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

Bénédicte Pincemin, Alexei Lavrentiev, Céline Guillot-Barbance

This is a pre-copyedited version of a contribution published in:
Domenica Fioredistella IEZZI, Damon MAYAFFRE, Michelangelo MISURACA (eds)
Text Analytics. Advances and Challenges
Studies in Classification, Data Analysis, and Knowledge Organization
Heidelberg : Springer, 2020, p. 127-143.
The definitive authenticated version is available online via
https://doi.org/10.1007/978-3-030-52680-1_11

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

Application to Establish and Define an Orality Gradient for Genres of Medieval French Texts

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

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

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

[3]Univ. Lyon, ENS Lyon, IHRIM UMR5317 – celine *dot* guillot *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. Orality prevails over textual features which could fit in a one-dimensional model as well, such as text form (verse vs prose) or time (composition century). 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 word forms, in order to get a more accurate and lexical description. This methodology could be transposed to other data with a potential gradient structure.

Keywords Textometry, Old French, Quoted speech, Spoken genres, Methodology, Correspondence analysis, 1D model, Data visualization, XML TEI, TXM software, DtmVic software.

1 Research Question and Scientific Background

The way one can describe written texts with respect to oral linguistic uses is of primary importance for diachronic linguistics and for studying ancient stages of languages. In order to investigate this issue on Medieval French (ninth to fifteenth centuries), we propose a new approach, combining corpus linguistics (McEnery and Hardie 2012), digital philology, textometric analysis (Lebart et al. 1998), and open-source tools and data. Our research is based on the assumption that a continuum between orality and scripturality must substitute for the old dichotomy between oral and written uses (Koch and Oesterreicher 2001). From that perspective, some written discourse may give us an indirect access to orality and to a hidden side of dead languages (Marchello-Nizia 2012).

Most of the previous scholarship dealing with orality in Medieval French and using corpora have focussed on quoted speech. Since quoted speech is explicitly given as spoken and is linguistically or graphically separated from the rest of the text, it can be analysed in a contrastive way with non spoken discourse. This is also the starting point of our study, following previous research (Guillot et al. 2013, 2015, Glikman and Mazziotta 2013). Thanks to updated and increased data, and thanks to new analytic tools, we would like to check if the opposition between quoted speech and plain text still constitutes the first dimension of variation in medieval written texts. In 2015, we ascertained this within a 2.5-million-token corpus : orality was the main variation dimension over time variation (Old French vs Middle French) and over five discourse domains. Since 2013, we have also observed that orality was associated with verbs (and related parts of speech: personal pronouns, adverbs…), as opposed to nouns (and related parts of speech: prepositions, determiners…). In this research, we investigate if this is confirmed with improved and augmented data.

We relate this orality dimension to generic features. Koch and Oesterreicher's study (2001) enables us to predict some correlations between text genres having affinities with orality and other genres, more written oriented. Here, we can distinguish 32 text genres into our 4-million-token corpus (69 % of the genres are represented by at least 2 texts, 44 % by 3 texts or more). We can also add another criterion, based on Koch (1993), distinguishing text genres that were destined to oral performance in the early stages of French (for instance, the *chanson de geste* or hagiographic texts). We thus make the hypothesis that features of orality may be related to quoted speech, and also to text genres, such as those which are intended for oral performance.

We detail in (Guillot-Barbance et al. 2017) the linguistic results of this new experiment, concerning orality features and text genres in Medieval French. The present paper puts the emphasis on methodological outputs: how correspondence analysis may be used in a more analytic way rather than synthetically, and how results can be better controlled with new and complementary tools.

Resorting to a statistical multidimensional approach in order to investigate orality and text genres draws our research closer to Biber's work. Indeed, Biber initially developed his method to study speech and writing variation among registers in a corpus of modern English language (Biber 1988). Texts are characterized by linguistic features (such as word length, first person pronoun, or *that* deletion). 67 features were used in the seminal study, and this set was slightly augmented later (Biber 2009). Those features are selected because they are pointed out as relevant in linguistic studies about text genres. Moreover, they have to be automatically identifiable in digital corpora with a tagging program. Then, a principal factor analysis (PFA) is computed on the normalized relative frequency table crossing texts and tagged features. The first dimensions revealed by the factorial analysis are then interpreted, thanks to both the identification of the most contributive features for the dimension as well as the scores of genres on the dimension (defined as the mean scores of the texts), especially genres that the dimension contrasts. This method was subsequently applied to various corpora in order to determine if any dimensions could be stable across languages (1995) and discourse domains (2009), and then be considered as linguistic universals. It was also applied to identify patterns of diachronic register variation (Biber 1995, Biber and Finegan 2001, Biber 2001). Thus, Biber's results (2014) concord with our first ones, as he points out the prominence of an oral versus literate dimension (which emerges as the first dimension in almost every study). Our study offers new insight not only because it deals with another language (Medieval French), but also because several technical choices are different, especially the three following ones: firstly, isolating quoted speech and processing it separately from the rest of the text systematically (as Biber and Finegan (2001) did for one genre, dealing with fiction dialog and fiction prose as two separate genres); secondly, describing texts through their part-of-speech frequencies instead of elaborated linguistic features which thirdly implies a correspondence analysis processing rather than a principal factor analysis, since our data compose a two-way table instead of a table of measurements (Lebart et al. 1998). We have chosen to base the analysis on part-of-speech tags for several reasons: the overall quality of this tagging is satisfying in our corpus, and we manage its strengths and weaknesses; it is more stable and appropriate to identify linguistic units than word forms, due to the spelling variants and evolution in the medieval period; parts of speech account for some morphosyntactic information; this representation is very close to text expression and does not introduce much subjectivity; it is less partial and less sparse than bigrams (Crossley and Louwerse 2007).

Correspondence analysis (Benzécri 1973, Husson et al. 2017) is a key tool to corpus characterization in the textual data analysis community (Lebart et al. 1998, Née et al. 2017, Poudat and Landragin 2017, Lebart et al. 2019) . It is also commonly applied to part-of-speech text characterizations since morphosyntactic taggers availability (see Brunet 2016 for instance). Nevertheless, in this field, correspondence analysis is almost entirely devoted to a "synthetic" use, through 2D visualizations (and experimentation of 3D ones, cf. Viprey 2006, Leblanc and Pé-

rès 2014). The synthetic use is also based on information compression (selection of the first dimensions) and noise reduction (elimination of the last dimensions). This synthetic approach is even required in case of Guttman effect on serially-structured corpora (Salem 1991). However we would like to adopt here a methodological point of view and to insist on how an analytic use of factorial tools can contribute. Biber's approach (1988, 2009) already illustrates such an analytic use, but this paper aims to demonstrate this kind of use in the textual data analysis context (with a correspondence analysis computed on a part-of-speech text characterization, in Section 3), we evaluate and discuss the limits of this use (Sect. 4), and we implement new tools to guide interpretation (Sect. 5).

2 Methodology and Preparation of Textual Data

Our corpus is composed of 137 texts (4 million tokens)[1], taken from the *Base de français médiéval* (http://txm.bfm-corpus.org). This 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. Every text is characterized with 28 metadata fields, including a 32-genre typology (Guillot-Barbance et al. 2017). In order to perform a textometric analysis (Lebart et al. 1998) on our XML-TEI annotated data, we used the TXM open-source corpus analysis platform (Heiden 2010) (http://textometrie.org) and DtmVic software (Lebart and Piron 2016) (http://www.dtmvic.com).

We have divided our corpus into 59 discourse units (DUs) obtained by splitting every genre into parts which represent spoken words on the one hand and the remaining parts on the other hand (some text genres have no spoken passages). Discourse unit labels, such as *q_rbrefLn*, combine four pieces of information: (i) the first letter is either *q* for quoted speech chunks, *sp* for speech turns, or *z* for the remaining (non-spoken) chunks; (ii) then there is the short name of the text genre (here, *rbref* means "récit bref", i. e. short narrative); (iii) the uppercase letter stands for the domain: literary (L), educational (D for "didactique"), religious (R), historical (H), legal (J for "juridique"), practical documents (P); and (iv) the last character indicates whether this DU is represented in our corpus by one (1), two (2) or more (*n*) texts.

We linguistically represented our texts with the Cattex POS tags[2] assigned by the TreeTagger tool (Schmid 1994), excluding punctuation, editorial markup, foreign words, and abbreviations.[3] The reliability of POS tags had been measured in a

[1] Since the first presentation of this work at the JADT 2018 conference in Roma, the corpus has been updated (better morphosyntactic tagging, new version for some texts). All experiments and results have been revised consequently.

[2] http://bfm.ens-lyon.fr/spip.php?article176

[3] CQL query: [fropos!="PON.*|ETR|OUT|RED|ABR"]

previous study (Guillot et al. 2015) for a subset of 7 texts in which the tags had been manually checked. For the present analysis, we have eliminated low-frequency POS tags (freq. < 1,700) that include many high error rate tags and do not carry much weight in the quantitative analysis. For the remaining high error rate tags (with more than 25% wrong assignments), we measured their influence on the correspondence analysis (CA) by checking their contribution to the first axis. Then, we removed the proper noun category (NOMpro) which showed both high error rate and high contribution to the first axis (14.67 %).

A new correspondence analysis enabled two additional improvements from a linguistic perspective. We decided to remove compound determiners (DETcom, PRE.DETcom, like *ledit*). Indeed, as they emerged late thirteenth century, they introduced a singular and substantial diachronic effect (high contributions to the first axis). Moreover, the second axis described 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 where the adjective is equivalent to a possessive determiner in other contexts. Therefore, we have merged the two categories (DETADJpos) and we finally obtained a contingency table crossing 59 DUs with 33 POS tags to explore with a CA.

3 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 or archive documents (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 experiments with English (Biber 1988), then numerous other languages (1995, 2014), with the same kind of linguistic features opposing verbs and clausal style with nouns and phrasal style (Table 1).

Then, as a second result, DUs can be sorted according to their degree of orality, from "less-oral" to "more-oral" (Table 2). POS tags prove to be as efficient as more specialized linguistic features to arrange DUs (and genres) in a sharp order. 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). Improvements made to the statistical processing in 2018 and to the data in 2019 strengthen the linguistic interpretation published in 2017. No significant change is observed on the gra-

dient given by the first axis, according to the four zones defined by the analysis, except for the spoken chunks of *Quatre livre des Rois* (q_histoireR1), which moves to the "less-oral" zone, because many unknown word forms (due to the anglo-normand dialect and to the peculiar use of diacritics in the edition) were erroneously tagged as NOMcom by the TreeTagger.

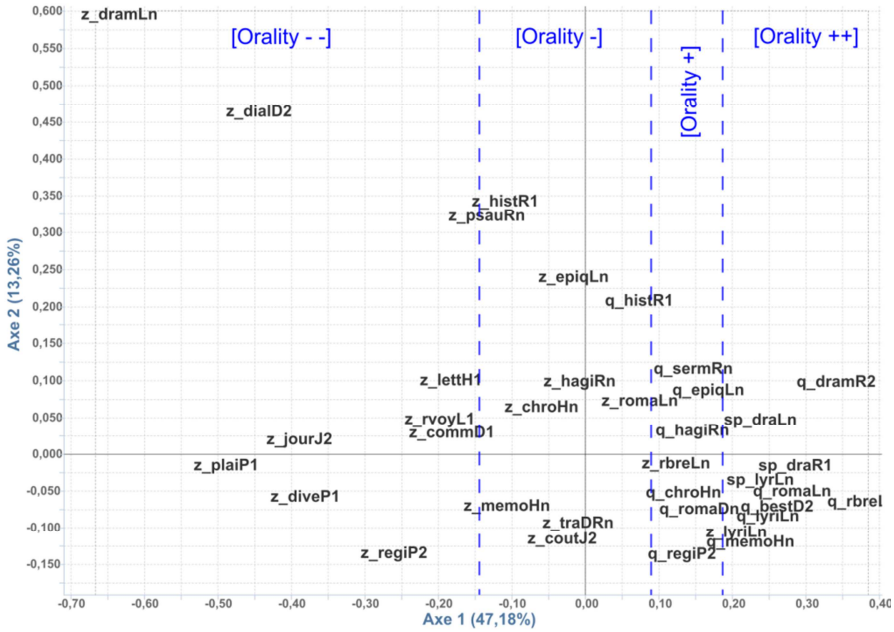

Fig. 1 CA map of the 59 DUs (TXM). 21 DUs with low representation quality (cosine squared to 1 × 2 plane < 0.3) and no significant contribution to this plane (ctrb1 < 2 % & ctrb2 < 2 %) have been filtered out (AFCWithStylesMacro.groovy), so that the figure is clearer. A few overlapping labels have been slightly shifted upwards or downwards, with no impact on axis 1 coordinate.

Table 1 The six POS with the highest contributions on the first axis (ctrb1 > 2 %), for both sides.

	Less-oral pole			More-oral pole	
Label	Part of Speech	Ctrb1	Label	Part of Speech	Ctrb1
PRE	preposition	15.22	PROper	personal pronoun	18.08
NOMcom	common noun	9.31	ADVgen	general adverb	7.91
VERppe	past participle	5.59	ADVneg	negative adverb	7.17
PRE.DETdef	preposition + definite determiner	4.91	VERcjg	finite verb	6.63
DETdef	definite determiner	4.28	PROadv	adverbial pronoun (*en*, *y*)	2.89
DETcar	cardinal determiner	2.84	DETADJpos	possessive determiner or adjective	2.26

Ctrb1 contribution to the first axis

Table 2 DU gradient on CA first axis

c1 rank	*DU*	*c1*	*Ctrb1*	*$Cos^2 1$*	*Q12*	*Mass*	*Word Nb*	*Ellipse size*
1	z_dramatiqueLn	-0.65	0.14	**0.30**	0.54	0.01	331	medium+
2	z_plaidP1	-0.50	**2.48**	**0.72**	0.72	0.28	10,030	small
3	z_dialogueD2	-0.45	0.05	0.20	0.40	0.01	266	medium+
4	z_journalJ2	-0.40	**19.81**	**0.86**	0.86	**3.48**	124,042	small
5	z_diversP1	-0.39	**1.38**	**0.70**	0.71	0.25	9002	small
6	z_registreP2	-0.27	**22.38**	**0.69**	0.83	**8.60**	306,586	small
7	z_rvoyageL1	-0.21	**1.03**	**0.34**	0.34	0.66	23,385	small
8	z_commentaireD1	-0.20	0.28	**0.54**	0.54	0.21	7511	small
9	z_lettreH1	-0.19	0.08	**0.35**	0.43	0.06	2245	medium
10	z_psautierRn	-0.15	0.99	0.09	0.48	**1.29**	46,062	small
11	sp_hagiographieR1	-0.14	0.30	0.14	0.15	0.44	15,760	small
12	z_manuelDn	-0.12	0.10	0.07	0.18	0.19	6887	small
13	z_memoiresHn	-0.12	**2.77**	**0.43**	0.52	**5.40**	192,524	small
14	z_histoireR1	-0.12	0.77	0.08	0.61	**1.49**	53,154	small
15	sp_dialogueD2	-0.10	0.25	0.17	0.17	0.66	23,424	small
16	z_chroniqueHn	-0.07	**1.68**	0.29	0.43	**9.23**	329,059	small
17	z_computD1	-0.07	0.06	0.04	0.04	0.39	14,082	small
18	z_lettreR1	-0.06	0.12	0.10	0.11	0.84	29,992	small
19	q_journalJ2	-0.06	0.00	0.01	0.06	0.03	1120	medium
20	z_coutumierJ2	-0.04	0.31	0.04	0.24	**4.38**	156,306	small
21	z_epiqueLn	-0.03	0.03	0.01	0.61	**1.23**	43,785	small
22	z_traiteDRn	-0.02	0.08	0.01	0.16	**4.94**	175,972	small
23	z_hagiographieRn	-0.02	0.05	0.03	0.53	**3.45**	122,923	small
24	z_sermonRn	-0.01	0.00	0.00	0.00	**3.28**	116,895	small
25	z_lapidaireD2	0.01	0.00	0.00	0.00	0.39	13,785	small
26	z_romanDn	0.01	0.02	0.02	0.03	**2.66**	94,934	small
27	z_rbrefsRn	0.03	0.06	0.03	0.16	**2.30**	81,894	small
28	q_commentaireD1	0.04	0.00	0.00	0.03	0.00	123	**large**
29	q_histoireR1	0.06	0.11	0.03	0.28	0.87	30,917	small
30	z_romanLn	0.06	**2.03**	0.23	0.45	**15.27**	544,536	small
31	q_traiteDRn	0.06	0.07	0.08	0.14	0.51	18,253	small
32	z_bestiaireD2	0.08	0.21	0.29	0.29	0.96	34,164	small
33	z_commentaireR1	0.09	0.33	0.13	0.14	**1.25**	44,671	small
34	q_dialogueD2	0.09	0.00	0.02	0.15	0.00	92	**large**
35	z_dramatiqueR2	0.09	0.02	0.10	0.26	0.07	2318	medium
36	q_sermentJ1	0.10	0.00	0.05	0.20	0.00	98	**large**
37	z_rbrefsLn	0.11	0.37	**0.38**	0.38	0.83	29,718	small

38	q_registreP2	0.12	0.05	0.22	0.38	0.09	3308	medium
39	q_chroniqueHn	0.12	0.89	**0.52**	0.58	**1.73**	61,524	small
40	q_sermonRn	0.13	0.22	0.21	0.34	0.35	12,495	small
41	q_hagiographieRn	0.13	0.60	**0.53**	0.54	0.96	34,152	small
42	q_romanDn	0.15	**3.18**	**0.65**	0.84	**4.11**	146,631	small
43	q_epiqueLn	0.15	0.85	**0.38**	0.46	**1.00**	35,791	small
44	q_coutumierJ2	0.16	0.06	0.14	0.25	0.07	2368	medium
45	q_rbrefsRn	0.16	0.50	0.27	0.28	0.57	20,257	small
46	z_preceptesD2	0.17	0.21	0.20	0.22	0.20	7188	small
47	q_manuelDn	0.18	0.51	0.29	0.29	0.44	15,786	small
48	z_lyriqueLn	0.19	**3.09**	**0.56**	0.65	**2.31**	82,224	small
49	q_preceptesD2	0.20	0.00	0.06	0.29	0.00	25	**large+**
50	q_memoiresHn	0.21	0.09	**0.44**	0.50	0.06	2151	medium
51	sp_dramatiqueLn	0.22	**1.05**	**0.65**	0.66	0.59	20,970	small
52	sp_lyriqueLn	0.22	0.04	**0.48**	0.49	0.02	885	medium
53	q_lyriqueLn	0.24	0.55	**0.67**	0.74	0.28	9945	small
54	q_bestiaireD2	0.25	0.10	**0.65**	0.69	0.05	1667	medium
55	q_romanLn	0.27	**26.58**	**0.89**	0.92	**10.47**	373,214	small
56	sp_dramatiqueR1	0.27	0.84	**0.78**	0.79	0.32	11,411	small
57	q_dramatiqueR2	0.33	0.04	**0.32**	0.34	0.01	405	medium+
58	q_lapidaireD2	0.33	0.00	0.05	0.05	0.00	11	**large+**
59	q_rbrefsLn	0.36	**2.13**	**0.76**	0.77	0.45	16,121	small

c1 coordinate on axis #1, *Ctrb1* contribution to axis #1, $Cos^2 1$ cosine squared to the axis #1, *Q12* representation quality in 1 × 2 plane (defined by cosine squared to 1 × 2 plane).
Values that are of most relevance for current interpretation are bolded: *Ctrb1* ≥ 1%, $Cos^2 1 \geq 0.3$, *Mass* ≥ 1%.

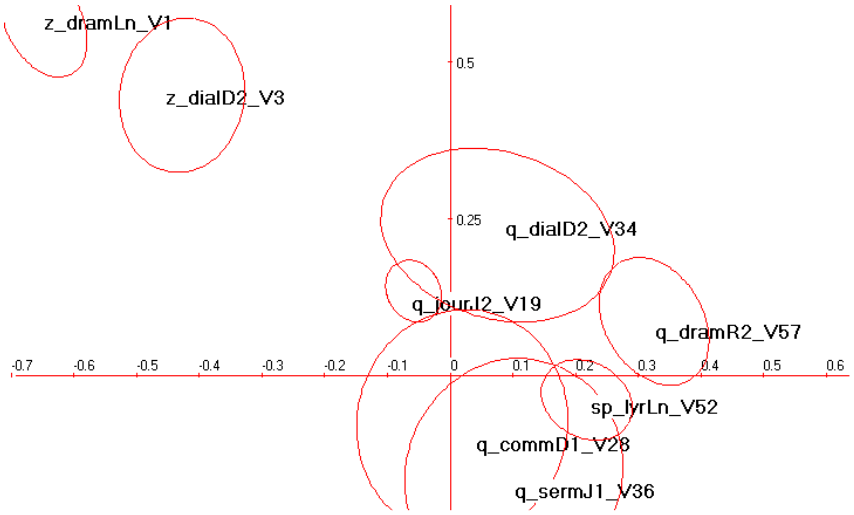

Fig. 2 CA map of the 10 DUs with the largest confidence ellipses (DtmVic). The two largest ones (q_proverbesD2, q_lapidaireD2) couldn't be drawn; the following three largest ones (q_commentaireD1, q_dialogueD2, q_sermentJ1) show that these DU positions cannot be interpreted; then other smaller ellipses indicate that the 54 remaining DU positions on axes 1 and 2 are stable.

A bootstrap validation (Lebart 2004, Dupuis and Lebart 2008) 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 (Fig. 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 one, which is more heterogeneous (the cosine squared values corroborate this). Yet, 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 illustrated the same kind of approach by using a 1D CA to reveal the hierarchy of characters in Racine's *Phèdre* (1981, p. 68). Biber's multidimensional analysis (1988, 2009) also adopted this kind of analytic use of factorial analysis, interpreting one dimension after the other. Our experiment emphasizes the analytical 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 illustra-

tion of a corpus structure in the textual data analysis field, CA is much more than a suggestive visualization or a quick sketch.

With that in mind, could any kind of feature be captured by this decomposition? If we look into our data, which cross genres and speech, one could question if orality could be the only available feature fitting a one-dimensional contrast: this would contest the prevalence of orality. Our next section is dedicated to this issue and tests if more fine-grained data, with more "degrees of freedom" and internal variations along bipolar or linear variables, could overtake orality as the main contrast.

4 Discussion: Can the First Axis Match any Data Feature, Is Orality Feature really the Main One?

4.1 Challenging with another bipolar textual feature: the verse/prose opposition

Every DU was divided according to the text form of the chunks: either prose (*p*), verse (*v*) or mixed form (*m*). The 4 smallest DUs, ranging from 18 to 27 tokens, were excluded. The CA was performed on these 76 new DUs characterized by the same 33 POS, and it showed the same overall opposition pattern between q_ and z_ DUs. To a lesser extent, the verse displays some affinity with orality. New results were obtained from the nine genres that were divided into different forms (chroniqueH, hagiographieR, lapidaireD, preceptesD, rbrefsL, romanD, romanL, sermonR, traiteDR). The typical configuration is illustrated by historical chronicle (Fig. 3a). The only significant z/q inversion occurs in didactic novel (Fig. 3b) but it can be accounted for by the heterogeneity of this genre: the didactic novel in verse, which is clearly marked with orality, is represented by different parts of the *Roman de la rose*. This text is quite different from the other didactic novels and can be considered as a genre on its own.

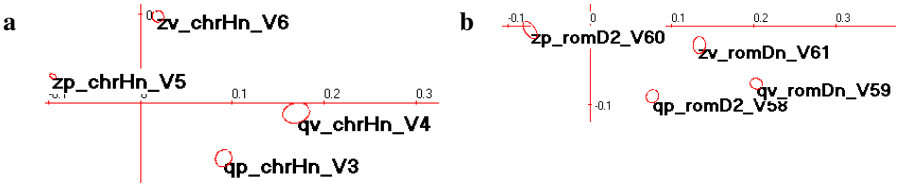

Fig. 3 Configuration of DUs for two genres: **a** Historical Chronicle, **b** Didactic Novel
(the CA is computed on the full 76 × 33 table but only the selected DUs are displayed) (DtmVic)

4.2 Challenging with the most fine-grained units: individual texts

The 247 DUs represent up to 3 parts for a text: quoted speech chunks (q), speech turns (sp), and remaining (non oral) chunks (z). The 9 smallest DUs, ranging from 10 to 27 tokens, were excluded. Once again, orality emerged from the first axis: we calculated that for the DUs of at least 1,000 words (which ensures that their position is stable), there were no more than 4 z/q inversions in 75 texts. These 4 negative differences between q_ and z_ first coordinates for a text are: -0.12 (rosem1), -0.11 (monstre), -0.03 (rosem3), -0.01 (tdechamp). So only two texts behave substantially in a peculiar way: the first volume of *Roman de la rose* by Jean de Meun (rosem1), strongly marked with orality, including non-spoken parts, and Monstrelet's *Chronicle* (monstre), a historical text where spoken parts display very clearly the features of the less-oral pole.

4.3 Challenging with a linear feature: time

We still considered the 247 text-level DUs, but for balance and homogeneity considerations, we focussed on the period from the twelfth to the fifteenth centuries and took out the ninth, tenth and eleventh centuries which were just represented by 9 DUs (0.3 % of words). In order to visualize the distribution of centuries on the factorial map, we used a grayscale coloring of the DU positions according to their century. It appears that orality prevails over diachrony (Fig. 4). Indeed, diachrony is essentially represented by the second dimension, whereas orality still remains the first dimension, as observed in § 4.2 on quite similar data. It should be noted that the diachrony dimension here does not account for the emergence of the compound determiners in the thirteenth century (*ledit*, etc.), since this part of speech was left out from the analysis (§ 2).

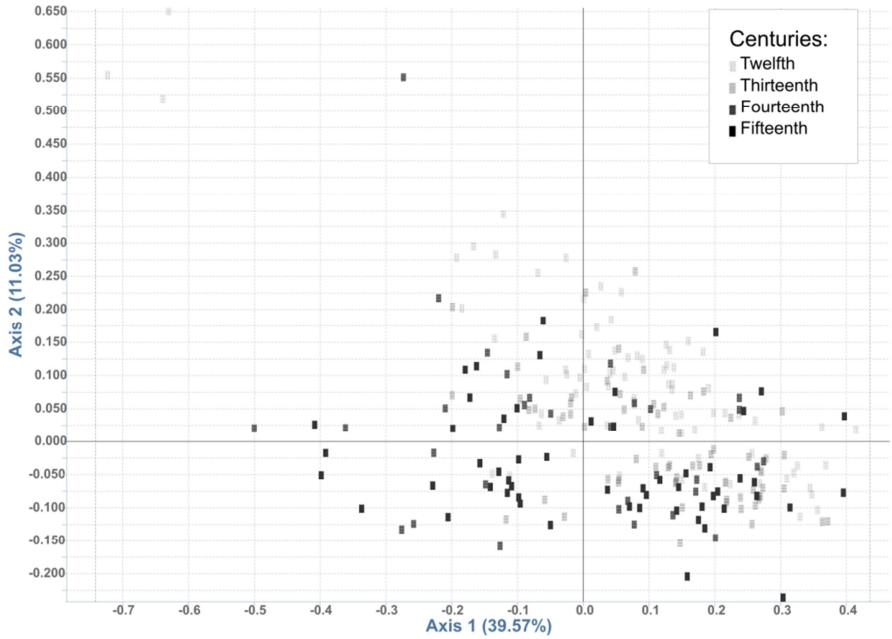

Fig. 4 Correspondence Analysis on text-level DUs (238 DU × 33 POS) colorized according to centuries (twelfth to fifteenth) (TXM with AFCWithStyles macro).

In their diachronic study on English from seventeeth to twentieth century, Biber and Finegan (2001) noted that contrast between written and spoken registers increases over time. Such pattern does not appear clearly on our data. However, one could wonder if, in Biber and Finegan's analysis, orality contrast could be partially represented for past periods, given the fact that the orality dimension they used was computed to maximize the contrast in the modern language, it may be less efficient to represent all of the orality variation in a different time. In our study, all centuries contribute to the axis composition.

4.4 Corpus scale effect

If the corpus is reduced to a few texts, the first axis may no longer represent orality. Our experiment is based on a previous research conducted by Mazziotta and Glickman (2019) on three texts with mixed genres.

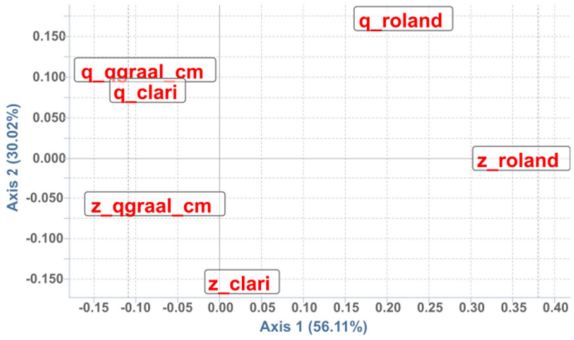

Fig. 5 CA on the three texts *Chanson de Roland* (roland), *Conquête de Constantinople* (clari) and *Queste del Saint Graal* (qgraal) (6 DUs × 33 POS) (TXM)

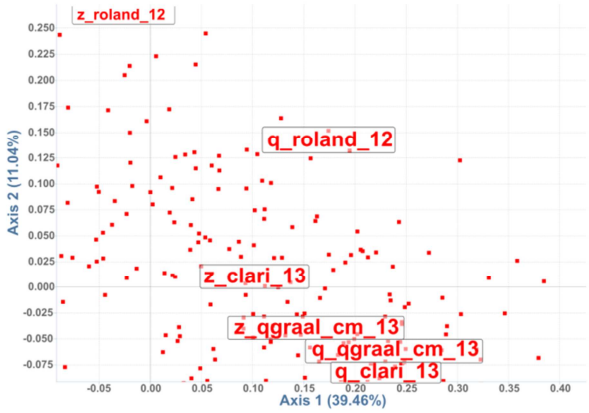

Fig. 6 Position of the DUs for the three texts *Chanson de Roland* (roland), *Conquête de Constantinople* (clari) and *Queste del Saint Graal* (qgraal) in the CA computed on the full corpus at the text level (247 DUs × 33 POS) (see § 4.2) (TXM)

The main contrast we notice within the six DUs is the opposition between texts (*Chanson de Roland* vs the two other texts) (Fig. 5), which concurs with the conclusions reached by Mazziotta and Glickman (2019) from an analysis of these texts based on their syntactic properties. Whereas in the entire corpus, the opposition highlighted between spoken DUs and other DUs remains predominant (Fig. 6). Nevertheless, from the POS perspective, in both cases, the main contrast still opposed common nouns (and determiners) to personal pronouns (and adverbs; finite verbs standing at an intermediate position in the 6-DU analysis).

In brief, orality stands as the first structural dimension when the corpus is large and diversified enough. In a small corpus composed of a few texts, the peculiarities of individual texts may overtake other general linguistic properties and achieve an overall structuring effect.

Our experiments lead us to conclude that the degree of orality proves to be a fundamental linguistic dimension which can summarize the main contrast arising from a large variety of texts. It should be related to a deep grammatical competition between common nouns, on the one hand, and verbs and personal pronouns, on the other hand. Biber (2014) also insisted on this grammatical expression of the first dimension through the opposition of a clausal style and a phrasal style. Brunet

(2016, p. 147-148) shared the same observations recurrently, pointing out the stricking polarization of discourse between the “verb clan” and the “noun clan”.

5 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.

5.1 Gradient-ordered Barplot

The first method we propose is to visualize the evolution of POS frequencies according to the orality gradient using a specificity barplot 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 more-oral or less-oral 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 (Fig. 7a), and one can easily find out which DUs have an outlying use of this POS. By contrast, a POS that is not correlated to the orality gradient, such as a demonstrative pronoun (Fig. 7b), presents a chart with no overall pattern.

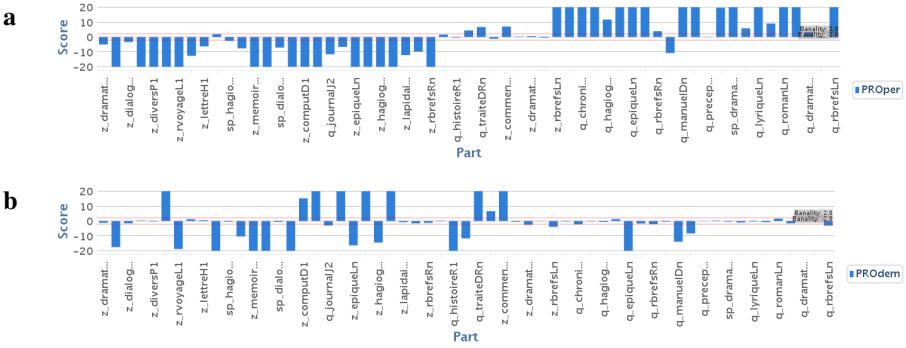

Fig. 7 Gradient-ordered specificity barplot (TXM) for: **a** Personal Pronoun, as example of a POS which is correlated to the first axis, and **b** Demonstrative Pronoun, as example of a POS which is not correlated to the first axis. For readability reasons, the height of specificity bars is capped at 20.

This kind of visualization enriches qualitatively the factorial analysis output. Whereas, the coordinate on the dimension is a unique quantitative score that tells us *whether* a POS (or a feature) is more-oral or less-oral oriented, and with which intensity, the gradient-oriented barplot accounts for *how* this polarity is implemented throughout the corpus.

5.2 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 selected 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 ($cos^2 1 > 0.29$) poles. The cos^2 thresholds were adjusted to get two balanced sets with enough different DUs to get an adequate representativeness. Less-oral pole comprises 10 DUs: z_journalJ2, z_plaidP1, z_diversP1, z_registreP2, z_commentaireD1, z_memoiresHn, z_lettreH1, z_rvoyageL1, z_dramatiqueLn, and z_chroniqueHn. More-oral pole comprises 12 DUs: q_romanLn, sp_dramatiqueR1, q_rbrefsLn, q_lyriqueLn, q_romanDn, q_bestiaireD2, sp_dramatiqueLn, z_lyriqueLn, q_hagiographieRn, q_chroniqueHn, sp_lyriqueLn, and q_memoiresHn. Then, a specificity computation (Lebart et al. 1998) revealed lexical features for each pole, showing typical words as they can be read in texts.

Our example sheds light on the uses of demonstrative pronouns, which are not related to the orality gradient as a category (Fig. 7b) but have strong associations with it at a lexical level (Table 3): speech chunks make much use of the demonstrative *cist*, which is more related to the speaker and more used as a deictic pronoun; the demonstrative *cil* is more frequent in less-oral pole in anaphoric and recognitional uses (Guillot-Barbance 2017, p. 335-338).

Viewing in context words which correspond to features associated with one pole or the other, as in Biber's methodology, is enlightening, even crucial. Indeed, it helps to understand in any text passage which concrete phenomena and expressions coincide with the more-orality or less-orality trend of this passage. But our method introduces an intermediary step, based on representative subcorpora, which makes it more powerful for two reasons. Firstly, thanks to the statistical computation, linguistic realizations are ordered: one immediately gets a synthetic view of the most characteristic word forms. And secondly, characterizations found at the lexical level are complementary to characterizations at the descriptive level (POS, or other features): characteristic words are not all implementations of characteristic features (see the example of demonstrative pronouns above), and conversely every implementation of a characteristic feature may not necessarily correspond to a polarized word form. For instance, *Sire* is a common noun, which is a POS identified as typically less-oral, that mainly occurs in DUs identified as more-oral and as a term of address introducing speech turns: so it would be somehow confusing to find *Sire* underlined in less-oral DUs (as it is a noun) and not underlined in more-oral DUs (despite being one of the most typical words in these DUs).

Table 3 Typical demonstrative pronouns for each pole (selected by highest specificity score)

Less-oral	**icellui** (141.4), ce (115.3), **icelle** (104.8), **iceulx** (84.8), **ceulx** (79.9), **icelles** (66.4), cen (28.3), **ycelle** (26.2), **ycellui** (23.6), CE (22.7), **chil** (22.5), **ceulz** (18.8), **yceulx** (16.3), chou (15.3), **ycelles** (15.0), **ciaus** (14.0), cecy (9.2), **ceaux** (8.5), Cen (6.8), **celuy** (6.5), **celles** (6.4), cela (5.9), **celx** (5.7), **chiax** (5.1)…
More-oral	Ce (84.8), C' (52.1), çou (36.0), c' (23.2), che (20.0), Che (18.5), Ch' (12.8), ice (7.7), **cist** (6.7), Ice (6.7), chou (6.1), **ceste** (5.9), **cestui** (5.8), **cesti** (5.4), ç' (5.3), cela (5.2)…

Each word is followed by its specificity score in brackets. Bolded forms are *cil* or *cist* variants.

6 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 (Sect. 2), validation process to identify where data is lacking (3), cross-checking genre-level results with finer text-level experiments (4), and refining the analysis based on morphosyntactic tags with lexical information (5). However, our main objective has been to establish a methodology in order to reveal and study any gradient-like deep structuration of data. A simple seriation (as illustrated in Dupuis and Lebart 2008) could provide the same results for the first step, as it generates the same ordered view of the data. Yet, CA gives much more information, qualifying the relation of each variable to the gradient with indicators such as contributions and cosines squared. Interpretation can go further: CA coordinates are controlled with bootstrap and confidence ellipses, gradient-ordered barplot visualizations are efficient to analyze the relationship of any individual variable to the overall gradient, and the gradient poles can be illustrated by words, which add a concrete textual account for the deep structure. Thus, in our corpus of French medieval texts, we discover that orality is the main contrastive dimension and it characterizes quoted speech as well as text genres. We discuss what contributes to the prevalence of such feature: its affordance to a one-dimensional representation, and its relevance at the corpus scale. The methodology could be applied to other data, and is entirely implemented using tools freely available to the scientific community.

We believe that the methodology we elaborated provides an inspiring framework for the analysis of textual data. It allowed us to account for fine-grain and gradual relations between linguistic units while revealing the fundamental morphosyntactic opposition between the noun on the one hand, and the verb and personal pronoun on the other hand. Brunet worded this ubiquitous internal contrast in corpora with a witty metaphor (2016, end of Chap. 11): "Put together peaceful and indistinct players on a plain and neutral ground and give them a round and smooth ball, then come back a few moments later. You will see teams fighting against each other."

This research has benefited from the PaLaFra ANR-DFG project (ANR-14-FRAL-0006). We would like to thank L. Lebart and C. Poudat for their inspiring comments on previous versions of this paper.

References

Benzécri, J.-P., et al. (1973). *L'Analyse des Données, tome 2. L'Analyse des Correspondances*. Paris: Dunod.

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

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

Biber, D. (1995). *Dimensions of register variation: a cross-linguistic comparison*. Cambridge University Press.

Biber, D. (2001). Dimensions of variation among eighteenth-century speech-based and written registers. In S. Conrad & D. Biber (Eds), *Variation in English: Multi-Dimensional studies*. (pp. 200-214). London: Longman.

Biber, D. (2009). Multidimensional approaches. In A. Lüdeling and M. Kytö (Eds), *Corpus linguistics: An international handbook* (pp. 822-855). Berlin: Walter de Gruyter.

Biber, D. (2014). The ubiquitous oral versus literate dimension: A survey of multidimensional studies. In J. Connor-Linton & L.W. Amoroso (Eds), *Measured language: Quantitative studies of acquisition, assessment, and variation* (pp. 1-20). Washington DC: Georgetown University Press.

Biber, D., & Finegan, E. (2001). Diachronic relations among speech-based and written registers in English. In S. Conrad & D. Biber (Eds), *Variation in English: Multi-Dimensional studies*. (pp. 66-83). London: Longman.

Brunet, É. (2016). *Tous comptes faits, Écrits choisis tome III, Questions linguistiques*. B. Pincemin (Ed.), Paris: Champion.

Crossley, S.A., Louwerse, M.M. (2007). Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics*, *12*, 453-478.

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

Glikman, J., Mazziotta, N. (2013). Représentation de l'oral et structures syntaxiques dans la prose de la Queste del saint Graal (1225-1230). In D. Lagorgette & P. Larrivée (Eds), *Représentations du sens linguistique 5* (pp. 43-64). Chambéry: Université de Savoie.

Guillot, C., Lavrentiev, A., Pincemin, B., Heiden, S. (2013). Le discours direct au Moyen Âge : vers une définition et une méthodologie d'analyse. In D. Lagorgette & P. Larrivée (Eds), *Représentations du sens linguistique 5* (pp. 17-41). Chambéry: Université de Savoie.

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 K. J. Kragh & J. Lindschouw (Eds), *Les variations diasystémiques et leurs interdépendances dans les langues romanes -Actes du Colloque DIA II* (pp. 15-28). Strasbourg: Éditions de linguistique et de philologie.

Guillot-Barbance, C. (2017). *Le démonstratif en français : étude de sémantique grammaticale diachronique (9ème-15ème siècles)*. Louvain: Peeters.

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 R. Otoguro R. et al. (Eds), *24th Pacific Asia Con-*

ference on Language, Information and Computation - PACLIC24 (pp. 389-398). Sendai: Institute for Digital Enhancement of Cognitive Development, Waseda University.

Husson, F., Lê, S., Pagès, J. (2017). *Exploratory Multivariate Analysis by Example Using R*, 2nd Edition. Boca Raton: Chapman and Hall/CRC.

Koch, P. (1993). Pour une typologie conceptionnelle et médiale des plus anciens documents/monuments des langues romanes. In M. Selig, B. Frank & J. Hartmann (Eds), *Le passage à l'écrit des langues romanes* (pp; 39-81). Tübingen: Narr,

Koch, P., & Österreicher, W. (2001). Gesprochene Sprache und geschriebene Sprache. Langage parlé et langage écrit. In G. Holtus, M. Metzeltin & C. Schmitt (Eds), *Lexikon der romanistischen Linguistik 1-2* (pp. 584-627). Tübingen: Niemeyer.

Lebart, L. (2004). Validation Techniques in Text Mining (with Application to the Processing of Open-ended Questions). In S. Sirmakessis S. (Ed.) *Text Mining and its Applications* (pp. 169-178). Berlin, Heidelberg: Springer Verlag.

Lebart, L., & Piron, M. (2016). *Exploring Numerical and Textual Data in Practice with Dtm-Vic*. L2C. http://www.dtmvic.com/06_ManualE.html. Accessed 22 October 2019.

Lebart, L., Salem, A., Berry, L. (1998). *Exploring Textual Data.* Dordrecht: Kluwer Academic.

Lebart, L., Pincemin, B., Poudat, C. (2019). *Analyse des données textuelles*. Québec: Presses de l'université du Québec.

Leblanc, J.-M., & Pérès, M. (2014). Modèles tridimensionnels pour la représentation de l'état des connaissances et propositions de visualisation pour l'analyse des corpus textuels. In E. Née, J.-M. Daube, M. Valette, S. Fleury (Eds), *Proceedings of the 12th International Conference on Textual Data statistical Analysis JADT 2014*,.(pp. 373-384). Paris: Université Paris 3 & INaLCO.

Marchello-Nizia, C. (2012). L'oral représenté : un accès construit à une face cachée des langues 'mortes'. In C. Guillot et al. (Eds), Le changement en français. Études de linguistique diachronique. Bern/Berlin/Bruxelles : Peter Lang, 247-264.

Mazziotta, N., & Glickman, J. (2019). Oral représenté et narration en ancien français. Spécificités syntaxiques dans trois textes de genres distincts. *Linx*, *78*.https://doi.org/10.4000/linx.3151.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice.* Cambridge University Press.

Née, É. (dir.) (2017). *Méthodes et outils informatiques pour l'analyse des discours*. Presses Universitaires de Rennes.

Poudat, C., & Landragin, F. (2017). *Explorer un corpus textuel*. Louvain-la-Neuve: De Boeck.

Salem, A. (1991). Les séries textuelles chronologiques. *Histoire et Mesure*, *6 (1)*, 149–175.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing* (pp. 44–49).

Viprey, J.-M. (2006). Ergonomiser la visualisation AFC dans un environnement d'exploration textuelle : une projection « géodésique ». In J.-M. Viprey et al. (Eds), *Proceedings of 8th international Conference on Textual Data statistical Analysis JADT'06*, vol. II (pp. 989-1000). Besançon: Presses universitaires de Franche-Comté.