Textual autocorrelation : formalism and illustrations

François Bavaud, Christelle Cocco, Aris Xanthos

University of Lausanne
{francois.bavaud, christelle.cocco, aris.xanthos}@unil.ch

Abstract

Textual autocorrelation is a broad and pervasive concept, referring to the similarity between nearby textual units: lexical repetitions along consecutive sentences, semantic association between neighbouring lexemes, persistence of discourse types (narrative, descriptive, dialogal...) and so on. Textual autocorrelation can also be negative, as illustrated by alternating phonological or morpho-syntactic categories, or the succession of word lengths.
This contribution proposes a general Markov formalism for textual navigation, and inspired by spatial statistics. The formalism can express well-known constructs in textual data analysis, such as term-document matrices, references and hyperlinks navigation, (web) information retrieval, and in particular textual autocorrelation, as measured by Moran's I relatively to the exchange matrix associated to neighbourhoods of various possible types.
Four case studies (word lengths alternation, lexical repulsion, parts of speech autocorrelation, and semantic autocorrelation) illustrate the theory. In particular, one observes a short-range repulsion between nouns together with a short-range attraction between verbs, both at the lexical and semantic levels.

Résumé

Le concept d'autocorrélation textuelle, fort vaste, réfère à la similarité entre unités textuelles voisines: répétitions lexicales entre phrases successives, association sémantique entre lexèmes voisins, persistance du type de discours (narratif, descriptif, dialogal...) et ainsi de suite. L'autocorrélation textuelle peut être également négative, comme l'illustrent l'alternance entre les catégories phonologiques ou morpho-syntaxiques, ou la succession des longueurs de mots.
Cette contribution propose un formalisme markovien général pour la navigation textuelle, inspiré par la statistique spatiale. Le formalisme est capable d'exprimer des constructions bien connues en analyse des données textuelles, telles que les matrices termes-documents, les références et la navigation par hyperliens, la recherche documentaire sur internet, et, en particulier, l'autocorélation textuelle, telle que mesurée par le I de Moran relatif à une matrice d'échange associée à des voisinages de différents types possibles.
Quatre cas d'étude illustrent la théorie: alternance des longueurs de mots, répulsion lexicale, autocorrélation des catégories morpho-syntaxiques et autocorrélation sémantique. On observe en particulier une répulsion à courte portée entre les noms, ainsi qu'une attraction à courte portée entre les verbes, tant au niveau lexical que sémantique.

Keywords: hyponymy, local variance, Markov transitions, Moran's I, semantic scores, textual attraction, textual dissimilarities, textual navigation, textual repulsion

1. Introduction

Reading can be formalised as navigating among textual positions. The present contribution indulges in conceiving text as a space rather as a time series. Borrowing concepts from spatial statistics such as "spatial neighbourhoods" and "spatial field" enables the computation of global and local measures of dispersion, whose comparison permits to measure the textual tendency for short- or long-range repetition or avoidance between lexical, syntactic or semantic categories.

Standard, linear reading proceeds by left-to-right transitions from the first position to the last one. Re-reading, zapping, switching to footnotes and references, or following hyperlinks breaks down the linearity of reading, and generates jumps and reversals, as well as multiple navigation possibilities. The simplest model accounting for non-linear, stochastic reading is a Markov chain on textual positions; assuming the chain to be finite and regular (which implies some kind of rewinding, trapping for ever the reader in the text), the long-term iterations of the process define the weights of textual position - the stationary distribution of the Markov chain.

Comparing the field differences among two independently sampled positions (ordinary variance) versus two positions sampled according to the Markov chain (local variance) permits to measure and test textual autocorrelation. Although Markov transitions for reading are of course highly asymmetric in general, the corresponding ordinary and local variances depend only upon the symmetric part of the Markov chains, defining in turn a *symmetric neighbourhood* around each textual position (see equation (1)).

The aim of this paper is twofold: first, to expose a quite general formalism for textual navigation, encompassing traditional book reading, hypertext browsing, and other possibilities which might or might not correspond to a presently existing practice. Thanks to its generality, the mathematical apparatus makes it possible to define diverse kinds of neighbourhoods and autocorrelation phenomena, and also to unify other important textual concepts and methods such as hard versus soft documents, term-document matrices, web search, semantic dissimilarities, correspondence analysis, latent semantic analysis, and multidimensional scaling.

The second part of the paper presents four illustrations of lexical, syntactic and semantic autocorrelation, essentially restricted (in contrast to the above generality) to neighbourhoods of increasing size in a single document, and limited to measuring the positional attraction and repulsion between textual categories.

2. General formalism

Textual navigation. Consider a corpus made out of n tokens, consisting of textual units such as words, characters, sentences, etc., and occurring at *positions* indexed by $i, j...$, taking on values from 1 to n. Some textual positions may be more prominent than other (e.g. through typographic or locational emphasis, or otherwise), as quantified by possibly varying *relative weights* $f_i > 0$ with $\sum_{i=1}^{n} f_i =: f_\bullet = 1$ [1].

Textual navigation is modelled by a non negative transition matrix $T = (t_{ij})$ (with $\sum_j t_{ij} = 1$), the probability that position j will be read after position i, as intended by the author, or as

[1] here and in the sequel, symbol "$\bullet$" denotes the summation over all values of the substituted index.

reflecting the readers effective behaviour, possibly by using hyperlinks. Linear navigation, i.e. usual reading, obtains as the particulary case $t_{ij} = 1(j = i + 1)$ [2].

In the fiction of an ever-reading agent[3], the position weight should reflect the long run visiting frequency, that is $f_j = \sum_i f_i t_{ij}$ (stationary distribution).

Textual field, neighbourhoods and autocorrelation. Let x_i be a numerical variable or *textual field* characterising the token situated at i, such as the presence/absence of a term or a category, the word length or frequency, or a "semantic score" resulting from MDS applied to a matrix of pairwise semantic dissimilarities (see Section 3.4). Its average over the corpus is $\bar{x} = \sum_i f_i x_i$ and its variance reads

$$\text{var}(x) = \sum_i f_i (x_i - \bar{x})^2 = \frac{1}{2} \sum_{ij} f_i f_j (x_i - x_j)^2$$

In the latter expression, $f_i f_j$ is the probability of independently selecting two positions i and j. Replacing it with $f_i t_{ij}$, the probability of reading the ordered bigram (i, j) defines the *local variance* (Lebart 1969; Bavaud 2008)

$$\text{var}_{\text{loc}}(x) = \frac{1}{2} \sum_{ij} f_i t_{ij} (x_i - x_j)^2 = \frac{1}{2} \sum_{ij} e_{ij} (x_i - x_j)^2 \qquad (1)$$

where $e_{ij} := \frac{1}{2}(f_i t_{ij} + f_j t_{ji})$ are the components of the symmetric *exchange matrix* E, non-negative and obeying $\sum_i e_{ij} = f_j$ (Berger and Snell 1957). The exchange matrix E characterizes *textual neighbourhoods*, and enters the definition of the ratio $\text{var}_{\text{loc}}(x)/\text{var}(x)$, known as Geary's c in spatial statistics (see e.g. Cressie 1991), and reducing to the Durbin-Watson test statistics for the linear neighbourhoods $e_{ij} = 1(j = i \pm 1)/(2n)$ (up to boundary corrections). *Moran's I* (Moran 1950) obtains as

$$I(x) := \frac{\text{var}(x) - \text{var}_{\text{loc}}(x)}{\text{var}(x)} \qquad -1 \leq I(x) \leq 1$$

and constitutes the standard measure of spatial autocorrelation (see e.g. Cressie 1991): high values of $I(x)$ characterise a field x whose local variations (that is, in the range of E) are of lesser magnitude than its overall variations across the whole corpus.

A multivariate generalisation considers the (squared) Euclidean dissimilarities $D_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 = \|x_i - x_j\|^2$ between pairs of positions located at i and j, where x_{ik} is the (conveniently standardised) value of the k-th feature associated to i. Extending the above definitions, one defines the *inertia* Δ, the *local inertia* Δ_{loc} and their relative difference $\delta \in [-1, 1]$ as

$$\Delta := \frac{1}{2} \sum_{ij} f_i f_j D_{ij} \qquad \Delta_{\text{loc}} := \frac{1}{2} \sum_{ij} e_{ij} D_{ij} \qquad \delta := \frac{\Delta - \Delta_{\text{loc}}}{\Delta} \ .$$

[2] here and in the sequel, $1(A)$ denotes the *indicator function* of event A, taking on the value 1 if A is true, and 0 otherwise.

[3] a more complete modelling, not pursued here, could consider the additional *outside text* state indexed by 0, with t_{i0} as the probability of exiting the text at position i, and t_{0i} as the probability of entering from the text at position i.

The width of neighbourhoods can be increased by considering the iterates

$$E^{(r)} = (e_{ij}^{(r)}) := \Pi W^r \qquad \Pi := \mathrm{diag}(f) \tag{2}$$

where W stands for the reversible Markov transition matrix with components

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

Testing textual autocorrelation. Under the null hypothesis H_0 of no textual autocorrelation, the values x_i are independent of their positions i, that is all the quantities of the form $I(\pi(x))$ are equiprobable, where $\pi(x_i) := x_{\pi(i)}$ is a permutation among positions. p-values obtain by comparing the observed value $I(x)$ to B permuted values $I(\pi(x))$ (among $n!$ possibilities), or, by extension, by comparing the relative difference δ to B permuted values $\delta(\pi(D))$.

Under H_0, the *neutral* or expected value of $I(x)$ is $I_0 = (\mathrm{trace}(T) - 1)/(n-1)$, which is non zero in general. In particular, $I_0 = -1/(n-1)$ for an off-diagonal exchange matrix.

Documents, references and hyperlinks. Documents are made of textual units occurring at specific positions, and can be defined by the membership or indicator matrix $Z = (z_{ig})$ with $z_{i\bullet} = 1$, where $z_{ig} = 1$ if position i belongs to document g, and $z_{ig} = 0$ otherwise. Soft generalisations with $z_{ig} \geq 0$ and $\sum_g z_{ig} = 1$, not pursued here, are also conceivable.

In all generality, a *reference* or a *hyperlink* intends to explain, clarify or detail a part (a paragraph, a chapter or the totality) of the current document g by highlighting a part of the referenced document h. Simple choices for representing references and hyperlinks are

- position-to-position transitions $p(i \to j) = t_{ij}$ (where $i \in g$ and $j \in h$) between documents, breaking down the linear textual structure
- position-to-document transitions $p(i \to h) = \sum_j t_{ij} z_{jh}$
- document-to-document transitions $p(g \to h) = \tau_{gh}$, with $\sum_h \tau_{gh} = 1$, as in the PageRank algorithm, allowing for self- and multiple references, and fixing the case of documents with no outlinks by allowing random jumps towards other documents (e.g. Langville and Meyer 2006). Consistency requires the transitions between documents to result from the addition of all positional transitions between units contained in the documents, that is

$$\tau_{gh} = \frac{1}{\rho_g} \sum_{ij} f_i z_{ig} t_{ij} z_{jh} \qquad \text{with} \qquad \rho_g = \sum_i f_i z_{ig} \tag{3}$$

In any case, ρ_g in (3) is the relative weight of document g, and obtains as the stationary distribution of $\mathcal{T} = (\tau_{gh})$, that is $\mathcal{T}'\rho = \rho$ where $\mathcal{T} = R^{-1} Z' \Pi T Z$ with $R := \mathrm{diag}(\rho)$.

Symmetric, normalised *document-document exchange matrices* $\epsilon = (\epsilon_{gh})$ obtain from the above as $\epsilon_{gh} = \sum_{ij} e_{ij} z_{ig} z_{jh}$, that is $\epsilon = Z'EZ$ with $\epsilon_{g\bullet} = \rho_g$.

Discrete document dissimilarities simply obtain as $D_{ij}^{\text{doc}} := \frac{1}{2}\sum_g (z_{ig} - z_{jg})^2$, taking on value 1 iff positions i and j refer to different documents (and 0 otherwise). The resulting inertia, local inertia and relative difference read

$$\Delta^{\text{doc}} = \frac{1}{2}(1 - \sum_g \rho_g^2) \qquad \Delta_{\text{loc}}^{\text{doc}} = \frac{1}{2}(1 - \sum_g \epsilon_{gg}) \qquad \delta^{\text{doc}} = \frac{\sum_g (\epsilon_{gg} - \rho_g^2)}{1 - \sum_g \rho_g^2} \ .$$

δ^{doc} takes on its maximal value 1 iff ϵ is diagonal, that is in absence of references or hyperlinks between different documents - in which case Δ^{doc} only contains intra-document contributions. Conversely, δ^{doc} can be shown to take on its minimal value $-1/(m-1)$ in the "completely hypertextual" (and completely unreadable) case consisting in systematically switching between m documents of same weight, without any intra-document navigation.

Discrete term dissimilarities. Consider the *term indicator matrix* $X = (x_{i\alpha})$ with $x_{i\alpha} = 1$ if term α occurs at place i, and $x_{i\alpha} = 0$ otherwise. By construction, $x_{i\bullet} = 1$ and $D^{\text{term}}_{ij} := \frac{1}{2}\sum_\alpha (x_{i\alpha} - x_{j\alpha})^2$ takes on the value 1 iff the terms at i and j differ, that is D^{term} is the *discrete metric* between terms. Proceeding as above, one finds

$$\Delta^{\text{term}} = \frac{1}{2}(1 - \sum_\alpha p_\alpha^2) \qquad \Delta^{\text{term}}_{\text{loc}} = \frac{1}{2}(1-\theta) \qquad \delta^{\text{term}} = \frac{\theta - \sum_\alpha p_\alpha^2}{1 - \sum_\alpha p_\alpha^2}$$

where $\theta := \sum_{ij\alpha} e_{ij} x_{i\alpha} x_{j\alpha}$ is the probability that a term repetition occurs among two neighbourly units. The corresponding relative difference δ^{term} takes on its minimum value on corpora free of lexical repetitions (within the range specified by the exchange matrix), and its maximum value 1 on corpora made of disconnected parts, each of which repeats a same term.

Term-document matrices and their generalisations. The above formalism permits, among other configurations, to characterise the traditional textual layout, consisting of weakly interacting documents avoiding nearby term repetitions, which favours a high δ^{doc} (lowered by the presence of references and hyperlinks) together with a low δ^{term} (raised by term repetitions).

Terms and documents can be directly related by marginalising over textual positions, as in the normalised *term-document matrix* $N := X'\Pi Z$ with components $n_{\alpha g} = \sum_i f_i x_{i\alpha} z_{ig}$, where $n_{\alpha\bullet} = \sum_i f_i x_{i\alpha} =: p_\alpha$ is the relative abundance of term α, $n_{\bullet g} = \rho_g$, and $n_{\bullet\bullet} = 1$.

A hierarchy of squared Euclidean distances between terms can be defined from the term-document matrix (Bavaud and Xanthos 2005); their first members are

$$D^0_{\alpha\beta} = (\frac{1}{p_\alpha} + \frac{1}{p_\beta})\, 1(\alpha \neq \beta) \qquad\qquad D^\chi_{\alpha\beta} = \sum_g \frac{1}{\rho_g}(\frac{n_{\alpha g}}{p_\alpha} - \frac{n_{\beta g}}{p_\beta})^2 \ . \tag{4}$$

Those distances define distances between positions as $D_{ij} := D_{\alpha(i)\beta(j)}$, where $\alpha(i)$ denotes the term occurring at position i. Also, the above discrete term dissimilarity D^{term} follows from $D_{\alpha\beta} = 1(\alpha \neq \beta)$; similar considerations apply to documents.

LSA, FCA, VSM and PageRank. Let us briefly make explicit a few well-known quantities expressible within the above formalism. First, *Latent Semantic Analysis* results from the singular decomposition of $n_{\alpha g}$, or weighted versions of it, taking terms or documents frequencies into account; the canonical case, *Factorial Correspondence Analysis* results from the singular decomposition of $n_{\alpha g}/\sqrt{p_\alpha \rho_g}$, or, equivalently, from the weighted classical MDS applied on D^χ in (4).

In the *Vector Space Model*, the *cosine similarity* $(\sum_{\alpha\in Q} n_{\alpha g})/(|Q|\sqrt{\sum_\alpha n_{\alpha g}^2})$ constitutes the standard (unweighted) measure of similarity between a document g and a query Q (defined as a set of terms). The latter numerator is a measure of *relevance* of document g to the query Q; the variant $\rho_g 1(\sum_{\alpha\in Q} n_{\alpha g} \geq 0)$ constitutes the basic relevance measure used in Google's *PageRank* algorithm (Langville and Meyer 2006).

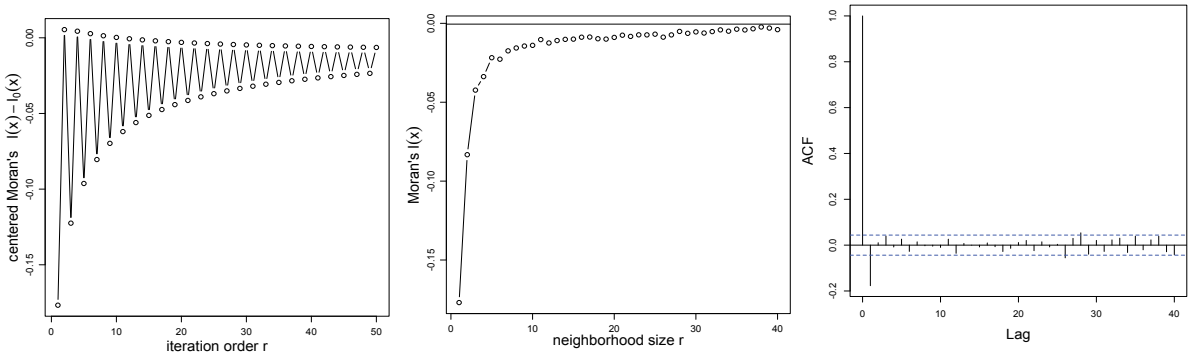

Figure 1: *word lengths alternation. Left: centered Moran's* $I^{(r)}(x) - I_0^{(r)}(x)$ *for* r*-iterated "jump" neighbourhoods, where* $I_0^{(r)}(x)$ *is the expected Moran index in absence of autocorrelation. Both* $I^{(r)}(x)$ *and* $I_0^{(r)}(x)$ *are zigzagging functions of* r*, and so is their difference, although with a smaller amplitude. Middle: Moran's* $I^{[r]}(x)$ *versus "window" neighbourhoods size* r*, together with the line depicting the expected value* $I^{[r]}(x) = -0.0005$*. Right: autocorrelation function* $\texttt{ACF}(r)$ *with confidence intervals, as produced by R.*

3. Case studies

As stated above, the following analyses are restricted to single documents (no hyper-links nor references). Permutation tests have been performed, with the expected result that large enough Moran indices and relative differences are significant at the conventional $\alpha = 5\%$ or $\alpha = 1\%$ levels whenever the text size n is large enough.

Two kinds of exchange matrices are considered, namely the r*-iterated neighbourhoods* $E^{(r)}$ and the r*-sized neighbourhoods* $E^{[r]}$, with $r \geq 1$. The former results from the iteration of the basic *bilateral jump* exchange matrix (section 2), corrected at the extremities, namely

$$e_{ij}^{(1)} = e_{ij} := \frac{1(j = i \pm 1) + 1(i = j = 1) + 1(i = j = n)}{2n}$$

producing uniform weights $f_i = 1/n$, and whose expected values under absence of autocorrelation $I_0^{(r)} = (\text{trace}(W^r) - 1)/(n-1)$ exhibit a complex dependence on r.

By contrast, r-sized neighbourhoods constitute usual *bilateral windows*, obtained by normalising the relation "to be at maximum distance r apart" as

$$e_{ij}^{[r]} := \frac{c_{ij}^{[r]}}{c_{\bullet\bullet}^{[r]}} \qquad\qquad c_{ij}^{[r]} := 1(|j-i| \leq r) \cdot 1(i \neq j)$$

whose weights $f_i^{[r]} := e_{i\bullet}^{[r]}$ are smaller at the document extremities than in the bulk, with constant expected values $I_0^{[r]} = -1/(n-1)$ under absence of autocorrelation.

3.1. Word lengths alternation

In a typical text, the interlacing between functional and content words produces a tendency to alternation between the lengths of consecutive words. That is, one expects the quantity x_i =

"length of the word occurring at the i-th position" to exhibit a negative short-range autocorrelation. The fact is confirmed by the two first plots in Figure 1, based upon the 2'000 first tokens of the novel *Notre-Dame de Paris* by Victor Hugo, containing a total of 180'610 tokens. The zigzagging appearing in r-iterated neighbourhoods is generated by self-interaction $e_{ii}^{(r)} > 0$ for even values of r - an expected logical artefact, although possibly confusing for the interpretation.

In the time series formalism, the *autocorrelation function* at lag r is defined as $\mathtt{ACF}(r) := \mathrm{corr}(x_t, x_{t+r})$ (Figure 1, right). As a matter of fact, $\mathtt{ACF}(r)$ constitutes another avatar of Moran's $I(x)$ for a particular exchange matrix, namely $e_{ij} = 1(j = i \pm r)/2n$ (up to boundary corrections). Also, $I^{(1)}(x) = I^{[1]}(x) = \mathtt{ACF}(1)$ by construction, with value -0.177 in the present case (see Figure 1).

3.2. Lexical repulsion

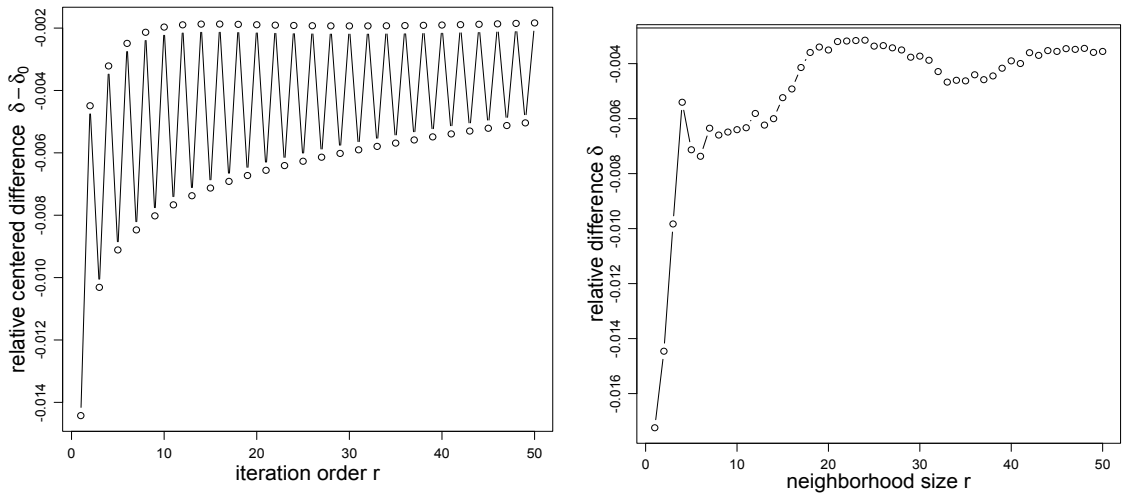

Figure 2: *lexical repulsion. Left: relative centered inertia difference* $\delta^{(r)} - \delta_0^{(r)}$ *for* r*-iterated neighbourhoods, where* $\delta_0^{(r)}$ *is the expected relative inertia difference in absence of autocorrelation. Right: relative inertia difference* $\delta^{[r]}$ *versus neighbourhoods size* r*, together with the line depicting the expected value* $\delta_0^{[r]} = -0.0027$.

Let us consider the spatial autocorrelation associated with the term dissimilarities D_{ij}^{term} on the $n = 371$ words (tokens), $v = 206$ being distinct (types) occurring in the *Atlantic Charter* text, an official statement issued by Britain and the United States in August 1941. The relative inertia differences are depicted in Figure 2. In any case, both plots betray *the avoidance of nearby term repetitions* (i.e. for small r), as expected in the usual documents.

3.3. Parts of speech autocorrelation

The third example considers the French text *Déclaration des droits de l'homme et du citoyen* (1789). Using TreeTagger, words have been classified into four part-of-speech (POS) groups, namely nouns, verbs, adjectives and adverbs, summing to a total of $n = 668$ occurrences - the other categories having been deleted. Four corresponding POS-dissimilarities matrices are considered (see above), for instance D_{ij}^{verb} taking on the value 1 iff the pair (i, j) consists of a

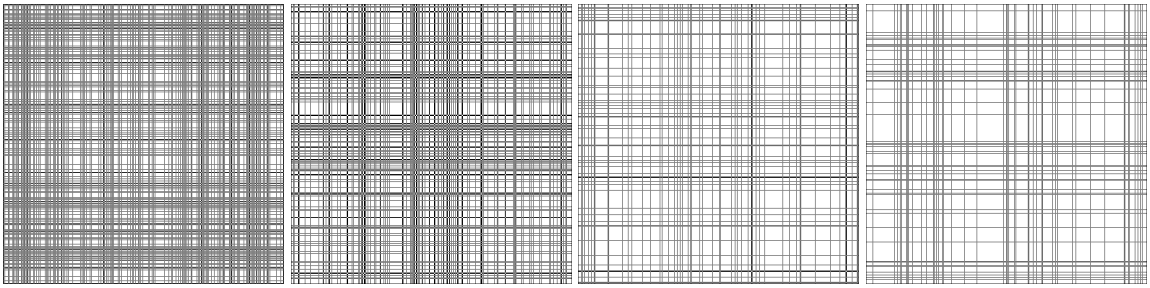

Figure 3: *parts of speech autocorrelation: the four POS dissimilarities as revealed by the four positional* 668×668 *matrices* D^{noun}, D^{verb}, $D^{adjective}$ *and* D^{adverb}, *coded 1=black and 0=white.*

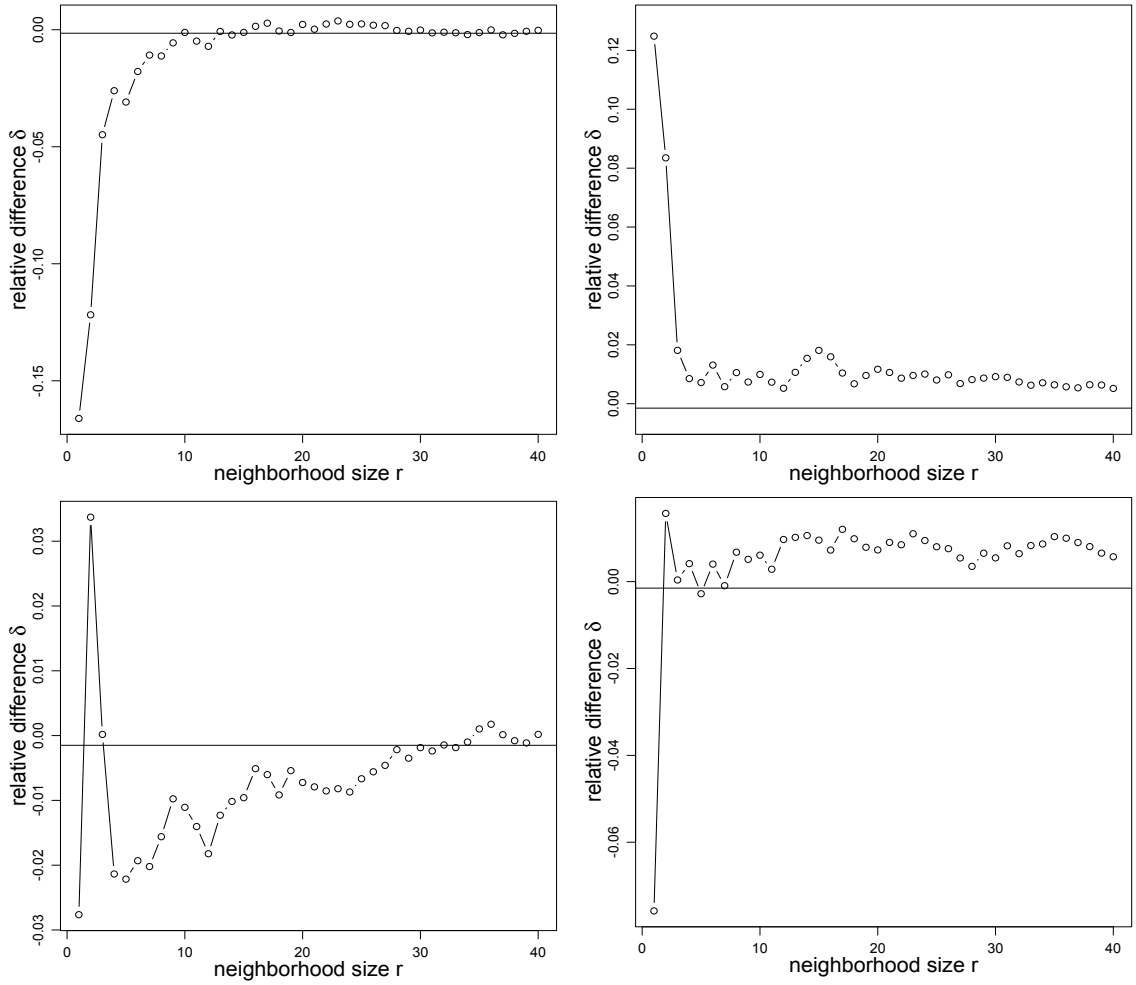

Figure 4: *parts of speech autocorrelation. Relative differences* $\delta^{[r]}$ *for* r*-sized neighbourhoods, for nouns (top left), verbs (top right), adjectives (bottom left) and adverbs (bottom right).*

verb and a non-verb, and 0 otherwise (i.e. verb and verb, or non-verb and non-verb). D^{noun}, $D^{\text{adjective}}$ and D^{adverb} are defined similarly (see figure 3).

The corresponding relative differences, as a function of the neighbourhood size r, are depicted in figure 4, and demonstrate a *short-term repulsion* between nouns together with a *short-term attraction* between verbs (typically, an auxiliary verb followed by a past participle, e.g. *"ont*

été" or *"est jugé"*, or a modal verb followed by an infinitive, e.g. *"peut être"* or *"doit obéir"*). Also, adjectives tend to repel each other in the mid-range, with the notable exception of neighbours at distance two (most examples consist of pairs of adjectives linked by a conjunction, e.g. *"libres et égaux"*, *"naturels et imprescriptibles"*, etc.). Finally, the presence of two consecutive adverbs is very unlikely (compare also with figure 3).

3.4. Semantic autocorrelation

Semantic dissimilarities.

In ontologies such as Wordnet, let $c_1 \leq c_2$ denote the *hyponymy* relation *"concept c_1 is an instance of concept c_2"*, and let $c_1 \vee c_2$ denote the *least general concept* subsuming both c_1 and c_2. For instance, `bicycle` $\leq$ `vehicle` and `bicycle` $\vee$ `car=` `vehicle` for nouns, and `(to listen)` $\leq$ `(to perceive)` and `(to listen)` $\vee$ `(to view)= (to perceive)` for verbs.

The probability $p(c)$ of a concept can be estimated as the relative number of words $n(w)$ (in some reference corpus, such as the Brown corpus) whose sense $C(w)$ is an instance of concept c, that is

$$p(c) \triangleq \frac{\sum_w n(w)\, 1(C(w) \leq c)}{\sum_w n(w)}$$

Resnik (1995) proposes a measure of similarity between concepts c_1 and c_2 as

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

By construction, $s(c_1, c_2) \geq \min\{s(c_1, c_3), s(c_2, c_3)\}$ for any triple of concepts. Also, $p(c_1 \vee c_2) \geq p(c_1)$ and $p(c_1 \vee c_2) \geq p(c_2)$, thus making the *dissimilarity*

$$D(c_1, c_2) := s(c_1, c_1) + s(c_2, c_2) - 2s(c_1, c_2) = \log \frac{p^2(c_1 \vee c_2)}{p(c_1)\, p(c_2)} \tag{5}$$

non negative.

If $c_1 \leq c_2$, define the *length* of the edge joining c_1 and c_2 as $\log(p(c_2)/p(c_1))$. In the general case when $c_1 \not\leq c_2$ and $c_2 \not\leq c_1$, define the *edge length* between c_1 and c_2 as the edge length from c_1 to $c_1 \vee c_2$ added to the length from c_2 to $c_1 \vee c_2$, with the result

$$\log \frac{p(c_1 \vee c_2)}{p(c_1)} + \log \frac{p(c_1 \vee c_2)}{p(c_2)} = D(c_1, c_2)$$

which demonstrates that $D(c_1, c_2)$ is a *tree dissimilarity*, and hence a *squared Euclidean distance* (e.g. Bavaud 2010), from which semantic coordinates can therefore be extracted by classical multidimensional scaling (MDS).

WordNet::Similarity: this Perl module, built by Pedersen et al. (2004), aims at extracting the similarities $s(c_i, c_j)$ defined above (option: "resnik"), from which the squared Euclidean dissimilarities (5) between nouns or verbs are computed. Similarities were extracted using the "first sense" of each concept, that is the most frequent.

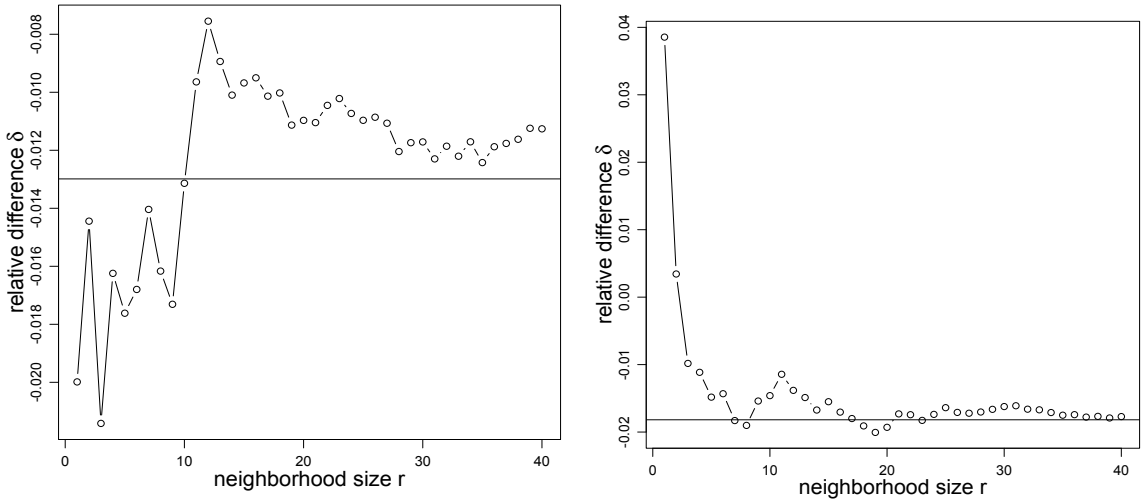

Figure 5: *semantic autocorrelation. Relative inertia difference as a function of the neighbourhod size* r*, for the* $n = 79$ *nouns (left) and the* $n = 56$ *verbs (right), together with the expected line* $\delta_0 = -1/(n-1)$.

Numerical investigations are based upon the *Atlantic Charter* text mentioned above, comprising $n = 79$ noun tokens and $n = 56$ verb tokens. The relative inertia differences are depicted in Figure 5. Clearly, semantically close nouns tend to *repel* each other at short range, while semantically close verbs tend to *attract* each other - a presumably original observation, calling for further empirical investigation on alternative corpora, as well for a linguistic interpretation.

Figure 6 exibits the MDS scree graphs. The corresponding first factorial coordinates $\{x_{i\alpha}\}$ are depicted in figures 7 (nouns) and 8 (verbs). Note the possibility (not pursued further) to analyse textual semantic autocorrelation separately in each dimension α by inspecting $I(x_\alpha)$.

Three groups of nouns appear in Figure 7, left: nouns in the north-east region (*"government", "country", "people", "nation", "labour" and "tyranny"*) are subsumed (in WordNet) by the concept of "group" or "grouping". Nouns in the north-west region l are subsumed by the concept of "attribute", defined in WordNet as "an abstraction belonging to or characteristic of an entity". The third group in the south-west region contains the other nouns, whose common hypernym is the most general concept of "entity" - the root in the nouns hierarchy.

Three groups of verbs appear as well in Figure 8, left: verbs in the south-east region (*"seek", "desire", "wish", "hope", and "want"*) belong to the concept of "want", "desire". Verbs in the south-west region (*"deem", "respect", "believe" and "lighten"*) are subsumed by the concept "think", "cogitate", "cerebrate". The group in the north-west region is semantically heterogeneous, and contains the other verbs without specific common hypernyms.

4. Conclusion

In this paper, we have proposed a unified formalism for assessing and testing textual autocorrelation, namely the tendency for values of a textual variable to be more similar (or more different) between neighbouring units than between units at arbitrary positions in the text. In this framework, based upon textual position, the topology of a text (or set of texts) is conve-

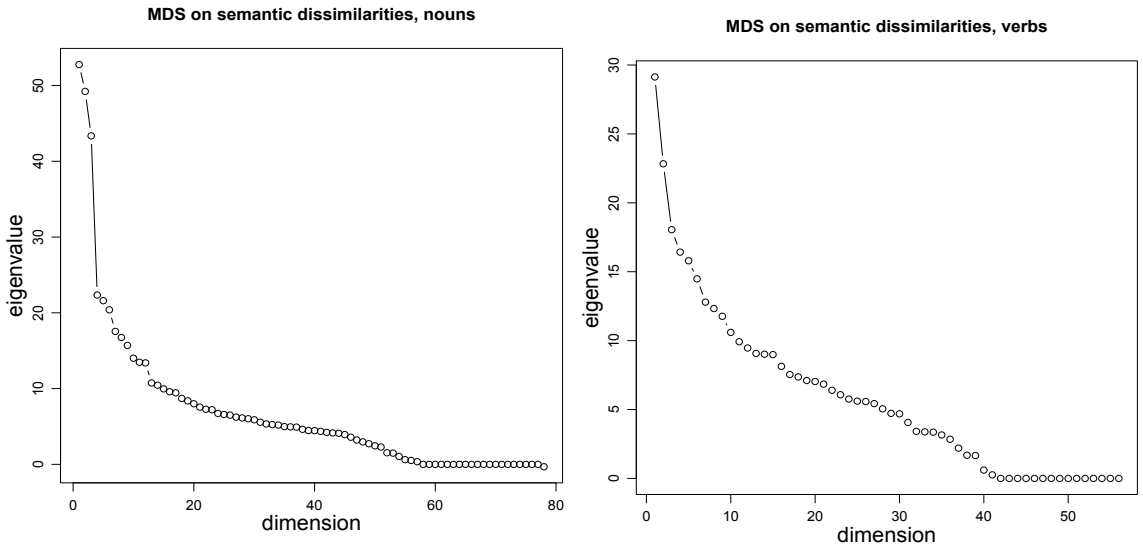

Figure 6: *semantic autocorrelation. Scree graphs of the MDS associated to the nouns (left) and verbs (right).*

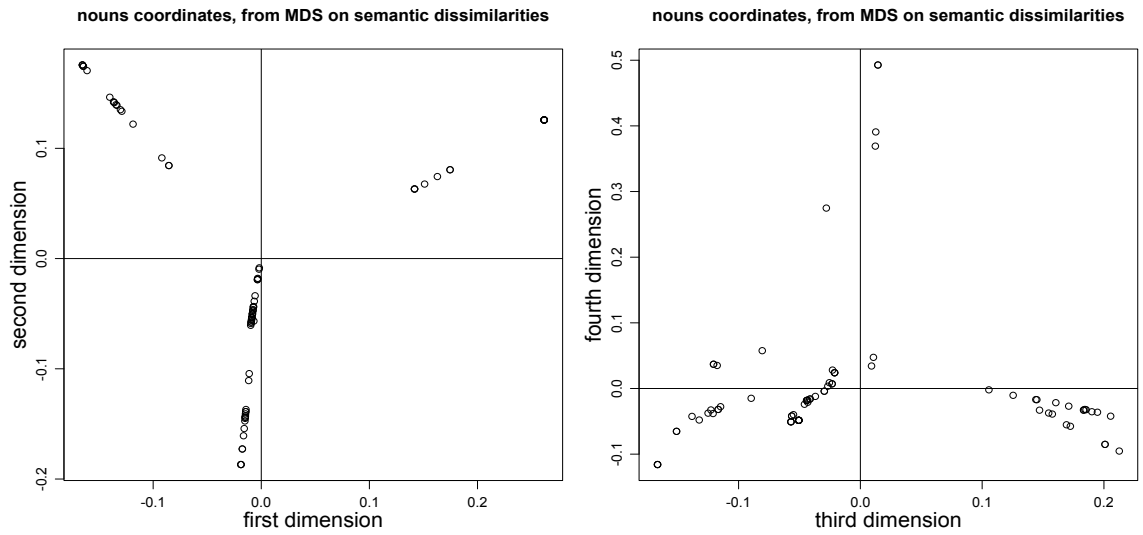

Figure 7: *semantic autocorrelation. Nouns coordinates extracted from MDS, first dimensions.*

niently represented by means of an *exchange matrix*, which makes it possible to formalize a wide variety of textual navigation scenarios, ranging from traditional linear reading to hypertextual hopping across documents. Formal relation with well-known constructs in textual data analysis such as term-document matrices, latent semantic analysis and correspondence analysis, or (web) information retrieval was also briefly touched upon.

This approach has been illustrated with examples pertaining to the lexical, morpho-syntactic and semantic structure of language. In particular, a short-range repulsion for nouns together with a short-range attraction for verbs, both at the lexical and semantic levels, has been evidenced. This effect is yet to be confirmed on larger corpora, as well as to be given a proper linguistic interpretation.

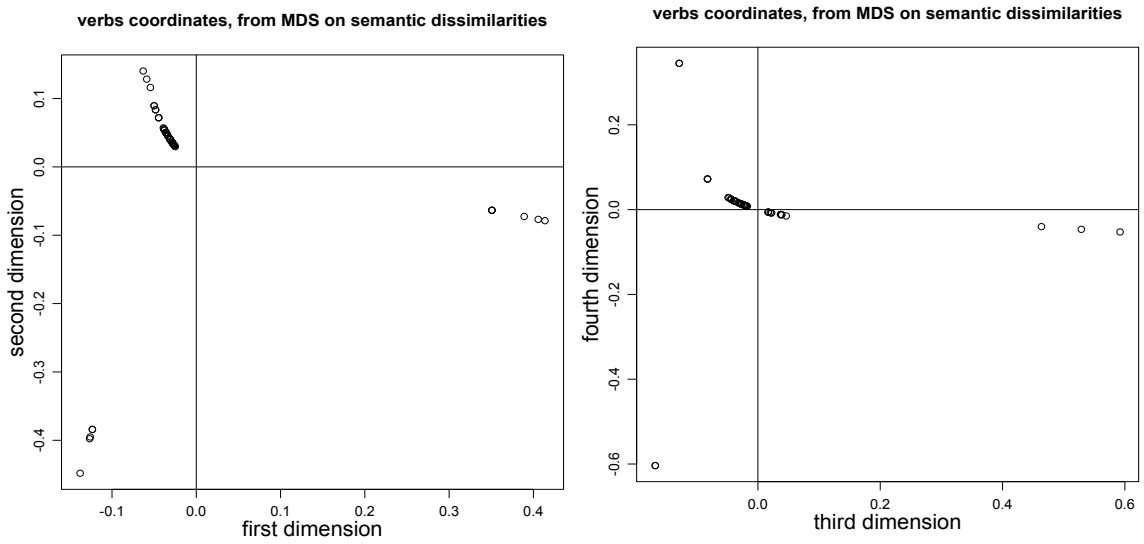

Figure 8: *semantic autocorrelation. Verbs coordinates extracted from MDS, first dimensions.*

It is the authors' belief that the proposed formalism is suitable for going beyond the traditional representation of documents and corpora in quantitative approaches of text data, and in particular for accounting for the various types of links that may occur within or between documents. This intuition remains to be further tested with data where such linking plays a significant role.

References

Bavaud, F., Xanthos, A. (2005). Markov associativities. *Journal of Quantitative Linguistics* 12, pp. 123–137

Bavaud, F. (2008). Local concentrations. *Papers in Regional Science* 87, pp. 357–370

Bavaud, F. (2010). Euclidean Distances, Soft and Spectral Clustering on Weighted Graphs. *Proceedings of the ECML PKDD'10*. Lecture Notes in Computer Science 6321, pp. 103–118. Springer.

Berger, J., Snell, J.L. (1957). On the concept of equal exchange. *Behavioral Science* 2, pp. 111–118

Cressie, N. (1991). *Statistics for Spatial Data*. Wiley

Langville, A.N. and Meyer, C.D. (2006). *Google Page Rank and Beyond*. Princeton University Press

Lebart, L. (1969). Analyse statistique de la contigüité. *Publication de l'Institut de Statistiques de l'Université de Paris* 18, pp. 81–112

Moran, P.A.P. (1950). Notes on continuous stochastic phenomena. *Biometrika* 37, pp. 17–23

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). WordNet::Similarity - Measuring the Relatedness of Concepts. *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics* pp. 38–41

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *International Joint Conference for Artificial Intelligence (IJCAI-95)*, pp. 448–453