Using the First Axis of a Correspondence Analysis as an Analytical Tool

Bénédicte Pincemin, Céline Guillot-Barbance, Alexei Lavrentiev

▶ **To cite this version:**

Bénédicte Pincemin, Céline Guillot-Barbance, Alexei Lavrentiev. Using the First Axis of a Correspondence Analysis as an Analytical Tool: Application to Establish and Define an Orality Gradient for Genres of Medieval French Texts. 14th International Conference on the Statistical Analysis of Textual Data / 14es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2018), DII– Department of Enterprise Engineering "Mario Lucertini" Tor Vergata University; DSS– Department of Statistical Sciences, Sapienza University, Rome, Jun 2018, Roma, Italy. pp.594-601. halshs-01759219

HAL Id: halshs-01759219
https://shs.hal.science/halshs-01759219

Submitted on 6 Apr 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Using the First Axis of a Correspondence Analysis as an Analytical Tool. Application to Establish and Define an Orality Gradient for Genres of Medieval French Texts

Bénédicte Pincemin[1], Céline Guillot-Barbance[2], Alexei Lavrentiev[3]

[1]Univ. Lyon, CNRS, IHRIM UMR5317 – benedicte *dot* pincemin *at* ens-lyon *dot* fr

[2]Univ. Lyon, ENS Lyon, IHRIM UMR5317 – celine *dot* guillot *at* ens-lyon *dot* fr

[3]Univ. Lyon, CNRS, IHRIM UMR5317 – alexei *dot* lavrentev *at* ens-lyon *dot* fr

Abstract

Our corpus of medieval French texts is divided into 59 discourse units (DUs) which cross text genres and spoken *vs* non spoken text chunks (as tagged with *q* and *sp* TEI tags). A correspondence analysis (CA) performed on selected POS tags indicates orality as the main dimension of variation across DUs. We then design several methodological paths to investigate this gradient as computed by the CA first axis. Bootstrap is used to check the stability of observations; gradient-ordered barplots provide both a synthetic and analytic view of the correlation of any variable with the gradient; a way is also found to characterize the gradient poles (here, more-oral or less-oral poles) not only with the POS used for the CA analysis, but also with words, in order to get a more precise and lexical description. This methodology could be transposed to other data with a potential gradient structure.

Keywords: textometry, Old French, represented speech, spoken genres, methodology, correspondence analysis, 1D model, data visualization, XML TEI, TXM software, DtmVic software.

1. Linguistic issue and preparation of textual data

We investigate spoken language features of Medieval French in a corpus composed of 137 texts (4 million tokens), taken from the Base de français médiéval[1]. The corpus is annotated with part-of-speech (POS) tags at the word level; speech quotation chunks and speech turns are marked up using TEI XML tags at an intermediate level between sentences and paragraphs; and every text can be situated in a 32-genre typology (Guillot et al., 2017). Our hypothesis is that the features of orality may be related to text chunks representing speech, and also to text genres, as for instance some text genres are intended for oral performance. In order to perform a textometric analysis (Lebart et al. 1998) on our XML-TEI annotated data, we use the TXM open-source corpus analysis platform (Heiden, 2010; Heiden et al., 2010)[2].

We divide our corpus into 59 discourse units (DUs) obtained by splitting every genre into parts which represent speech on the one hand, and the remaining parts on the other hand (some text genres have no spoken passages). Discourse unit labels, like *q_rbrefLn* for instance, combine four pieces of information: (i) the first letter is either *q* for quoted speech

[1] Base de français médieval: http://bfm.ens-lyon.fr

[2] Textometry Project and TXM software: http://textometrie.org

chunks, *sp* for speech turns, or *z* for remaining (non oral) chunks; (ii) then we have the short name of the text genre (here, *rbref* means "récit bref", i. e. short narrative); (iii) the uppercase letter stands for the domain[3]; (iv) the last character indicates whether this DU is represented in our corpus by one (1), two (2) or more (*n*) texts.

We linguistically represent our texts with the POS tags[4] they use[5]. The reliability of POS tags was measured in a previous study (Guillot et al., 2015) for a subset of 7 texts in which tags had been manually checked. For the present analysis, we eliminate low-frequency POS tags (freq. < 1 500), which include many high error rate tags and do not carry much weight into the quantitative analysis. For the remaining high error rate tags (with more than 25% wrong assignments), we measure their influence on the correspondence analysis (CA) by checking their contribution to the first axis. Then we remove the proper nouns category (NOMpro) which shows both high error rate and high contribution to the first axis (14.66 %).

A new correspondence analysis enables two additional improvements from a linguistic perspective. We remove compound determiners (DETcom, PRE.DETcom, like *ledit*) as they emerged at the end of the 13th century, so that they introduce a singular and substantial diachronic effect (high contributions on the first axis). Moreover, the second axis describes mainly the association between psalms (z_psautierRn) and possessive adjectives (ADJpos): this corresponds to very specific phrases with some distinctive nouns (*la meie aneme*, *li miens Deus*, *la tue misericorde*), and the adjective is equivalent to a possessive determiner in other contexts, so we merge the two categories (DETADJpos). We finally get a contingency table crossing 59 DUs with 33 POS tags to explore with a CA.

2. Linguistic and methodological results from correspondence analysis

Our study reveals that the first axis can in fact be interpreted as an orality gradient. The factorial map (Fig. 1) shows z_ DUs on the left hand side of the first axis, opposed to q_ and sp_ DUs on the right hand side. Some genres intended for oral performance go to the right with speech chunks (especially plays –dramatiqueL, dramatiqueR), whereas genres related to written processing (especially practical acts (P): charters, etc.) go to the left with out-of-speech chunks. As this opposition matches the first axis, orality appears as the first contrastive dimension for Old French (as regards POS frequencies), as it is in Biber's experiences with English (Biber, 1988), with the same kind of linguistic features (Table 1). Then, as a second result, DUs can be sorted according to their degree of orality, from "less oral" to "more oral" (see Appendix[6]). Peculiar positions (for didactic dialogs or psalms for instance) can be explained by a formal use of language given by the rules of the genre. The linguistic analysis of the DU gradient is detailed in (Guillot-Barbance et al., 2017)[7].

[3] There are 6 domains: literature (L), education (D for "didactique"), religion (R), history (H), law (J for "juridique"), practical acts (P).

[4] We use the Cattex2009 tagset, designed for Old French: http://bfm.ens-lyon.fr/spip.php?article176.

[5] We exclude punctuations, editorial markup and foreign words. CQL query: [fropos!="PON.*|ETR|OUT|RED"]

[6] Appendix is available online as a related file of this paper in HAL archive: https://halshs.archives-ouvertes.fr/halshs-01759219

[7] Improvements made to the statistical processing in 2018 (management of the second axis with ADJpos and DETpos merging, confidence ellipses) strengthen the linguistic interpretation published in 2017, no significant

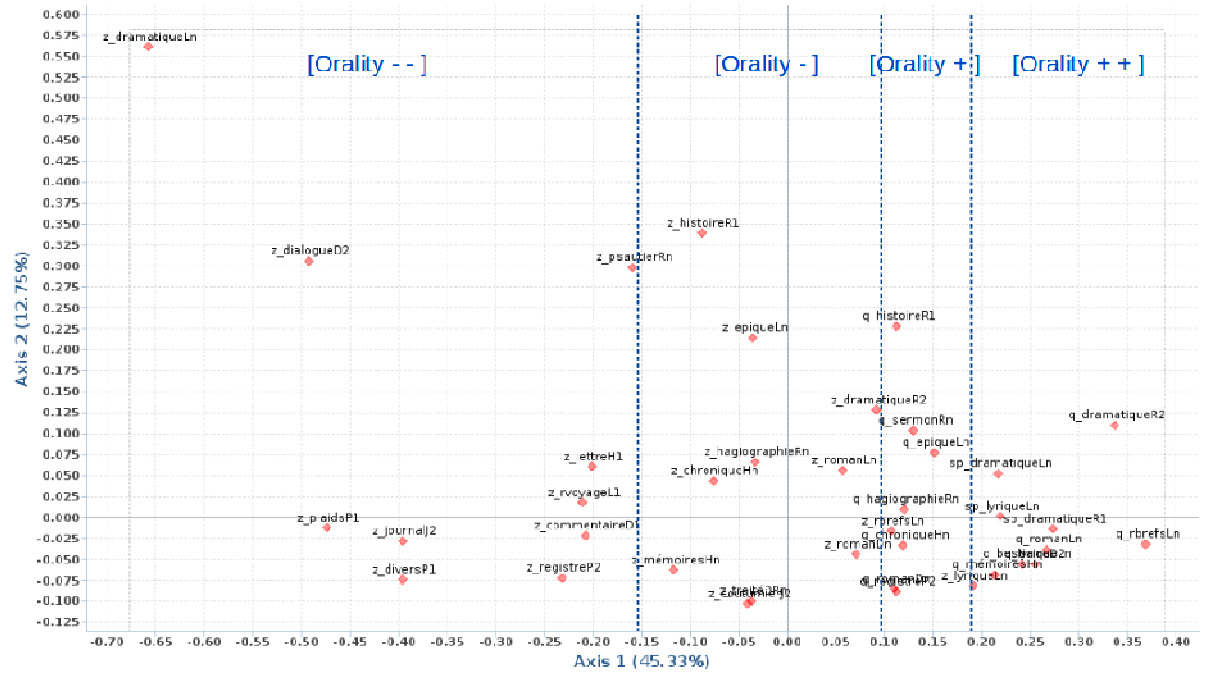

***Figure 1**. CA map of the 59 DUs (TXM). 21 DUs with low representation quality (cosine squared to 1 × 2 plane < 0.3) and no significant contribution to this plane (ctrb1 < 2 % & ctrb2 < 2 %) have been filtered out (macro CAfilter.groovy), so that the figure is clearer.*

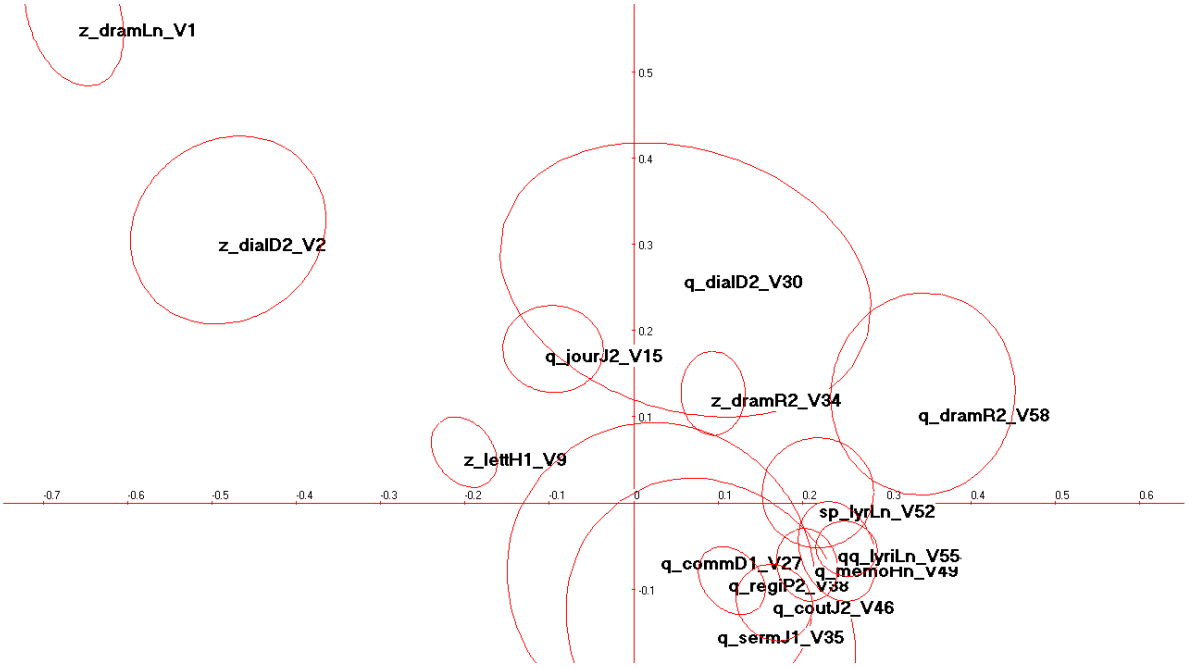

***Figure 2**. CA map of the 17 DUs with the largest confidence ellipses (DtmVic). The two largest ones (q_proverbesD2, q_lapidaireD2) couldn't be drawn; the following three largest ones (q_commentaireD1, q_dialogueD2, q_sermentJ1) show that these DU positions cannot be interpreted; then other smaller ellipses indicate that the 54 remaining DU positions on axes #1 and 2 are stable.*

change is observed on gradient given by the first axis, according to the four zones defined by the analysis, except for a few points which are not related to this axis (low cosine squared).

Table 1. *The eight POS with the highest contributions on the first axis, for both sides.*

"Less oral" pole		*"More oral" pole*	
PRE	preposition	PROper	personal pronoun
NOMcom	common noun	ADVgen	general adverb
PRE.DETdef	preposition + definite determiner	ADVneg	negative adverb
VERppe	past participle	VERcjg	finite verb
DETdef	definite determiner	PROadv	adverbial pronoun (*en*, *y*)
DETcar	cardinal determiner	DETADJpos	possessive determiner or adjective
VERppa	present participle	CONsub	subordinating conjunction
CONcoo	coordinating conjunction	VERinf	infinitive verb

A bootstrap validation (Dupuis & Lebart, 2008, Lebart & Piron, 2016) is applied to evaluate the stability of DU positions on the first axis (Figure 2). Sizes of ellipses in the 1×2 map are correlated to sizes of DUs: the fewer the words there are in the DU, the less data the statistics process, and the greater is the confidence ellipse (Table 1). Only five DUs are ascribed a big ellipse which shows their uncertain position (Figure 2): all of them are DUs from about ten words to about a hundred words, which are DUs for very singular linguistic usages, and are neither representative nor relevant for this overall linguistic analysis. The orality gradient is then confirmed throughout a statistic validation on our data.

The 2D factorial map provides a synthetic and efficient visualization. The second axis display reveals that the "more oral" pole is more compact, more consistent, than the "less oral" pole, which is more heterogeneous (the cosine squared values corroborate this). But what we want to stress in this methodological paper, is that the main linguistic result is uniquely provided by the interpretation of the first axis. Benzécri has illustrated the same kind of approach by using a 1D CA to reveal the hierarchy of characters in Racine's *Phèdre* (1981 : 68). This method emphasizes the analytic power of CA, which separates the data (by the mathematical means of Singular Value Decomposition) into "deep" components (factors), just as a prism breaks light up into its constituent spectral colors. Despite its main use as a 2D illustration of a corpus structure in the textual data analysis field, CA is much more than a suggestive visualization or a quick sketch.

3. Complementary tools to analyse 1D gradient in textual data

We now test new means to gain insight into the causation of this gradient in our data.

3.1. Gradient-ordered barplot

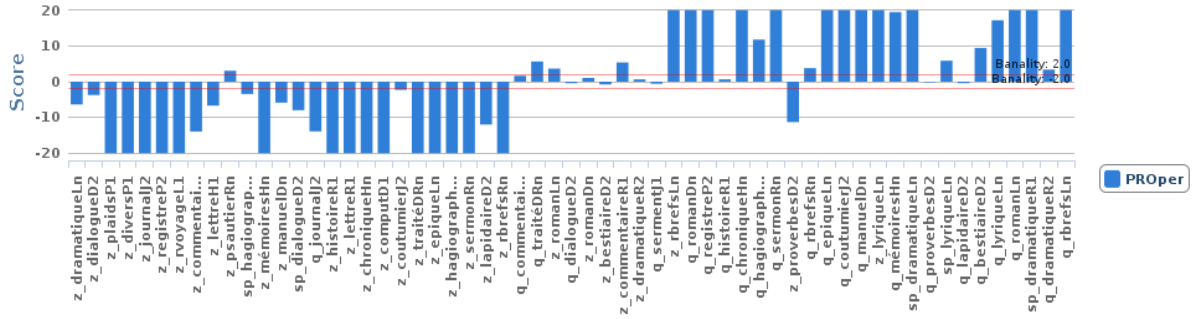

Figure 3*. Gradient-ordered specificity barplot for Personal Pronoun, as example of a POS which is correlated to the first axis. For readability reasons, the height of specificity bars is limited to 20.*

The first method we propose is to visualize the evolution of POS frequencies according to the orality gradient using a specificity bar-plot chart where the DU order on the x-axis is given by

the DU order on the first CA axis: this display visually reveals how much a POS is correlated with speech or non speech features, and details its affinity with each DU. For instance, personal pronouns are typical for the more-oral pole: this is displayed as a rising profile (Figure 3), and one can easily find out which DU have an outlying use of this POS. Whereas a POS like adjectives (Figure 4), which is not correlated to the orality gradient, gets a chart with no overall pattern.

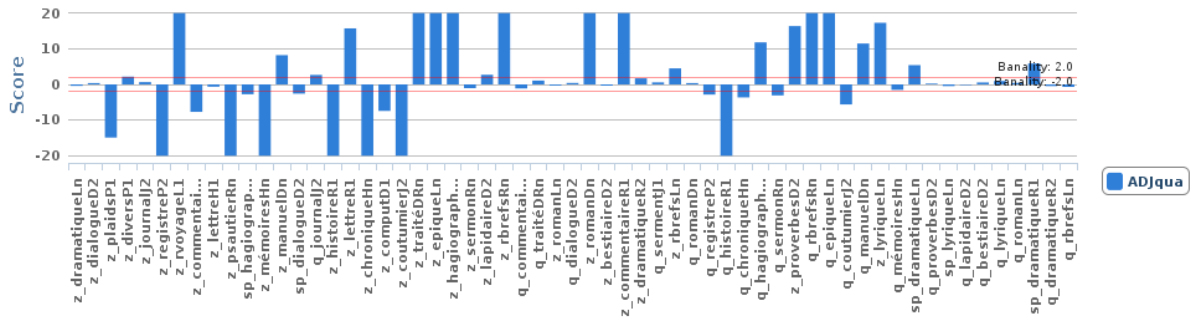

Figure 4*. Gradient-ordered specificity barplot for adjectives, as example of a POS which is not correlated to the first axis. For readability reasons, the height of specificity bars is limited to 20.*

3.2. Back-to-text close reading by getting representative words for each side of the first axis

The second methodological innovation concerns obtaining lexical information about orality characteristics in our texts. We select two sets of DUs based on their cosine squared scores for the first CA axis in order to represent the more-oral ($cos^2 1 > 0.4$) and less-oral ($cos^2 1 > 0.35$) poles (Table 2). The cos^2 thresholds are adjusted to get two balanced sets with enough different DUs to get an adequate representativeness. Then, a specificity computation, which statistically characterizes the distribution of words into these two sets, reveals lexical features for more oral and less oral poles, showing typical words as they can be read in texts. Light is thus shed on the quantitative result throughqualitative observations.

Table 2.
Representative DUs

Less-oral pole	***More-oral pole***
z_journalJ2	q_romanLn
z_plaidsP1	sp_dramatiqueR1
z_commentaireD1	q_rbrefLn
z_diversP1	q_bestiaireD2
z_registreP2	sp_dramatiqueLn
z_lettreH1	q_lyriqueLn
z_dialogueD2	z_lyriqueLn
z_rvoyageL1	q_chroniqueHn
	sp_lyriqueLn
	q_hagiographieRn
	q_romanDn
	q_mémoiresHn

Table 3a*. Adjectives typical for the less-oral subcorpus*

fropos	***word***	***F***	***f_z***	***S+_z***
ADJqua	ladite	709	709	275
ADJqua	presens	432	430	162
ADJqua	Cedit	331	331	128
ADJqua	maistre	344	329	105
ADJqua	Saint	530	442	89
ADJqua	oudit	207	207	80
ADJqua	present	195	191	67
ADJqua	certeinnes	122	122	47
ADJqua	frans	239	205	46
ADJqua	VIIJ	105	105	41
ADJqua	feu	132	126	40
ADJqua	Petit	144	131	36
ADJqua	parisis	92	92	36
ADJqua	sains	242	192	33
ADJqua	Porel	68	68	26
ADJqua	GRACE	67	67	26
ADJqua	Saincte	90	83	24
ADJqua	royaulx	73	70	23
ADJqua	Perrin	64	63	23
ADJqua	yceulx	68	66	23

Table 3b*. Adjectives typical for the more-oral subcorpus*

fropos	***word***	***F***	***f_q&sp***	***S+_q&sp***
ADJqua	grant	3288	2594	129
ADJqua	bele	344	344	79
ADJqua	granz	288	281	54
ADJqua	biax	197	197	45
ADJqua	Biax	196	196	45
ADJqua	bone	198	195	40
ADJqua	voir	224	214	36
ADJqua	douce	128	128	29
ADJqua	Biaus	127	127	29
ADJqua	biaus	125	125	29
ADJqua	biau	122	122	28
ADJqua	sage	204	189	27
ADJqua	Bele	111	111	25
ADJqua	mal	166	156	24
ADJqua	Biau	98	98	22
ADJqua	mortel	97	97	22
ADJqua	bel	250	217	21
ADJqua	boen	88	88	20
ADJqua	meillor	87	87	20
ADJqua	las	100	98	20

Our example sheds light on the uses of adjective: whereas adjectives are not related to the orality gradient as a category (Figure 4), they have strong associations at a lexical level (Table 3). Represented speech makes much use of terms of address introducing speech turns (*bel*, *douz* – and their formal variants: *biaus*, *biax*, etc.), and evaluative adjectives (*grant*, *mal*, *boen*). For the less-oral pole, there are more POS tagging errors; adjectives are more diverse

and often associated with a subset of DUs, for instance *present*, *saint*, *maistre* are typical of two texts.

4. Conclusion

In this contribution, we have shown several ways to take into account the limits of real data, especially textual data: managing the POS tags reliability (§1), validation process to identify where data is lacking (§2), refining morphosyntatic based analysis with lexical information (§3). But our main objective is to establish a methodology in order to reveal and study any gradient-like deep structuration of data. A simple seriation (as illustrated in Dupuis & Lebart, 2008) could provide the same results for the first step, as it generates the same ordered view of the data. But CA gives much more information, qualifying the relation of each variable to the gradient with indicators like contributions and cosines squared. Interpretation can go further: CA coordinates are controlled with bootstrap and confidence ellipses, gradient-ordered barplot visualizations are efficient to analyse in detail the relationship of any individual variable to the overall gradient, and the gradient poles can be illustrated by words, which add a concrete and textual account for the deep structure. Thus, on our corpus of French medieval texts, we discover that orality is the main contrastive dimension and that it characterizes represented speech as well as text genres. The methodology could be applied to other data, and is already entirely implemented using tools freely available to the scientific community.

This research has benefited from the PaLaFra ANR-DFG project (ANR-14-FRAL-0006), for corpus extension and POS evaluation. We are also very grateful to Ludovic Lebart, for his inspiring comments on a preliminary presentation of this research, and for DtmVic software, which has evolved in order to take into account the quantitative particularities of our data.

References

Benzécri J.-P. et al. (1981). *Pratique de l'Analyse des données, tome 3. Linguistique & lexicologie*. Dunod, Bordas, Paris.

Biber D. (1988). *Variation across speech and writing*. Cambridge University Press.

Dupuis F., Lebart L. (2008). Visualisation, validation et sériation. Application à un corpus de textes médiévaux. In Heiden S. and Pincemin B., eds, *Actes JADT 2008*, Presses univ. de Lyon: 433-444.

Guillot C., Heiden S., Lavrentiev A., Pincemin B. (2015). L'oral représenté dans un corpus de français médiéval (9e-15e) : approche contrastive et outillée de la variation diasystémique. In Kragh K. J. and Lindschouw J., eds, *Les variations diasystémiques et leurs interdépendances dans les langues romanes -Actes du Colloque DIA II*, Éd. de linguistique et de philologie, Strasbourg : 15-28.

Guillot-Barbance C., Pincemin B., Lavrentiev A. (2017). Représentation de l'oral en français médiéval et genres textuels, *Langages*, 208: 53-68.

Heiden S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Otoguro R. et al., eds, *PACLIC24*, Waseda Univ., Sendai : 389-398.

Heiden S., Magué J.-Ph., Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In Bolasco S. et al., eds, *Statistical Analysis of Textual Data -Proceedings of JADT 2010*, Edizioni Univ. di Lettere Economia Diritto, Rome : 1021-1031.

Lebart L., Piron M. (2016). *Pratique de l'Analyse de Données Numériques et Textuelles avec Dtm-Vic*. L2C, http://www.dtmvic.com.

Lebart L., Salem A., Berry L. (1998). *Exploring Textual Data*. Kluwer academic pub., Boston.