Preprint. Final version (book chapter): Bavaud, F., C.Cocco and A.Xanthos. "Textual navigation and autocorrelation", pp. 35-56. In "Sequences in Language and Text", Mikros, G. (Ed.) & Macutek, J. (Ed.). Berlin, Boston: De Gruyter Mouton, 2015.

Textual navigation and autocorrelation

François Bavaud, Christelle Cocco, Aris Xanthos

1. Introduction

Much work in the field of quantitative text processing and analysis has adopted one of two distinct symbolic representations: in the so-called *bag-of-words* model, text is conceived as an urn from which units (typically words) are independently drawn according to a specified probability distribution; alternatively, the *sequential* model describes text as a categorical time series, where each unit type occurs with a probability that depends, to some extent, on the context of occurrence.

The present contribution pursues two related goals. First, it aims to generalise the standard sequential model of text by decoupling the order in which units *occur* from the order in which they are *read*. The latter can be represented by a Markov transition matrix between positions in the text, which makes it possible to account for a variety of ways of navigating the text, including in particular non-linear and non-deterministic ones. Markov transitions thus define *textual neighbourhoods* as well as *positional weights* – the stationary distribution or *prominence* of textual positions.

Building on the notion of textual neighbourhood, the second goal of this contribution is to introduce a unified framework for *textual autocorrelation*, namely the tendency for neighbouring positions to be more (or less) similar than randomly chosen positions with respect to some observable property – for instance whether the same unit types tend to occur, or units of similar length, or consisting of similar sub-units, and so on. Inspired from spatial analysis (see e.g. Cressie 1991; Anselin 1995; Bavaud 2013), this approach relates the above mentioned transition matrix (specifying *neighbourhoodness*) with a second matrix specifying the *dissimilarity* between textual positions.

The remainder of this contribution is organised as follows. Section 2 introduces the foundations of the proposed formalism and illustrates them with toy examples. Section 3 presents several case studies intended to show how the formalism and some of its extensions apply to more realistic research problems involving, among others, morphosyntactic and semantic dissimilarities computed in literary or hypertextual documents. Conclusion briefly summarises the key ideas introduced in this contribution.

2. Formalism

2.1. Textual navigation: positions, transitions and exchange matrix

A text consists of n *positions* $i = 1, \ldots, n$. Each position contains an occurrence of a *type* (or *term*) $a = 1, \ldots, v$. Types may refer to characters, words (possibly lemmatised), sentences, or units from any other level of analysis. The occurrence of an instance of type a at position i is denoted $o(i) = a$.

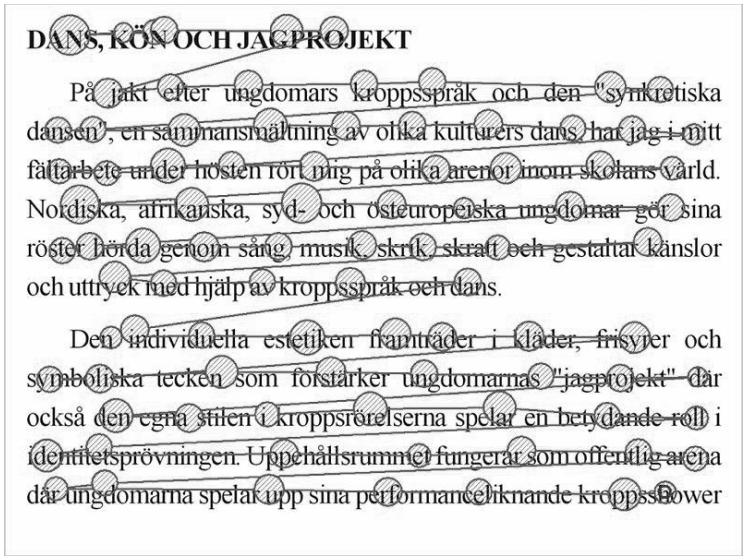

Fig. 1: Fixations and ocular saccades during reading, illustrating the skimming navigation model. Source: Wikimedia Commons, from a study of speed reading made by Humanistlaboratoriet, Lund University (2005).

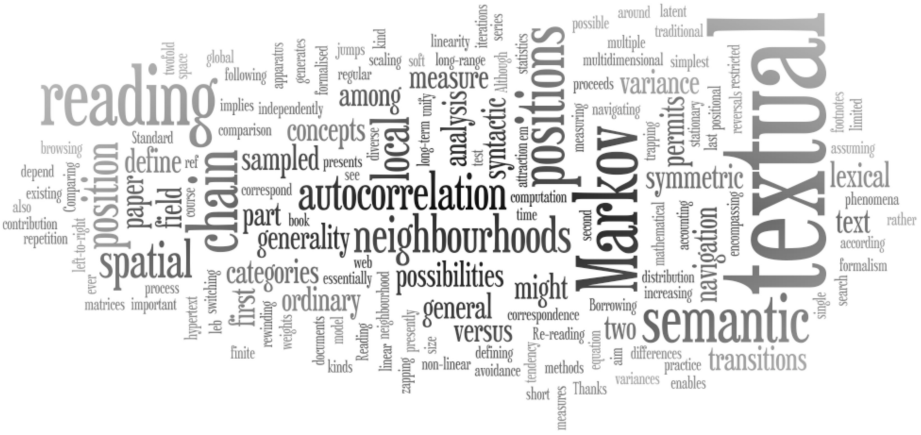

Fig. 2: Word cloud (generated by http://www.wordle.net), aimed at illustrating the free or bag-of-words navigation model with differing positional weights f.

A simple yet non-trivial model of textual navigation can be specified by means of a regular $n \times n$ Markov transition matrix $T = (t_{ij})$, obeying $t_{ij} \geq 0$ and $\sum_j t_{ij} = 1$. Its stationary distribution $f_i > 0$, obeying $\sum_i f_i t_{ij} = f_j$ and $\sum_i f_i = 1$, characterises the *prominence* of position i. For instance, neglecting boundary effects, [1]

- $t_{ij} = 1(j = i + 1)$ corresponds to *standard linear reading*[2]
- $t_{ij} = 1(i < j < i + r + 1)/r$ describes *(directed) skimming* with jumps of maximum length $r \geq 1$, as suggested by Figure 1
- $t_{ij} = f_j$ defines the *free* or *bag-of-words navigation*, consisting of independent random jumps towards weighted positions (Markov chain of order 0), as suggested by Figure 2.

Define the $n \times n$ exchange matrix $E = (e_{ij})$ as

[1] Boundary effects are, if necessary, dealt with the usual techniques (reflecting boundary, periodic conditions, addition of a "rest state", etc.), supporting the formal fiction of an ever-reading agent embodied in stationary navigation.

[2] Here and in the sequel, $1(A)$ denotes the indicator function of event A, with value $1(A) = 1$ if A is true, and $1(A) = 0$ otherwise.

$$e_{ij} := \frac{1}{2}(f_i t_{ij} + f_j t_{ji}) \tag{1}$$

By construction, E is symmetric, non-negative, with margins $e_{i\bullet} = e_{\bullet i} = f_i$, and total $e_{\bullet\bullet} = 1$. [3] The exchange matrix constitutes a symmetrical measure of positional interaction, reading flow, or neighbourhoodness between textual positions i and j.

In particular, the following three exchange matrices correspond to the three transition matrices defined above (still neglecting boundary effects):

- $e_{ij} = [1(j = i - 1) + 1(j = i + 1)]/2n$ for *standard linear reading*
- $e_{ij} = \frac{1}{2rn} 1(|i - j| \leq r)\ 1(i \neq j)$ for *skimming* with (undirected) jumps of maximum length r
- $e_{ij} = f_i f_j$ for the *free* or *bag-of-words* exchange matrix.

Toy example 1: consider a text consisting of four positions, either *(periodically) linearly read* as ...123412341... (A) or *freely read* (B). The previously introduced quantities are then

$$T^A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad f^A = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix} \quad E^A = \frac{1}{8}\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \tag{2}$$

$$T^B = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \quad f^B = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix} \quad E^B = \frac{1}{16}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \tag{3}$$

Note that (2) can also be conceived as an example of (periodic) skimming with jumps of maximum length $r = 1$, which is indeed equivalent to (periodic) linear reading. Similarly, free navigation is equivalent to skimming with jumps of maximum size n, with the single difference that the latter forbids jumping towards the currently occupied position.

2.2. Autocorrelation index

Let us now consider a matrix of dissimilarities D_{ij} between pairs of positions (i, j). Here and in the sequel, we restrict ourselves to *squared Euclidean dissimilarities*, i.e. of the form $D_{ij} = \| x_i - x_j \|^2$, where $\|.\|$ denotes the Euclidean norm and x_i, x_j are p-dimensional vectors, for some $p \geq 1$.

The average dissimilarity between a pair of randomly chosen positions defines the *(global) inertia* Δ, while the average dissimilarity between a pair of neighbouring positions defines the *local inertia* Δ_{loc} :

$$\Delta := \frac{1}{2}\sum_{ij} f_i\, f_j D_{ij} \qquad\qquad \Delta_{\text{loc}} := \frac{1}{2}\sum_{ij} e_{ij}\, D_{ij} \tag{4}$$

A local inertia much smaller than the global inertia reflects the presence of *textual autocorrelation*: closer average similarity between neighbouring positions than between randomly chosen positions. Conversely, *negative* autocorrelation characterizes a situation

[3] Here and in the sequel, the notation "•" denotes summation over the replaced index, as in $a_{\bullet j} := \sum_i a_{ij}$, $a_{i\bullet} := \sum_j a_{ij}$ and $a_{\bullet\bullet} := \sum_{ij} a_{ij}$.

where neighbouring positions are more dissimilar than randomly chosen ones[4]. Textual autocorrelation can be measured by the *autocorrelation index*

$$\delta := \frac{\Delta - \Delta_{\mathrm{loc}}}{\Delta}$$

generalizing *Moran's I index* of spatial statistics (Moran 1950). The latter holds in the one-dimensional case $D_{ij} = (x_i - x_j)^2$, where x_i, x_j are scalars, $\Delta = \mathrm{var}(x)$, and $\Delta_{\mathrm{loc}} = \mathrm{var}_{\mathrm{loc}}(x)$ (e.g. Lebart 1969).

Under the null hypothesis H_0 of absence of textual autocorrelation, the expected value of the autocorrelation index is not zero in general, but instead

$$E_0(\delta) = \frac{\mathrm{trace}(W) - 1}{n - 1}\ , \tag{5}$$

where $W = (w_{ij})$ is the transition matrix of a reversible Markov chain defined as $w_{ij} := e_{ij}/f_i$ (so that E_0 reduces to $-1/(n-1)$ for off-diagonal exchange matrices). Similarly, under normal approximation, the variance reads

$$\mathrm{Var}_0(\delta) = \frac{2}{n^2 - 1}\left[\mathrm{trace}(W^2) - 1 - \frac{(\mathrm{trace}(W) - 1)^2}{n - 1}\right],$$

(e.g. Cliff and Ord 1981; Bavaud 2013), thus making the autocorrelation index significant at level α if

$$\left|\frac{\delta - E_0(\delta)}{\sqrt{\mathrm{Var}_0(\delta)}}\right| \geq u_{1-\frac{\alpha}{2}}\ , \tag{6}$$

where u_α denotes the α-th quantile of the standard normal distribution.

Toy example 1, continued: suppose that the types occurring at the four positions of example 1 are the following trigrams: $o(1) = \alpha\beta\gamma$, $o(2) = \alpha\delta\epsilon$, $o(3) = \epsilon\zeta\eta$, and $o(4) = \alpha\beta\gamma$. Define the (squared Euclidean) dissimilarity D_{ij} as the number of characters by which trigrams $o(i)$ and $o(j)$ differ:

$$D = \begin{pmatrix} 0 & 2 & 3 & 0 \\ 2 & 0 & 2 & 2 \\ 3 & 2 & 0 & 3 \\ 0 & 2 & 3 & 0 \end{pmatrix}$$

Under linear periodic navigation (2), one gets the values $\Delta^A = 3/4$, $\Delta^A_{\mathrm{loc}} = 7/8$ and $\delta^A = -1/6$, higher than $E_0(\delta^A) = -1/3$: the dissimilarity between immediate neighbours under linear periodic navigationis on average (after subtracting the bias $E_0(\delta^A)$) smaller than the dissimilarity between randomly chosen pairs – although not significantly by the above normal test.

By contrast, free navigation (3) yields $\Delta^B = \Delta^B_{\mathrm{loc}} = 3/4$ and $\delta^B = 0$, since local and global inertia here coincide by construction. Note that $E_0(\delta^B) = 0$ and $\mathrm{Var}_0(\delta^B) = 0$ in case of free navigation, regardless of the values of D.

2.3. Type dissimilarities

In most cases, the dissimilarity D_{ij} between positions i and j depends only on the types $o(i) = a$ and $o(j) = b$ found at these positions. Thus, the calculation of the autocorrelation index can be based on the $v \times v$ *type dissimilarity* matrix D_{ab} rather than on the $n \times n$ position

[4] up to the bias associated to the contribution of self-comparisons: see (6).

dissimilarity matrix – which makes it possible to simplify both computation (since $v < n$ in general) and notation (cf. sections 3.2 and 3.3).

Here are some examples of *squared Euclidean* type dissimilarities, i.e. of the form $D_{ab} = \| x_a - x_b \|^2$ where $x_a \in \mathbb{R}^p$ are the p-dimensional coordinates of type a, recoverable by multidimensional scaling (see section 3.4):

a) $D_{ab} = (x_a - x_b)^2$ where x_a characterises type a, e.g. x_a = "length of a" or $x_a = 1(a \in A)$ (*presence-absence dissimilarity* with respect to property *A*)

b) $D_{ab} = 1(a \neq b)$, the *discrete metric*

c) $D_{ab} = (\frac{1}{\pi_a} + \frac{1}{\pi_b})1(a \neq b)$, *the weighted discrete metric*, where $\pi_a > 0$ is the relative proportion of type a, with $\sum_a \pi_a = 1$ (Le Roux and Rouanet 2004; Bavaud and Xanthos 2005)

d) $D_{ab} = \sum_{k=1}^{p} \frac{n_{\bullet k}}{n_{\bullet\bullet}} (\frac{n_{ak} n_{\bullet\bullet}}{n_{a\bullet} n_{\bullet k}} - \frac{n_{bk} n_{\bullet\bullet}}{n_{b\bullet} n_{\bullet k}})^2$, the *chi-square dissimilarity*, used for composite types made of distinguishable features, where n_{ak} is the *type-feature matrix* counting the occurrences of feature k in type a.

In order to compute the autocorrelation index using a type dissimilarity matrix, a $v \times v$ *type exchange* matrix can be defined as

$$\epsilon_{ab} = \sum_i \sum_j e_{ij}\, 1(o(i) = a) 1(o(j) = b) \ ,$$

whose margins specify the relative distribution of types: $\pi_a = \epsilon_{a\bullet} = \epsilon_{\bullet a} = \sum_i f_i \ 1(o(i) = a)$. Global and local inertias (4) can then be calculated as

$$\Delta = \frac{1}{2} \sum_{a,b} \pi_a \, \pi_b D_{ab} \qquad\qquad \Delta_{\text{loc}} = \frac{1}{2} \sum_{a,b} \epsilon_{ab} \, D_{ab} \tag{7}$$

The Markov transition probability from term a to term b is now ϵ_{ab}/π_a. Following (5), the autocorrelation index $\delta = 1 - \Delta_{\text{loc}}/\Delta$ has to be compared with its expected value under independence

$$E_0(\delta) = \frac{\sum_{a=1}^{v} \frac{\epsilon_{aa}}{\pi_a} - 1}{v - 1} \tag{8}$$

which generally *differs* from (5). Indeed, the permutation-invariance implied by the null hypothesis H_0 of absence of textual autocorrelation relies on permutations of *positions* $i = 1, \ldots, n$ in (5), while it considers permutations of *terms* $a = 1, \ldots, v$ in (8) – two natural although not equivalent assumptions. In the following, the position permutation test will be adopted by default, unless explicitly stated otherwise.

Toy example 1, continued: recall that the set of types occurring in the text of example 1 is $\{\alpha\beta\gamma, \alpha\delta\epsilon, \epsilon\zeta\eta\}$. The type dissimilarity D_{ab} corresponding to the position dissimilarity D_{ij} previously used is defined as the number of characters by which trigrams a and b differ:

$$D = \begin{pmatrix} 0 & 2 & 3 \\ 2 & 0 & 2 \\ 3 & 2 & 0 \end{pmatrix}$$

Under linear periodic navigation (2), the type exchange matrix and type proportions are

$$(\epsilon_{ab}) = \frac{1}{8} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \quad \pi = \frac{1}{4} \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} ,$$

yielding inertias (7) $\Delta^A = 3/4$ and $\Delta^A_{\text{loc}} = 7/8$ with $\delta = -1/6$, as already obtained.

3. Case studies

The next sections present several case studies involving in particular chi-squared dissimilarities between composite types such as play lines, hypertext navigation, and semantic dissimilarities (further illustrations, including Markov iterations $W \to W^r$, may be found in Bavaud *et al.* 2012). Unless otherwise stated, we use the "skimming" navigation model defined in section 2.1 (slightly adapted to handle border effects) and let the maximum length of jumps vary as $r = 1,2,3, \ldots$, yielding autocorrelation indices $\delta^{[r]}$ for neighbourhoods of size r, i.e. including r neighbours to the left and r neighbours to the right of each position. In particular, $\delta^{[1]}$ constitutes a generalisation of the Durbin-Watson statistic.

3.1. Autocorrelation between lines of a play

The play *Sganarelle ou le Cocu imaginaire* by Molière (1660) contains $n = 207$ lines declaimed by feminine or masculine characters (coded $s_i = 1$ or $s_i = 0$ respectively). Each line i is also characterised by the number of occurrences n_{ik} of each part-of-speech (POS) tag $k = 1, \ldots, p$ as assigned by *TreeTagger* (Schmid 1994); here $p = 28$. The first few rows and columns of the data are represented on Table 1.

Position	***Gender***	***# interjections***	***# adverbs***	***# verbs (present)***	***…***
1	1	1	2	1	…
2	0	1	11	20	…
3	1	1	0	0	…
4	0	3	15	15	…
…	…	…	…	…	…

Tab. 1: First rows and columns of the Sganarelle data.

The following distances are considered:

- the length dissimilarity $D_{ij}^{\text{length}} = (l_i - l_j)^2$, where $l_i := \sum_k n_{ik}$ is the total count of POS tags for line i
- the gender dissimilarity $D_{ij}^{\text{gender}} = (s_i - s_j)^2 = 1(s_i \neq s_j)$
- the chi-square dissimilarity D_{ij}^{χ} associated to the 207×28 contingency table n_{ik} (see section 2.3 d for the corresponding type dissimilarity)
- the reduced chi-square dissimilarity $D_{ij}^{R\chi}$ obtained after aggregating all POS tag counts into two supercategories, namely verbs and non-verbs.

The length autocorrelation index (Figure 3 left) reveals that the length of lines tends to be strongly autocorrelated over neighbourhoods of size up to 5: long (short) lines tend to be surrounded at short range by long (short) lines. This might reflect the play structure, which comprises long passages declaiming general considerations on human condition, and more action-oriented passages, made of shorter lines.

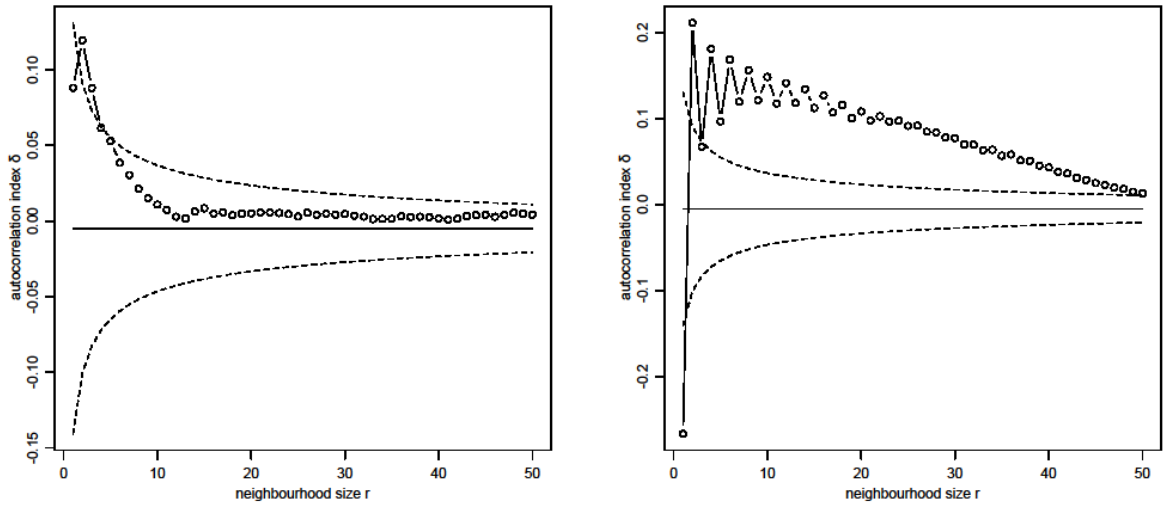

Fig. 3: Autocorrelation index $\delta^{[r]}$ (circles), together with expected value (5) (solid line), and 95% confidence interval (6) (dashed lines). Left: length dissimilarity D^{length}. Right: gender dissimilarity D^{gender}.

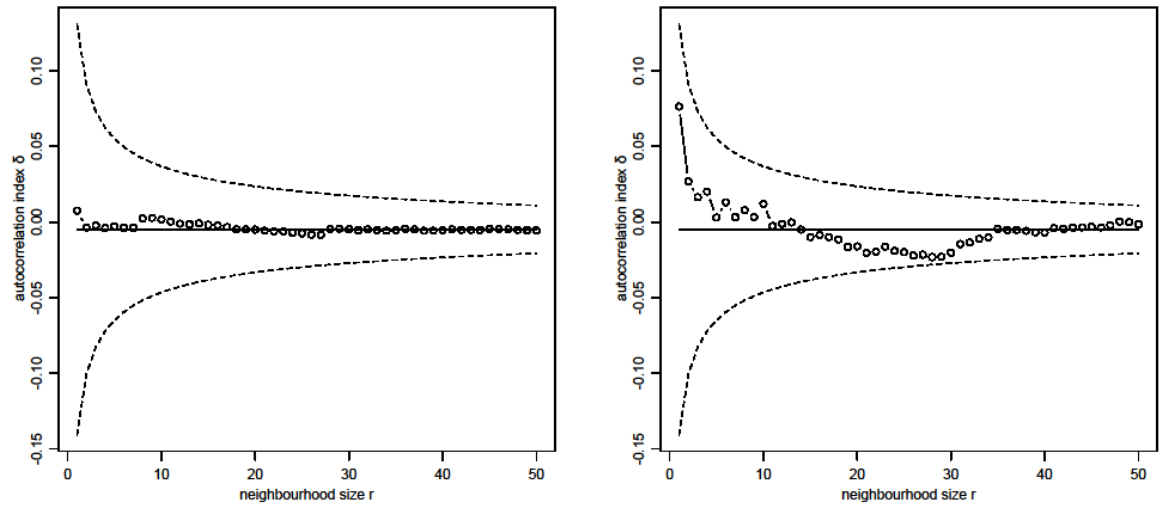

Fig 4: Autocorrelation index $\delta^{[r]}$ (same setting as figure 3) for the chi-square dissimilarity D^{χ} (left) and the reduced chi-square dissimilarity $D^{R\chi}$ (right).

The strong negative gender autocorrelation observed on figure 3 (right) for a neighbourhood size of 1 shows that lines declaimed by characters of a given gender have a clear tendency to be immediately followed by lines declaimed by characters of the other gender, and vice-versa. The significant positive autocorrelation for neighbourhoods of size 2 seems to be a logical consequence of this, as well as the globally alternating pattern of the curve. Interestingly, the autocorrelation is always positive for larger neighbourhood sizes, which can be explained by two observations: (i) overall, masculine lines are considerably more frequent than feminine lines (64.7% vs. 35.3%); (ii) the probability of being followed by a line of differing gender is much lower for masculine lines than for feminine ones (44.4% vs. 82.2%). These factors concur to dominate the short-range preference for alternation.

The POS tag profile of lines tends to exhibit no autocorrelation, although the alignments observed in Figure 4 (left) are intriguing. The proportion of verbs (Figure 4 right) tends to be positively (but not significantly, presumably due to the small size $n = 207$ of the sample) autocorrelated up to a range of 10, and negatively autocorrelated for a range between 20 and 30 – an observation whose interpretation requires further investigation.

3.2. Free navigation within documents

Let the text be partitioned into documents $g = 1, \dots, m$; $i \in g$ denotes the membership of position i to document g and $\rho_g := \sum_{i \in g} f_i$ is the relative weight of document g. Consider now the free textual navigation *confined within each document*

$$t_{ij} = f_j^{g[i]} := \frac{1(j \in g[i])\, f_j}{\rho_{g[i]}} \ , \tag{9}$$

where $g[i]$ denotes the document to which position i belongs. The associated exchange matrix $e_{ij} := \sum_{g=1}^{m} \rho_g \, f_i^g f_j^g$ is reducible, i.e. made out m disconnected submatrices. Note that f obtains here as the margin of E, rather as the stationary distribution of T, which is reducible and hence not regular. In this setup, the local inertia is nothing but the within-groups inertia

$$\Delta_{\text{loc}} = \sum_g \rho_g \, \Delta_g \ =: \Delta_W \qquad\qquad \Delta_g := \frac{1}{2} \sum_{ij} f_i^g \, f_j^g D_{ij}$$

and hence $\delta = \Delta_B / \Delta \geq 0$, where $\Delta_B = \Delta - \Delta_W = \sum_g \rho_g \, D_{g0}$ is the between-groups inertia, and D_{g0} is the dissimilarity between the centroid of the group g and the overall centroid 0. Here δ, always non negative, behaves as a kind of generalised F-ratio.

In practical applications, textual positional weights are uniform, and the free navigation within documents involves the familiar *term-document matrix*

$$n_{ag} := n_{\bullet\bullet} \sum_i f_i \, 1(o(i) = a) 1(i \in g) \tag{10}$$

with

$$f_i = \frac{1}{n_{\bullet\bullet}} \qquad f_i^g = \frac{1(i \in g)}{n_{\bullet g}} \qquad \rho_g = \frac{n_{\bullet g}}{n_{\bullet\bullet}} \ . \tag{11}$$

In particular, Δ, Δ_{loc} and δ can be computed from (7), where

$$\epsilon_{ab} = \sum_g \frac{n_{ag} \, n_{bg}}{n_{\bullet\bullet} \, n_{\bullet g}} \qquad \pi_a = \frac{n_{a\bullet}}{n_{\bullet\bullet}} \quad \text{(free within–documents navigation).} \tag{12}$$

The significance of $\delta = (\Delta - \Delta_{\text{loc}})/\Delta$ can be tested by (6), where $\text{trace}(W^2) = \text{trace}(W) = m$ for the free within-documents navigation under the position permutation test (5).

When $D_{ab} = (\frac{n_{\bullet\bullet}}{n_{a\bullet}} + \frac{n_{\bullet\bullet}}{n_{b\bullet}}) 1(a \neq b)$ is the weighted discrete metric (section 2.3 c), the autocorrelation index turns out to be

$$\delta = \frac{\chi^2}{n_{\bullet\bullet}(v - 1)} \tag{13}$$

where v is the number of types and χ^2 the term-document chi-square.

Toy example 2: consider $v = 7$ types represented by greek letters, whose $n = n_{\bullet\bullet} = 20$ occurrences possess the same weight $f_i = 1/20$, and are distributed among $m = 4$ documents as "$\beta\beta\gamma\delta$", "$\alpha\alpha\gamma\epsilon$", "$\alpha\alpha\beta\beta$" and "$\alpha\alpha\alpha\alpha\epsilon\zeta\zeta\eta$" (Figure 5). The term-document matrix, term

weights and document weights read

$$(n_{ag}) = \begin{pmatrix} & \boldsymbol{g=1} & \boldsymbol{g=2} & \boldsymbol{g=3} & \boldsymbol{g=4} \\ \boldsymbol{\alpha} & 0 & 2 & 2 & 4 \\ \boldsymbol{\beta} & 2 & 0 & 2 & 0 \\ \boldsymbol{\gamma} & 1 & 1 & 0 & 0 \\ \boldsymbol{\delta} & 1 & 0 & 0 & 0 \\ \boldsymbol{\epsilon} & 0 & 1 & 0 & 1 \\ \boldsymbol{\zeta} & 0 & 0 & 0 & 2 \\ \boldsymbol{\eta} & 0 & 0 & 0 & 1 \end{pmatrix} \quad \pi = \frac{1}{20}\begin{pmatrix} 8 \\ 4 \\ 2 \\ 1 \\ 2 \\ 2 \\ 1 \end{pmatrix} \quad \rho = \frac{1}{5}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 2 \end{pmatrix} \tag{14}$$

Consider three type dissimilarities, namely the "vowels" presence-absence dissimilarity D^A (section 2.3 a, with $A = \{\alpha, \epsilon\}$ and $A^c = \{\beta, \gamma, \delta, \zeta, \eta\}$):

$$D^A = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix},$$

the discrete metric D^B (section 2.3 b):

$$D^B = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix},$$

and the weighted discrete metric D^C (section 2.3 c):

$$D^C = \begin{pmatrix} 0 & 7.5 & 12.5 & 22.5 & 12.5 & 12.5 & 22.5 \\ 7.5 & 0 & 15 & 25 & 15 & 15 & 25 \\ 12.5 & 15 & 0 & 30 & 20 & 20 & 30 \\ 22.5 & 25 & 30 & 0 & 30 & 30 & 40 \\ 12.5 & 15 & 20 & 30 & 0 & 20 & 30 \\ 12.5 & 15 & 20 & 30 & 20 & 0 & 30 \\ 22.5 & 25 & 30 & 40 & 30 & 30 & 0 \end{pmatrix}.$$

The corresponding values of global inertias (7), local inertias (12) and textual autocorrelation δ are given in Table 2 below.

Sganarelle, continued: consider the distribution of the 961 nouns and 1'204 verbs of the play Sganarelle among the $m = 24$ scenes of the play, treated here as documents (section 3.1). The autocorrelation index (13) for nouns associated to the weighted discrete metric takes on the value $\delta^{\text{nouns}} = 0.0238$, lower than the expected value (5) $E_0(\delta^{\text{nouns}}) = 0.0240$. For verbs, one gets $\delta^{\text{verbs}} = 0.0198 > E_0(\delta^{\text{verbs}}) = 0.0191$. Although not statistically significant, the sign of the differences reveals a lexical content within scenes more homogeneous for verbs than for nouns. Finer analysis can be obtained from *Correspondence Analysis* (see e.g. Greenacre 2007), performing a spectral decomposition of the chi-square in (13).

3.3. Hypertext navigation

Consider a set G of electronic documents $g = 1, \ldots, m$ containing *hyperlinks* attached to a set A of *active terms* and specified by a function $\gamma[a]$ from A to G, associating each active term a to a target document $g = \gamma[a]$. A simple model of hypertext navigation consists in clicking at each

position occupied by an active term, thus jumping to the target document, while staying in the same document when meeting an inactive term; in both cases, the next position i is selected as f_i^g in (11). This dynamics generates a document to document transition matrix $\Phi = (\varphi_{gh})$, involving the term-document matrix n_{ag} (10), as

$$\varphi_{gh} := \sum_a \frac{n_{ag}}{n_{\bullet g}} \tau_{(ag)h} \tag{15}$$

where $\tau_{(ag)h}$ is the probability to jump from term a in document g to document h and obeys $\tau_{(ag)\bullet} = 1$. In the present setup, $\tau_{(ag)h} = 1(h = \gamma[a])$ for $a \in A$ and $\tau_{(ag)h} = 1(h = g)$ for $a \notin A$. Alternative specifications taking into account clicking probabilities, or contextual effects (of term a relatively to its background g) could also be cast within this formalism. The document-to-document transition matrix obeys $\varphi_{gh} \geq 0$ and $\varphi_{g\bullet} = 1$, and the broad Markovian family of hypertext navigations (15) generalizes specific proposals such as the free within-documents setup, or the Markov chain associated to the PageRank algorithm (Page 2001).

By standard Markovian theory (e.g. Grinstead and Snell 1998), each document belongs to a single "communication-based" equivalence class, which is either transient, i.e. consisting of documents eventually unattainable by lack of incoming hyperlinks, or recurrent, i.e. consisting of documents visited again and again once the chain is entered into the class. The chain is regular iff it is aperiodic and consists of a single recurrent class, in which case its evolution converges to the stationary distributions of Φ obeying $\sum_g s_g \varphi_{gh} = s_h$, which differs in general from the document weights $\rho_g = n_{\bullet g}/n_{\bullet\bullet}$.

In the regular case, textual autocorrelation for type dissimilarities (section 2.3) can be computed by means of (7), where (compare with (12))

$$\epsilon_{ab} = \eta_{(a\bullet)(b\bullet)} \qquad \eta_{(ag)(bh)} := \frac{1}{2}\frac{n_{ag}}{n_{\bullet g}}\frac{n_{bh}}{n_{\bullet h}}[\tau_{(ag)h}s_g + \tau_{(bh)g}s_h]$$

and

$$\pi_a = \eta_{(a\bullet)(\bullet\bullet)} = \sum_g \frac{n_{ag}}{n_{\bullet g}} s_g \neq \frac{n_{a\bullet}}{n_{\bullet\bullet}} \qquad \text{(hypertextual navigation).}$$

Toy example 2, continued: let the active terms be $A = \{\alpha, \beta, \gamma, \delta\}$, with hyperlinks $\gamma[\alpha] = 1$, $\gamma[\beta] = 2$, $\gamma[\gamma] = 3$ and $\gamma[\delta] = 4$ (Figure 5). The transition matrix (15) turns out to be regular. From (14), the document-document transition probability, its stationary distribution and the document weights are

$$\Phi = \begin{pmatrix} 0 & 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 \end{pmatrix} \qquad s = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/6 \\ 1/6 \end{pmatrix} \qquad \rho = \begin{pmatrix} 1/5 \\ 1/5 \\ 1/5 \\ 2/5 \end{pmatrix} .$$

In a certain sense, hyperlink navigation *magnifies the importance of each document* g by a factor s_g/ρ_g, respectively equal to $5/3$, $5/3$, $5/6$ and $5/12$ for the $m = 4$ documents of toy example 2. Similary, the *term magnification factor* $n_{\bullet\bullet}\pi_a/n_{a\bullet}$ is 1.04 for α, 0.83 for β, 1.67 for γ, 1.67 for δ, 1.04 for ϵ, 0.42 for ζ and 0.42 for η.

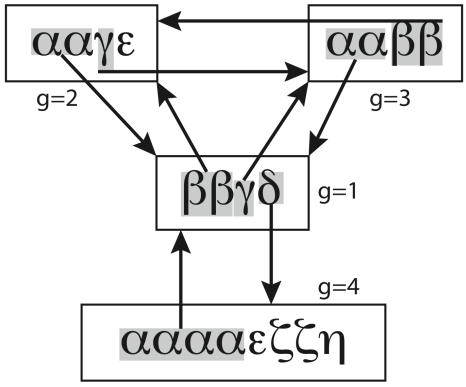

Fig 5: Hypertextual navigation (toy example 2) between $m = 4$ documents containing $|A| = 4$ active terms α, β, γ, and δ.

	free within-document navigation				hypertextual navigation			
	Δ	Δ_{loc}	δ	$E_0(\delta)$	Δ	Δ_{loc}	δ	$E_0(\delta)$
D^A	0.25	0.18	0.28	0.16	0.25	0.36	-0.47	-0.10
D^B	0.38	0.31	0.20	0.16	0.39	0.48	-0.24	-0.10
D^C	6.00	4.94	0.18	0.16	6.15	6.90	-0.12	-0.10

Tab. 2: (Toy example 2) terms autocorrelation is positive under free within document navigation, but negative under hypertextual navigation. Here $E_0(\delta)$ refers to the term permutation test (8).

Table 2 summarizes the resulting textual autocorrelation for the three term dissimilarities already investigated: systematically, hyperlink navigation strongly increases the terms heterogeneity as measured by the local inertia, since each of the active terms α, β, γ and δ point towards documents not containing them.

Textual autocorrelation in "WikiTractatus": strict application of the above formalism to real data requires full knowledge of a *finite, isolated network* of m hyperlinked documents.

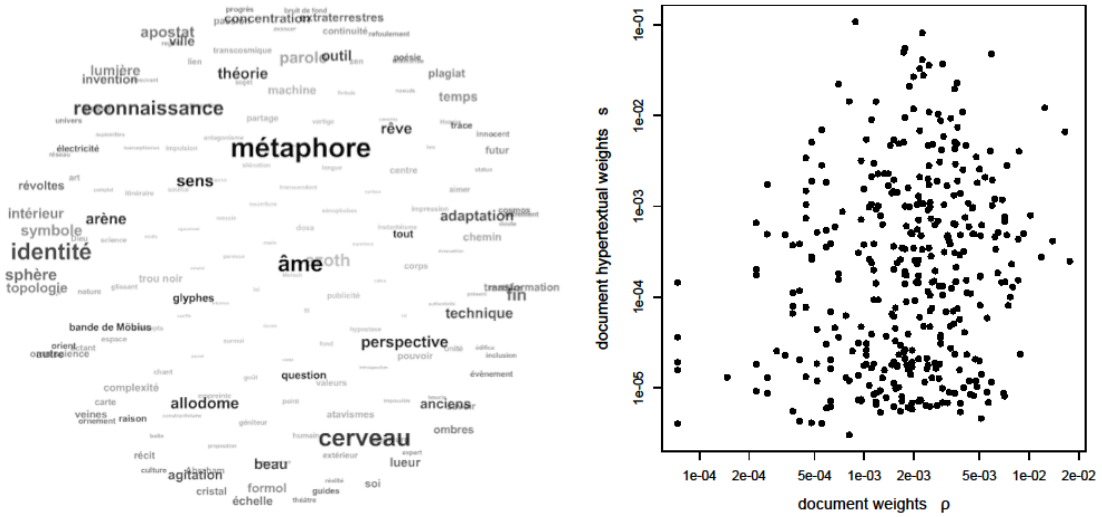

Fig. 6: Entry page of "WikiTractatus", constituting the document "les mots" (left). Free-within navigation weights ρ_g versus hypertextual weights s_g, logarithmic scale (right).

The site "WikiTractatus" is a "network of aphorisms" (in French) created by André Ourednik (2010), containing $n = 27'172$ words (tokens) distributed among $v = 5'487$ terms and $m = 380$ documents (Figure 6). Each document is identified by a distinct active term, as in a dictionary notice, and hyperlinks connect each active term in the corpus to the corresponding document. Also, each document contains hyperlinks pointing towards distinct documents, with the exception of the document "clavier" which contains no active terms. As a consequence,

"clavier" acts as an absorbing state of the Markov chain (15), and all remaining documents are transient – as attested by the study of Φ^r, converging for r large towards a null matrix except for a unit column vector associated to the document "clavier". Suppressing document "clavier" together with its incoming hyperlinks makes the Markov chain regular.

In contrast to Table 2, terms are positively autocorrelated under hypertextual navigation on "WikiTractatus": with the discrete metric D and the term permutation test (8), one finds $\Delta = 0.495$, $\Delta_{\text{loc}} = 0.484$, $\delta = 0.023$ and $E_0(\delta) = 0.014$. In the same setup, the free within-documents navigation yields very close results, namely $\Delta = 0.496$, $\Delta_{\text{loc}} = 0.484$, $\delta = 0.024$ and $E_0(\delta) = 0.015$. Here both types of navigation have identical effects on textual autocorrelation per se, but quite different effects on the document (and type) relative weights (Figure 6 right).

3.4. Semantic autocorrelation

Semantic similarities have been systematically investigated in the last two decades, using in particular reference word taxonomies expressing "ontological" relationships (e.g. Resnik 1999). In WordNet (Miller *et al.* 1990), words, and in particular nouns and verbs, are grouped into *synsets*, i.e. cognitive synonyms, and each synset represents a different concept. *Hyponomy* expresses inclusion between concepts: the relation *"concept c_1 is an instance of concept c_2"* is denoted $c_1 \leq c_2$, and $c_1 \vee c_2$ represents the *least general concept* subsuming both c_1 and c_2. For instance, in the toy ontology of Figure 7, `cat` $\leq$ `animal` and `cat` $\vee$ `dog` = `animal`.

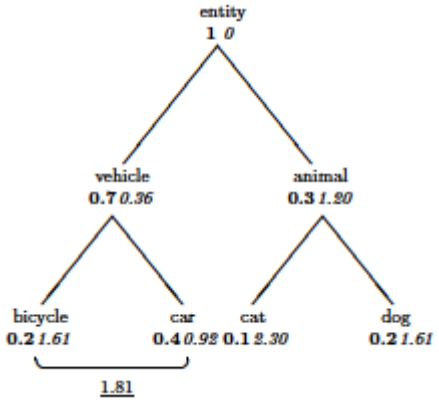

Fig. 7: Toy noun ontology made up of 7 concepts: numbers in bold are probabilities (16), numbers in italic are similarities (17), and the underlined number is the dissimilarity between `bicycle` *and* `car` *according to (18).*

Based on a reference corpus (hereafter the Brown corpus, Kučera and Francis 1967), the probability $p(c)$ of concept c can be estimated as the proportion of word tokens whose sense $C(w)$ is an instance of concept c. Thus, representing the number of occurrences of word w by $n(w)$,

$$p(c) := \frac{\sum_w n(w)\, 1(C(w) \leq c)}{\sum_w n(w)} \tag{16}$$

Following Resnik (1999), a measure of *similarity* between concepts can then be defined as:

$$s(c_1, c_2) := -\log p(c_1 \vee c_2) \geq 0 \tag{17}$$

from which a squared Euclidean *dissimilarity* between concepts can be derived as (Bavaud *et al.* 2012):

$$D(c_1, c_2) := s(c_1, c_1) + s(c_2, c_2) - 2s(c_1, c_2) \tag{18}$$

For instance, based on the probabilities given in Figure 7,

$$
\begin{aligned}
D(\texttt{bicycle}, \texttt{car}) &= s(\texttt{bicycle}, \texttt{bicycle}) + s(\texttt{car}, \texttt{car}) \\
&\quad -2s(\texttt{bicycle}, \texttt{car}) \\
&= -\log(0.2) - \log(0.4) + 2\log(0.7) \\
&= 1.81
\end{aligned}
$$

According to *TreeTagger* (Schmid 1994), the short story *The Masque of the Red Death* by Edgar Allan Poe (1842) contains 497 positions occupied by nouns and 379 positions occupied by verbs. Similarities between nouns and between verbs can be obtained using the *WordNet::Similarity* interface (Pedersen *et al.* 2004) – systematically using, in this case study, the most frequent sense of ambiguous concepts. Autocorrelation indices (for neighbourhoods of size r) calculated using the corresponding dissimilarities exhibit no noticeable pattern (Figure 8).

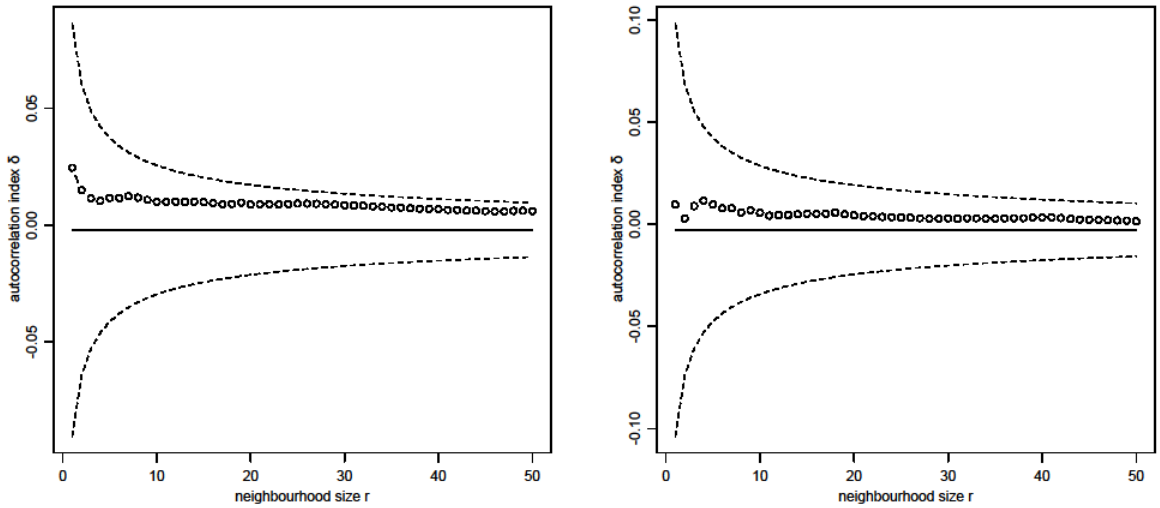

Fig. 8: Autocorrelation index $\delta^{[r]}$ (same setting as Figure 3) in "The Masque of the Red Death" for the semantic dissimilarity (18) for nouns (left) and for verbs (right).

This being said, the p-dimensional coordinates x_a entering in any squared Euclidean distance $D_{ab} = \| x_a - x_b \|^2$ can be recovered by (weighted) *multidimensional scaling* (MDS) (e.g. Torgeson 1958; Mardia *et al.* 1979), yielding orthogonal factorial coordinates $x_{a\alpha}$ (for $\alpha = 1, \ldots, v-1$) whose low-dimensional projections express a maximum proportion of (global) inertia.

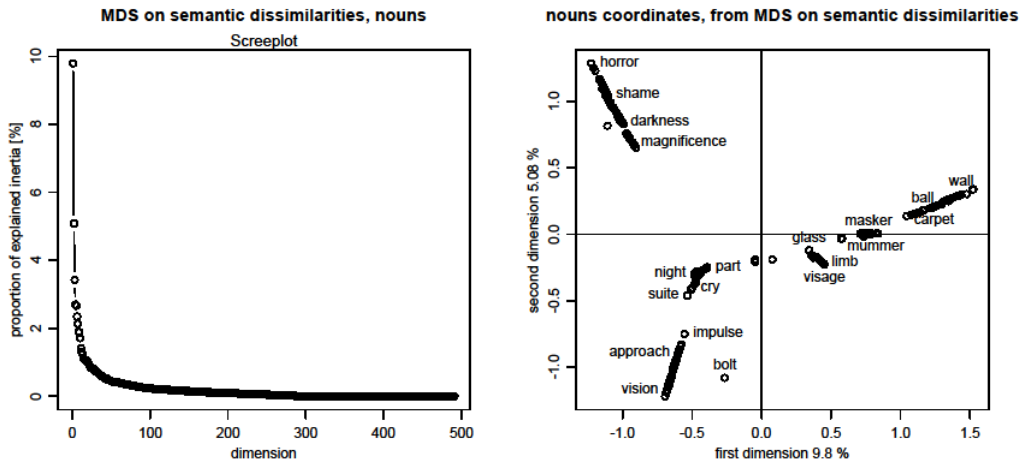

Fig. 9: Screeplot for the MDS on semantic dissimilarities for nouns (left). Factorial coordinates $x_{\alpha\alpha}$ and proportion of explained inertia for $\alpha = 1{,}2$ (right).

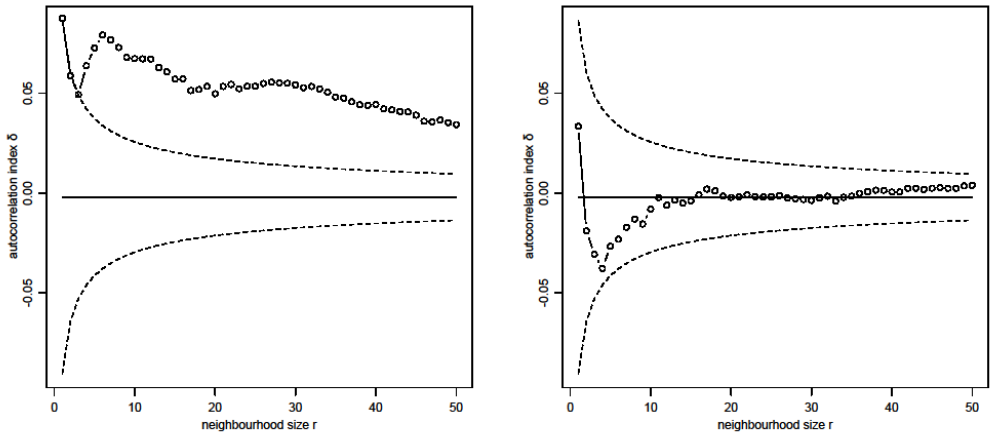

Fig. 10: Autocorrelation index $\delta^{[r]}$ for nouns in the first semantic dimension (left) and in the second semantic dimension (right).

The first semantic coordinate for nouns in *The Masque of the Red Death* (Figure 9) clearly contrasts *abstract entities* such as *horror*, *pestilence*, *disease*, *hour*, *mean*, *night*, *vision*, or *precaution*, on the left, with *physical entities* such as *window*, *roof*, *wall*, *body*, *victim*, *glass*, or *visage*, on the right, respectively defined in WordNet as "a general concept formed by extracting common features from specific examples" and "an entity that has physical existence". Figure 10 (left) shows this first coordinate to be strongly autocorrelated, echoing long-range semantic persistence, in contrast to the second coordinate (Figure 10 right), whose interpretation is more difficult.

For verbs (Figure 11), the first semantic coordinate differentiates *stative verbs*, such as *be*, *seem*, or *sound*, from all other verbs, while the second semantic coordinate differentiates the verb *have* from all other verbs. Figure 12 reveals that the first coordinate is strongly autocorrelated, while the second coordinate is negatively autocorrelated for neighbourhood ranges up to 2. Although the latter result is not significant for $\alpha = 0.05$ according to (6), it is likely due to the use of *have* as an auxiliary verb in past perfect and other compound verb tenses.

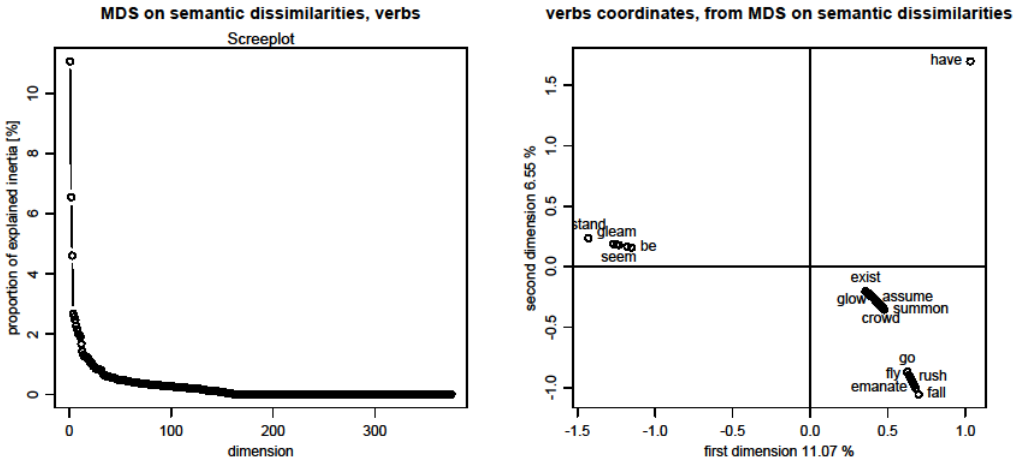

Fig. 11: Screeplot for the MDS on semantic dissimilarities for verbs (left). Factorial coordinates $x_{\alpha\alpha}$ and proportion of explained inertia for $\alpha = 1{,}2$ (right).

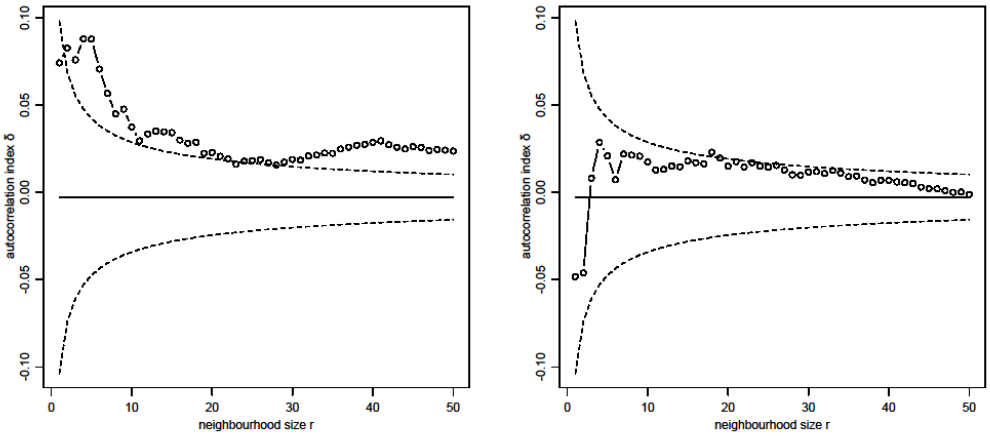

Fig. 12: Autocorrelation index $\delta^{[r]}$ for verbs in the first semantic dimension (left) and in the second semantic dimension (right)

4. Conclusion

In this contribution, we have introduced a unified formalism for *textual autocorrelation*, i.e. the tendency for neighbouring textual positions to be more (or less) similar than randomly chosen positions. This approach to sequence and text analysis is based on two primitives: (i) *neighbourhoodness* between textual positions, as determined by a Markov model of navigation, and formally represented by the exchange matrix E; and (ii) *(dis-)similarity* between positions, as encoded in the (typically squared Euclidean) dissimilarity matrix D.

By varying E and or D, the proposed formalism recovers and revisits well-known statistical objects and concepts, such as the F-ratio, the chi-square and Correspondence Analysis. It also gives a unified account of various representations commonly used for textual data analysis, in particular the sequential and bag-of-words models, as well as the term-document matrix. It can also be extended to provide a model of hypertext navigation, where hyperlinks act as magnifying (or reducing) glasses, modifying the relative weights of documents, and altering (or not) textual autocorrelation.

This approach is applicable to any form of sequence and text analysis that can be expressed in terms of dissimilarity between positions (or between types). The presented case studies have aimed at illustrating this versatility by addressing lexical, morphosyntactic, and semantic properties of texts. As shown in the latter case, squared Euclidean dissimilarities can be visualised and decomposed into factorial components by multidimensional scaling; the textual autocorrelation of each component can in turn be analysed and interpreted – yielding in particular a new means of dealing with semantically related problems.

References

Anselin, Luc (1995) Local Indicators of Spatial Association. *Geographical Analysis* 27. 93–115.

Bavaud, François (2013) Testing Spatial Autocorrelation in Weighted Networks: the Modes Permutation Test. *Journal of Geographical Systems* 14. 233–247.

Bavaud, François, Christelle Cocco & Aris Xanthos (2012) Textual autocorrelation : formalism and illustrations. In Anne Dister, Dominique Longrée & Gérald Purnelle (eds.), *11èmes*

Journées internationales d'analyse statistique des données textuelles, 109–120. Liège : Université de Liège.
Bavaud, François & Aris Xanthos (2005) Markov associativities. *Journal of Quantitative Linguistics* 12. 123–137.
Berger, Joseph & James Laurie Snell (1957) On the concept of equal exchange. *Behavioral Science* 2. 111–118.
Cliff, Andrew David & John Keith Ord (1981) *Spatial Processes: Models & Applications*. London: Pion.
Cressie, Noel (1991) *Statistics for Spatial Data*. New York: John Wiley & Sons.
Greenacre, Michael (2007) *Correspondence Analysis in Practice*, 2nd edn. London: Chapman and Hall/CRC Press.
Grinstead, Charles Miller & James Laurie Snell (1998) *Introduction to Probability*. American Mathematical Soc.
Kučera, Henry & Winthrop Nelson Francis (1967) *Computational Analysis of Present-day American English*. Providence: Brown University press.
Lebart, Ludovic (1969) Analyse statistique de la contigüité. *Publication de l'Institut de Statistiques de l'Université de Paris* 18. 81–112.
Le Roux, Brigitte & Henry Rouanet (2004) *Geometric Data Analysis*. Kluwer: Dordrecht.
Mardia, Kanti V., John T. Kent & John M. Bibby (1979) *Multivariate Analysis*. New York: Academic Press.
Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine Miller (1990) WordNet: An on-line lexical database. *International Journal of Lexicography* 3. 235–244.
Moran, Patrick & Alfred Pierce (1950) Notes on continuous stochastic phenomena. *Biometrika* 37. 17–23.
Ourednik, André (2010) Wikitractatus `http://wikitractatus.ourednik.info/` (consulted in December 2012).
Page, Lawrence. (2001) Method for node ranking in a linked database. *U.S. Patent No 6,285,999.*
Pedersen, Ted, Siddharth Patwardhan & Jason Michelizzi (2004) WordNet::Similarity – Measuring the Relatedness of Concepts. In Susan Dumais, Daniel Marcu & Salim Roukos (eds.), *Proceedings of HLT-NAACL 2004: Demonstration Papers*, 38–41. Boston: Association for Computational Linguistics.
Resnik, Philip (1999) Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research* 11. 95–130.
Schmid, Helmut (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
Torgeson, Warren S. (1958) *Theory and Methods of Scaling*. John Wiley & sons, New York.