Untangling Associations
Advances in Collocation and Keyword Analysis
Université Paul Valéry – Montpellier 3, 22 septembre 2023

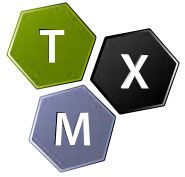
The Specificity Measure in Textometry

A Hermeneutic Use of the Fisher's Exact Test

Bénédicte PINCEMIN
(Univ. Lyon, CNRS, IHRIM UMR5317 – TXM Team)

Outline

- Why (still) choose the Fisher Yates Exact Test?
  - Assets; Criticisms; Alternatives
- How are Specificities implemented and used?
  - A tour of design features illustrated by a research on the SHOAH corpus powered by the TXM software



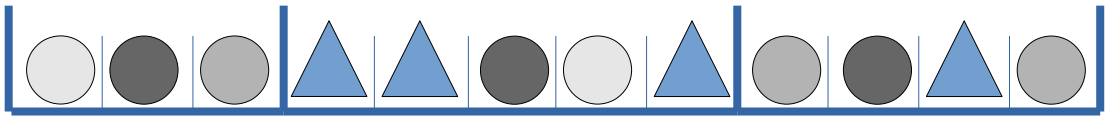
- Computational Text Analysis Methods
  - Corpus Linguistics, Text Mining, Distant Reading, Textometry…: A diversity of approaches serving +/- different expectations and aims?

The Fisher Yates Exact Test

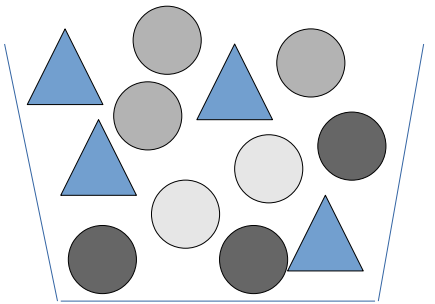
- Textometric **Specificities** (S+) (Lafon 1980), or characteristic elements (Lebart et al. 1998), implement a Fisher Yates Exact Test (FYE) (Pedersen 1996; Stefanowitsch & Gries 2003; Evert 2004;...).
- Not out of date, even fully relevant
- 2 main assets:
  - **Clear**: transparent, meaningful – a direct translation of the linguistic question into a mathematical model
    - It is not a question of being the best(?), but of understanding what is measured
  - **Reliable**: “exact” = nonparametric = no external assumption about the underlying probability distribution or about the value of a parameter
    - No validity limit for low frequencies – the full range of frequencies is managed
    - The measure embeds a statistical evaluation (no need for confidence intervals)

The mathematical model

- The corpus is a set of parts (containers) filled with words (content).



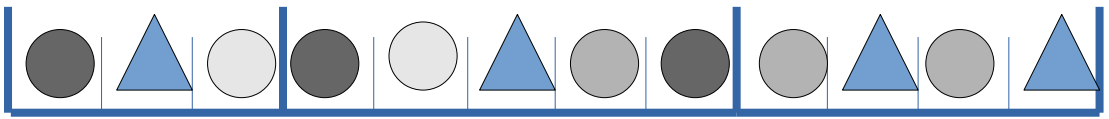
- Let's remove words from parts:
  (the total frequency for each word and part sizes are retained)



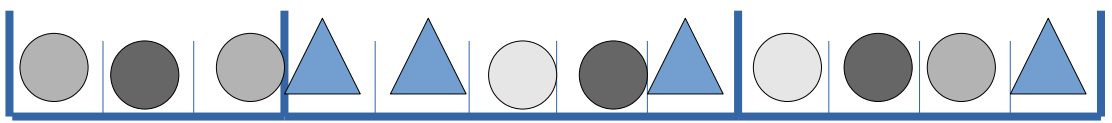
Corpus Vocabulary

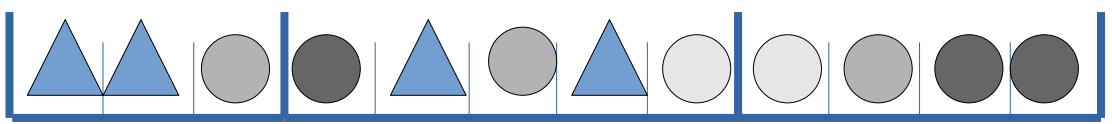
The mathematical model

- Then re-allocate the words randomly (only allocation changes: the total frequency for each word and part sizes are retained)



- Let's generate many such instances, actually all possible combinations of words into parts:





Etc.

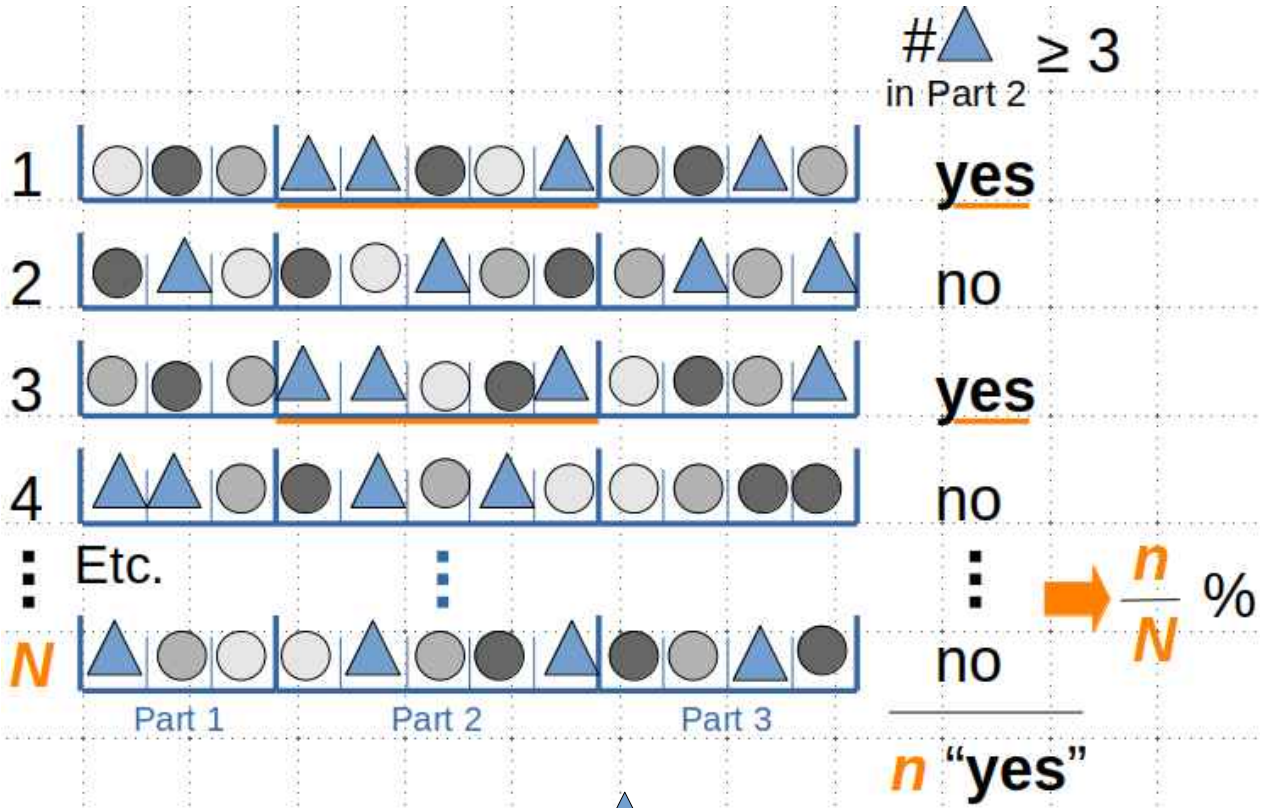
The mathematical model

- Then, for any given word in any given part,

  you can measure how rare/surprising it would be (on a random basis)

  to reach the frequency you observe

  by the **proportion of all instances with such a frequency** (or more).



Example for the word in Part 2 with an observed frequency of 3.

The mathematical model

- The proportion can be directly computed with a formula
  - A rather complex formula... :-(
  - ...Available open-source implementations :-)
    - R: Hypergeometric Distribution, used in textometry package, corpora package,...)

- It requires 4 parameters:
  - *T*: total size of the corpus (number of tokens)
  - *t*: size of the part
  - *F*: total frequency of the word (absolute number of occurrences)
  - *f*: frequency of the word in the part

From Proportion (%) to Score notation (S+, S-)

- For better readability (since proportions here are often very low), proportions are converted to their **order of magnitude** (Log_{10}):

  If $n/N = 0.1$ then $S+ = |Log_{10}(0.1)| = |Log_{10}(10^{-1})| = 1$
  if $n/N = 0.01$ then $S+ = |Log_{10}(0.01)| = |Log_{10}(10^{-2})| = 2$, etc.
  if $n/N = 0.15$ then $S+ = |Log_{10}(0.15)| = |Log_{10}(10^{-0.8})| = 0.8$

- Positive or negative **sign**: In case the frequency is *less* than the frequency for equal distribution, we compute a *negative* Specificity from the proportion of cases with the observed frequency *or less*.
  - Negative specificities denote words with especially low frequencies (that is, less than statistically expected)
  - Positive specificities denote words with especially high frequencies (that is, greater than statistically expected)

A clear and meaningful measure

- The interpretation of a specificity score is straightforward, for instance:
  - A Specificity of **+4** means that, if words were distributed randomly, there would be a 1 in 1**0,000** chance to get this frequency or **more** (0.01 % of all possible word allocations reach this frequency).
  - A Specificity of **-2** means that, if words were distributed randomly, there would be a 1 in 1**00** chance to observe this frequency or **less** (1 % of all word allocations with such a low frequency in the part).

A gauge rather than a predictive model

- The model is mathematically exact,
  but it is not linguistically realistic: the aim is
  - not to model language
    - word occurrences are not independent events, since there are obvious contextual, syntactic and semantic interconnections
    - scores cannot be understood as lexical or linguistic probabilities
  - but to get a clear benchmark, a measuring tool

A gauge rather than a predictive model

- However, scores are used as both absolute and relative indicators
  - Absolute threshold: ~3 (min. 2)
    - words with a score less than 2 ($n/N=p$=0.01, 1%) or even 3 ($n/N=p$=0.001, 1 ‰) are poor candidates, since their frequency can be due to common fluctuations;
    - the p=0.05 (5 %) usual statistical threshold is inadequate because
      - language doesn't work randomly (too many words would be identified as outliers)
      - problem of multiple comparisons: raising the threshold is a way to deal with this problem.
  - Relative ranking
    - sort in descending score and focus on top words.

Criticisms about the FYE test (1/4)

- Bias toward high frequency words, correlation to frequency
  - This is a natural consequence of a statistical approach:
    the more occurrences you observe,
    the more confident you are in your judgment,
    the lower the probability can be when a deviation is observed.
  - This must be taken into account when interpreting results:
    S does not replace *F*, *f*, *t*, etc.
    - A kind of tupleization (Gries 2019)
    - Not a one-fits-all measure: Scores cannot be directly compared and do not provide any absolute qualification.

Criticisms about the FYE test (2/4)

- Conversely, low frequency words are penalized
  - Low frequency words may not be able to reach the significance threshold even in the case of a notable effect size
  - There is a (good) reason for this:
    the measure embeds useful statistical considerations, that is, are there enough occurrences to make a quantitative judgment.
    - For few occurrences or short pieces of text, one cannot exclude that a relatively high (or low) frequency would be due to **common fluctuations**
  - Here again, completing FYE results with frequencies is useful (tupleization).

Criticisms about the FYE test (3/4)

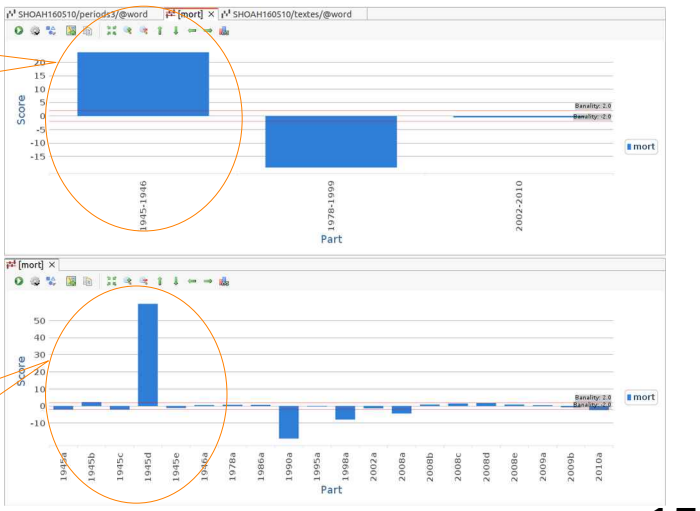
- Complex and intensive computation
  - The formula is complex:
    - However for users' hermeneutic concerns, the main thing is to understand the *model*, not necessarily the details of the formula?
  - Open-source efficient implementations are available

  - The advancing of current hardware power makes this calculation accessible
    - What used to be a challenge 40 years ago is no longer an impediment

Criticisms about the FYE test (4/4)

- A bag-of-word model – the Dispersion of a word inside a part is not taken into account
  - Here again, tupleization!: considering both Specificity score and a Dispersion measure, even if only the number of texts in the part in which the word occurs for instance.
  - Another solution consists in **recursively applying** the Specificity computing at different scales, typically on the *part* level then on the *text* level.
    - Note that this solution is not a change in the model, rather in the way of using it.

1 part



S+ for only 1 in 6 texts

Alternative Association measures

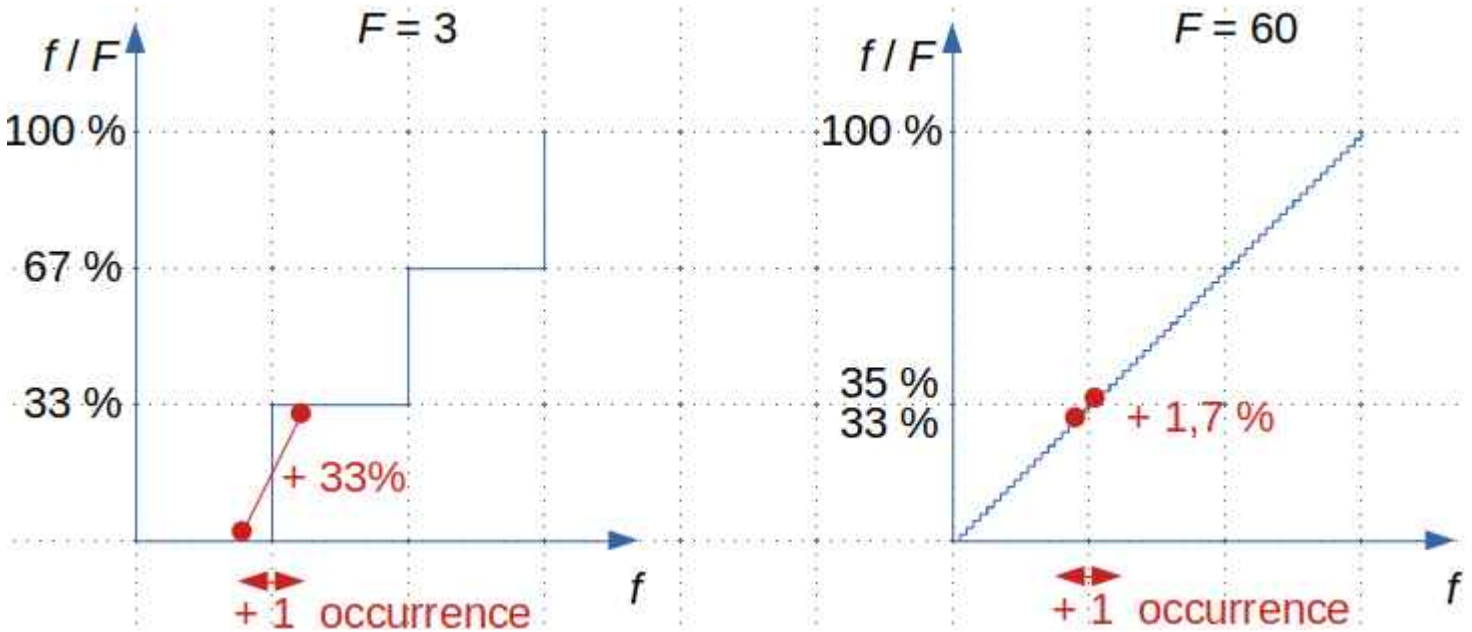
- *Parametric* statistical measures
  - Log-Likelihood, Chi-squared, z-score...
- Focus on measures that compare relative frequencies
  - Odds Ratio and variants
    - Relative frequency comparison is equivalent to coverage comparison:
      $(f / t) / (F / T) = (f \times T) / (t \times F) = (f / F) / (t / T)$
  - Clear and meaningful – intuitive
    - + simple formula
    - another way of defining overuse as related to density

Hesitations about relative frequencies (1/3)

- Proportionality assumption
  - *f* in *t* (the short text) = 2*f* in 2*t* (the long text)
  - this cannot be the case actually, because in usual lexical distributions, about half of the text's words (word types) occur once (hapaxes)
    - If every word frequency is multiplied, then no word has a unique occurrence -no hapaxes
  - This hypothesis favors word occurrences in short texts

Hesitations about relative frequencies (2/3)

- “Stairs” phenomenon for low frequency words
  - every occurrence is heavily loaded → the measure is very responsive to the addition or withdrawal of a single occurrence



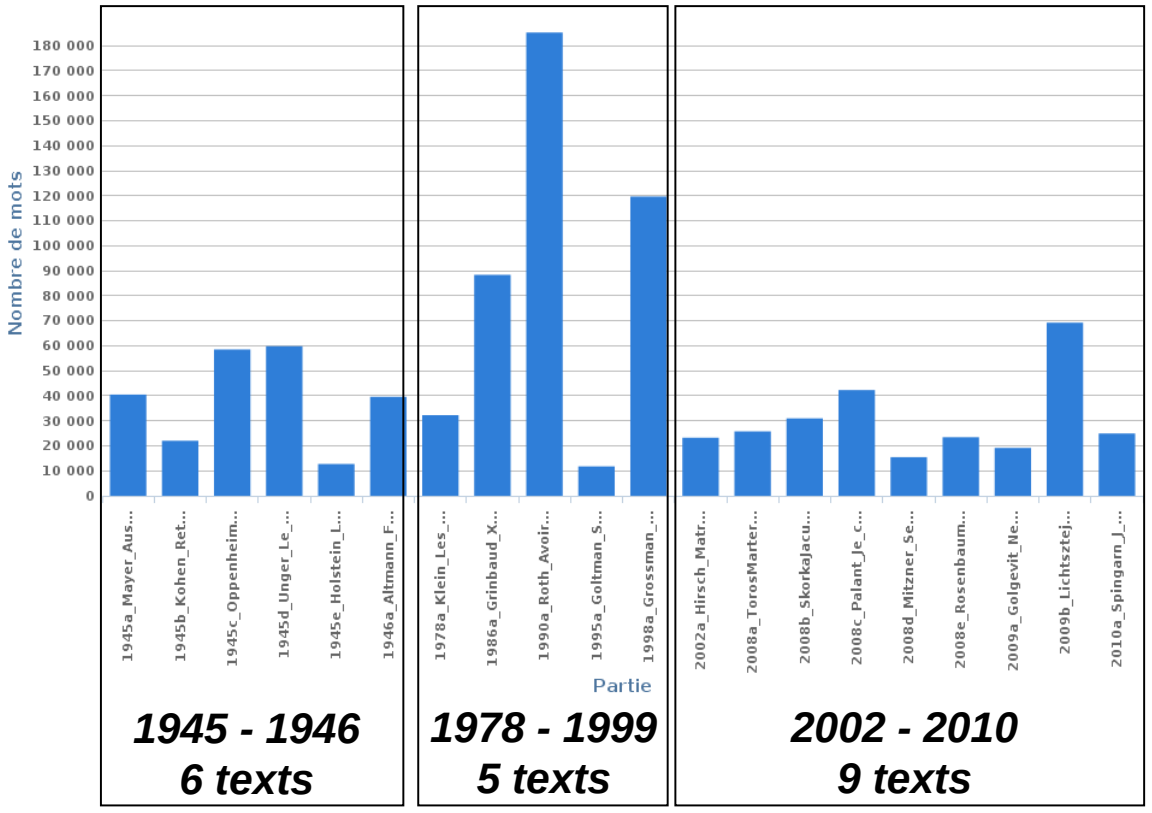
Hesitations about relative frequencies (3/3)

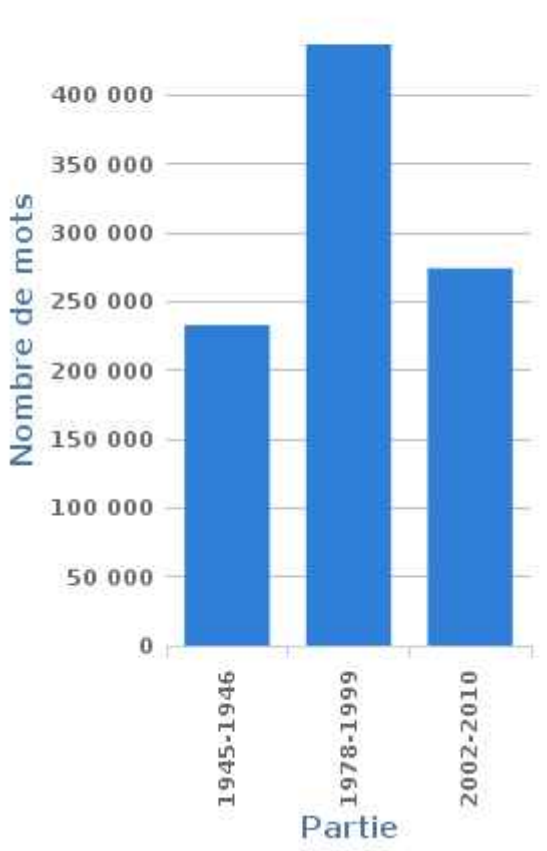
- Maybe too intuitive, actually?
  - somehow on the surface, grasps what you see
    - ex. words that occur in only 1 part (document hapaxes)
  - whereas a statistic measure may reveal something that is not so much visible
    - ex. a tendency for a frequent word to be less used
    - ex. words that do not occur and for which this absence could draw our attention (nullax)
  - This depends on what you are looking for, what kind of associations interest you
    - do not miss visible, ex. keyword extraction?
      =/= detect invisible, ex. authorship attribution?

Specificities in textometric practice

- Tupleization in:
  - Ergonomy of the graphical user interface (what is displayed together, hyperlinks).
  - Interpretative paths that users build: how several results are associated in an analysis.
- Illustration:
  - A collective and interdisciplinary analysis (History, Linguistics, Psychology)
    - with Damon Mayaffre, Serge Heiden, Philippe Weyl ; two publications: 2016, 2018.
  - on 20 Shoah testimonials from Auschwitz camp (French language, 1 million tokens)
  - about memory: memory evolution; collective vs personal memory (Matrice project)
  - carried out with open-source textometric software (TXM, IRaMuTeQ)
    - TXM: search engine = CQP, statistics = R IRaMuTeQ: statistics = R
  - Specificities are extensively used to identify lexical features of testimonials, depending on whether they were written temporally close or distant to the narrated events.

The SHOAH corpus: 20 texts, 3 periods



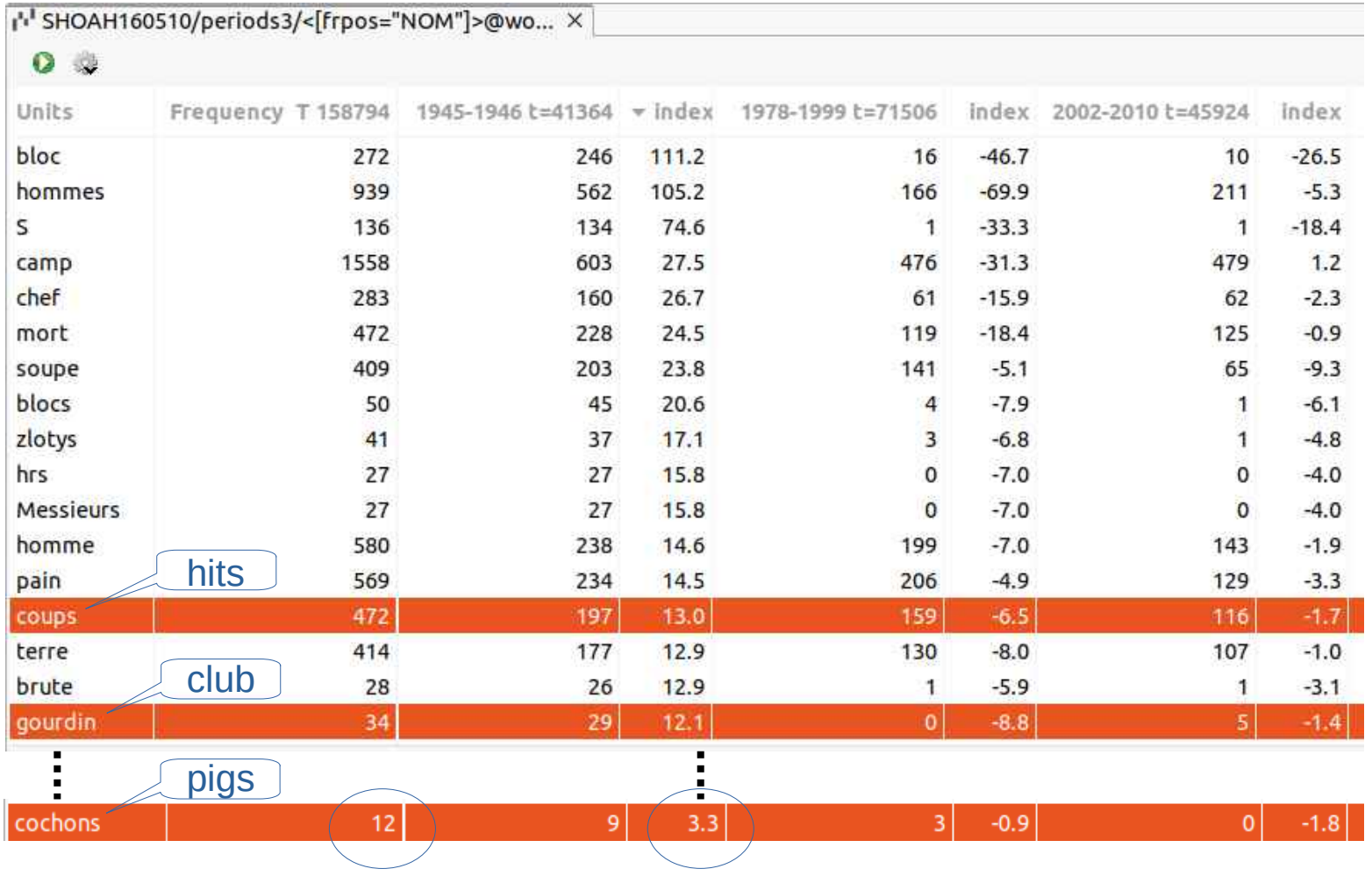


A systematic tour of textometric features rather than a complete analytical path

- Considering not only *S* but also ***F***, ***f***, ***t***: table display of Specificity results
- Examining **words in context**: a core textometric functionality
- Multiple ways of using associations, that complement one another:
  - Words for a Part (column sorting), Parts for a Word (bar chart visualization), Words common to all Parts
  - Words contextually associated to a Word (collocations), Word groups (~ topics)
  - (Not in our SHOAH study: Building the characteristic Period for a Word (chronological Specificities), Words in parallel relationship with a Word (resonance))
- Managing **dispersion** through recursive specificity calculation
- Advanced text encoding and **digital philology**: any structure or feature encoded in the corpus is available to build **accurate selections**
  - What is the reference for frequencies (*T*)
  - How can the corpus be divided into parts (*t*)
  - What lexical types do you consider and how can complex lexical units be defined (*F*)
  - Many available parameters for collocations too
- Combined use of **Correspondence Analysis** (CA) and Specificities to disambiguate visual proximity

Considering not only *S* but also *F*, *f*, *t*: table display

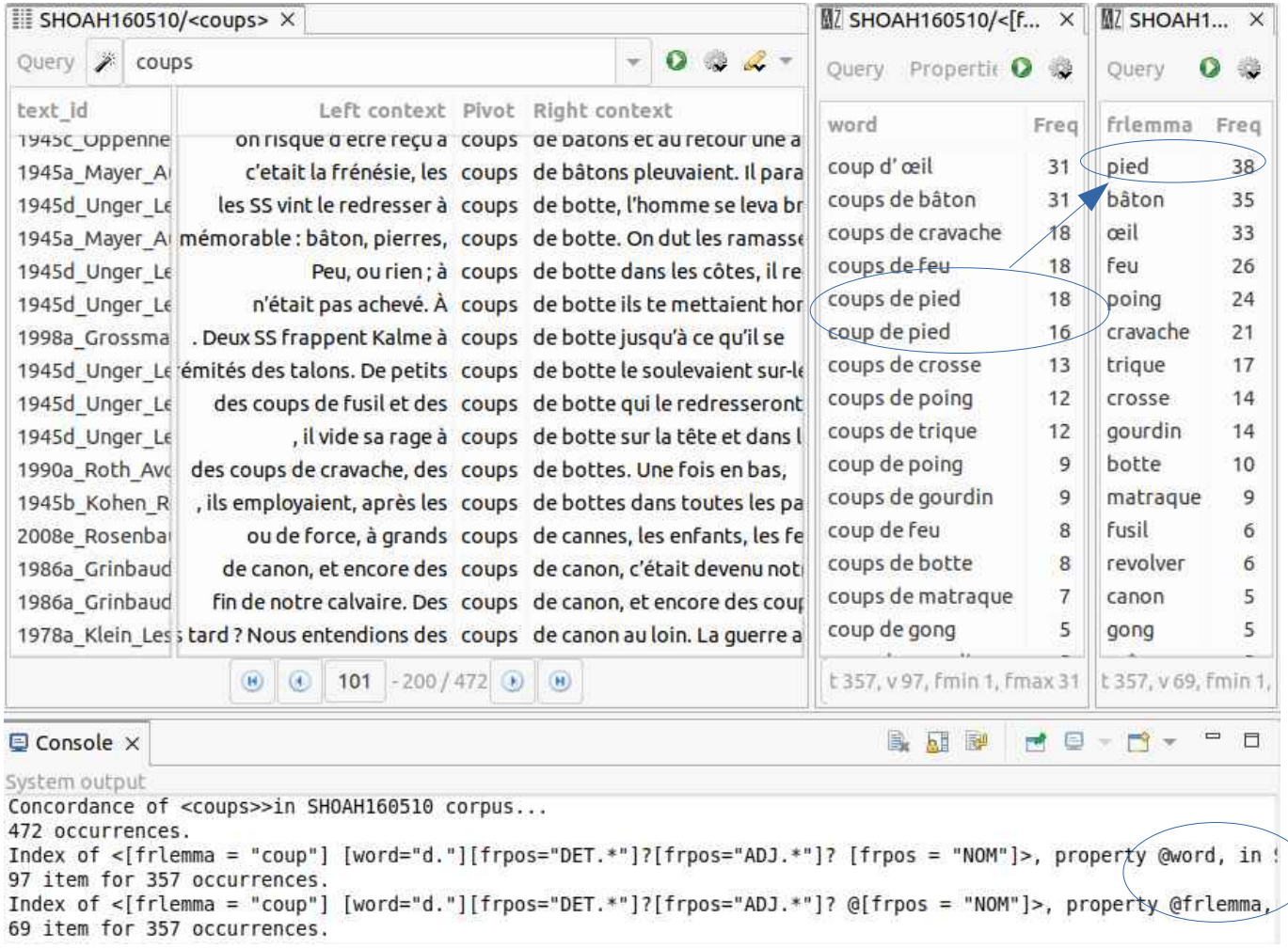
- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+



Units	Frequency T 158794	1945-1946 t=41364	▾ index	1978-1999 t=71506	index	2002-2010 t=45924	index	
bloc	272	246	111.2	16	-46.7	10	-26.5	
hommes	939	562	105.2	166	-69.9	211	-5.3	
S	136	134	74.6	1	-33.3	1	-18.4	
camp	1558	603	27.5	476	-31.3	479	1.2	
chef	283	160	26.7	61	-15.9	62	-2.3	
mort	472	228	24.5	119	-18.4	125	-0.9	
soupe	409	203	23.8	141	-5.1	65	-9.3	
blocs	50	45	20.6	4	-7.9	1	-6.1	
zlotys	41	37	17.1	3	-6.8	1	-4.8	
hrs	27	27	15.8	0	-7.0	0	-4.0	
Messieurs	27	27	15.8	0	-7.0	0	-4.0	
homme	580	238	14.6	199	-7.0	143	-1.9	
pain	569	234	14.5	206	-4.9	129	-3.3	
coups	472	197	13.0	159	-6.5	116	-1.7	*(41%)*
terre	414	177	12.9	130	-8.0	107	-1.0	
brute	28	26	12.9	1	-5.9	1	-3.1	
gourdin	34	29	12.1	0	-8.8	5	-1.4	*(85%)*
⋮			⋮					
cochons	12	9	3.3	3	-0.9	0	-1.8	*(75%)*

Words in context: a core textometric functionality

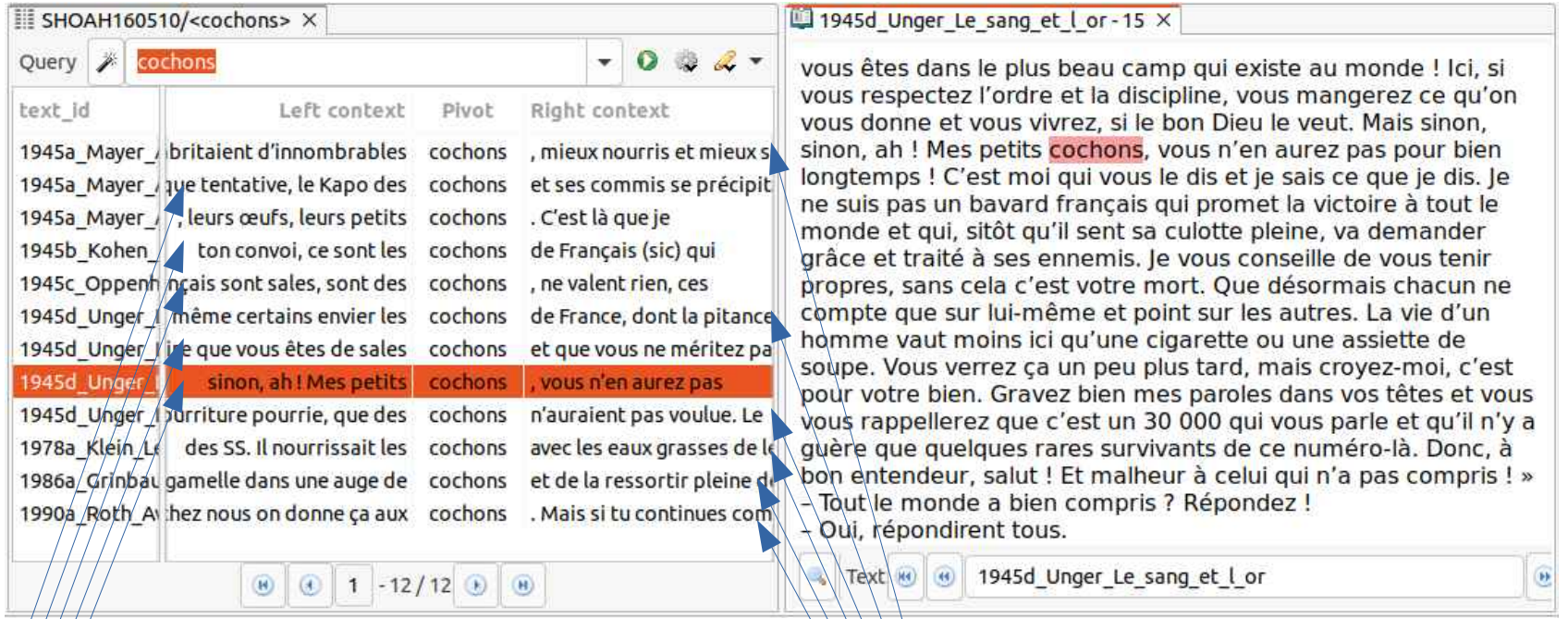
- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+



foot (kick)
stick
eye (glance)
fire (gunshot)
fist (punch)
whip
cane
riffle butt
...

Words in context: a core textometric functionality

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+



Insult: Humans are called pigs

2 main semantic uses can be identified

Pigs' food is even better than humans' one.

Double-click on a concordance line to read the passage if more context is needed

Words for a Part (column sorting)

Negative specificities for the first period

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+

SHOAH160510/periods3/<[frpos="NOM"]>@wo...

Units	Frequency T 158794	1945-1946 t=41364	▲ index	1978-1999 t=71506	index	2002-2010 t=45924	index
père	727	34	-52.3	406	8.5	287	9.2
mère	656	35	-43.7	231	-6.7	390	58.8
frère	422	13	-37.0	238	5.8	171	6.7
parents	427	14	-36.5	205	0.9	208	17.3
famille	567	32	-36.5	310	5.6	225	7.6
ans	784	72	-32.2	358	0.4	354	21.4
années	297	11	-24.4	190	10.4	96	1.0
fille	402	29	-21.6	188	0.6	185	12.5
rue	355	23	-21.0	140	-1.7	192	22.5
yiddish	162	1	-19.5	58	-2.0	103	19.2
sœur	227	8	-19.2	122	2.3	97	5.2
oncle	140	0	-18.4	80	2.6	60	3.5
guerre	593	74	-15.5	282	0.9	237	8.3
époque	298	23	-15.3	154	1.9	121	5.0
tante	144	3	-14.5	72	0.9	69	5.9
déportés	216	12	-14.5	101	0.5	103	8.4
mari	140	3	-14.0	64	0.3	73	8.2

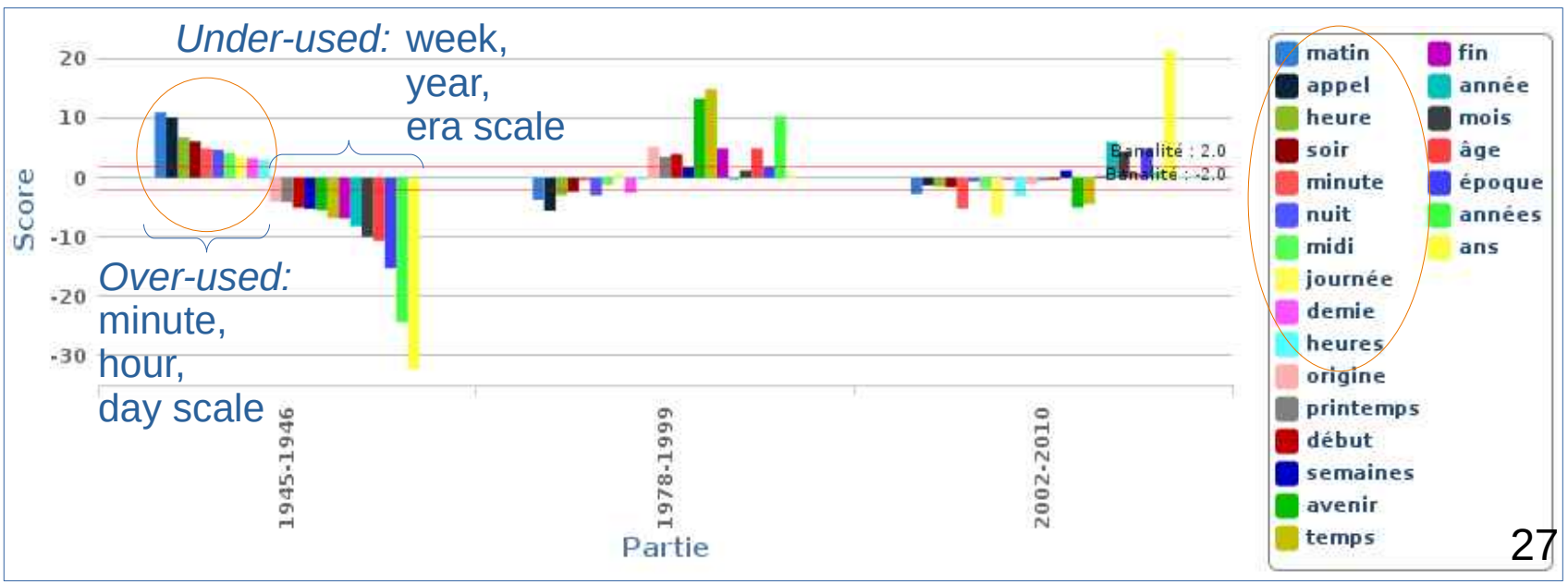
Parts for a Word (bar chart visualization)

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+

Units	Frequency T 158794	1945-1946 t=41364	index	1978-1999 t=71506	index	2002-2010 t=45924	index
gourdin	34	29	12.1	0	-8.8	5	-1.4
â	19	19	11.1	0	-4.9	0	-2.8
matin	644	246	11.1	245	-3.7	153	-2.7
boue				7	-5.9	8	-1.7
secrétaire				9	-7.1	13	-1.2
bourreaux	57	39	10.6	11	-4.4	7	-2.6
chefs	53	37	10.5	7	-6.1	9	-1.5
appel	271	120	10.2	85	-5.5	66	-1.3
départ	263	117	10.1	99	-2.0	47	-4.6

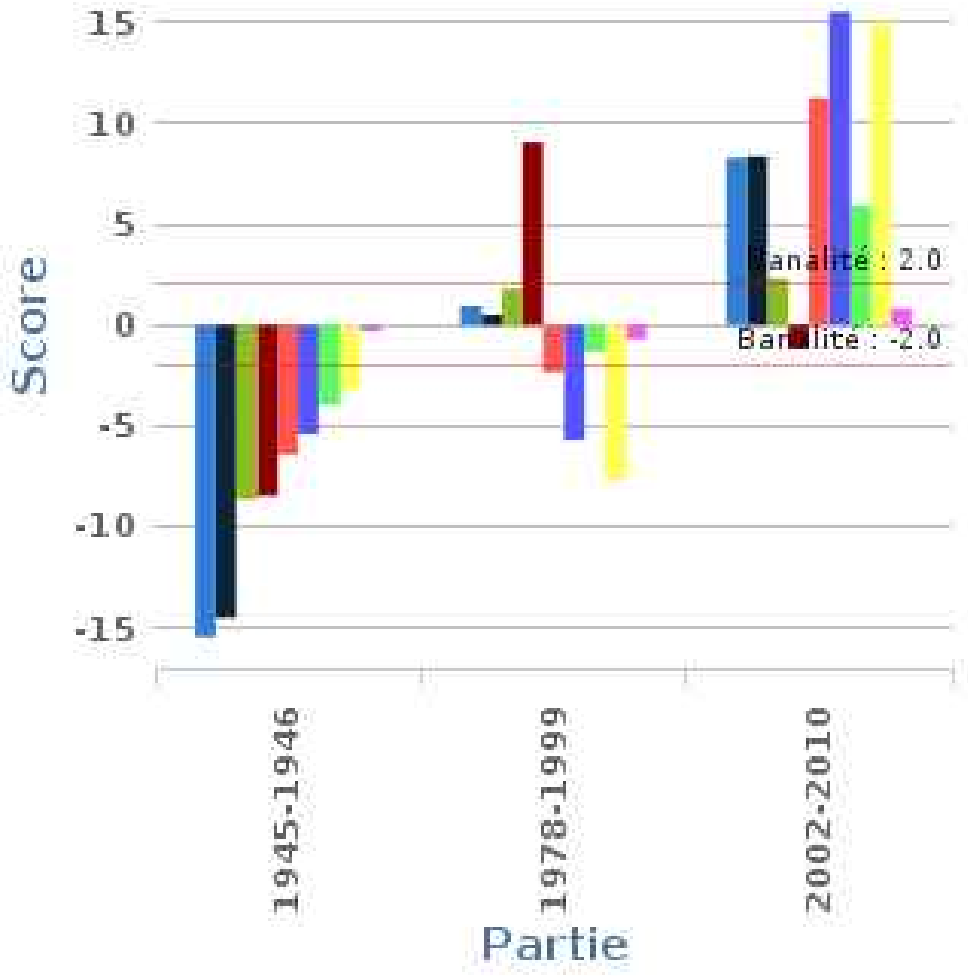
Compute the histogram of the selected lines

Copy



Parts for a Word (bar chart visualization)

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+



war

roundup

Historical Vocabulary

Words not specific in any Part: Basic Vocabulary

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+

	A	B	C	D	E	F	G	H	I
1	unit	F	score_max	1945-1946	score	1978-1999	score	2002-2010	score
2	jour	1129	0.976	313	0.976	493	-0.7313	323	-0.3736
3	monde	604	1.4116	177	1.4116	257	-0.9299	170	-0.4491
4	place	518	1.3041	152	1.3041	219	-0.9522	147	-0.3831
5	tête	436	1.3175	98	-1.3175	210	0.9899	128	0.3585
6	suite	421	0.5327	115	0.5327	185	-0.4621	121	-0.308
7	un	387	0.6619	108	0.6619	170	-0.4553	109	-0.4027
8	lendemain	355	1.2298	106	1.2298	152	-0.6652	97	-0.562
9	coup	348	1.3423	104	1.2247	158	0.3323	86	-1.3423
10	Block	326	1.1355	97	1.1355	136	-0.9017	93	-0.3322
11	lieu	313	0.3442	82	0.3031	142	0.3241	89	-0.3442
12	bras	295	0.8104	85	0.8104	129	-0.4579	81	-0.5025
13	groupe	293	0.5724	76	-0.2894	127	-0.5208	90	0.5724
14	faim	282	0.8107	80	0.6883	118	-0.8107	84	0.4025
15	bout	275	0.5304	76	0.5304	122	-0.36	77	-0.4014
16	wagon	260	0.5596	63	-0.5568	117	-0.2826	80	0.5596
17	air	260	1.4289	81	1.4289	109	-0.7639	70	-0.5814
18	vêtements	247	0.8794	67	0.4282	117	0.6029	63	-0.8794
101	extermination	71	0.8248	18	-0.2931	28	-0.69	25	0.8248
138	peuple	56	0.666	15	0.3003	28	0.5701	13	-0.666
276	paix	34	0.8059	9	0.2653	12	-0.7787	13	0.8059

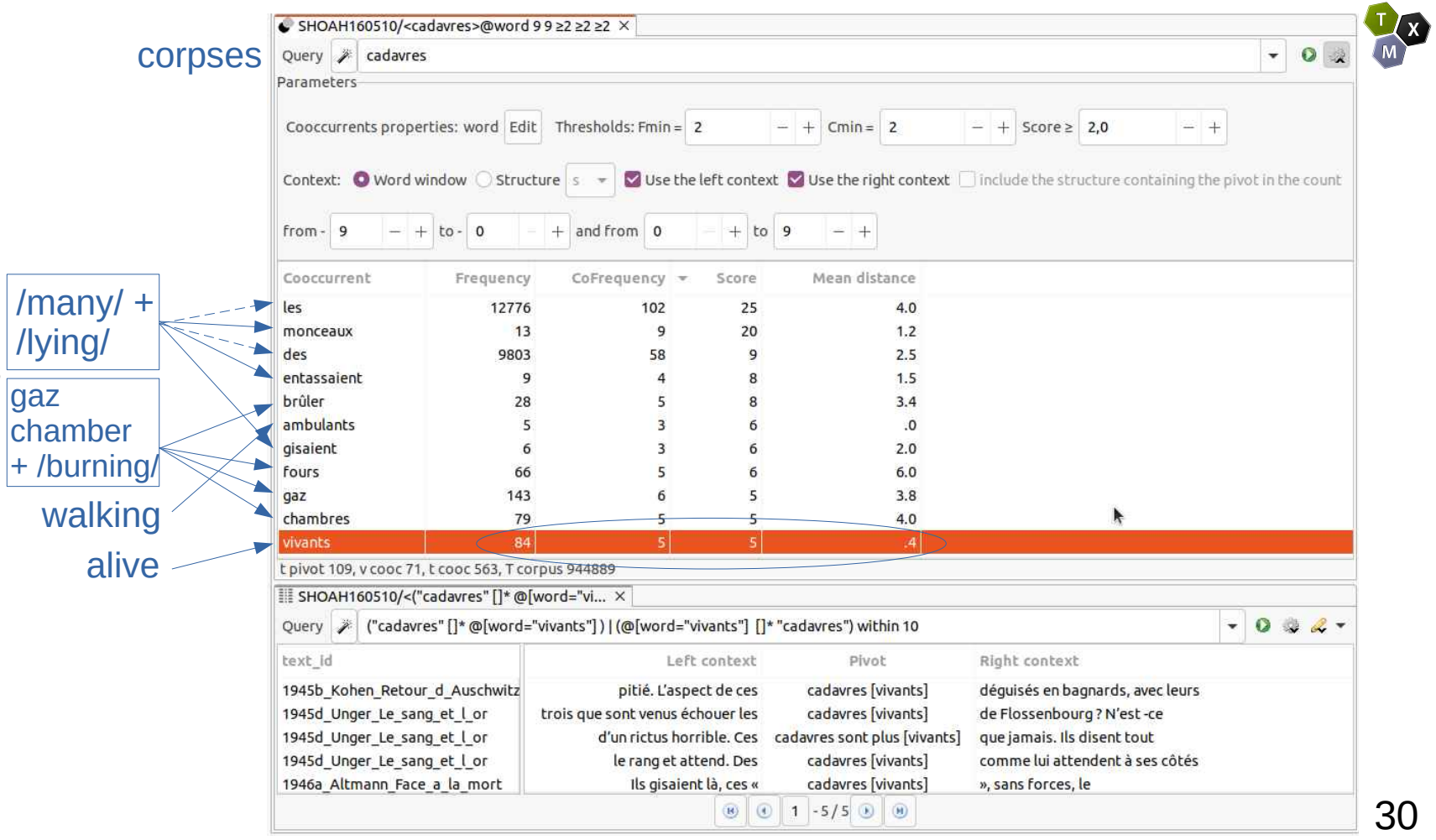
-1.5 < S < +1.5 in every Part

Some historical vocabulary is common to all periods

nation

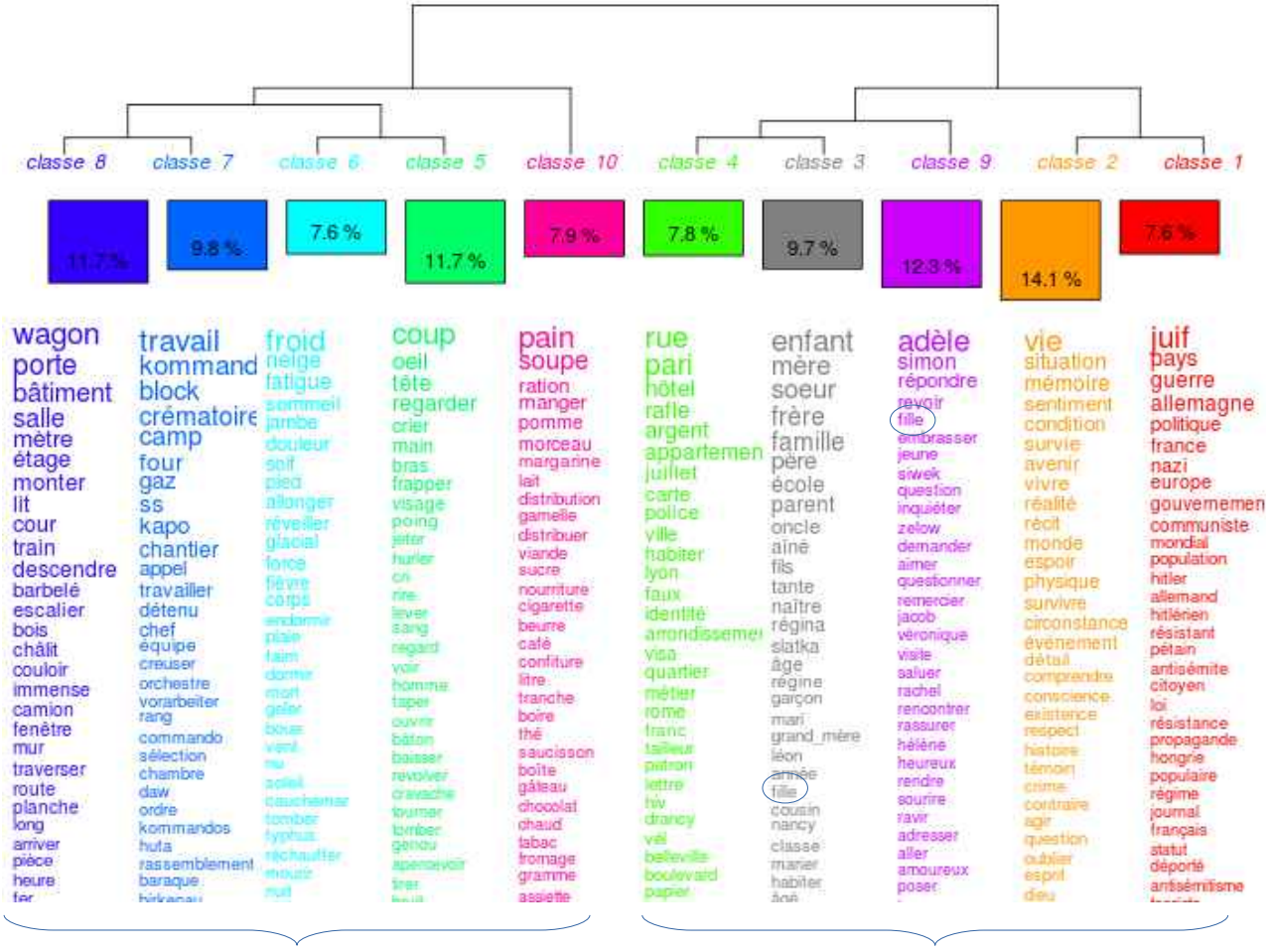
Words contextually associated to a Word (collocations)

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+



Word groups: a thematic summary of the corpus

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+



Reinert Method

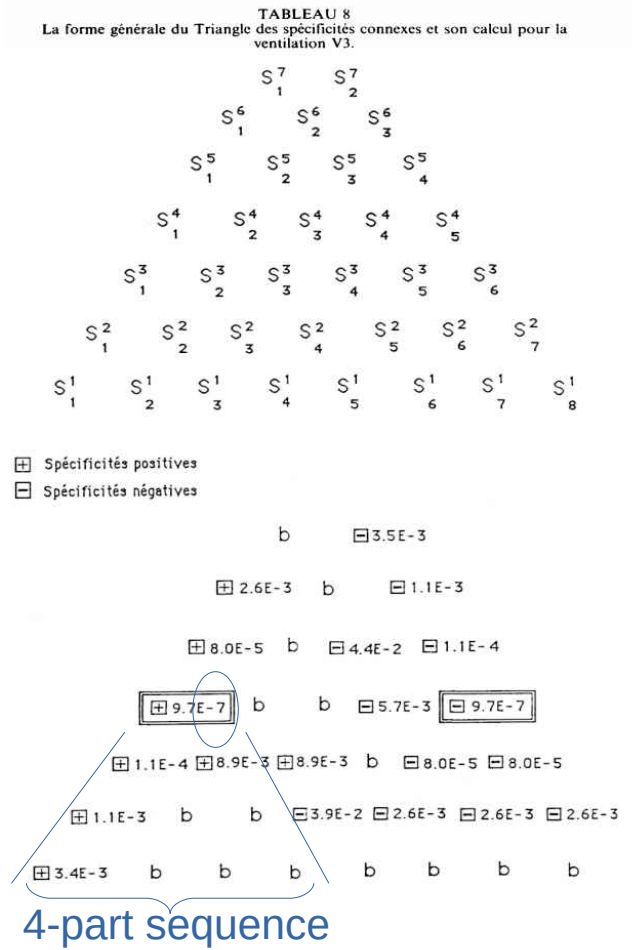
(Reinert 1983, 1990)

- Corpus is split in **context segments** (~40-word-long here)
- Unsupervised **classification of segments**
- **Specific words** are computed for each segment class (here with a Chi-squared test)

IRaMuTeQ

Building the characteristic Period for a Word

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+



(Figure from **Salem 1991**, 163; blue annotations are added)

Term	F	f	S	Parts
the	148431	47834	+51	"01washington" - "43bush"
1898	54	45	+51	"25mckinley"
commercial	676	657	+51	"02adams" - "31hoover"
together	674	321	+51	"36johnson" - "43bush"
in	38058	21715	+51	"19hayes" - "39carter"
areas	263	0	+51	"19hayes" - "44obama"
iraqi	53	0	+51	"40reagan" - "44obama"
terrorists	88	0	+51	"32roosevelt" - "44obama"
be	18560	18320	+51	01washington" - "42clinton"
1946	76	76	+51	"33truman" - "42clinton"
is	16650	1035	+51	"29harding" - "30coolidge"
percent	349	0	+51	"16lincoln" - "44obama"
goals	125	0	+51	"27taft" - "44obama"
which	12335	12294	+51	01washington" - "42clinton"
as	11988	9074	+51	1washington" - "29harding"
health	669	133	+51	"31hoover" - "44obama"
t	347	0	+51	"16lincoln" - "44obama"
families	344	57	+51	"33truman" - "44obama"
1980	89	0	+51	"32roosevelt" - "44obama"
notes	313	310	+51	"04madison" - "28wilson"
cuba	312	104	+51	"25mckinley"
on	9172	797	+51	"03jefferson" - "05monroe"
ve	262	0	+51	"19hayes" - "44obama"
gentlemen	79	46	+51	01washington" - "02adams"
are	8719	3616	+51	"26roosevelt" - "44obama"
technology	131	1	+51	"33truman" - "44obama"

(Figure from **Lebart et al. 2019**;
Corpus = *State of the Union* addresses 1790-2008)

Chronological Specificities

(Salem 1991; *Adjacent characteristic elements* in Lebart et al. 1998)

For **diachronic** or **sequentially ordered corpora**

Words in parallel relationship with a Word

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+

Resonance
(Salem 2004)

For **parallel corpora**.
Various kinds of possible alignments:

- Language A / Language B
- Adult speech turn / Child speech turn
- Political debate: for each topic/question, answer from Candidate A / Candidate B
- Etc.

(Figures from **Salem 2004**)

migrants *(persons)*

immigration *(concept)*

F. Mitterrand :

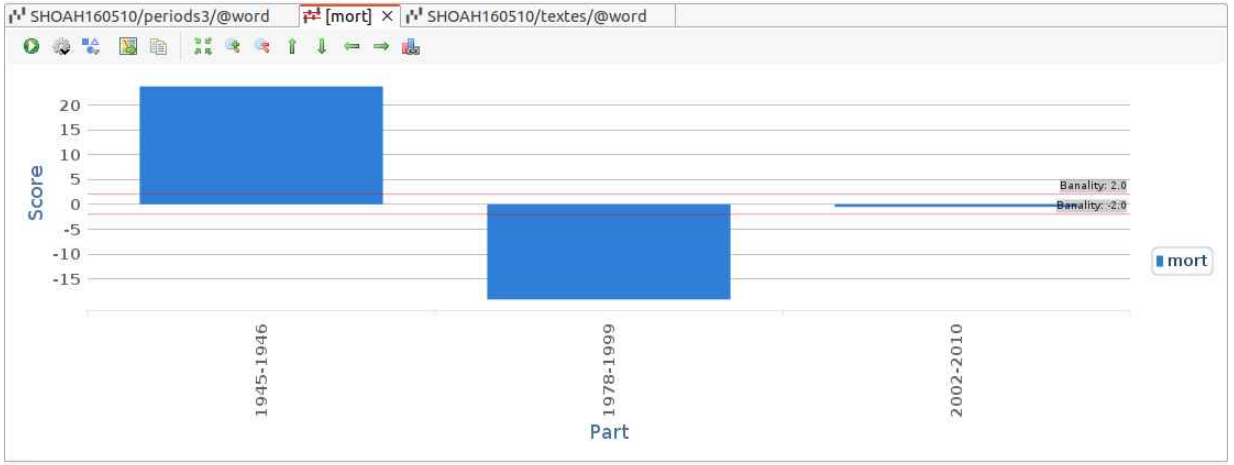
il faut d' abord distinguer , c' est un problème qui a été vraiment exagéré et compliqué à plaisir. il y a plusieurs catégories de personnes visées par le débat actuel . il y a d - abord ceux qui ne sont pas des **immigrés** , qui sont les enfants d ' **immigrés** et qui sont nés sur notre sol . ceux - là ont vocation . ils sont français , sauf s ' ils en décident autrement à l' âge de dix - huit ans . il y a , ensuite , les naturalisés ; ce sont les **immigrés** qui désirent devenir français , là , l' administration étudie leur cas et il aboutit à reconnaître le droit à la naturalisation , selon son propre rythme . je n ' insiste pas. et puis il y a les **immigrés** _ ceux qui n ' ont pas envie de devenir français, qui veulent rester attachés à leur pays d ' origine _ de deux catégories : il y a les clandestins , et il y a ceux qui sont reconnus parce qu ' ils ont un contrat de travail et une carte de séjour . ceux qui sont clandestins , il n' y a qu ' une seule loi possible : il faut _ c ' est malheureux pour eux , mais c' est la nécessité _ il faut qu' ils rentrent chez eux et les dispositions doivent être prises , et elles ont été prises pour ceux - là , pour qu ' ils rentrent chez eux . et puis il y a ceux qui sont là avec leur contrat de travail et leur carte de séjour . est - ce qu ' il y en a trop ? ce que je sais , c' est que , dans les années qui ont précédé 1981 , il y a eu une formidable aspiration à faire venir chez nous des **immigrés** _ sans doute parce qu' on les payait moins bien que les autres , moins bien que les français , que les travailleurs français . /…/

J. Chirac :

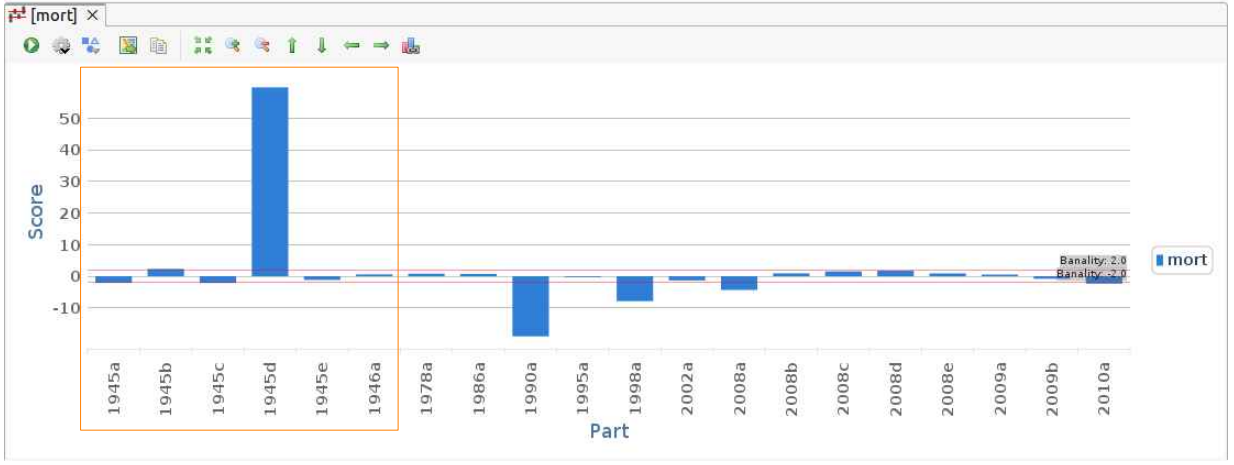
je voudrais répondre, moi , très clairement en m' appuyant sur mon bilan dans cette affaire ; parce que c' est très gentil de faire des promesses, mais enfin, encore faut il qu ' elles soient rendues crédibles par un bilan. s ' agissant de l ' **immigration** tout court , il faut la stopper , parce que nous n ' avons plus les moyens de donner du travail à des étrangers. aussi , naturellement , en supposant quelques souplesses naturellement, mais il faut la stopper . s ' agissant de l ' **immigration** clandestine, il faut évidemment lutter contre cette **immigration** avec beaucoup d ' énergie et reconduire les intéressés à la frontière ou les expulser. ils ont pris leurs risques en venant chez nous de façon illégale, ils sont le vivier naturel _ non pas en raison de leurs origines naturellement, mais parce que ce sont des marginaux et qui se cachent _ ils sont le vivier naturel des délinquants, voire des criminels : il faut donc les expulser. en 1981 - 82 - 83 , vous en avez régularisés 130000 : erreur capitale, car ça a été immédiatement un appel équivalent et même beaucoup plus large. nous , nous avons refoulé, en deux ans, plus de 130000 personnes, ce qui fait tout de même deux cents par jour, et je considère que ce n' est pas suffisant . nous le faisons , naturellement, en nous entourant de toutes les exigences de l' humanisme, de respect des droits de l' homme , mais c ' est une nécessité impérieuse. et puis nous devons nous protéger contre ces entrées. alors , je voudrais simplement poser une question. moi, j ' ai fait voter des lois pour la sécurité _ mais j' imagine que nous y viendrons tout à l' heure _ et contre l' **immigration** et notamment l' **immigration** clandestine, en particulier /…/

Managing Dispersion via recursive specificity calculation

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+



Specificities for “mort” (*death*) at **Part level**



Specificities for “mort” (*death*) at **Text level**

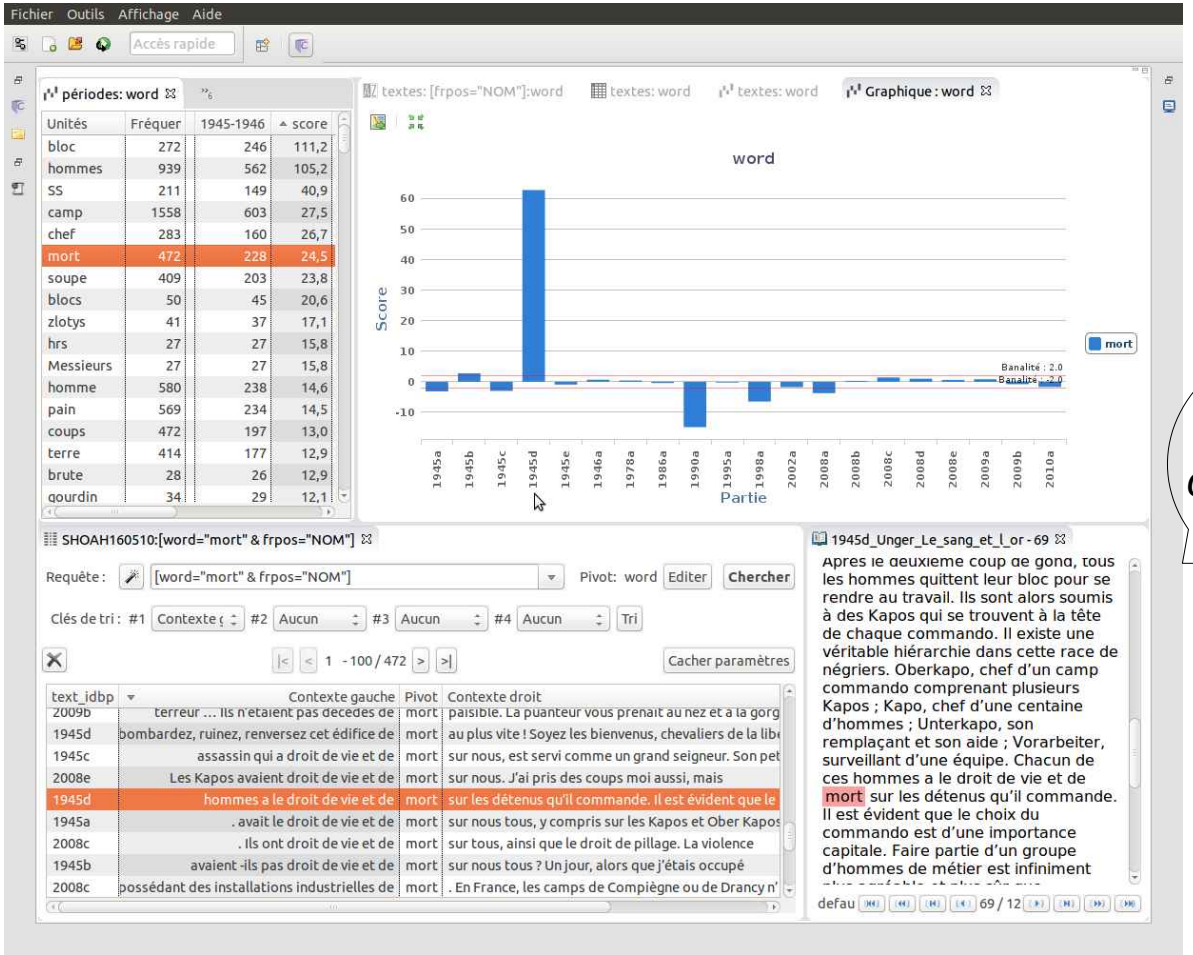
Managing Dispersion via recursive specificity calculation

A Graphical User Interface to work on Specificity results (Mayaffre *et al*. 2018):

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+

1. Specif. for a **Part**

3. Word in Context (**KWIC**)



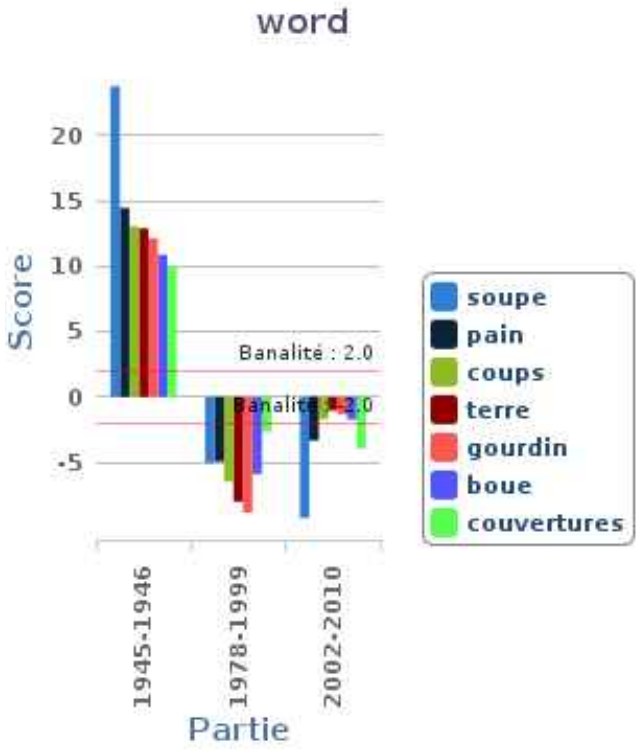
2. Overview of **Specif.** at **Text** level

A graphical instantiation of textometric tupleization?

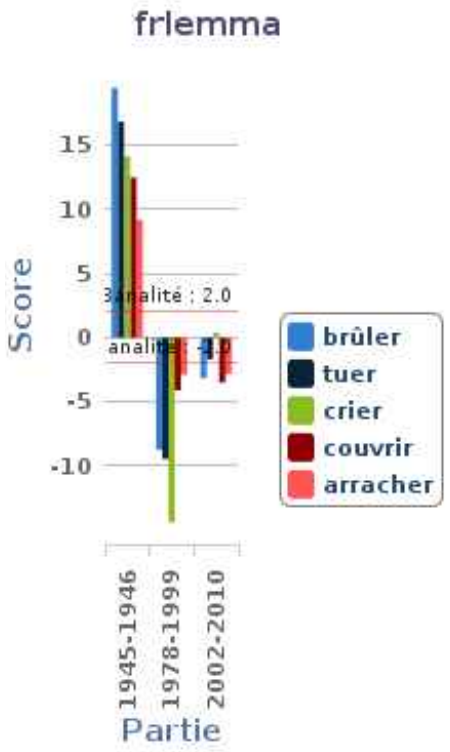
4. Word in **full Text**

Choosing the reference for frequencies (*T*)

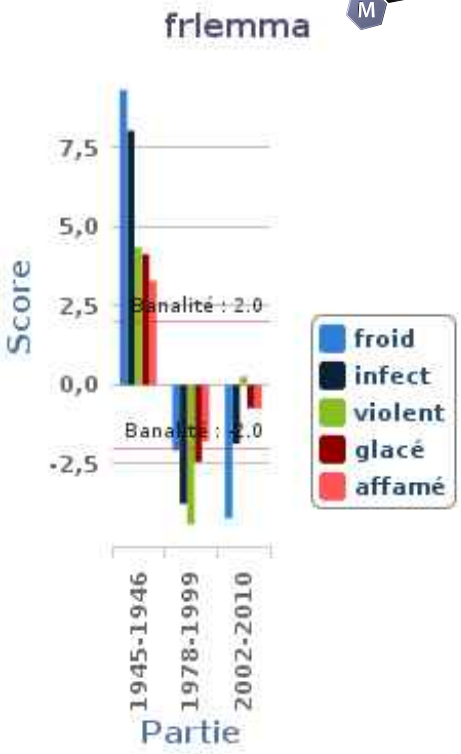
- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+



Nouns



Verbs

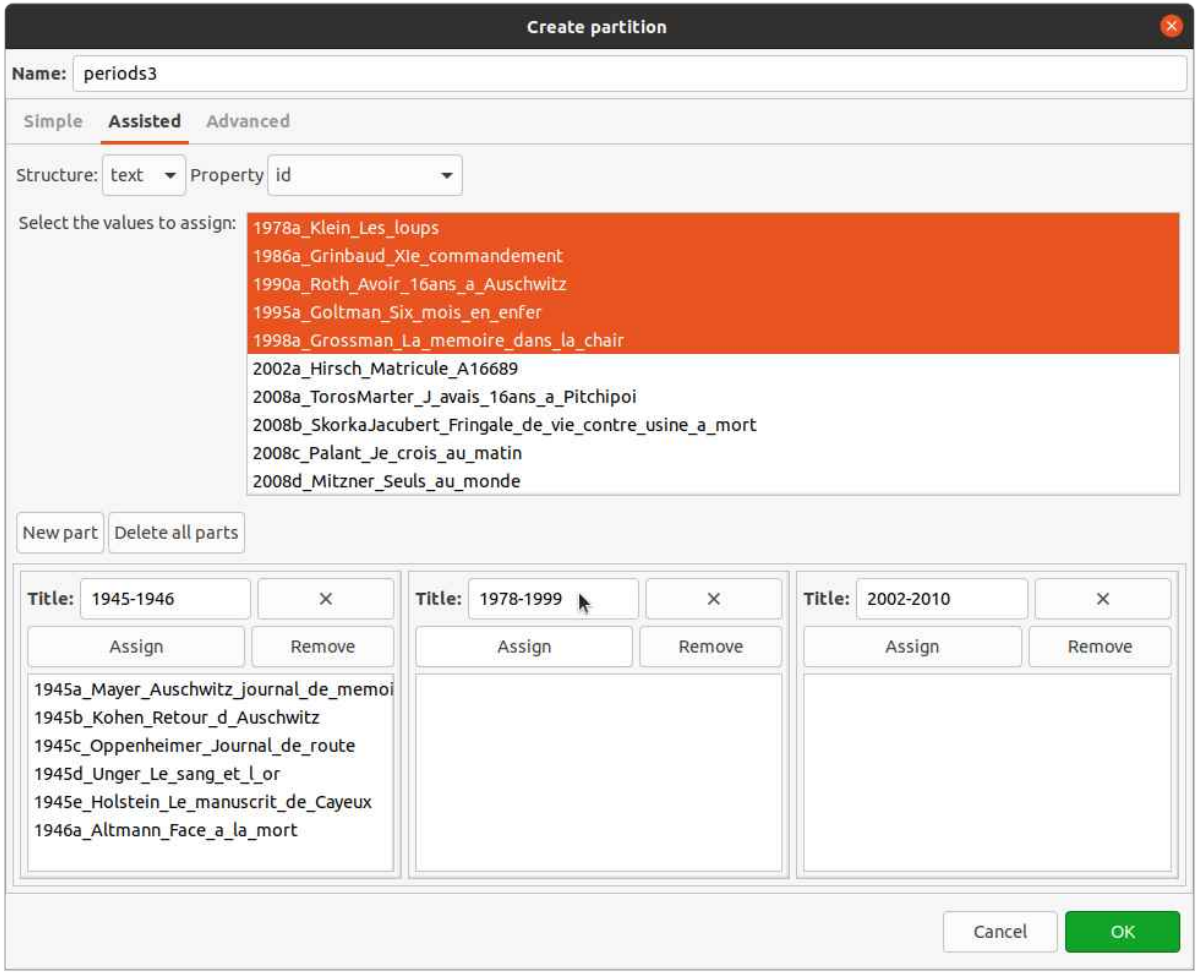


Adjectives

Noun frequencies are compared to frequencies of other nouns only
– no possible impact of stylistic effects of noun/verb balance variation for instance.

Choosing/building a set of parts (*t*)

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+



Powerful and flexible ways to define sets of parts.

(Text level / encoded structure level / word level)

Choosing the lexical type (to group tokens) (*F*)

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+

Verb lemmas

SHOAH160510/periods3/<[frpos="VER.*"]>@...

Units	Frequency T 148532	1945-1946 t=35666	index
falloir	976	366	20.5
brûler	145	87	19.5
faire	4128	1244	19.3
tuer	176	95	16.9
crier	184	93	14.2
couvrir	97	57	12.5
conduire	211	95	10.7
chasser	52	34	9.5
arracher	85	47	9.2
devoir	1643	500	8.8
taire	45	29	8.0
ranger	84	44	7.7
mourir	384	140	7.5
frapper	199	82	7.2
toucher	132	59	6.8
manger	479	163	6.3
distribuer	184	74	6.1
trembler	55	30	6.0
coucher	192	76	5.9
pouvoir	2677	748	5.9

Verb lemmas+POS

SHOAH160510/periods3/<[frpos="VER.*"]>@...

Units	Frequency T 148532	1945-1946 t=35666	index
avoir_VER:ppre	185	104	20.2
falloir_VER:pres	399	166	14.2
faire_VER:simp	292	129	13.5
crier_VER:pres	53	37	11.6
venir_VER:simp	99	54	10.2
mourir_VER:infi	133	63	8.4
devoir_VER:simp	57	34	8.0
devoir_VER:pres	592	203	8.0
falloir_VER:simp	53	32	7.7
être_VER:simp	1245	383	7.5
couvrir_VER:pper	62	35	7.3
brûler_VER:pres	31	22	7.3
brûler_VER:pper	50	30	7.2
donner_VER:simp	51	30	6.9
parvenir_VER:simp	18	15	6.7
recevoir_VER:simp	58	32	6.5
aller_VER:simp	43	26	6.4
traîner_VER:pres	29	20	6.4
sortir_VER:infi	230	89	6.3
tuer_VER:pres	10	10	6.2

Verb POS (→ tense)

SHOAH160510/periods3/<[frpos="VER.*"]>@...

Units	Frequency T 148872	1945-1946 t=35822	index
VER:simp	6493	2186	70.7
VER:infi	25405	6620	15.6
VER:subi	201	83	7.3
VER:ppre	4283	1178	7.1
VER:futu	2156	619	6.5
VER:cond	1780	489	3.3
VER:impf	29559	7215	1.2
VER:impe	15	5	0.5
VER:subp	1081	251	-0.6
VER:pper	32395	7101	-24.3
VER:pres	45504	10075	-30.5

Console

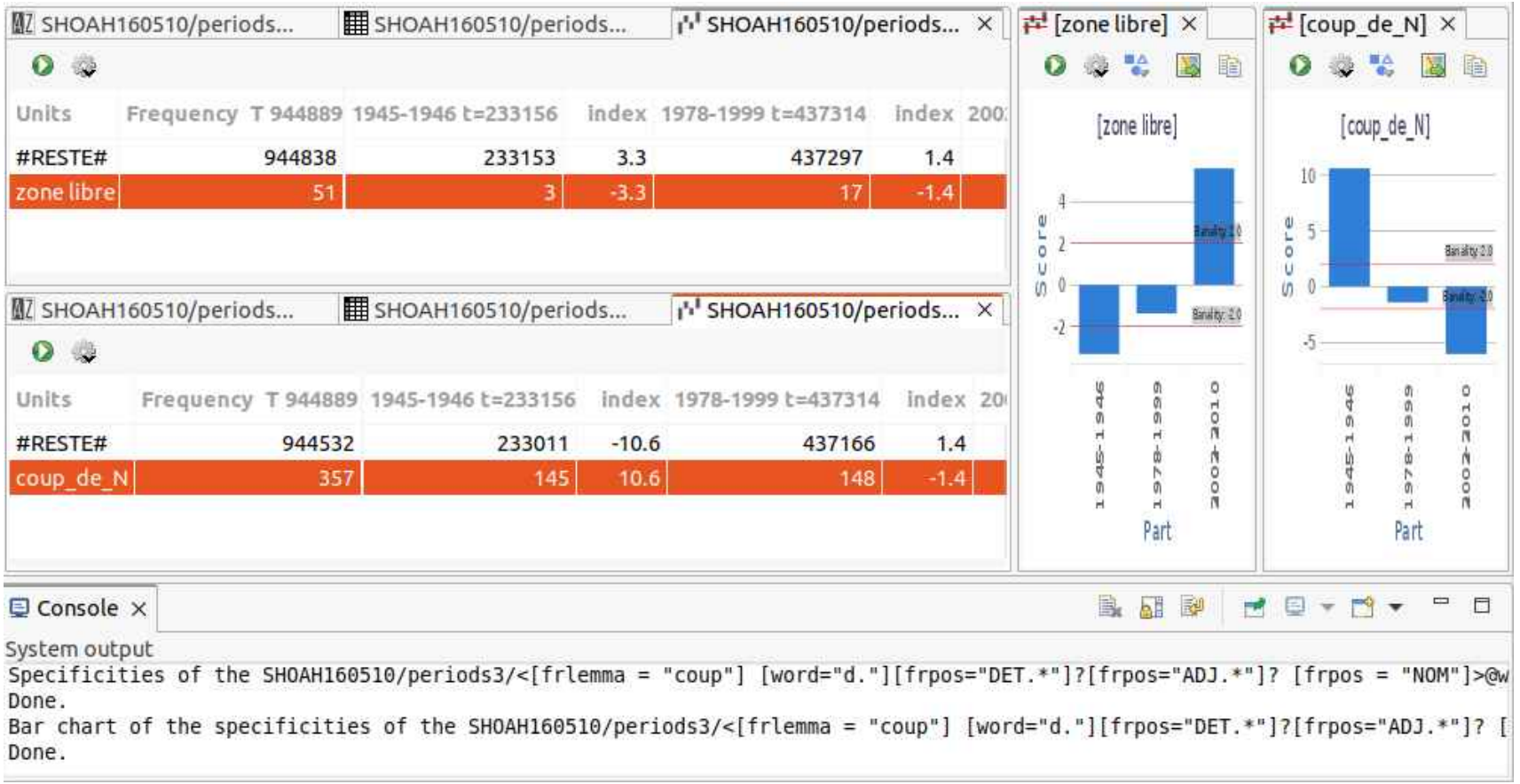
System output

```
Specificities of the SHOAH160510/periods3/<[frpos="VER.*"]>@frlemma ≤944,889 /944,889/@frlemma /944,889 lexical table...
Done.
Specificities of the SHOAH160510/periods3/<[frpos="VER.*"]>@frlemma_frpos ≤944,889 /944,889/@frlemma /944,889 lexical table...
Done.
Specificities of the SHOAH160510/periods3/<[frpos="VER.*"]>@frpos ≤944,889 /944,889/@frpos /944,889 lexical table...
Done.
```

Defining complex lexical units (*F*)

Specificity computation is possible for phrases (*zone libre*) or patterns (*coup de* + NOUN):

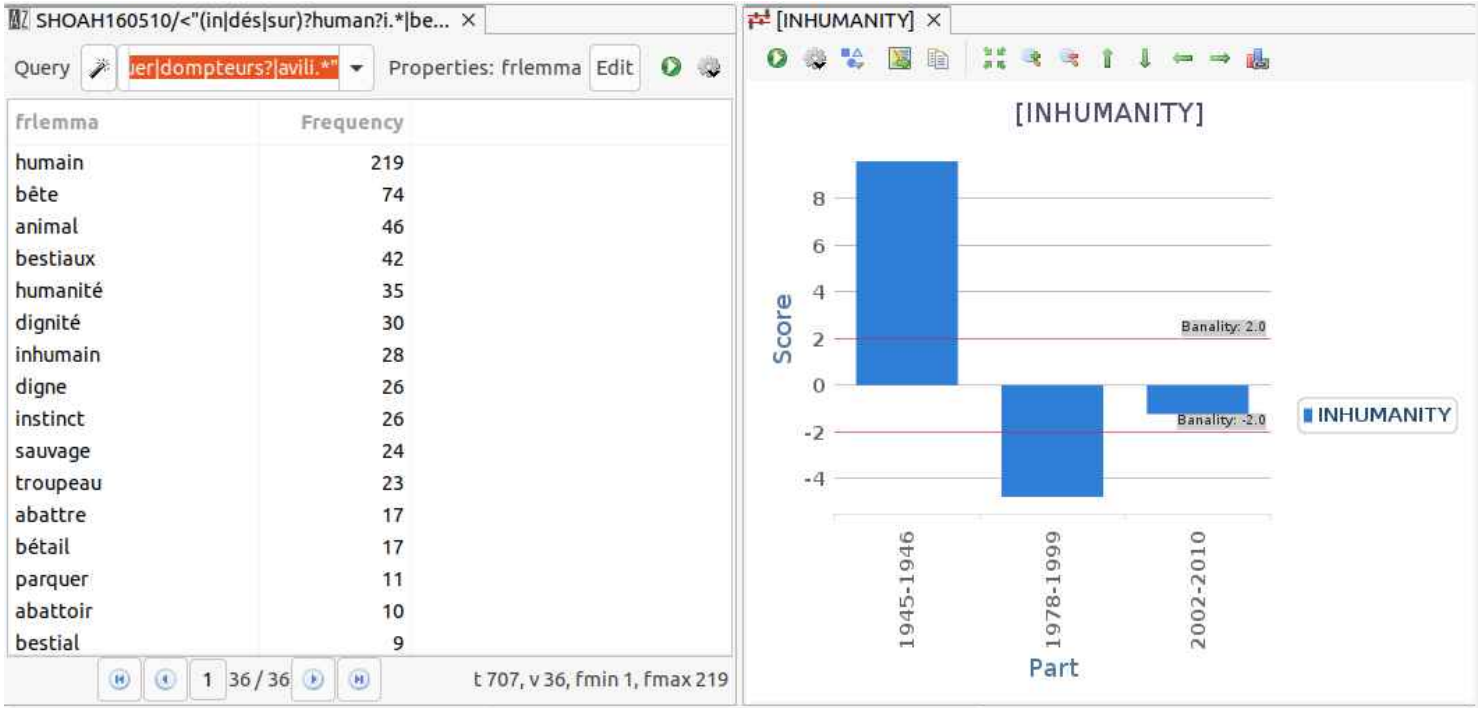
- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+



Note: TXM is powered with CQP search engine.

Defining thematic units and generalized types Tgen (*F*)

Table display (F, f, t)
Words in context
Part → Words
Word → Parts
Part set → Words
Word → Words
Thematic word groups
Word → Part sequ.
Word → // Words
Dispersion
Selection of T
Building parts t
Type & tokens F
Collocation param.
CA & S+

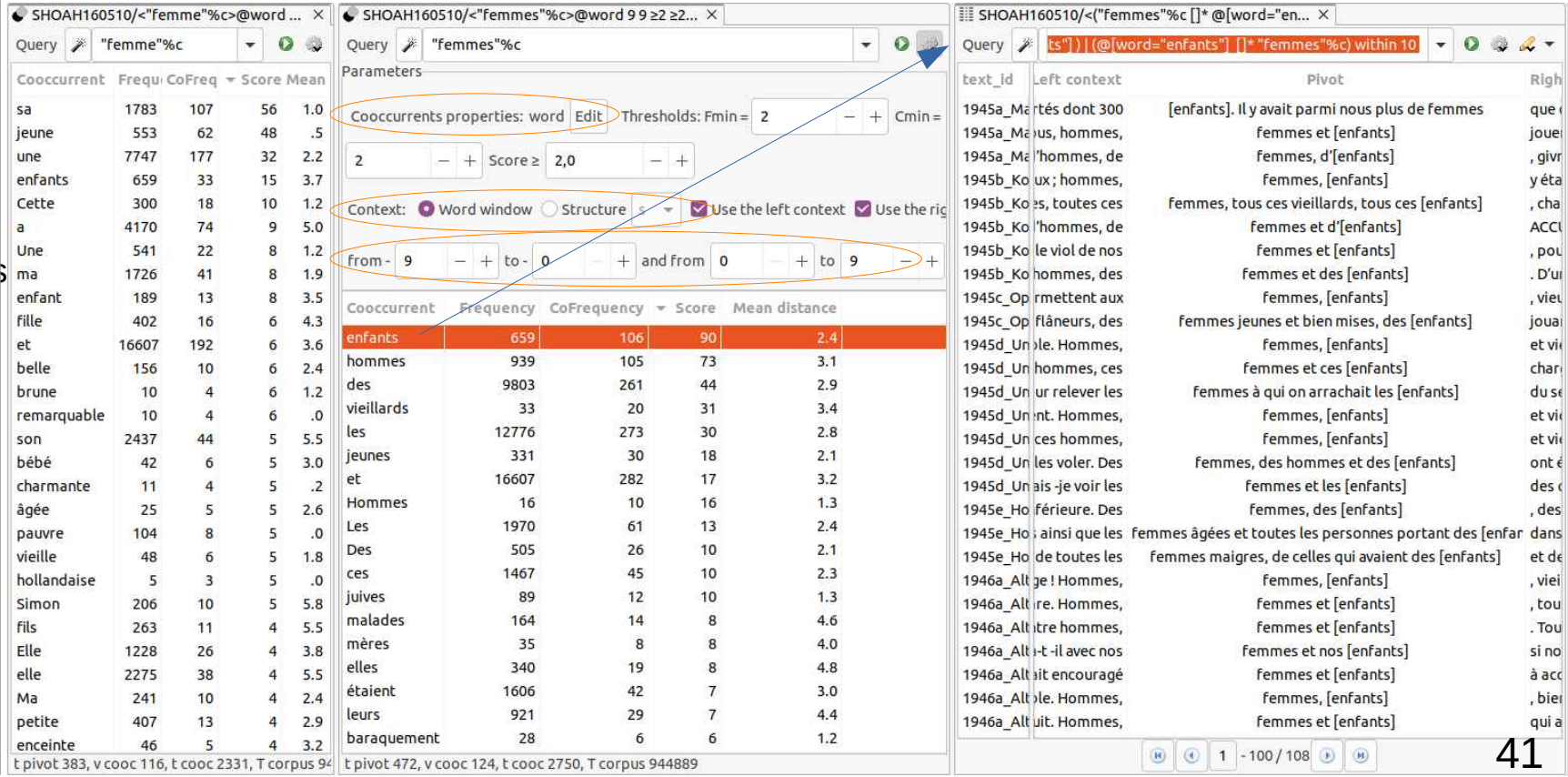


Many available settings for collocations too

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+

Sing. = wife, woman → family story

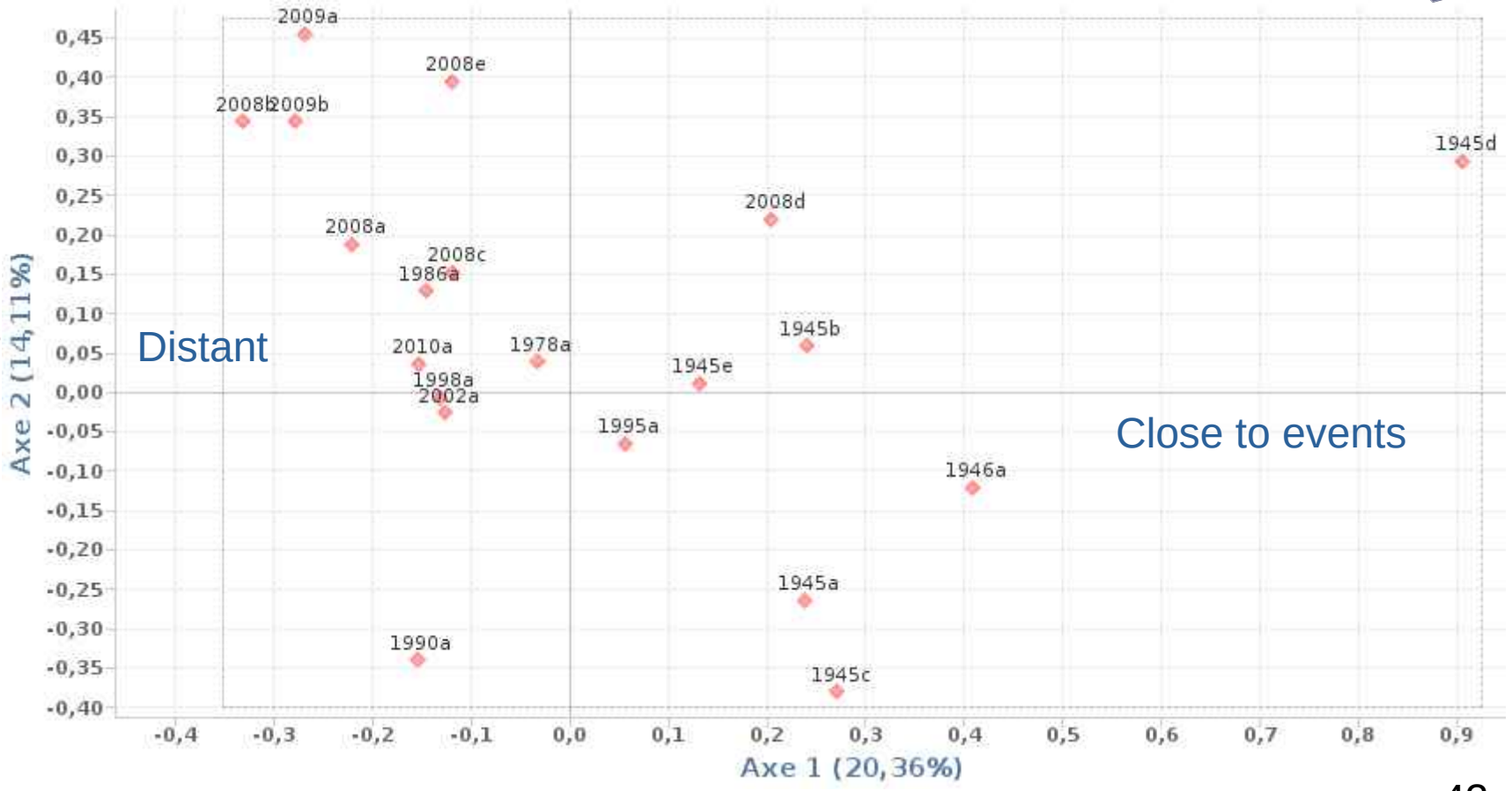
Plural = women → categories (men, children, elderly people)



Combined use of CA and S to disambiguate visual proximity

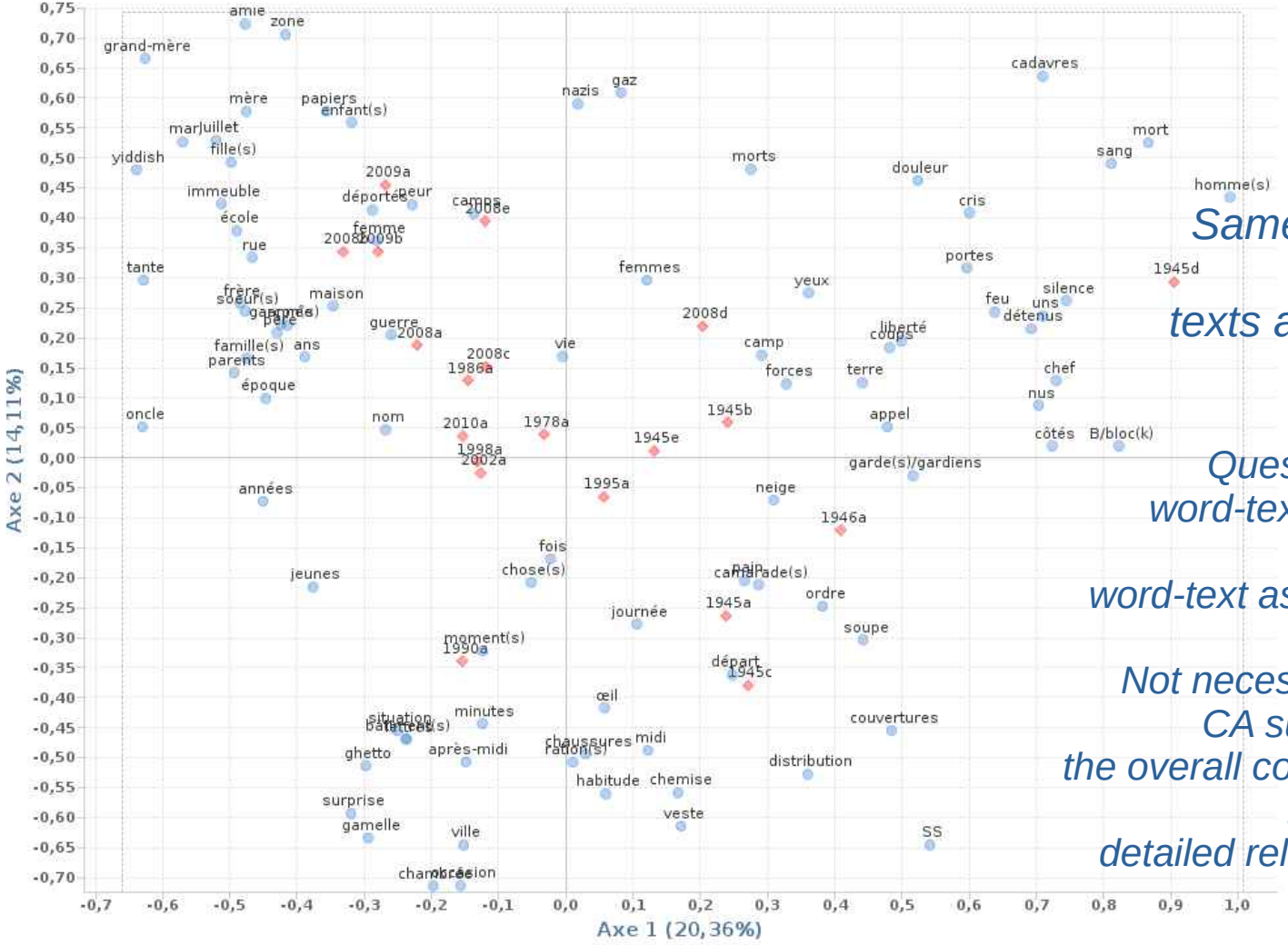
First step: a Text-only plot

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+



Combined use of CA and S to disambiguate visual proximity

Table display (F, f, t)
Words in context
Part → Words
Word → Parts
Part set → Words
Word → Words
Thematic word groups
Word → Part sequ.
Word → // Words
Dispersion
Selection of T
Building parts t
Type & tokens F
Collocation param.
CA & S+



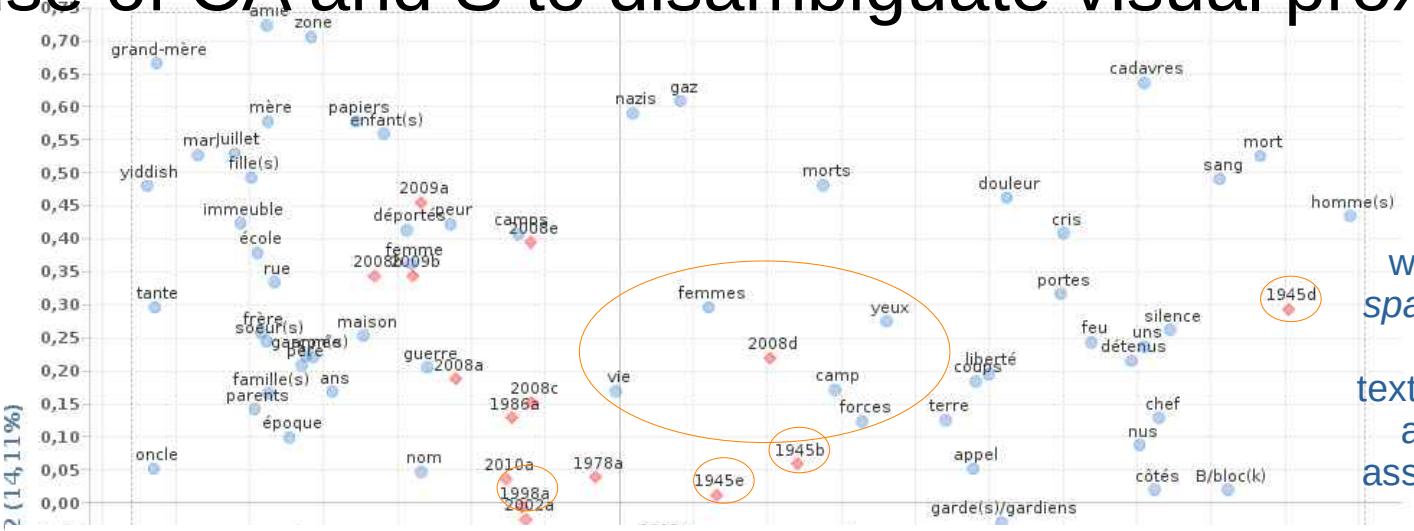
Step #2: Same analysis with both texts and words plotted

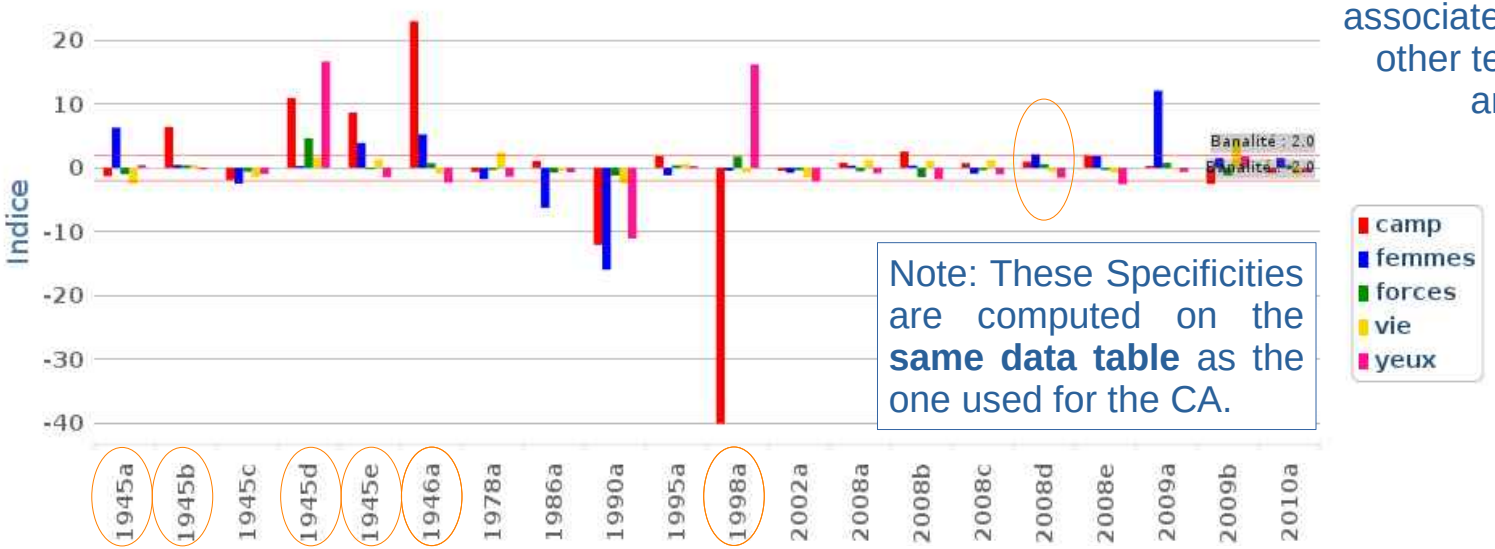
Question: Does word-text proximity reflects word-text association?

Not necessarily true: CA summarizes the overall configuration rather than detailed relationships

Combined use of CA and S to disambiguate visual proximity

- Table display (F, f, t)
- Words in context
- Part → Words
- Word → Parts
- Part set → Words
- Word → Words
- Thematic word groups
- Word → Part sequ.
- Word → // Words
- Dispersion
- Selection of T
- Building parts t
- Type & tokens F
- Collocation param.
- CA & S+



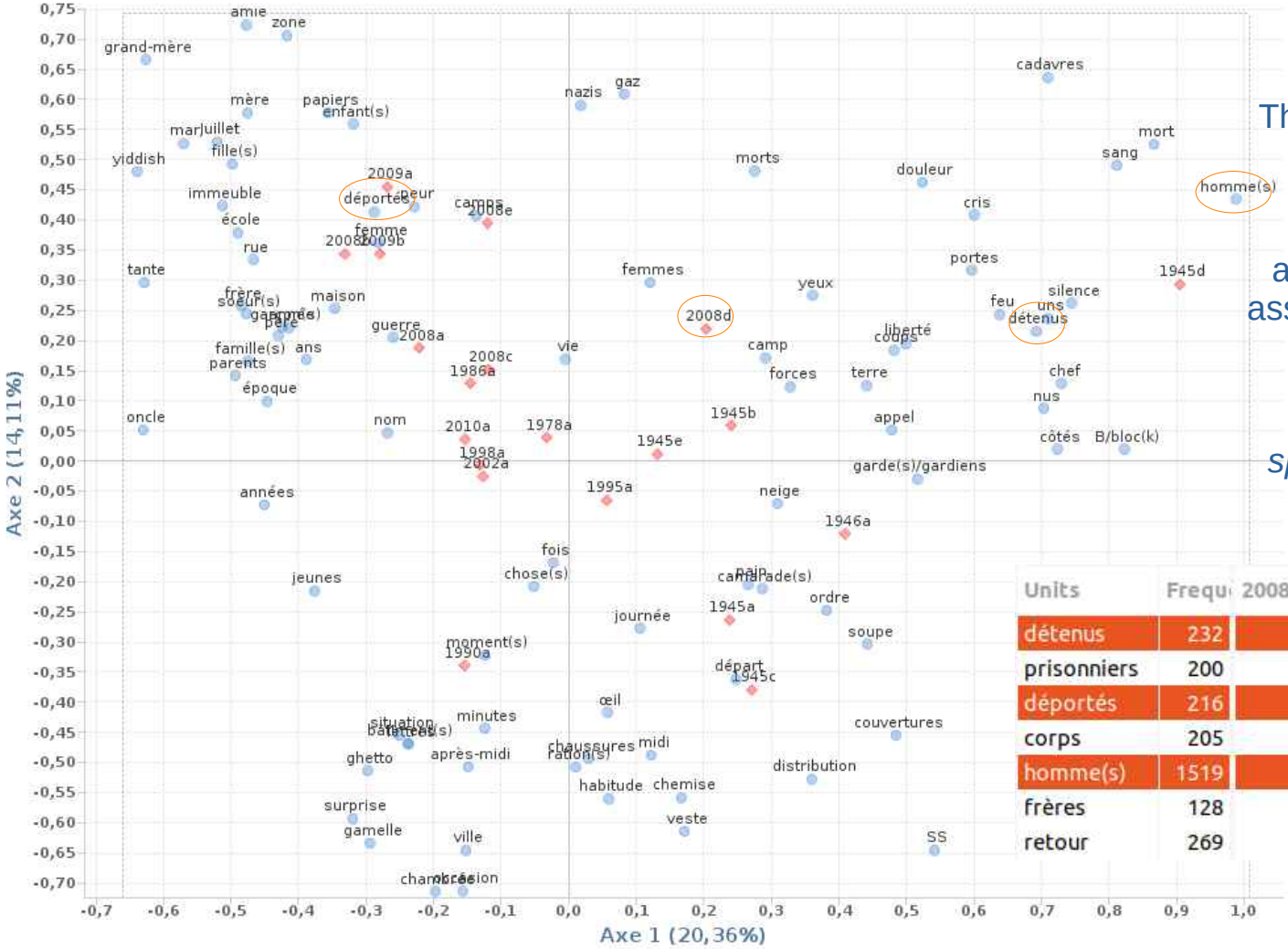


Here we observe that words that are *spatially* closest to the **2008d** text do not carry any statistical association with it (they are associated with other texts all around)

Note: These Specificities are computed on the **same data table** as the one used for the CA.

Combined use of CA and S to disambiguate visual proximity

Table display (F, f, t)
Words in context
Part → Words
Word → Parts
Part set → Words
Word → Words
Thematic word groups
Word → Part sequ.
Word → // Words
Dispersion
Selection of T
Building parts t
Type & tokens F
Collocation param.
CA & S+



Units	Frequ	2008d t=1492	index
détenus	232	26	12.5
prisonniers	200	19	8.1
déportés	216	19	7.5
corps	205	15	5.1
homme(s)	1519	50	4.0
frères	128	10	3.8
retour	269	14	3.2

The Specificity measure indicates which words are especially associated with **2008d** (we see they may *not* be *spatially* close to the text position)

Textual Data Analysis

- Corpus Linguistics, Text Mining, Distant Reading, Textometry…
  - Textual **data**
  - Similar **measures**: Chi-squared, FYE test, z-score, Mutual Information…
  - Different approaches?

Textual Data Analysis: Which Aim?

- **Automatic** analysis
- Textual information **extraction** out of the text
- **Generalization** (i.e. language description)
- Processing = calibrated and optimized **engine**

- Interpretative and interactive **exploration**
- Go **back to** text, **read**, read again
- Characterization, differentiation, what is special or **unique** in this corpus
- Processing = **tool** in user's hands

Textual Data Analysis: Which Corpora?

- Text as **resource**
- Data
- Representativeness (of something else)

- Text**s** as **sources**

  (cf. Valette 2016)
- Work, creation
- Entity (itself)

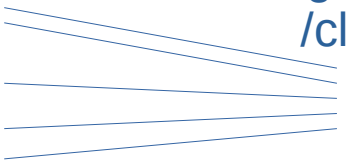
Textual Data Analysis: Which Evaluation?

- Right / Wrong, unique answer, Gold standard → **benchmark**
- Objective processing
- Final aggregated score (clear order, **hierarchy**)
- Measure = **final** and stand-alone result

- Multiple interpretations, **contextualization**
- Subjective analysis
- Multiple dimensions, qualitative considerations
- Measure = a **step** among others in a pathway

Evaluation and scientific value in Textometry

- *Form criteria*: transparency, reproducibility of results. Precise description of the process, explanation of choices...
- *Substance criteria*: rigor, consistency, quality and depth of reasoning...
- Importance given to **convergent** observations, multiple clues that go in the same direction – not a proof but a substantiated conjecture



faisceau (FR) ~ (convergent) beam (EN)?

- **Intention** (understanding of principles) rather than **extension** (validation on enumeration of cases
  → because no definite set of right answers)

Summary

1. The textometric **Specificity** measure is a **Fisher-Yates exact test**
2. It implements a **transparent modeling**
   (portion in all possible random word allocations)
   → *asset for hermeneutic concerns: users can fully **understand** what scores mean*
3. In the textometric approach, the specificity measure is involved in **interactive analytical paths** combining various pieces of information
   → a *kind of "dynamic" tupleization*
4. **Textual Data Analysis** actually embraces a diverse range of practices with sometimes divergent expectations and objectives, concerning for instance the place given to **text**, the extent of **automation**, and the nature of **evaluation**; this substantiates a variety of orientations and choices.

References (1/4)

- **Evert, Stefan (2004)**. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, URN urn:nbn:de:bsz:93-opus-23714.
- **Evert, Stephanie (2022)**. *corpora: Statistics and Data Sets for Corpus Frequency Data*. R package version 0.6.
- **Gries, Stefan Th. (2019)**. "15 years of collostructions: some long overdue additions / corrections (to / of actually all sorts of corpus-linguistics measures)". *International Journal of Corpus Linguistics*, 24(3), 385-412.
- **Gries, Stefan Th. (2021)**. "A new approach to (key) keywords analysis: using frequency, and now also dispersion". *Research in Corpus Linguistics*, 9(2), 1-33.
- **Heiden, Serge (2010)**. "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme". *In* Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto, & Yasunari Harada (eds.), *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*. Tokyo: Waseda University, Institute for Digital Enhancement of Cognitive Development, 389-398.

References (2/4)

- **Lafon, Pierre (1980)**. "Sur la variabilité de la fréquence des formes dans un corpus". *Mots*, 1, 127-165.
- **Lebart, Ludovic, Pincemin, Bénédicte & Poudat, Céline (2019)**. *Analyse des données textuelles*. Québec: Presses de l'Université du Québec.
- **Lebart, Ludovic, Salem, André, & Berry, Lisette (1998)**. *Exploring textual data*. Dordrecht: Kluwer Academic.
- **Loiseau, Sylvain, Vaudor, Lise, Decorde, Matthieu, Heiden, Serge (2022)**. *textometry: Textual Data Analysis Package Used by the TXM Software*. R package version 0.1.6.
- **Mayaffre, Damon, Pincemin, Bénédicte, Heiden, Serge, & Weyl, Philippe (2018)**. "L'évolution de la mémoire de la Shoah au prisme de la statistique textuelle". In Denis Peschanski & Brigitte Sion (eds), *La vérité du témoin. Mémoire et mémorialisation*. Paris: Hermann Éditeurs, Paris & Bry-sur-Marne: Institut National de l'Audiovisuel, 93-124.

References (3/4)

- **Pedersen, Ted (1996)**. "Fishing for exactness". In *Proceedings of the South-Central SAS Users group Conference*. Austin, TX, 188-200.
- **Pincemin, Bénédicte, Mayaffre, Damon, Heiden, Serge, Weyl, Philippe (2016)**. "Génétique mémorielle. Shoah, mémoire et ADT". In Damon Mayaffre, Céline Poudat, Laurent Vanni, Véronique Magri, & Peter Follette (eds.), *JADT 2016 – Statistical Analysis of Textual Data*, Nice: Presses de FacImprimeur, 697-706.
- **Reinert, Max (1983)**. "Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte". *Les cahiers de l'analyse des données*, 8(2), 187-198.
- **Reinert, Max (1990)**. "Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval". *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 26(1), 24–54.

References (4/4)

- **Salem, André (1991)**. "Les séries textuelles chronologiques". *Histoire et Mesure*, 6 (1), 149-175.
- **Salem, André (2004)**. "Introduction à la résonance textuelle". In Gérald Purnelle, Cédric Fairon, & Anne Dister (eds), *JADT 2004 – Actes des 7e Journées internationales d'analyse statistique des données textuelles*. Louvain-la-Neuve: Presses Universitaires de Louvain, 986-992.
- **Stefanowitsch, Anatol & Gries, Stefan Th. (2003)**. "Collostructions: investigating the interaction between words and constructions". *International Journal of Corpus Linguistics*, 8(2). 209-243.
- **Valette, Mathieu (2016)**. "Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée". In Damon Mayaffre, Céline Poudat, Laurent Vanni, Véronique Magri, & Peter Follette (eds.), *JADT 2016 – Statistical Analysis ofTextual Data*, Nice: Presses de FacImprimeur, 697-706.

Appendix: Formulae

	Part	Rest	Total
Word	**f**	(F-f)	**F**
Rest	(t-f)	(T-F) -(t-f)	(T-F)
Total	**t**	(T-t)	**T**

1. Probability for one frequency *f*:

$$p(k=f)=p_f=\frac{\binom{F}{f}\binom{T-F}{t-f}}{\binom{T}{t}}$$

where $\binom{n}{m}$ is the binomial coefficient:

$$\binom{n}{m}=\frac{n!}{m!\,(n-m)!}$$

$$f>F\frac{t}{T}$$

$$f\leq F\frac{t}{T}$$

2. Cumulated probabilities:

$$p(k\geq f)=\sum_{i=f}^{min(F,T)} p_i$$

$$p(k\leq f)=\sum_{i=0}^{f} p_i$$

3. Conversion to Specificity score:

$$S+=\left|\log_{10}(p(k\geq f))\right|$$

$$S-=-\left|\log_{10}(p(k\leq f))\right|$$