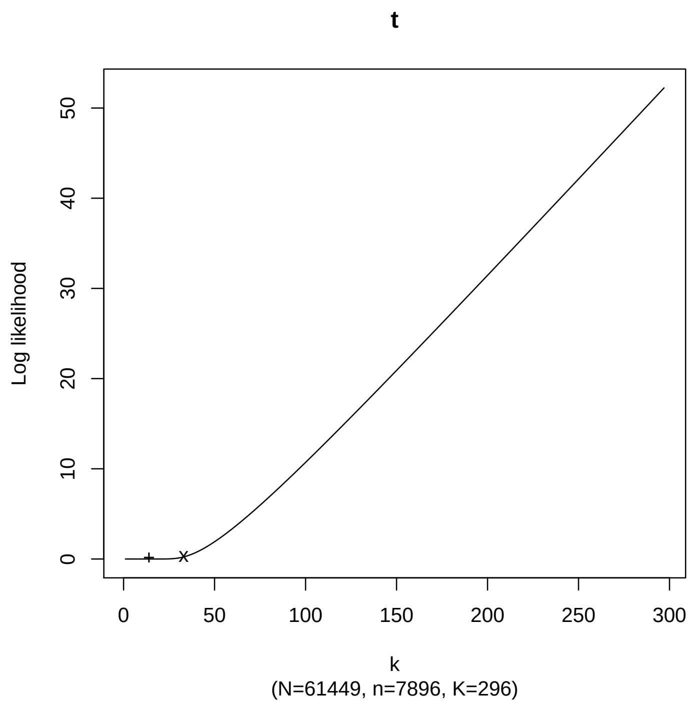
```
[1] 5.801351e-08
```

```
> wam.t(N, n, K, mode);
```

```
[1] 0.2861072
```

Graph of the function :

```
> plot(wam.t(N, n, K, allk),
+       type="l", xlab="k", ylab="Log likelihood",
+            main="t",
+            sub="(N=61449, n=7896, K=296)")
> points(k, wam.t(N, n, K, k), pch="+")
> points(mode, wam.t(N, n, K, mode), pch="x")
```




## 2.6 chisq

```
> wam.chisq(N, n, K, k);
```

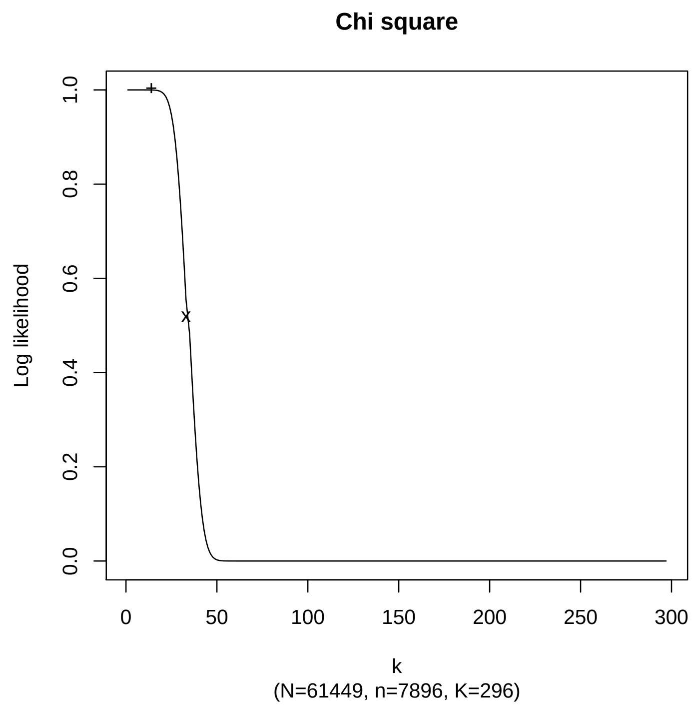
```
[1] 0.9997298
```

```
> wam.chisq(N, n, K, mode);
```

```
[1] 0.5182654
```

Graph of the function :

```
> plot(wam.chisq(N, n, K, allk),
+      type="l", xlab="k", ylab="Log likelihood",
+           main="Chi square",
+           sub="(N=61449, n=7896, K=296)")
> points(k, wam.chisq(N, n, K, k), pch="+")
> points(mode, wam.chisq(N, n, K, mode), pch="x")
```




## 2.7 fisher

```
> wam.fisher(N, n, K, k);
```

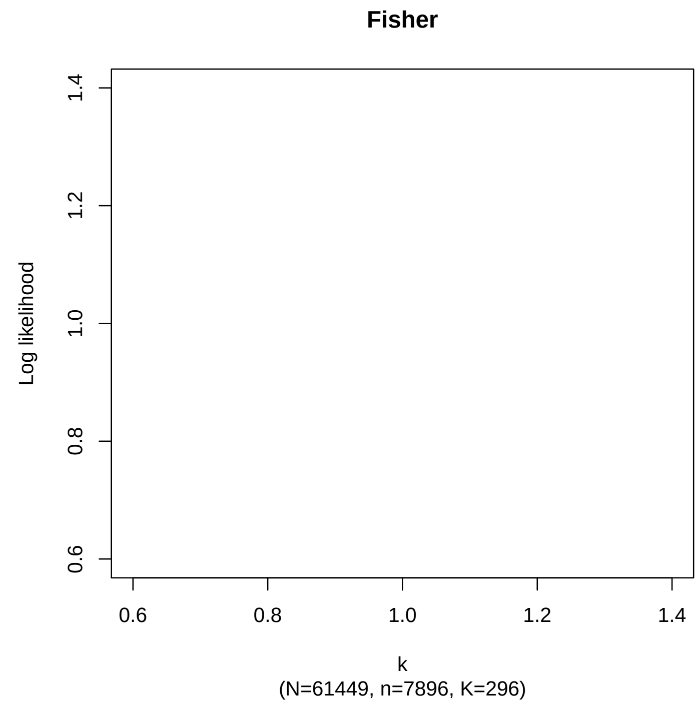
```
[1] 0.0001353407
```

```
> wam.fisher(N, n, K, mode);
```

```
[1] 1
```

Graph of the function :

```
> plot(wam.fisher(N, n, K, allk),
+       type="l", xlab="k", ylab="Log likelihood",
+            main="Fisher",
+            sub="(N=61449, n=7896, K=296)")
> points(k, wam.fisher(N, n, K, k), pch="+")
> points(mode, wam.fisher(N, n, K, mode), pch="x")
```




## 3 High level interface

The function *wam* provide a more high level interface to the actual functions.

This function allows for computing several indicators at the same time.

It takes also as argument the subcorpus name and lexical types.

```
> rm(list=ls())
> data(robespierre, package="wam")
> attach(robespierre)
> wam.res <- wam(N, n, K, k, measure=c("loglikelihood", "specificities"),
+ types=types, parts=parts)
```

function allows for retrieving the basic information (see the manual page for WordAssociation):

```
> head(k(wam.res));
```

```
[1] 464 365 281 227 200 188
```

```
> head(association(wam.res));
```

```
     loglikelihood specificities
[1,]     2.5758482     2.1848933
[2,]     0.8133694    -0.9840740
[3,]     0.3408187    -0.5627793
[4,]     0.2069236    -0.3991217
[5,]     1.4940429     1.4588965
[6,]     0.9787803     1.0712966
```

```
> indicator.name(wam.res)
```

```
[1] "loglikelihood" "specificities"
```

The print function allows for an easy to use reading of the results. See :

```
> wam.res <- wam(N, n, K, k, measure=c("loglikelihood", "specificities"),
+ types=types, parts=parts)
> print(wam.res, from=1, to=10, parts="D4");
```

```
Printing association measure for 1 part(s); from: 1 to: 10
Corpus size: 61449
Sorted by: loglikelihood
-----------------------------------------------------------------------------
word                | sub freq | tot freq |   loglikelihood |   specificities
-----------------------------------------------------------------------------
.............................................................................
Part name: D4
Part size: 6903 tokens.
Positive specificities printed: 10
Negative specificities printed: 0
bourdon             |       20 |       20 |           87.50 |           43.25
nous                |        3 |      430 |           79.45 |          -40.51
vous                |        6 |      424 |           63.13 |          -32.58
salut               |       33 |       91 |           39.00 |           21.12
comité              |       35 |      103 |           37.31 |           20.20
fabre               |       13 |       19 |           34.59 |           18.43
convention          |       37 |      126 |           30.52 |           16.88
public              |       32 |      108 |           26.84 |           14.98
ils                 |       21 |      419 |           20.14 |          -11.35
il                  |      105 |      605 |           20.12 |           11.70
```

# 4 Bibliographie


Chaudhari, D. L., Damani, O. P. & Laxman, S. 2011. ”Lexical co-occurrence, statistical significance, and word association”. In: 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

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

http://acl.ldc.upenn.edu/J/J93/J93-1003.pdf

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." In: Proceedings, ALLC-ACH '96. Bergen, Norway.

http://www.cse.iitb.ac.in/~shwetaghonge/prec_recall.pdf

Lafon P. 1980. "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.