European/International Joint PhD
in Social Representations and Communication
International Lab Meeting - Spring Session 2015

European Commission REA-Research Executive Agency
FP7 - PEOPLE Initial Training Networks
So.Re.Com. Joint-IDP
(PITN-GA-2013-607279)

Funded by the European Union

The “Anthropological”, “Narrative”, “Dialogical” and “Subjective” paradigmatic approaches to Social Representations

at the European/International Joint PhD in Social Representations & Communication

Research Center and Multimedia LAB

SAPIENZA
UNIVERSITÀ DI ROMA

An introduction to IRAMUTEQ &
Research examples of IRAMUTEQ application

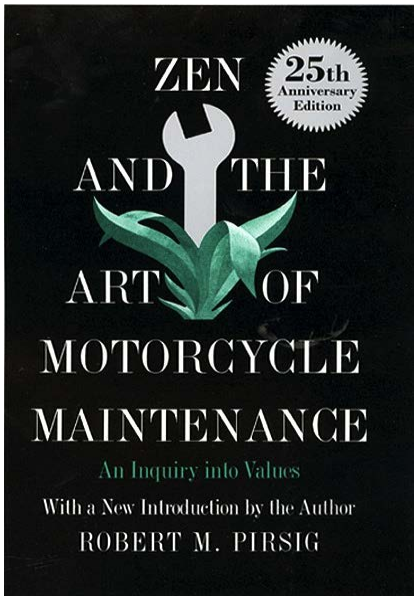
SAPIENZA
UNIVERSITÀ DI ROMA

Mauro Sarrica April 29th 2015

Statistics - Data Analysis - Software

*"The test of the machine is the **satisfaction** it gives you. There isn't any other test. If the machine produces tranquility it's right. **If it disturbs you it's wrong until either the machine or your mind is changed**."*

***"When analytic thought, the knife, is applied to experience, something is always killed in the process**."*



Content analysis – Lexicometric analysis

*"content analysis as the use of **replicable** and **valid** method for making specific inferences from text to other states or properties of its source"* (KRIPPENDORFF 1969, p.103).

*"a research technique for making replicable and valid **inferences from data to their context**"*

Inferences about who / why / how / effects / whom

Usual chain:

Categories – Text – (*new cat.*) – Frequencies – Statistics – Inferences

Content analysis – Lexicometric analysis

Lexicometric analysis – Automatic treatment of text

Qualitative *and* quantitative analysis of contents, properties and characteristics of texts

No need to read the texts --- really?

Objective measures --- or at least uniform results

Constructs models of the meanings present in a corpus

Content analysis – Lexicometric analysis

Benjamin Bourdon, *1888, Uni. Rennes, Exodus,*

- *keep the relevant words (excluding 'empty words)*
- *counts frequencies*
- *thematic classification*

Stenographers

- *F. W. Kaeding,1898, frequency of graphemes, sillables and words in german*
- *J. B. Estoup, france, define the notion of Rank (important for speed)*

Psychologists

- *A. Busemann, 1925, language in child.*
  Active (verbs) – Qualitative (adjective and adverbs)
- *D.P. Boder, '40, "adjective-verb ratio" index of affective instability*

1929 – George Zipf: Rank * Frequency = K

Giuliano, L., & La Rocca, G. (2008). *L'analisi automatica e semi-automatica dei dati testuali.* Milano: LED.

Content analysis – Lexicometric analysis

G. Yule, *1939, Literary texts*

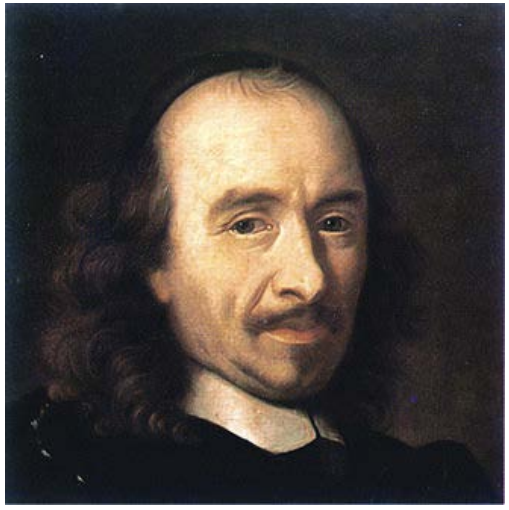
- *Authors can be identified looking at the characteristics of the corpus*

P Guiraud, *'50,*

- *relationships between length of the text (dimension i.e. n. of occurrences & extension of the vocabulary)*
- *Few words cover more than 50% of the occurrences "mots-outils"*
- *Concentration: high frequency words, keywords*
- *Extension: low frequency, variety, eccentricity*

Giuliano, L., & La Rocca, G. (2008). *L'analisi automatica e semi-automatica dei dati testuali.* Milano: LED.

Content analysis – Lexicometric analysis



Centre etudes Besançon

Automatic data processing of Corneille,

- *1957 - Trésor de la Langue Française,*
- *Texts as a representative sample*
- *From corpora to language*
- *External resources*
- *Specificities and Peculiar language*

Roberto Busa: *Index Tomisticus*,

lexicon of San Tommaso D'Aquino,

started in 1946 - 30 years - 56 volumes (Busa, 1974-1980)

IBM ==> Linguistic informatics begins

Giuliano, L., & La Rocca, G. (2008). *L'analisi automatica e semi-automatica dei dati testuali.* Milano: LED.

Content analysis – Lexicometric analysis

‘60 - Jean-Paul Benzécri

Multidimensional analysis
Analyse des Correspondances
French school of ***Analyse des Données***

’80, Ludovic Lebart - Alain Morineau (1985)
SPAD (Système Portable pour l’Analyse des Données)
With André Salem, Mónica Bécue (**SPAD-T**),

now open source: **DTM (Data Text Mining)**

Words as graphic element ==> Reconstructing the meaning afterward

Giuliano, L., & La Rocca, G. (2008). *L’analisi automatica e semi-automatica dei dati testuali.* Milano: LED.

What role for psychology

What role for psychology – (Kvale postmodern psy?)

Modern psychology, whether in the naturalist or the humanist version has become an intellectual second-hand store,
Displaying a variety of collections from last year's fashion of the neighbouring disciplines - 'you name it, we have it'

.

Statistics - Data Analysis - Software

Why these data

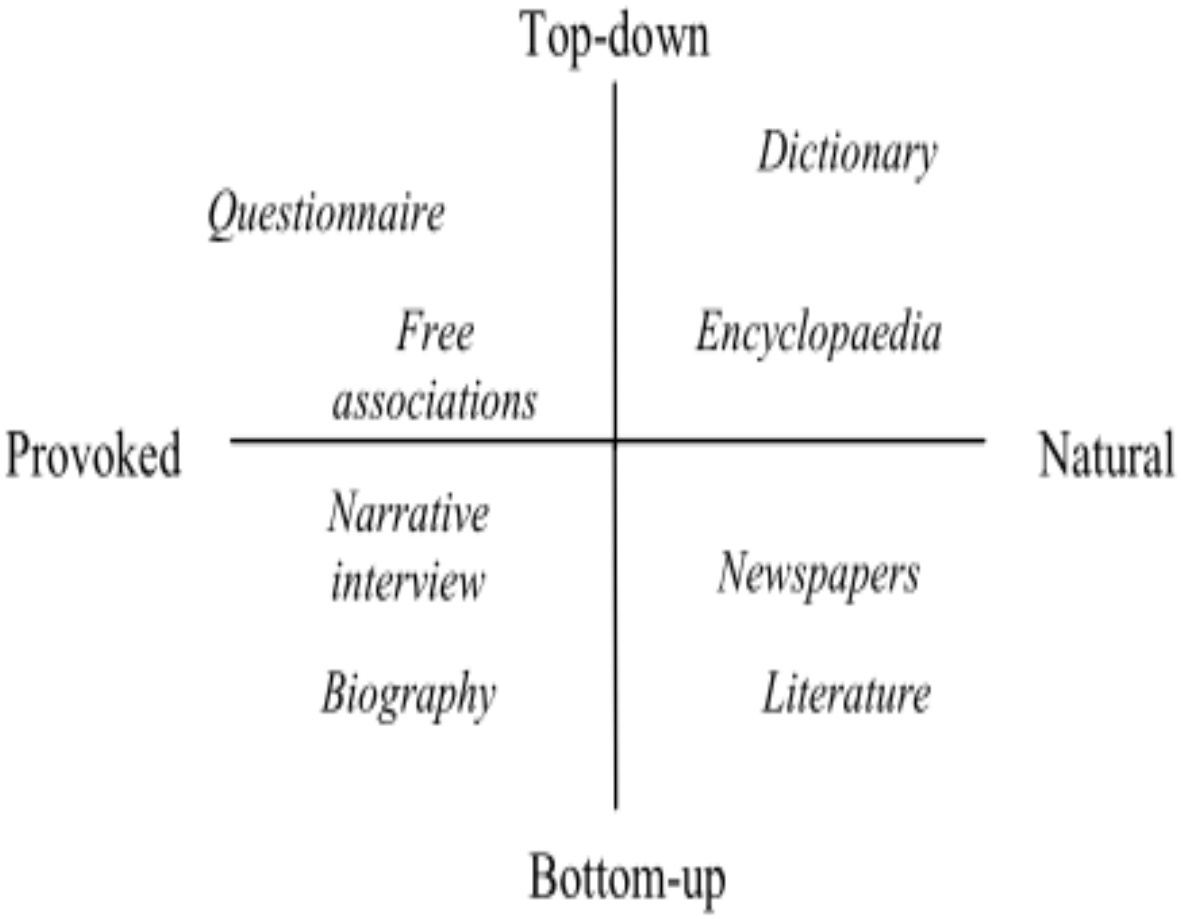
Why these analyses

Why this procedure

Why this software

Who is the audience

Corpus



(Chartier & Meunier, 2011)

Corpus

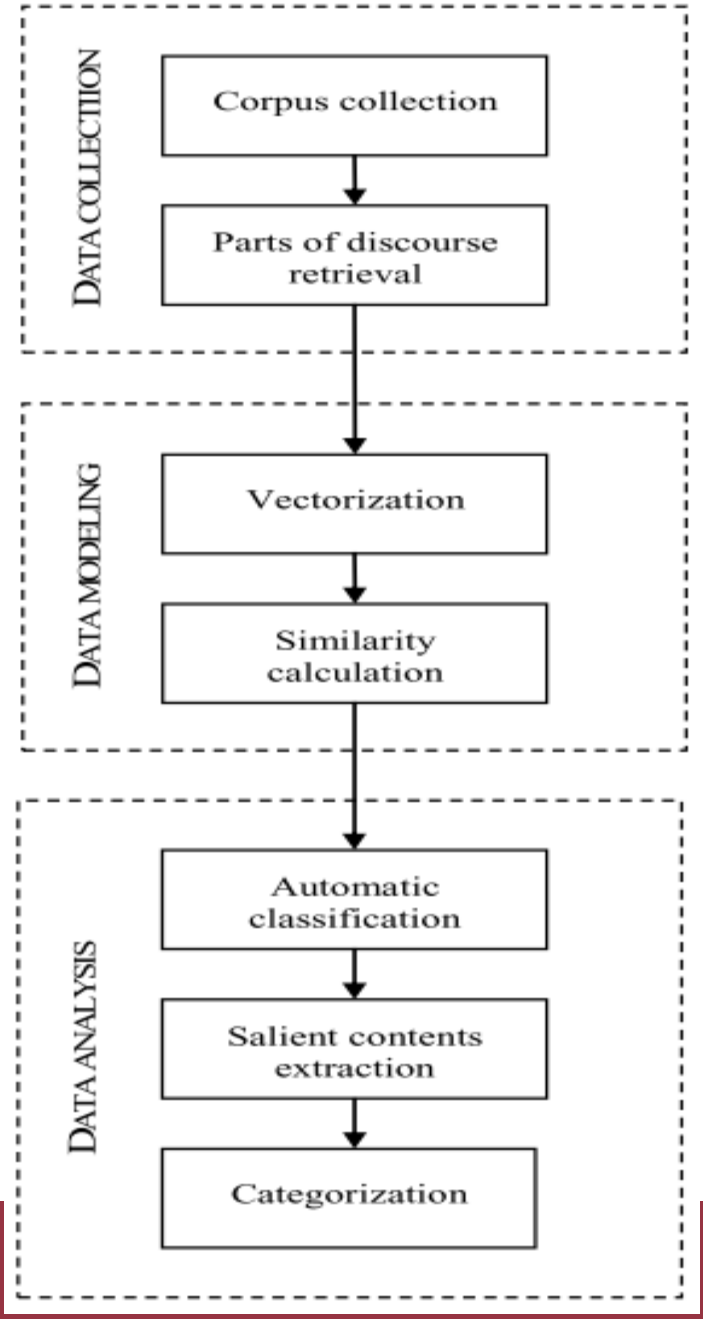
http://onlinebooks.library.upenn.edu/

"The Online Books Page is a digital library project directed by John Mark Ockerbloom, researcher at the University of Pennsylvania. It contains over 1 million volumes, having placed on the Web the sources of various universities and various projects, including the Gutemberg and Manuzio ones. It offers therefore volumes in all the languages of the world."

Other resources e.g.
Internet Sacred Text Archive – sacred tradition, mythology

Chamber of Deputies archives

Phases



(Chartier & Meunier, 2011)

Corpus

Collection of units which are pertinent and coherent

- *Homogeneity*
  - *context of production*
  - *Lexicometric characteristics*

Corpus > Texts > Fragments

Type / Token / Lemma

In affirming the intrinsically social nature of self, a social psychology of community would extend social psychology as a discipline.

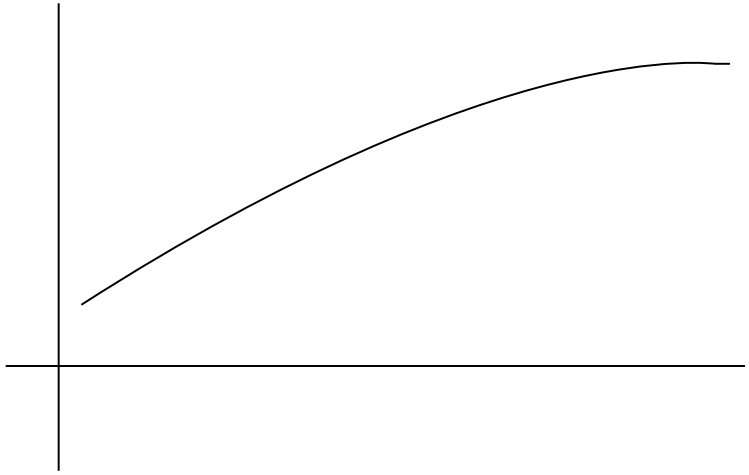
	Dimens.-Occurrences	*Vocabulary- Formes*	
Text	**id-Token**	**id-Type**	**Lemma**
In	1	1	*in*
Affirming	2	2	*to_affirm*
The	3	3	*the*
Intrinsically	4	4	*intrinsic*
Social	5	5	*social*
Nature	6	6	*nature*
Of	7	7	*of*
Self	8	8	*self*
A	9	9	*a*
Social	10	5	*social*
Psychology	11	10	*psychology*
Of	12	7	*of*
Community	13	11	*community*
Would	14	12	*to_will*
Extend	15	13	*to_extend*
Social	16	5	*social*
Psychology	17	10	*psychology*
As	18	14	*as*
A	19	9	*a*
Discipline	20	15	*discipline*

Characteristics of the Corpus

25.000 < Corpus < 1.000.000 – 5.000.000 occurrences

Type/Token
Formes/Occurrences ratio < 20%

Hapax < 50% Formes



Lemmatization

List based / Algorithm based

IRAMUTEQ – List based
/Applications/iramuteq.app/Contents/Resources/Dictionnaires

Now added ‘import from txm’
Uses treetagger (R) – context-based disambiguation

a_confronto	*a_confronto*	*sw*
a_cui	*a_cui*	*sw*
abachi	*abaco*	*nom*
abaco	*abaco*	*nom*
abacà	*abacà*	*nom*
abate	*abate*	*nom*
abati	*abate*	*nom*
abatini	*abatino*	*nom*
abatino	*abatino*	*nom*
abbacchi	*abbacchiare*	*ver*
abbacchiare	*abbacchiare*	*ver*
abbacchiato	*abbacchiato*	*adj*
abbacchiatura	*abbacchiatura*	*nom*
abbacchio	*abbacchio*	*nom*
abbacinamento	*abbacinamento*	*nom*
abbacinante	*abbacinare*	*ver*
abbacinanti	*abbacinare*	*ver*
abbacinare	*abbacinare*	*ver*
abbacinati	*abbacinare*	*ver*
abbacinato	*abbacinare*	*ver*

Résumé | Formes actives | Formes supplémentaires | Total | Hapax

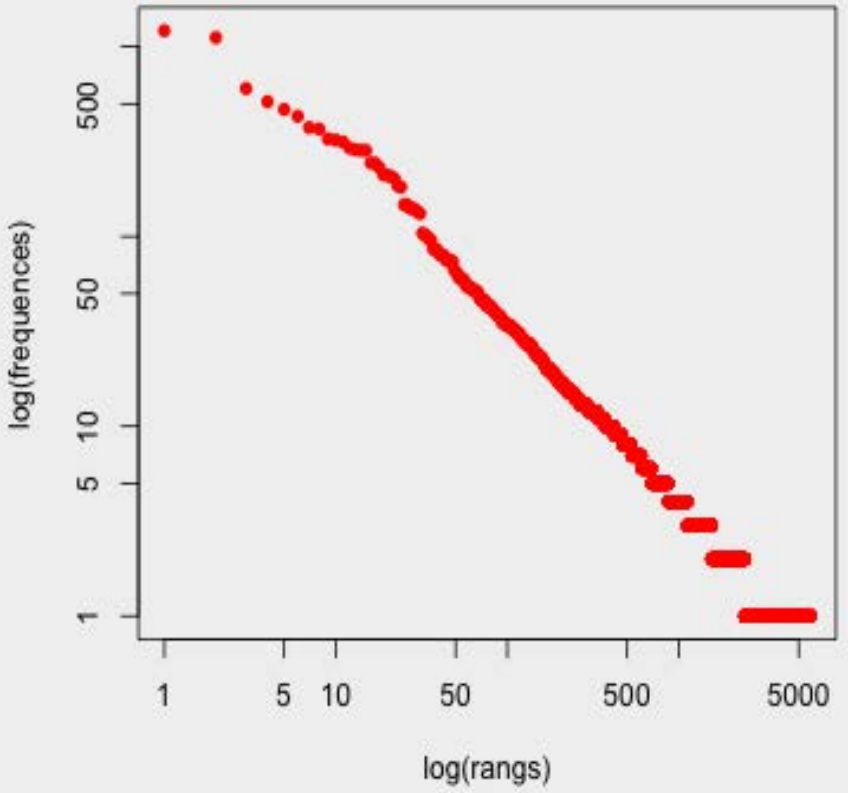
Résumé
Nombre de textes : 12
Nombre d'occurrences : 29439
Nombre de formes : 5827
Nombre d'hapax : 3400 (11.55%des occurrences - 58.35% des formes)
Moyenne d'occurrences par texte : 2453.25

BEFORE LEMMATIZATION

Formes 5827

Occurrences 29439 *ratio =19.8*

Hapax 3400 = 58.36 % Formes



Résumé | Formes actives | Formes supplémentaires | Total | Hapax

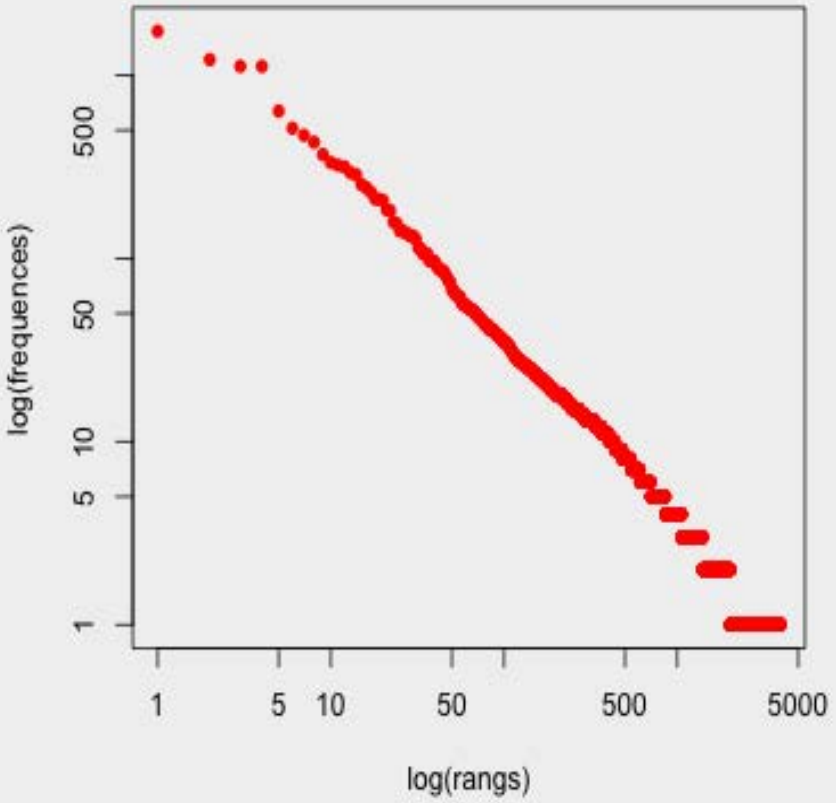
Résumé
Nombre de textes : 12
Nombre d'occurrences : 29439
Nombre de formes : 3927
Nombre d'hapax : 1902 (6.46%des occurrences - 48.43% des formes)
Moyenne d'occurrences par texte : 2453.25

AFTER LEMMATIZATION

Formes 3927

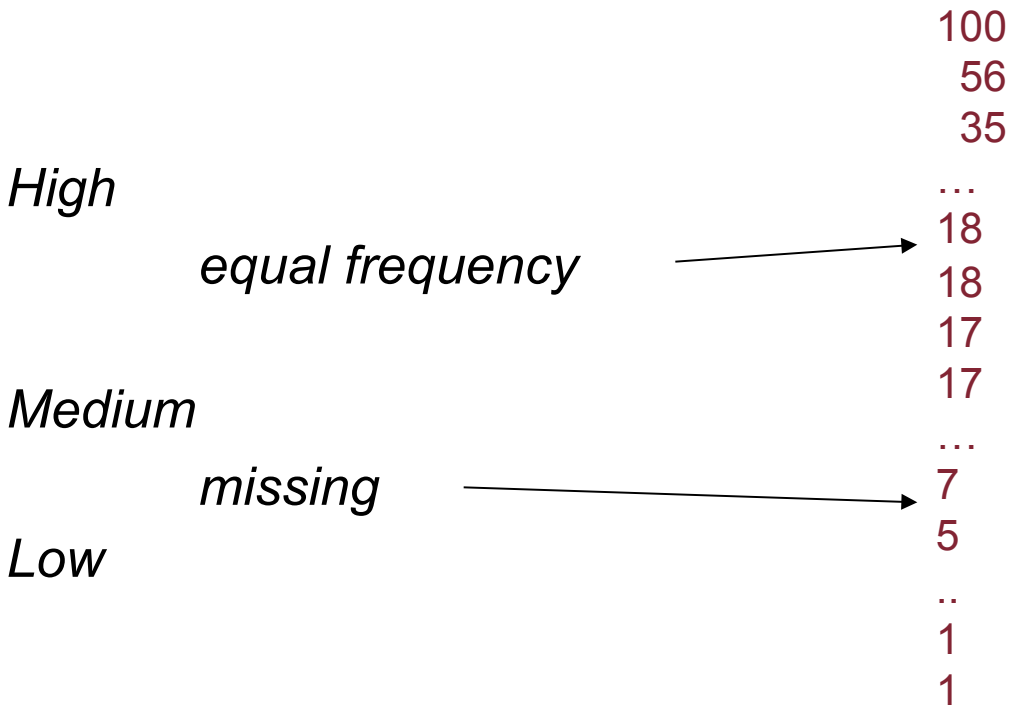
Occurrences 29439 *ratio =13.3*

Hapax 1902 = 48.43 % Formes



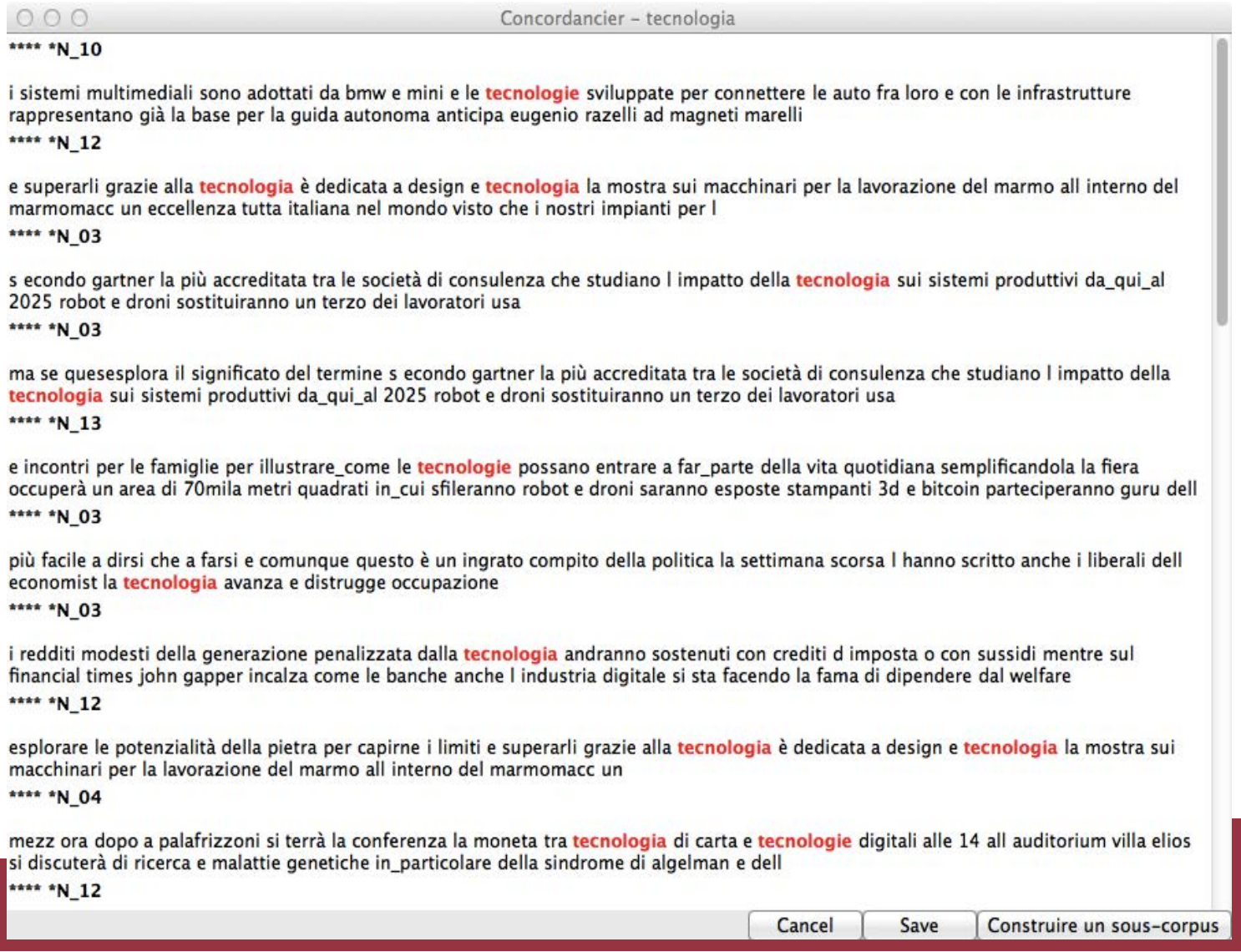
Lists of Words

Empty / Full words … interpretation depends on the goals



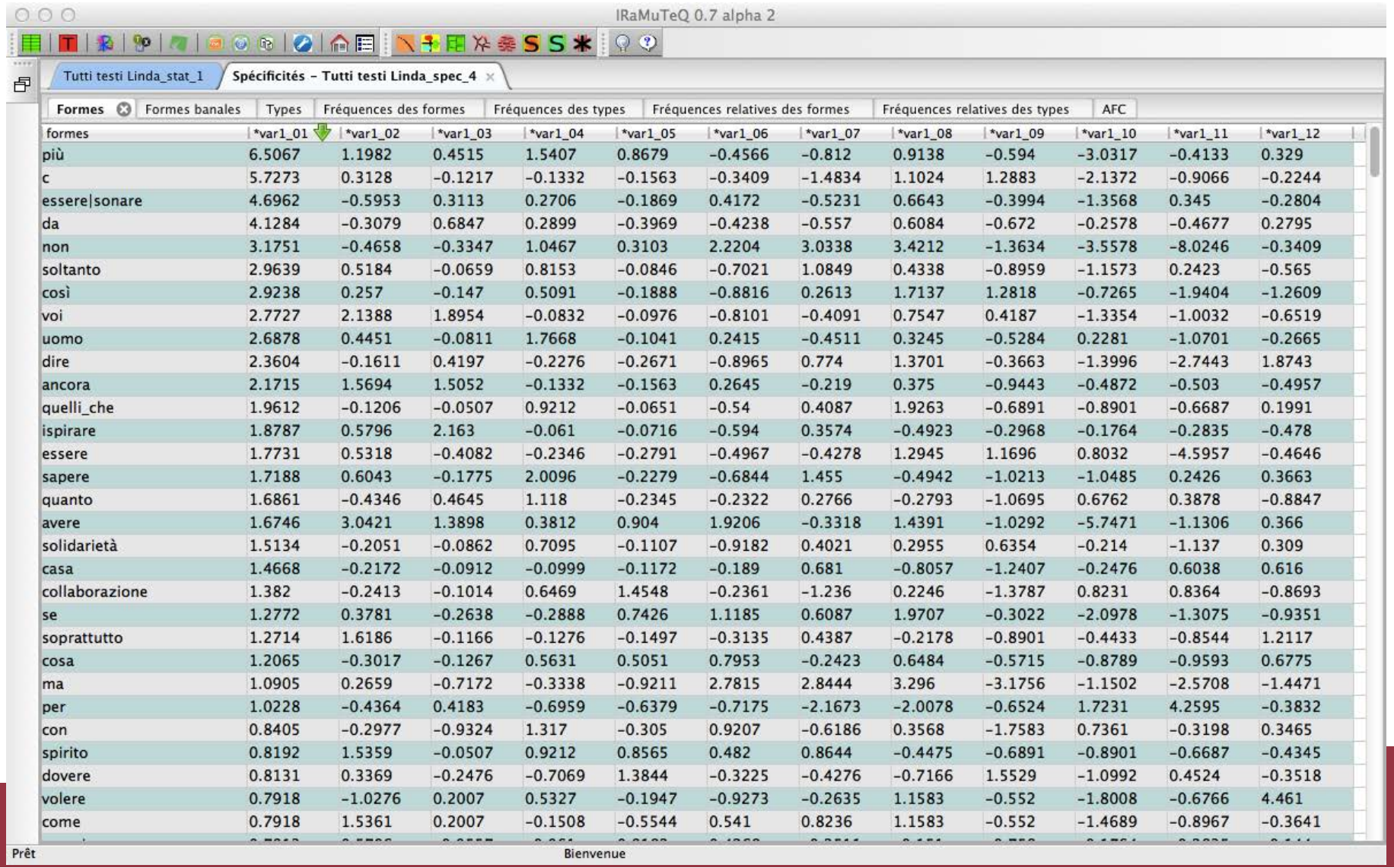
Concordances

The 'target' word in its local context



Specificities

Comparing texts - Lexicon over/under represented



formes	*var1_01	*var1_02	*var1_03	*var1_04	*var1_05	*var1_06	*var1_07	*var1_08	*var1_09	*var1_10	*var1_11	*var1_12
più	6.5067	1.1982	0.4515	1.5407	0.8679	-0.4566	-0.812	0.9138	-0.594	-3.0317	-0.4133	0.329
c	5.7273	0.3128	-0.1217	-0.1332	-0.1563	-0.3409	-1.4834	1.1024	1.2883	-2.1372	-0.9066	-0.2244
essere|sonare	4.6962	-0.5953	0.3113	0.2706	-0.1869	0.4172	-0.5231	0.6643	-0.3994	-1.3568	0.345	-0.2804
da	4.1284	-0.3079	0.6847	0.2899	-0.3969	-0.4238	-0.557	0.6084	-0.672	-0.2578	-0.4677	0.2795
non	3.1751	-0.4658	-0.3347	1.0467	0.3103	2.2204	3.0338	3.4212	-1.3634	-3.5578	-8.0246	-0.3409
soltanto	2.9639	0.5184	-0.0659	0.8153	-0.0846	-0.7021	1.0849	0.4338	-0.8959	-1.1573	0.2423	-0.565
così	2.9238	0.257	-0.147	0.5091	-0.1888	-0.8816	0.2613	1.7137	1.2818	-0.7265	-1.9404	-1.2609
voi	2.7727	2.1388	1.8954	-0.0832	-0.0976	-0.8101	-0.4091	0.7547	0.4187	-1.3354	-1.0032	-0.6519
uomo	2.6878	0.4451	-0.0811	1.7668	-0.1041	0.2415	-0.4511	0.3245	-0.5284	0.2281	-1.0701	-0.2665
dire	2.3604	-0.1611	0.4197	-0.2276	-0.2671	-0.8965	0.774	1.3701	-0.3663	-1.3996	-2.7443	1.8743
ancora	2.1715	1.5694	1.5052	-0.1332	-0.1563	0.2645	-0.219	0.375	-0.9443	-0.4872	-0.503	-0.4957
quelli_che	1.9612	-0.1206	-0.0507	0.9212	-0.0651	-0.54	0.4087	1.9263	-0.6891	-0.8901	-0.6687	0.1991
ispirare	1.8787	0.5796	2.163	-0.061	-0.0716	-0.594	0.3574	-0.4923	-0.2968	-0.1764	-0.2835	-0.478
essere	1.7731	0.5318	-0.4082	-0.2346	-0.2791	-0.4967	-0.4278	1.2945	1.1696	0.8032	-4.5957	-0.4646
sapere	1.7188	0.6043	-0.1775	2.0096	-0.2279	-0.6844	1.455	-0.4942	-1.0213	-1.0485	0.2426	0.3663
quanto	1.6861	-0.4346	0.4645	1.118	-0.2345	-0.2322	0.2766	-0.2793	-1.0695	0.6762	0.3878	-0.8847
avere	1.6746	3.0421	1.3898	0.3812	0.904	1.9206	-0.3318	1.4391	-1.0292	-5.7471	-1.1306	0.366
solidarietà	1.5134	-0.2051	-0.0862	0.7095	-0.1107	-0.9182	0.4021	0.2955	0.6354	-0.214	-1.137	0.309
casa	1.4668	-0.2172	-0.0912	-0.0999	-0.1172	-0.189	0.681	-0.8057	-1.2407	-0.2476	0.6038	0.616
collaborazione	1.382	-0.2413	-0.1014	0.6469	1.4548	-0.2361	-1.236	0.2246	-1.3787	0.8231	0.8364	-0.8693
se	1.2772	0.3781	-0.2638	-0.2888	0.7426	1.1185	0.6087	1.9707	-0.3022	-2.0978	-1.3075	-0.9351
soprattutto	1.2714	1.6186	-0.1166	-0.1276	-0.1497	-0.3135	0.4387	-0.2178	-0.8901	-0.4433	-0.8544	1.2117
cosa	1.2065	-0.3017	-0.1267	0.5631	0.5051	0.7953	-0.2423	0.6484	-0.5715	-0.8789	-0.9593	0.6775
ma	1.0905	0.2659	-0.7172	-0.3338	-0.9211	2.7815	2.8444	3.296	-3.1756	-1.1502	-2.5708	-1.4471
per	1.0228	-0.4364	0.4183	-0.6959	-0.6379	-0.7175	-2.1673	-2.0078	-0.6524	1.7231	4.2595	-0.3832
con	0.8405	-0.2977	-0.9324	1.317	-0.305	0.9207	-0.6186	0.3568	-1.7583	0.7361	-0.3198	0.3465
spirito	0.8192	1.5359	-0.0507	0.9212	0.8565	0.482	0.8644	-0.4475	-0.6891	-0.8901	-0.6687	-0.4345
dovere	0.8131	0.3369	-0.2476	-0.7069	1.3844	-0.3225	-0.4276	-0.7166	1.5529	-1.0992	0.4524	-0.3518
volere	0.7918	-1.0276	0.2007	0.5327	-0.1947	-0.9273	-0.2635	1.1583	-0.552	-1.8008	-0.6766	4.461
come	0.7918	1.5361	0.2007	-0.1508	-0.5544	0.541	0.8236	1.1583	-0.552	-1.4689	-0.8967	-0.3641

Specificities

Are based on hypergeometric law. It measures a probability

Range 0 -1. Often low, such as: 0,0000617

Scientific: 6,17E-5, that is 6,17 x 10^{-5}

The tables report the exponent , e.g 5

P<.05 becomes 0,05

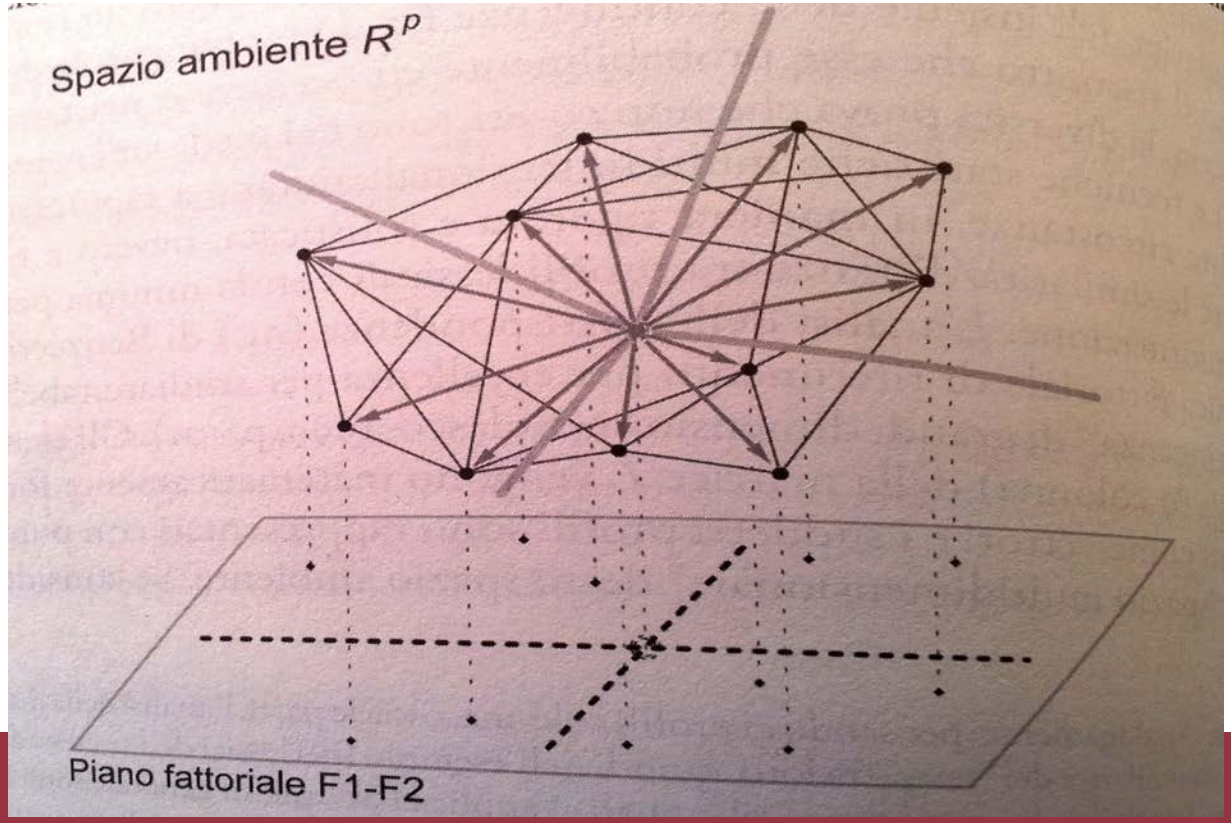
p< 5 E-2

That is 2 or more!!

Correspondence factor analysis

Reducing the variability of large co-occurence tables into low dimensional space

Bolasco, 2013, p.224



Corpus

http://onlinebooks.library.upenn.edu/

"The Online Books Page is a digital library project directed by John Mark Ockerbloom, researcher at the University of Pennsylvania. It contains over 1 million volumes, having placed on the Web the sources of various universities and various projects, including the Gutemberg and Manuzio ones. It offers therefore volumes in all the languages of the world."

Other resources e.g.
Internet Sacred Text Archive – sacred tradition, mythology

Chamber of Deputies archives

Alceste algorithm

Re-constructing worlds of meanings

Clusters of discourses

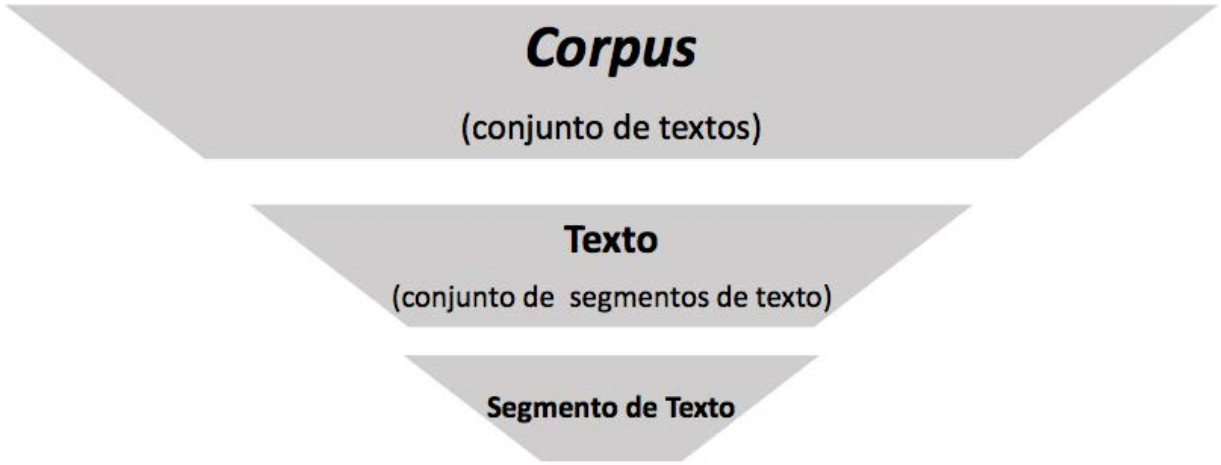


Figura 2: Noções de *corpus*, texto, segmento de texto

Figure from Tutorial by Brigido Viceu Camargo

Reinert, M. (1996). Un logiciel d'analyse lexicale: ALCESTE. Les Cahiers de L'analyse Des Données, XI, 471–484.

[Thematic content analysis] est, sur le plan théorique, généralement rejetée aujourd'hui, même si, dans de nombreuses études, elle est toujours utilisée, faute a'outils mieux appropriés. Elle ne pose d'ailleurs pas seulement des **problèmes théoriques, mais aussi des difficultés pra- tiques, les règles de décodage étant souvent difficiles à expliciter , ce qui ne facilite pas les prises de décision et occasionne une grande perte de temps tout en ne permettant pas une objectivation complète des procédures d'analyse.**

La méthodologie que nous proposons porte la marque de cette double expérience (**approche formelle, catégorisation conceptuelle**) et si les difficultés rencontrées nous ont éloignés de l'analyse de contenu, pour nous rapprocher d'un type d'analyse plus lexical, **nous en avons cependant conservé certains schèmes méthodologiques comme par exemple, la notion "d'unité de contexte".** Aussi nous l'avons dénommée "Analyse Lexicale par Contexte" (ou A.L.C.).

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. Les Cahiers de l'Analyse Des Données, 8, 187–198.

Par exemple, si **les indicateurs sont des mots, leur sens dépend, pour une part, du contexte**. Aussi, nous avons recherché une procédure qui, dans un premier temps, permette à l'utilisateur de **regrouper ces indicateurs polysémiques dans des classes caractéris- tiques de certains contextes**. Ces regroupements lui permettront a- lors de constituer ce que nous avons appelé des **"indicateurs de contexte"** dont le sens est plus précis et qui, par conséquent , tont susceptibles d'être associés à des hypothèses de contenu mieux définies.

Cette première condensation des données une fois effectuée, il est loisible à cet utilisateur **d'étudier les interrelations entre indicateurs de contexte, ou leurs relations avec d'autres caracté- ristiques de la population.**

**** *art_444 *00_05_cq *libération *quotidien *autres *2004 *moyen
il faudra un vrai courage politique pour que l'art retrouve la place que l'education nationale lui avait accordée. l'art à l'école, voie de démocratie djian jean_michel pour ceux qui sont traversés par le doute quant aux vertus de l'éducation artistique à l'école, le dernier film de gérard jugnot les choristes tombe à pic. jamais le cinéma ne rendra un tel hommage à cette pratique, d'autant que l'histoire est vraie, comme l'est, d'une autre manière, celle de ces jeunes de banlieues qui, dans l'esquive, le film d'abdelatif kechiche mettent en scène marivaux dans le jeu de l'amour et du hasard.
...
**** *art_445 *00_05_cq *libération *quotidien *arts_cul *2004 *moyen
annoncée moribonde, la scène française n'a pas dit son dernier mot. la preuve au printemps de bourges, qui s'ouvre aujourd'hui. le rap bouge encore binet stéphanie a la sortie de l'album revoir un printemps en septembre, les marseillais d'iam portaient sur leurs épaules tous les espoirs du rap français. après l'explosion des ventes en 1998, la médiatisation nationale via la radio skyrock, le rap français devient à l'entrée du millénaire médiocre, uniforme, enfermé dans ses clichés matérialistes machos racailleux.
...

**** *art_444 *00_05_cq *libération *quotidien *autres *2004 *moyen
il faudra un vrai courage politique pour que l'art retrouve la place que l'education nationale lui avait accordée. l'art à l'école, voie de démocratie djian jean_michel pour ceux qui sont traversés par le doute quant aux vertus de l'éducation artistique à l'école, le dernier film de gérard jugnot les choristes tombe à pic. jamais le cinéma ne rendra un tel hommage à cette pratique, d'autant que l'histoire est vraie, comme l'est, d'une autre manière, celle de ces jeunes de banlieues qui, dans l'esquive, le film d'abdelatif kechiche mettent en scène marivaux dans le jeu de l'amour et du hasard.
...

4 U.C.E

**** *art_445 *00_05_cq *libération *quotidien *arts_cul *2004 *moyen
annoncée moribonde, la scène française n'a pas dit son dernier mot. la preuve au printemps de bourges, qui s'ouvre aujourd'hui. le rap bouge encore binet stéphanie a la sortie de l'album revoir un printemps en septembre, les marseillais d'iam portaient sur leurs épaules tous les espoirs du rap français. après l'explosion des ventes en 1998, la médiatisation nationale via la radio skyrock, le rap français devient à l'entrée du millénaire médiocre, uniforme, enfermé dans ses clichés matérialistes machos racailleux.
...

- La méthode ALCESTE : particularité de la classification (Reinert, 1983, 1990)
  La classification est menée sur deux tableaux binaires (0 / 1) croisant Unités de Contexte (en ligne) et formes actives (en colonne).

  Unité de Contexte = ensemble d'U.C.E nécessaires pour atteindre x formes actives.

  - Par exemple, dans les paramètres par défaut d'ALCESTE, les deux tableaux sont contruits pour regrouper 10 formes actives pour le premier tableau et 12 pour le second.

	Forme 1	Forme 2	Forme 3	Forme i	
Uc1 (uce1+uce2)	0	1	1	...	
Uc2 (uce3+uce4)	1	0	1	...	
...	...	...	...	...	

	Forme 1	Forme 2	Forme 3	Forme i	
Uc1 (uce1+uce2+uce3)	1	1	1	...	
Uc2 (uce4+uce5)	0	0	1	...	
...	...	...	...	...	

Concluding

The raw results of TM usually fall into one of five categories: the trivial, the classic, the unexpected, the artefact, and the residue (Schonhardt-Bailey, Yager, & Lahlou, 2012).

***The trivial** are those which are so obvious as to be uninteresting (although in the case of SR, still, they are usually worth to mention since SR are precisely common sense).*

***The classic** are the ones that are consistent with previous research.*

***The unexpected** are new findings that the analyst can back up as "solid" with some other source of explanation or data.*

***The artefact** is what is due to technical issues with the data processing.*

*Finally **the residue** is what the analyst is unable to interpret.*

Therefore a good analysis requires understanding the underlying theoretical framework of the technique and the software, and awareness of the abductive processes at work in interpretation.

Links

Software & Tutorials

http://www.iramuteq.org/