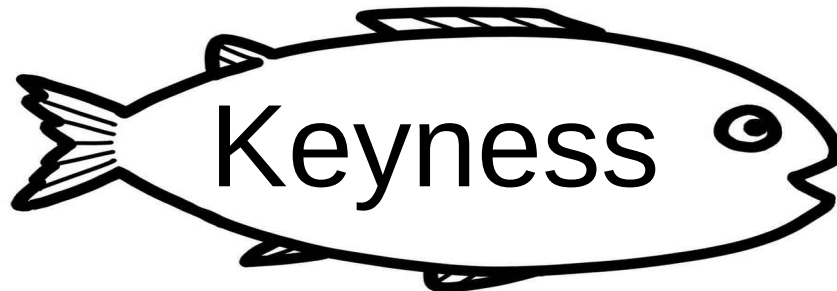




**Using and Developing Software for Keynes Analysis**  
Workshop of the [project Zeta and Company](#)  
Trier University, February 27-28, 2023.

# Fishing for Keynes



## The Specificity Measure in Textometry

Bénédicte PINCEMIN

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



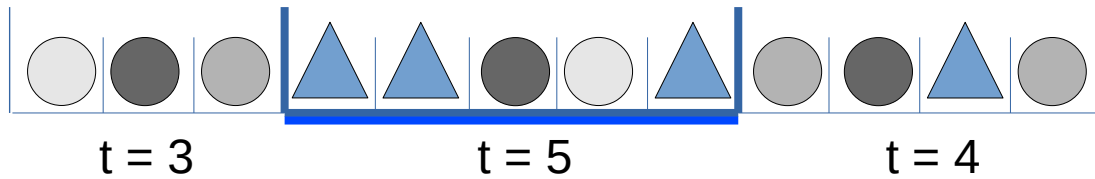
This work is licensed under the Creative Commons Attribution 4.0 International License.  
<http://creativecommons.org/licenses/by/4.0/>

# Outline

- What is the Specificity measure?
  - A clear model → complies with hermeneutic needs
  - Characteristics, strengths, weaknesses
- Specificities in practice (from **TXM** experience)
  - Synergy with other text analysis functionalities
  - Advanced settings which open and refine analytical possibilities

# The underlying statistical model

Example of a corpus with 3 parts (for instance 3 texts)  
and 4 different words:

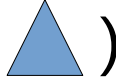


$T = 12$  *total size of the corpus*

$t = 5$  *size of the part*

$F = 4$  *total frequency of the word*

$f = 3$  *freq. of the word in the part*

We want to evaluate the frequency of one word (the blue triangle ) in one part (the central part).

**We compare it with random word allocations** (all possible word sequences with equal probability)

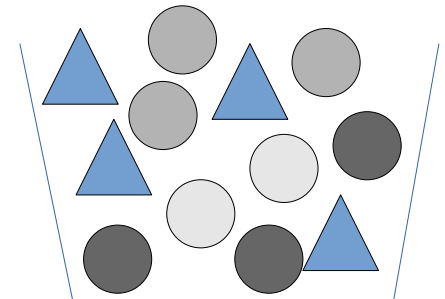


$T = 12$

$t = 5$

$F = 4$

$f = ?$



# The underlying statistical model

$f$	%	$p$
0	7	0.07
1	35	0.35
2	43	0.43
<b>3</b>	14	0.14
4	1	0.01

The probability of a frequency is given by the **proportion** of random allocations with this frequency.

3 is greater than the frequency for equal distribution ( $4 \times (5/12) = 1.6$ ), so we compute a **positive** Specificity.

We compute the **cumulative probability** of a frequency of 3 or more (one-tailed  $p$ -value):

$$p(f \geq 3) = p(f=3) + p(f=4) = 0.14 + 0.01 = 0.15$$

The Specificity is the **order of magnitude** (typically  $\text{Log}_{10}$ ) of the probability:

$$\text{If } p=0.1 \text{ then } S^+ = |\text{Log}_{10}(0.1)| = |\text{Log}_{10}(10^{-1})| = 1$$

$$\text{if } p=0.01 \text{ then } S^+ = |\text{Log}_{10}(0.01)| = |\text{Log}_{10}(10^{-2})| = 2, \text{ etc.}$$

$$\text{if } p=0.15 \text{ then } S^+ = |\text{Log}_{10}(0.15)| = |\text{Log}_{10}(10^{-0.8})| = 0.8$$

# The underlying statistical model

- In case the frequency is *less* than the frequency for equal distribution, we compute a *negative* Specificity that adds probabilities for the observed frequency *or less*.
- Examples of result interpretation:
  - A Specificity of **+4** means that, if words were distributed randomly, there would be a 1 in **10,000** 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 **100** chance to observe this frequency or less (1 % of all word allocations with such a low frequency in the part).

# The underlying statistical model

	Part	Rest	Total
Word	<b>f</b>	(F-f)	<b>F</b>
Rest	(t-f)	(T-F) -(t-f)	(T-F)
Total	<b>t</b>	(T-t)	<b>T</b>

	Word J	No Word J	Total
Word I	$n_{ij}$	$(n_i - n_{ij})$	$n_i$
No Word I	$(n_j - n_{ij})$	$(N + n_{ij} - n_i - n_j)$	$(N - n_i)$
Total	$n_j$	$(N - n_j)$	$N$

Data = **any contingency table**

The Specificity measure can be applied to:

- Keywords
- Collocations
- Links between values in two paradigms (sets)

	Feature A	All other cases: B, C, D... ( $\neg A$ )	Total
Feature $\alpha$	$n_{\alpha a}$	$(n_\alpha - n_{\alpha a})$	$n_\alpha$
All other cases: $\beta, \chi, \delta... (\neg \alpha)$	$(n_a - n_{\alpha a})$	$(N + n_{\alpha a} - n_\alpha - n_a)$	$(N - n_\alpha)$
Total	$n_a$	$N - n_a$	$N$

# The underlying model – Recap

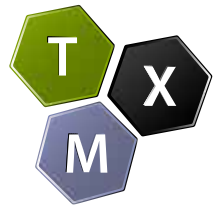
- The Specificity score evaluates whether the **frequency** of a word in a corpus part is noteworthy.
- It compares the original data to a **random** allocation of words.
- For every frequency, an exact probability is computed from the **proportion** of cases carrying out the frequency.
- Probabilities are **cumulated**: probability for the frequency *or more* (resp. *or less*) → how rare it is to **reach** such a high (resp. low) frequency.
- The Specificity score converts the probability into its **order of magnitude**.

# Dissemination in the scientific community

- In the field of Textometry
  - **Lafon 1980**, *MOTS* : Seminal paper (in French)
  - **Lebart, Salem & Berry 1998** : Handbook (in English) (Specificity score = “characteristic element diagnostic”).
  - Specificity computing is a core feature in textometric software



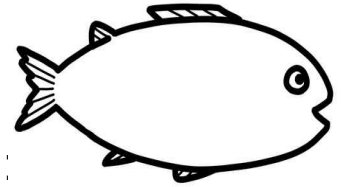
**Le Trameur**  
CLESTHIA EA-7345 - Université Sorbonne nouvelle Paris 3





# Dissemination in the scientific community

- At international level, the statistical model is known as the **Fisher's exact test** (or Fisher-Yates) – at a glance:
  - **Pedersen 1996** (“**Fishing for exactness**”, SCSUG SAS users conf.) : promotes the Fisher's exact test for the detection of word pairs
  - **Evert 2004** (*PhD*) and **2009** (*Corpus Linguistics. An international Handbook*, chap. 58 on Collocations): the Fisher's exact test is used as reference point for the comparison of measures for collocation.
  - **McEnery, Xiao & Tono 2006** (*Corpus-based language studies: An advanced resource book*): the Fisher's exact test is recommended when the expected frequency in a cell of a contingency table has a value less than 5.
  - **Gries 2012** (in the *Encyclopedia of applied linguistics*), **2014** (“Quantitative corpus approaches to linguistic analysis: seven or eight levels of resolution and the lessons they teach us”): the Fisher's exact test is applied to collexeme analysis (attraction or repulsion between a word and a slot in a syntactic construction), collocation, colligation.



# A partial success: designers' technical criticisms – discussion

- **Computationally expensive** (factorial formula, combinatorial problem)
  - *less acute with the advancing of current hardware power?*
- **Sophisticated implementation** (no straight application of formulae since this can meet computational boundaries)
  - *possibility of relying on available implementations for instance the [R package textometry](#).*
- **Limited usefulness** (results converge with those of simpler measures when frequencies and part sizes are large enough, as soon as expected frequencies are greater than 5)
  - *cases of expected frequencies less than 5 happen, and it could be clearer to consider a unified model for the whole frequency range.*



# Partial success: users' hermeneutic criticisms – discussion

- **Unclear formula** (not a concise and pragmatic ratio or percentage for instance)
  - *but the underlying model is transparent, which might be the key for hermeneutic considerations?*
- **Bias** towards high frequency words
  - *This is a straightforward consequence of a statistical approach: the more occurrences you observe, the more confident you are in your judgement, the lower the probability can be when a deviation is observed.*
  - *This is not to be corrected (the measure does what it is designed to) but this has to be taken into account in uses and may be complemented with other descriptive tools.*

# 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
    - words with a score less than 2 ( $p=0.01$ , 1%) or even 3 ( $p=0.001$ , 1 ‰) are poor candidates, since their frequency can be due to common fluctuations;
    - the  $p=0.05$  (5 %) usual 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.

# Useful mathematical properties

- **No validity threshold:** the model is exact and can deal with the full range of frequencies.
- **No need for bootstrap confidence intervals:** low frequencies that could generate unreliable/unstable results get low scores inherently
- Frequency may provide **more nuanced information than Boolean** presence/absence (document frequency)
  - However this comes with correlates that have to be understood:
    - the Specificity measure is hardly effective on short texts or small subcorpora
    - the Specificity measure is more responsive to high frequency words, that include many grammatical words
- **The  $\text{Log}_{10}$  notation** provides an efficient scale to read and rank results
  - Many probability values are very low and way below conventional threshold, without this scale conversion they may appear mixed-up or blended into a unique very low probability set (see the evolution of the [Calc corpus calculator](#) of the Czech national corpus)
- The measure must be fed with plain original word counts – **no relative frequencies**

# Not an all-in-one tool

- Measuring keyness? => Specificity measures **one aspect** (frequency variation) that can be related to keyness
- The textometric approach builds **analytical paths** that associate complementary views on data – for instance here:
  - **Word-in-context** features (such as Concordance KWIC view) are crucial to interpret the Specificity results
    - May suggest more relevant linguistic units (phrases, patterns, topics, etc.)
  - View on syntagmatic text **progression** that gives focus to **word sequence** throughout texts and corpus

VOEUX/presidents/@word

Property word

Units	Freq	DeGaulle	Index	Pompidou	index	Giscard	index	Mitterrand	index
on	100	18	-0.6	0	-2.4	11	-0.2	56	12.6
Chers	39	0	-4.1	0	-0.9	0	-2.0	27	9.3
Compatriotes	39	0	-4.1	0	-0.9	0	-2.0	27	9.3
Peur	8	0	-0.8	0	-0.2	0	-0.4	8	5.2
droit	36	2	-2.0	0	-0.9	1	-1.1	20	4.8
aura	26	2	-1.2	0	-0.6	1	-0.7	16	4.7
Europe	141	18	-2.2	2	-1.8	0	-7.2	52	4.3

VOEUX/<Compatriotes>

Query Compatriotes

text_loc, text_annee	Left context	Pivot	Right context
mitterrand, 1983	Mes Chers	Compatriotes	, A vous qui êtes réunis en
mitterrand, 1983	er cette chance. Mes Chers	Compatriotes	, voilà pour nous de grand
mitterrand, 1984	Mes Chers	Compatriotes	, Ce soir, partout en France
mitterrand, 1984	ulente. Eh bien ! Mes Chers	Compatriotes	, un pays est comme une fa
mitterrand, 1984	gue de l'espoir. Mes Chers	Compatriotes	, Ma mission est de dire la
mitterrand, 1985	heureuse année, mes Chers	Compatriotes	, Une année qui finit, une a

**VOEUX Corpus:**  
54 New Year addresses to the Nation  
by French presidents  
(1960-2013, 60 k words)



Synergy of **Specificities** (table and chart), **KWIC Concordance** and **Progression** view.



# Tackling the bag-of-word frontier

- **Progression** view is one answer to render the internal content of a part
- Another solution consists in **recursively applying** the **Specificity** computing at different scales, typically on the *part* level then on the *text* level.

54 New Year addresses to the Nation  
by French presidents  
(1960-2013, 60 k words)

File Edit Corpus Tools Utilities View Help

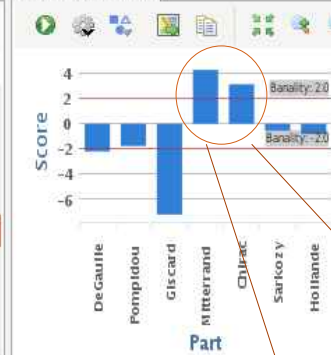


VOEUX/presidents/@frlemma

Property frlemma

Units	Freq	DeGaulle	index	Pompi	index	Giscard	index	Mitterrand	index
on	120	22	-0.6	1	-2.0	14	0.3	65	13.6
Chers	37	0	-3.9	0	-0.9	0	-1.9	27	10.2
Compatriotes	37	0	-3.9	0	-0.9	0	-1.9	27	10.2
droit	56	6	-1.5	0	-1.3	1	-2.0	28	5.3
<b>Europe</b>	<b>141</b>	<b>18</b>	<b>-2.2</b>	<b>2</b>	<b>-1.8</b>	<b>0</b>	<b>-7.2</b>	<b>52</b>	<b>4.3</b>
se	417	84	-0.5	23	0.3	40	-0.7	127	4.3
Nouvel	12	1	-0.6	1	0.3	0	-0.6	9	3.8
aimer	17	0	-1.8	0	-0.4	2	0.2	11	3.7

[Europe]



VOEUX/&lt;[frlemma="Europe"]&gt;

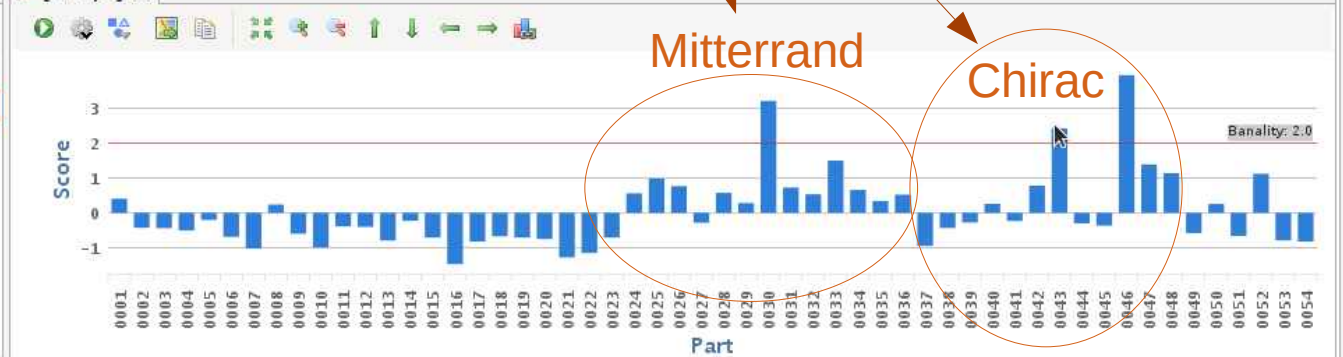
Query	[frlemma="Euro		
text_id, text_loc,	Left context	Pivot	Right context
0042, chirac, 2000	st exprimée. L'	Europe	s'est mise en
0042, chirac, 2000	mmment à Berlin	Europe	est notre nouv
0043, chirac, 2001	onnaie. C'est l'	Europe	qui avance. C'e
0043, chirac, 2001	avance. C'est l'	Europe	qui progresse.
0043, chirac, 2001	ogresse. Cette	Europe	, nous la regar
0043, chirac, 2001	combien il imp	Europe	s'affirme, qu'e
0043, chirac, 2001	e victoire de l'	Europe	. Après un sièc
0043, chirac, 2001	acon d'être en	Europe	de vivre l'Eur

VOEUX/texts/@frlemma

Property frlemma

Units	Freq	0001	index	0002	index
<b>Europe</b>	<b>141</b>	<b>2</b>	<b>0.4</b>	<b>2</b>	<b>-0.4</b>
européen	41	0	-0.2	0	-0.4
Européen	2	0	-0.0	0	-0.0
Européens	5	0	-0.0	0	-0.1
Eurydice	1	0	-0.0	0	-0.0
eux	34	1	0.6	0	-0.3
eux-	1	0	-0.0	1	1.6
évacuation	1	0	-0.0	0	-0.0

[Europe]



Mitterrand

Chirac

Specificities at the president level [top] and at the annual address text level [bottom]

# Settings: score ranges

- Positive specificity scores (top scores &  $S^+ \geq 3$ )
  - to detect words with unusually high frequency (that could be keyword candidates, *inter alia*)
- Negative specificity scores
  - Avoidances, taboos, other lexical choice...
  - *nullax* particular case ( $S^- \leq -3$  &  $f=0$ ): the word doesn't occur *and that is statistically bizarre* in the context of the corpus → a tool for the detection of interesting absences
- Low specificity scores in all parts  
( $|S| < \delta$  with typically  $\delta \in [0.5, 2]$  , & top frequencies)
  - *Basic vocabulary* (words that are common in every part)
  - Low frequencies are less interesting because they cannot get a high specificity score anyway  
→ the low score is simply a consequence of low frequency
  - Implementations: `specificities/banal` forms command in [IRaMuTeQ](#),  
`stats/BasicVocabulary` utility in [TXM](#)



# Settings: score ranges

## Illustrations with the VOEUX corpus

**Nullax** : De Gaulle : *compatriote, chômage, croissance...*

Pompidou : *compatriote, sans, république...*

Giscard : *Europe, vif...*

Mitterrand : *valeur...*

Chirac : *bonheur...*

Sarkozy : *progrès...*

Hollande (only 1 text) : *pouvoir*

No end like  
"Vive la France"

**Basic Vocabulary** : Corpus divided in Presidents, max. score  $\delta = 1.5$

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	unit	F	score_ma	DeGaulle	score	Pompidou	score	Giscard	score	Mitterrand	score	Chirac	score	Sarkozy	score	Hollande	score
2	du	1073	1,417	222	-0,4982	49	-0,8759	138	1,417	250	0,6924	285	0,3247	110	-0,7477	19	-0,9588
3	la	170	0,6545	36	-0,28	11	0,5145	20	0,3688	42	0,6164	40	-0,6545	16	-0,5563	5	0,4221
4	comme	115	0,731	22	-0,484	5	-0,384	15	0,5314	30	0,731	26	-0,6856	13	0,2786	4	0,5349
5	entre	114	1,4904	16	-1,4904	5	-0,375	16	0,7094	22	-0,5741	38	1,2138	13	0,2897	4	0,5428
6	grand	110	0,6545	26	0,5063	6	0,2651	14	0,4733	25	0,3141	25	-0,6545	12	-0,2673	2	-0,2894
7	aller	101	1,1886	17	-0,7884	7	0,5243	12	0,3486	17	-0,9335	34	1,1886	13	0,4708	1	-0,5158
8	où	95	1,2295	22	0,4314	3	-0,6138	16	1,2295	23	0,447	20	-0,8517	10	-0,2996	1	-0,4718
9	savoir	95	1,3982	13	-1,3982	6	0,3958	7	-0,7991	24	0,5658	30	0,814	13	0,5829	2	-0,2168
10	vie	89	1,1643	16	-0,577	6	0,4636	13	0,7313	19	-0,3167	18	-0,9532	15	1,1643	2	-0,19
11	falloir	83	0,8668	16	-0,4176	4	-0,2691	6	-0,7655	23	0,848	24	0,4651	10	0,3457	0	-0,8668
12	paix	80	1,4118	18	0,354	4	-0,2442	10	0,4019	25	1,4118	18	-0,594	5	-0,9788	0	-0,8354
13	dont	79	1,2125	14	-0,578	8	1,2125	11	0,587	17	-0,2966	21	0,275	8	-0,3264	0	-0,825
14	effort	60	0,4501	14	0,3964	3	-0,2243	6	-0,3056	12	-0,3878	15	-0,3305	8	0,4501	2	0,3784

life

peace

effort

# Settings: which “words”?

- Choice in **analytical property**
  - e.g. word form, lemma, lemma+POS,...
  - This choice sets linguistic expectations (what kind of types are wanted) and rules the type/token relationship (how tokens are assigned to types).
- Possibility of defining **complex lexical units**
  - Word set as a whole (typically representing a topic)
  - N-grams, morphosyntactic patterns...
  - Two tracks:
    - Focused search on a (sophisticated and precise) predefined linguistic unit (*a priori*), or
    - Overall search according to which the corpus is “tokenized” in some way (*more inductive*)

# Settings: which parts (or texts, documents)?

- Specificity applied to a “partition” (set of parts – no overlap, no exclusion)
  - Comparative analysis of  $N$  entities (authors, genres, etc.)
- Specificity applied to a “subcorpus” (1 part in a corpus)
  - Characterizing an entity within a corpus that serves as reference
- Various levels of corpus granularity: a part may be
  - a set of texts based on text metadata values
  - a set of intratextual chunks: for instance in a play, the speech turns of a character
  - a set of words: for instance, words that are not in direct speech chunks

# Settings: Table margins

- Adjusting a paradigmatic subset of variation
  - ex. searching of characteristic verbs *only within the verb set* of the corpus (in order to cancel out the overall style variation towards preference for nouns or for verbs) (Mayaffre 2006, JADT Conf.)

# Table margins: use case in Mayaffre JADT 2006

1) Specific words for Giscard are mostly nouns and related POS.

Lemmes	Occ. dans le corpus	Occ chez Giscard	Écarts réduits
1 - Actuel (adj.)	1247	684	29,7
2 - Situation (nom)	1663	749	24,3
3 - Heure (nom)	1643	686	21,0
4 - Problème (nom)	2870	1041	20,2
5 - Énergie (nom)	474	284	20,0
6 - Événement (nom)	97	96	18,3
7 - Un (déterminant)	44862	11036	18,2
8 - De (déterminant)	95156	22347	17,9
<b>9 - Indiquer (verbe)</b>	<b>351</b>	<b>208</b>	<b>16,6</b>
10 - Question (nom)	1866	689	16,5

*Tableau 1 : Les 10 premières spécificités lexicales de Giscard dans le corpus présidentiel 1958-2002*



# Table margins: use case in Mayaffre JADT 2006

2) Actually, Giscard underuses verbs (he overuses nouns and other related POS).

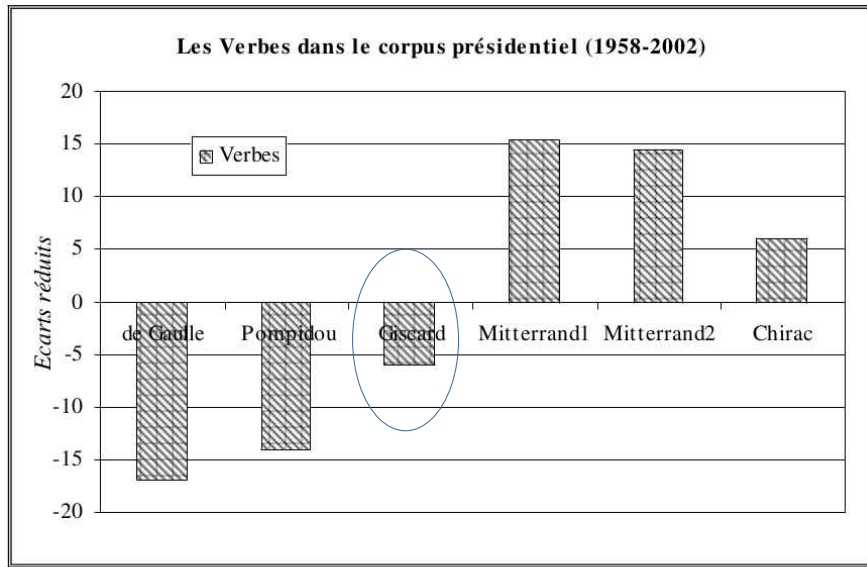


Figure 3 : Répartition des verbes dans le corpus présidentiel (1958-2002)

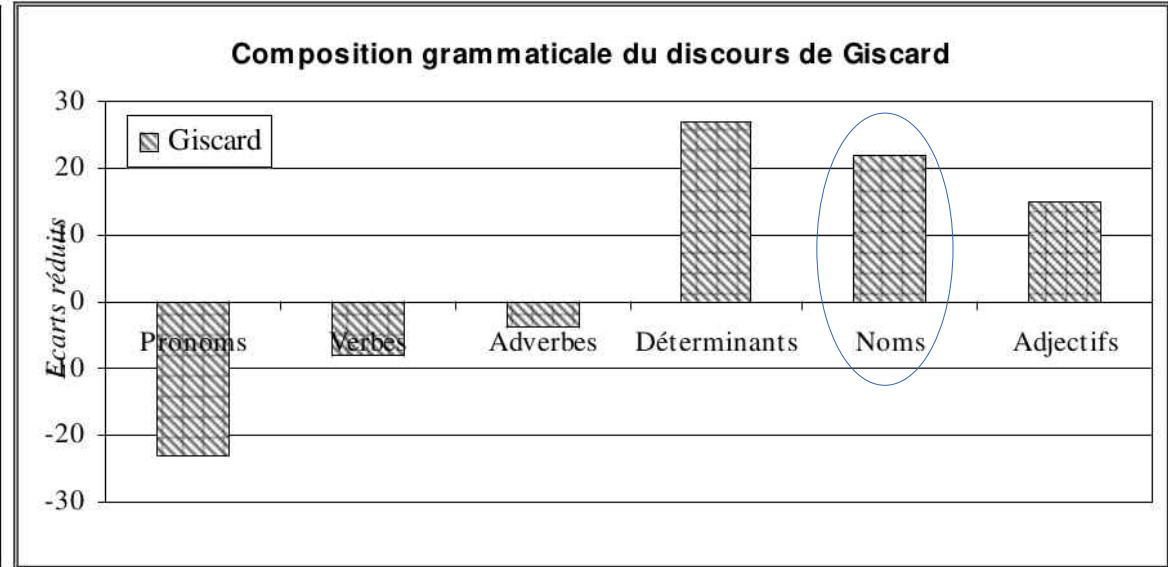
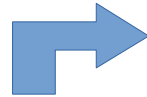


Figure 4 : Composition grammaticale du discours de Giscard

# Table margins: use case in Mayaffre JADT 2006

3) To study characteristic verbs → compute specificities considering only verbs (not all words)



	De Gaulle	Pompidou	Giscard
Forme 1	k(f1, Gaul)	k(f1, Pomp)	k(f1, Gis)
Forme 2	k(f2, Gaul)	...	...
...	...	...	...
"indiquer"	23	13	208
...	...	...	...
Forme n	k(fn, Gaul)	...	...
<b>Total t</b>	<b>224119</b>	<b>237536</b>	<b>410855</b>

	De Gaulle	Pompidou	Giscard	Mitterrand1	Mitterrand2	Chirac	<b>Total T(v)</b>
Verbe 1	k(v1, Gaul)	k(v1, Pomp)	k(v1, Gis)	k(v1, Mit1)	k(v1, Mit2)	k(v1, Chir)	K(v1)
Verbe 2	k(v2, Gaul)	...	...	...	...	...	...
...	...	...	...	...	...	...	...
"indiquer"	23	13	208	36	16	55	351
...	...	...	...	...	...	...	...
Verbe n	k(vn, Gaul)	...	...	...	...	...	K(vn)
<b>Total t(v)</b>	<b>30361</b>	<b>32740</b>	<b>59736</b>	<b>58247</b>	<b>53462</b>	<b>51928</b>	<b>286.474</b>

Tableau 3 : Matrice des données pour le calcul des spécificités grammaticalisées

"indiquer"	23	13	208	36	16	55	351
...	...	...	...	...	...	...	...
Forme n	k(fn, Gaul)	...	...	...	...	...	K(fn)
<b>Total t</b>	<b>224119</b>	<b>237536</b>	<b>410855</b>	<b>370182</b>	<b>340364</b>	<b>340926</b>	<b>1923982</b>

Tableau 2 : Matrice des données pour le calcul des spécificités traditionnelles

# Settings: Table margins

- Adjusting a paradigmatic subset of variation
  - In TXM, many possibilities:
    - Direct specificity computing → entire corpus as reference
    - From Index word selection to rows in Lexical table, two options:
      - Calculate the margins from the frequencies of all the words of the corpus → entire corpus as reference
      - Calculate the margins only from the frequencies of the elements of the index → selected word set as reference
    - Specificity computing for an input of any (externally built) contingency table → table content as reference
    - Full customized parameters values: the `r/PlotSpecif` utility computes  $S$  from any input values for  $f$ ,  $F$ ,  $t$ ,  $T$  parameters → customized selection underlying  $T$  as reference

# Settings for Word, Parts and Table margins: use case in Guillot et al. 2013



- Field = Diachronic linguistics – evolution of French language
- Corpus = taken from the *Base de Français Médiéval (BFM)*
- Word = patterns with an infinitive verb
- Part = direct speech (DD) vs other words
- Margins = infinitives

	F totale	f dans le non DD	Spécif pour le non DD	f dans le DD	Spécif pour le DD
Pouvoir + infinitif	1240	628	-13	612	13
Devoir + infinitif	521	179	-34	342	34
Vouloir + infinitif	417	200	-8	217	8
Aller + infinitif	223	163	5	60	-8
Vb + infinitif	4267	2288	-22	1979	22
Taille	T=364967	t=221814		t=143153	

# Settings: Table margins

- Effective impact on meaning and interpretation
  - There may be **several relevant** choices for margins but they implement **different meanings**.
  - Ex.: S+ for personal pronouns (PP)
    - When margins = all words: “Are some PP overused? Which ones?”
      - There may be none.
      - There may be all of them.
    - When margins = all PP: “Is there a balanced use of PP (reflecting their overall mean use in the corpus)? When a PP is used, is there a noticeable preference towards some of them?”
      - There may be a preference for a PP that is an underused word of the corpus, if the part globally underuses PP.
      - It cannot happen that all PP are overused (resp. underused), because what is considered is the balance within the PP paradigm.

# Personal Pronouns Specificities in the VOEUX corpus

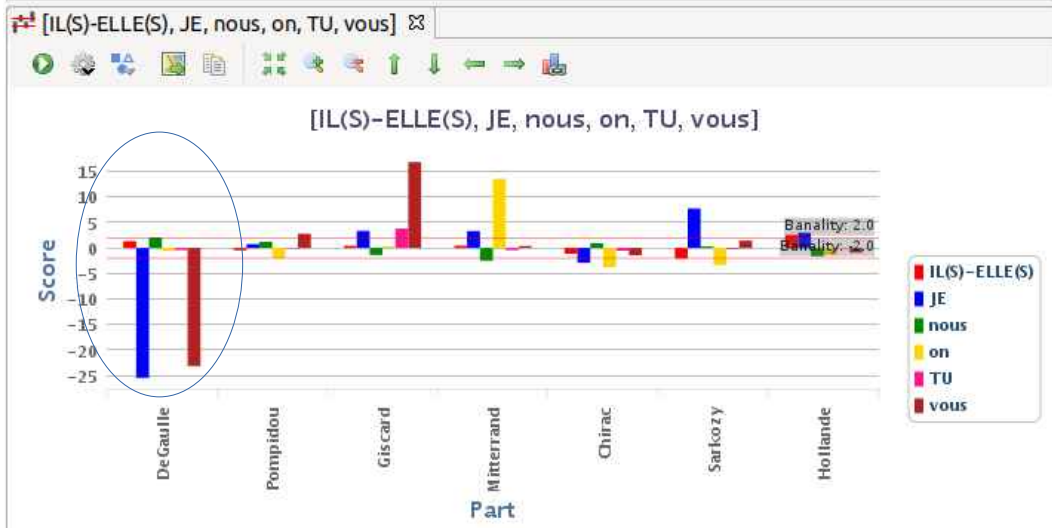


VOEUX/presidents/<[frlemma="je|me|moi|t...]

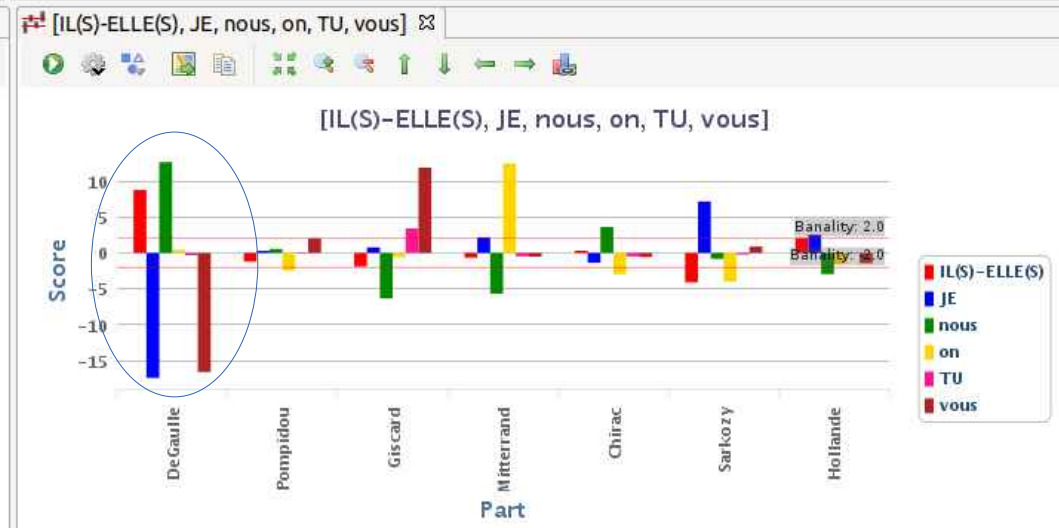
Units	Frei	DeGaulle	t=13054	index	Pompi	index	Giscar	index	Mitter	index	Chirac	index
#RESTE#	58390		12586	9.9	3111	-1.6	6401	-6.5	12918	-2.0	15505	3.0
nous	825		205	2.1	55	1.2	76	-1.4	151	-2.5	234	1.0
IL(S)-ELLE(S)	725		174	1.4	36	-0.5	84	0.4	166	0.5	174	-1.1
JE	708		49	-25.5	44	0.8	108	3.4	195	3.3	152	-2.9
vous	425		18	-23.2	38	2.8	110	16.9	97	0.4	96	-1.4
on	120		22	-0.6	1	-2.0	14	0.3	65	13.6	15	-3.8
TU	4		0	-0.4	0	-0.1	4	3.8	0	-0.4	0	-0.5

VOEUX/presidents/<[frlemma="je|me|moi|t...]

Units	Frei	DeGaulle	t=468	index	Pompi	index	Giscar	index	Mitte	index	Chirac	index
nous	825		205	12.6	55	0.6	76	-6.3	151	-5.7	234	3.0
IL(S)-ELL	725		174	8.8	36	-1.2	84	-1.9	166	-0.7	174	0.0
JE	708		49	-17.4	44	0.3	108	0.8	195	2.2	152	-1.1
vous	425		18	-16.6	38	2.0	110	11.9	97	-0.5	96	-0.4
on	120		22	0.5	1	-2.4	14	-0.6	65	12.5	15	-2.9
TU	4		0	-0.3	0	-0.1	4	3.4	0	-0.5	0	-0.4



Margins = all words



Margins = PP only

# Summary

1. The specificity measure is a **Fisher's exact test**, a powerful statistical test dedicated to contingency tables, supplemented with a convenient notation  
→ *it has sound theoretical foundations*
2. It evaluates word **frequency** variation among parts in a reference corpus  
→ *no semantic claim, no direct keyness -but maybe a piece of it?*
3. It implements a **transparent modeling** (portion in all random word allocations)  
→ *asset for hermeneutic concerns:  
users can fully **understand** what scores mean*
4. **Available packages** overcome the computational complexity of this test  
→ *affordable implementation*

# Summary

5. In the textometric approach, the specificity measure is involved in **interactive analytical paths** combining various computations and views
  - *a descriptive tool in a toolbox rather than an efficient and direct integrated measure*
6. The **bag-of-word** frontier may be addressed by subsequently processing what is in the bag
  - *from that perspective, the problem may not come from the measure itself but from the way it is used*
7. **Settings** (especially from textual data to contingency table input) open lots of analytical ways and have a large impact on the meaning of outputs
  - *this potential is worth exploring and managing*



If this talk was a unique keyword?

?

# If this talk was a unique keyword?

**The Statistical  
Model**

*seeking meaning*

**Parameter  
settings**  
Potential  
and impact

**interpretation**

*understanding*

*making sense*

**Results**

with KWIC and other  
complementary  
analyses and views

# References (1/3)

- **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, Stefan (2009).** "Corpora and collocations". In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics. An International Handbook*, vol. 2, article 58, 1212-1248. Berlin: Mouton de Gruyter.
- **Gries, Stefan Th. (2012).** "Corpus linguistics: quantitative methods". In Carol A. Chapelle (ed.), *The encyclopedia of applied linguistics*. Oxford: Wiley-Blackwell, 1380-1385.
- **Gries, Stefan Th. (2014).** "Quantitative corpus approaches to linguistic analysis: seven or eight levels of resolution and the lessons they teach us". In Irma Taavitsainen, Merja Kytö, Claudia Claridge, & Jeremy Smith (eds.), *Developments in English: expanding electronic evidence*. Cambridge: Cambridge University Press, 29-47.

# References (2/3)

- **Guillot, Céline, Lavrentiev, Alexei, Pincemin, Bénédicte, & Heiden, Serge (2013).** "Le discours direct au Moyen Âge : vers une définition et une méthodologie d'analyse". In Dominique Lagorgette & Pierre Larrivée (eds), *Représentations du sens linguistique 5*. Chambéry: Université de Savoie, 17-41.
- **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.
- **Lafon, Pierre (1980).** "Sur la variabilité de la fréquence des formes dans un corpus". *Mots*, 1, 127-165.
- **Lebart, Ludovic, Salem, André, & Berry, Lisette (1998).** *Exploring textual data*. Dordrecht: Kluwer Academic.

# References (3/3)

- **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.
- **McEnery, Tony, Xiao, Richard, & Tono, Yukio (2006).** *Corpus-based language studies: An advanced resource book*. Abingdon: Routledge.
- **Mayaffre, Damon (2006).** "Faut-il prendre en compte la composition grammaticale des textes dans le calcul des spécificités lexicales ? Tests logométriques appliqués au discours présidentiel sous la Vème République". In Jean-Marie Viprey (ed.), *JADT 2006 – Actes des 8e Journées internationales d'analyse statistique des données textuelles*. Besançon: Presses Universitaires de Franche-Comté, 677-685.
- **Pedersen, Ted (1996).** "Fishing for exactness". In *Proceedings of the South-Central SAS Users group Conference*. Austin, TX, 188-200.

# Appendix: Formulae

	Part	Rest	Total
Word	<b>f</b>	(F-f)	<b>F</b>
Rest	(t-f)	(T-F) -(t-f)	(T-F)
Total	<b>t</b>	(T-t)	<b>T</b>

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:

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

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:

$$S_+ = |\log_{10}(p(k \geq f))|$$

$$S_- = -|\log_{10}(p(k \leq f))|$$

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

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