Taking into account semantic similarities in correspondence analysis

Mattia Egloff
Department of Language and Information Sciences
University of Lausanne
megloff1@unil.ch

François Bavaud
Department of Language and Information Sciences
Institute of Geography and Sustainability
University of Lausanne
fbavaud@unil.ch

Abstract

Term-document matrices feed most distributional approaches to quantitative textual studies, without consideration for the semantic similarities between terms, whose presence arguably reduce the content variety. This contribution presents a formalism remedying this omission, and makes an explicit use of the semantic similarities as extracted from WordNet. A case study in similarity-reduced correspondence analysis illustrates the proposal.

Introduction The term-document matrix $N = (n_{ik})$ counts the occurrences of n terms in p documents, and constitutes the privileged input of most distributional studies in quantitative textual linguistics: chi2 dissimilarities between terms or documents, distance-based clustering of terms or documents, *multidimensional scaling* (MDS) on terms or documents; and, also, *latent clustering* by *non-negative matrix factorization* (e.g. Lee and Seung, 1999) or *topic modeling* (e.g. Blei, 2012); as well as nonlinear variants resulting from transformations of the independence quotients, as in the Hellinger dissimilarities, or transformations of the chi2 dissimilarities themselves (e.g. Bavaud, 2011).

When using the term-document matrix, the semantic link between words is only indirectly addressed through the celebrated "distributional hypothesis", postulating an association between distributional similarity and meaning similarity (Harris, 1954) (see also e.g. Sahlgren, 2008; McGillivray et al., 2008). Largely accepted and much documented at it is, the distributional hypothesis seems hardly tackled in an explicit way, for lack of formal measure of semantic similarity, precisely. By contrast, the present study distinguishes both kind of similarities. It also yields a new measure of textual variety taking explicitly into account the semantic similarities between terms.

Data After manually extracting the paragraphs of each of the $p = 11$ chapters of Book I of "An Inquiry into the Nature and Causes of the Wealth of Nations" by Adam Smith (Smith, 1776), we tagged the parts of speech and lemma for each word of the corpus using the nlp4j tagger (Choi, 2016). Subsequently we created a lemma-chapter matrix, retaining only the type of words serving a specific task, such as verbs. Terms i, j present in the chapters were then associated to their *first conceptual senses* c_i, c_j, that is to their first WordNet synsets (Miller, 1995). We inspected several similarity matrices $\hat{s}_{ij} = \hat{s}(c_i, c_j)$ between pairs of concepts c_i and c_j.

Semantic similarities The classical similarities $\hat{s}(c_i, c_j)$ between two concepts c_i and c_j computed on WordNet take on different forms. The conceptually easiest is the path similarity, defined from the number $\ell(c_i, c_j) \geq 0$ of edges of the shortest-path (in the WordNet hierarchy) between c_i and c_j as follows:

$$\hat{s}^{\text{path}}(c_i, c_j) = \frac{1}{1 + \ell(c_i, c_j)} \qquad (1)$$

The Leacock Chodorow similarity (Leacock and Chodorow, 1998) is based on the same principle but considers also the maximum depth $D = \max_i \ell(c_i, 0)$ (where 0 represents the *root* of the hierarchy, occuped by the concept subsuming all

Egloff, M. and Bavaud, F. *Taking into account semantic similarities in correspondence analysis*. In: Piotrowski, M. (Ed.) *Proceedings of the Workshop on Computational Methods in the Humanities 2018* (COMHUM 2018) pp. 45–51. CEUR Workshop Proceedings http://ceur-ws.org/Vol-2314/ urn:nbn:de:0074-2314-0

the others) of the concepts in the WordNet taxonomy:

$$\hat{\mathbb{s}}^{\text{lch}}(c_i, c_j) = -\log \frac{\ell(c_i, c_j)}{2D}$$

The Wu-Palmer similarity (Wu and Palmer, 1994) is based on the notion of *lowest common subsumer* $c_i \vee c_j$, that is the *least general concept* in the hierarchy that is a hypernym or ancestor of both c_i and c_j:

$$\hat{\mathbb{s}}^{\text{wup}}(c_i, c_j) = \frac{2\ell(c_i \vee c_j, 0)}{\ell(c_i, 0) + \ell(c_j, 0)}$$

The following similarities are further based on the concept of Information Content, proposed by Resnik (Resnik, 1993a,b). The Information Content of a concept c is defined as $-\log(p(c))$, where $p(c)$ is the probability to encounter a concept c in a reference corpus. The Resnik similarity (Resnik, 1995) is defined as:

$$\hat{\mathbb{s}}^{\text{res}}(c_i, c_j) = -\log p(c_i \vee c_j)$$

The Lin similarity (Lin et al., 1998) is defined as:

$$\hat{\mathbb{s}}^{\text{lin}}(c_i, c_j) = \frac{2 \cdot \log p(c_i \vee c_j)}{\log p(c_i) + \log p(c_j)}$$

Finally, the Jiang Coranth similarity (Jiang and Conrath, 1997) is defined as:

$$\hat{\mathbb{s}}^{\text{jch}}(c_i, c_j) = \frac{1}{-\log p(c_i) - \log p(c_j) + 2 \cdot \log p(c_i \vee c_j)}$$

and obeys $\hat{\mathbb{s}}^{\text{jch}}(c_i, c_i) = \infty$.

Among the above similarities, the path, Wu-Palmer and Lin similarities obey the conditions

$$\hat{\mathbb{s}}_{ij} = \hat{\mathbb{s}}_{ji} \geq 0 \qquad \text{and} \qquad \hat{\mathbb{s}}_{ii} = 1 \ . \qquad (2)$$

In what follows, we shall use the path similarities when required.

A similarity-reduced measure of textual variety Let $f_i \geq 0$ be the relative frequency of term i, normalized to $\sum_{i=1}^{n} f_i$. *Shannon entropy* $H = -\sum_i f_i \ln f_i$ constitutes a measure of relative textual variety, ranging from 0 (a single term repeats itself) to $\ln n$ (all terms are different). Yet, the entropy does not take into account the possible similarity between the terms, in contrast to the *reduced entropy* R (our nomenclature) defined as

$$R = -\sum_{i=1}^{n} f_i \ln b_i \quad \text{where} \quad b_i = \sum_{j=1}^{n} \hat{\mathbb{s}}_{ij} f_j \ . \qquad (3)$$

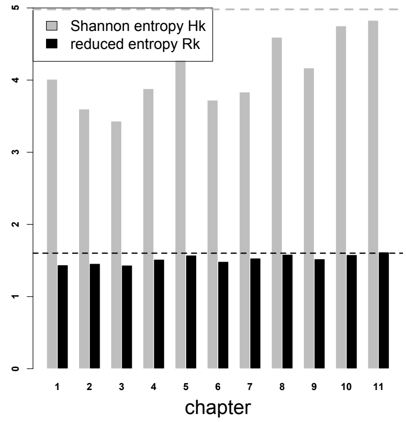

Figure 1: Entropies H_k and reduced entropies R_k for each chapter k; dashed lines depict H and R.

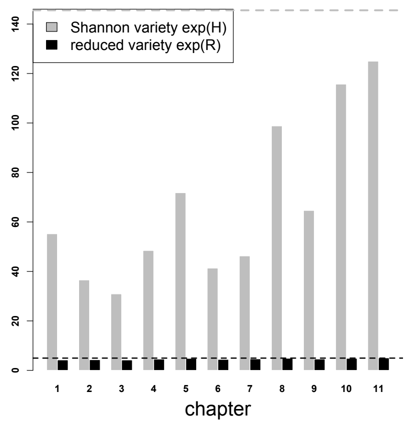

Figure 2: Shannon varieties $\exp(H_k)$ and reduced varieties $\exp(R_k)$ for each chapter k; dashed lines depict $\exp(H)$ and $\exp(R)$.

In Ecology, b_i is the *banality* of species i, measuring its average similarity to other species (Marcon, 2016), proposed by Leinster and Cobbold (2012), as well as by Ricotta and Szeidl (2006). By construction, $f_i \leq b_i \leq 1$ and thus $R \leq H$: the larger the similarities, the lower the textual variety as measured by the reduced entropy, as requested.

Returning to the case study, we have, out of the 643 verb lemmas initially present in the corpus, retained the $n = 234$ verb lemmas occurring at least 5 times ("be" and "have" excluded). Overall term weights f_i, chapter weights ρ_k and term weights f_i^k within a chapter obtain from the $n \times p = 234 \times 11$ term-document matrix $N = (n_{ik})$ as

$$f_i = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \qquad \rho_k = \frac{n_{\bullet k}}{n_{\bullet\bullet}} \qquad f_i^k = \frac{n_{ik}}{n_{\bullet k}} \qquad (4)$$

The corresponding entropies and reduced entropies read $H = 4.98 > R = 1.60$. For each

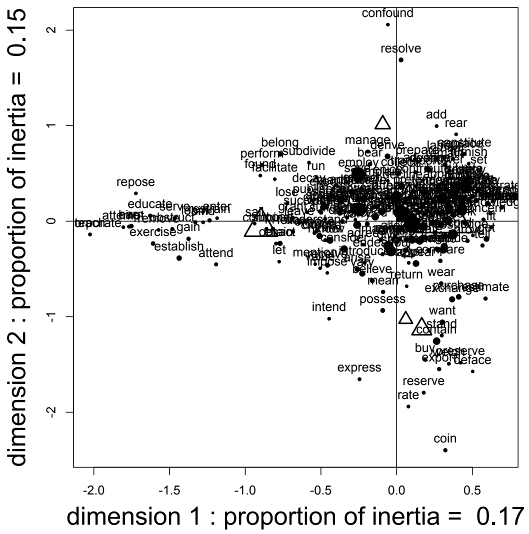

Figure 3: Biplot of the 234×11 term-document matrix. Circles depict terms and triangles depict documents.

chapter, the corresponding quantities are depicted in figure 1. One can observe the so-called *concavity property* $H > \sum_k \rho_k H_k$ and $R > \sum_k \rho_k R_k$, which says that the variety of the whole is larger than the average variety of its constituents.

Shannon variety $N_{\text{Shannon}} = \exp(H) \leq n$ represents the equivalent number of distinct types in a uniformly constituted corpus of same richness or diversity (in the entropy sense) as the currently examined corpus. Likewise, the *reduced variety* $N_{\text{reduced}} = \exp(R) \ll N_{\text{Shannon}}$ measures the equivalent number of types if the latter were uniformly distributed and *completely dissimilar* (that is $s_{ij} = 0$ for $i \neq j$): see figure 2.

Ordinary correspondence analysis (recall) *Correspondence analysis* (CA) permits a simultaneous representation of terms and documents in the so-called *biplot* (figure 3). CA results from weighted multidimensional scaling (MDS) applied to the *chi2 dissimilarities* D^{χ}_{kl} between documents k and l

$$D^{\chi}_{kl} = \sum_{i=1}^{n} f_i (q_{ik} - q_{il})^2 \quad \text{where} \quad q_{ik} = \frac{n_{ik} n_{\bullet\bullet}}{n_{i\bullet} n_{\bullet k}} \tag{5}$$

or equivalently, on MDS applied to the chi2 dissimilarities between terms. Note the q_{ik} in (5) to constitute the *independence quotients*, that is the ratio of the observed counts to their expected value under independence. Figure 3 constitutes the two-dimensional projection of a weighted Euclidean configuration of $\min(234-1, 11-1) = 10$ dimensions, expressing a maximal proportion of $0.17 + 0.15 = 32\%$ of dispersion or *inertia* $\Delta = \frac{1}{2}\sum_{kl} \rho_k \rho_l D^{\chi}_{kl}$.

Similarity-reduced correspondence analysis In the case where documents k and l, differing by the presence of *distinct terms*, contain *semantically similar* terms, the "naive" chi2 dissimilarity (5), which implicitly assumes distinct terms to be completely dissimilar, arguably overestimates their difference. The latter should be downsized accordingly, in a way both reflecting the amount of shared similarity between k and l, and still retaining the squared Euclidean nature of their dissimilarity – a crucial requirement for the validity of MDS. This simple idea leads us to propose the following *reduced* squared Euclidean distance $\hat{D}_{kl}$ between documents, taking into account both the *distributional* and *semantic* differences between the documents, namely

$$\tilde{D}_{kl} = \sum_{ij} \tilde{\mathbb{t}}_{ij} (q_{ik} - q_{il})(q_{jk} - q_{jl}) \tag{6}$$

where $(q_{ik} - q_{il})(q_{jk} - q_{jl})$ captures the distributional contribution, and

$$\tilde{\mathbb{t}}_{ij} = \frac{f_i f_j \hat{\mathbb{s}}_{ij}}{\sqrt{b_i b_j}} \qquad \text{where} \quad b = \hat{\mathbb{S}} f \quad \text{is the banality}$$

captures the semantic contribution. Matrix $\tilde{\mathbb{T}} = (\tilde{\mathbb{t}}_{ij})$ has been designed so that

- $\tilde{\mathbb{T}} = \mathrm{diag}(f)$ for "naive" similarities $\hat{\mathbb{S}} = I$ (where I is the identity matrix), in which case $\tilde{D}$ is the usual chi2-dissimilarity
- $\tilde{\mathbb{T}} = ff'$ for "confounded types" $\hat{\mathbb{S}} = J$ (where J is the unit matrix filled with ones), in which case $\tilde{D}$ is identically zero.

Also, one can prove $\tilde{D}$ in (6) to be a squared Euclidean dissimilarity iff $\mathbb{S}$ is positive semi-definite, that is iff all its eigenvalues are non-negative, a verified condition for path dissimilarities (see the Appendix). Figure 4 depicts the corresponding MDS.

Semantic MDS on terms Positive semi-definite semantic similarities $\hat{\mathbb{S}}$ of the form (2), such as the path similarities, generate squared Euclidean dissimilarities as

$$\hat{\mathbb{d}}_{ij} = 1 - \hat{\mathbb{s}}_{ij} \tag{7}$$

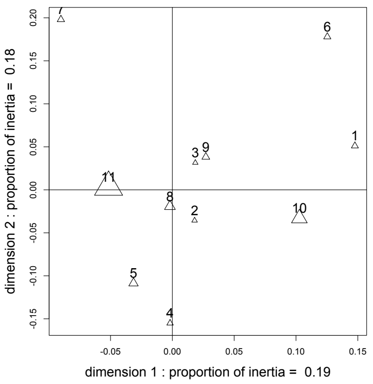

Figure 4: Weighted MDS of the document reduced dissimilarities $\tilde{D}$ (6), displaying the optimal two-dimensional projection of the reduced inertia $\tilde{\Delta} = \frac{1}{2}\sum_{kl} \rho_k \rho_l \tilde{D}_{kl} = 0.025$, which is roughly 50 times smaller than the ordinary inertia $\Delta = \frac{1}{2}\sum_{kl} \rho_k \rho_l D^{\chi}_{kl} = 1.156$ of usual CA (figure 3).

(see the Appendix), and this circumstance allows a weighted MDS on semantic dissimilarities between terms, aimed at depicting an optimal low-dimensional representation of the semantic inertia

$$\hat{\mathbb{\Delta}} = \frac{1}{2}\sum_{ij} f_i f_j \hat{\mathbb{d}}_{ij} \quad , \tag{8}$$

irrespectively of the distributional term-document structure (figures 5 and 6).

A family of similarities interpolating between totally distinct types and confounded types The exact form of similarities $\hat{\mathbb{S}}$ between terms fully governs the similarity-reduction mechanism investigated so far. Yet, little systematic investigation seems to have been devoted to the formal properties of similarities (by contrast to the study of the dissimilarities families found e.g. in Critchley and Fichet (1994) or Deza and Laurent (2009), which may obey much more specific properties than (2). In particular, $\hat{\mathbb{s}}^{\alpha}_{ij}$ satisfies (2) for $\alpha \geq 0$ if $\hat{\mathbb{s}}_{ij}$ does, and varying α permits to interpolate between the extreme cases of "naive" similarities $\hat{\mathbb{S}} = I$ and "confounded types" $\hat{\mathbb{S}} = J$.

Lists of synonyms[1] yield binary similarity matrices $\mathbb{s}_{ij} = 0$ or 1. More generally, $\mathbb{S}$ can be defined as a convex combination of binary synonymy relations, insuring its non-negativity, symmetry, positive definiteness, with $\mathbb{s}_{ii} = 1$ for all terms i. A family of such semantic similarities indexed by the *bandwidth parameter* $\beta > 0$ obtains as

$$\mathbb{s}_{ij} = \exp(-\beta\, \hat{\mathbb{d}}_{ij}/\hat{\mathbb{\Delta}}) \tag{9}$$

[1] e.g. http://www.crisco.unicaen.fr/des/

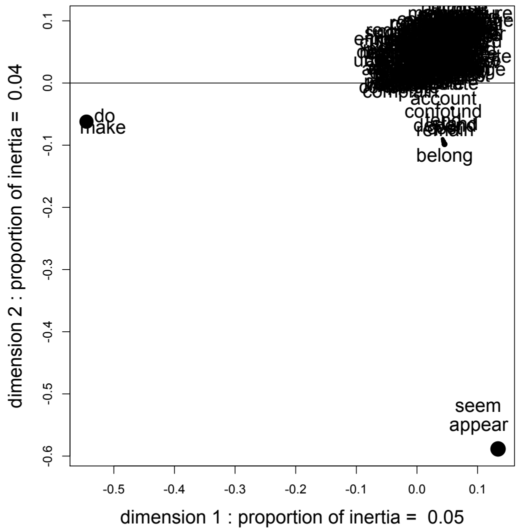

Figure 5: Weighted MDS on the term semantic dissimilarities (7) for the 234 retained verbs. The first dimension opposes `do` and `make` (whose similarity is 1) to the other verbs. The second dimension opposes `appear` and `seem` (with similarity 1) to the other verbs.

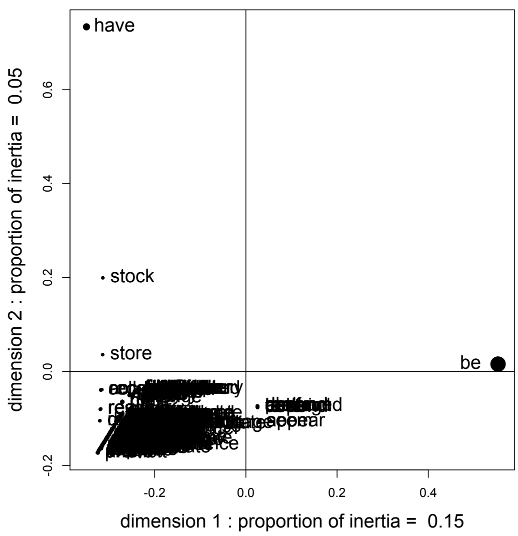

Figure 6: Weighted MDS on the term semantic dissimilarities (7) for the 643 verbs initially present in the corpus, emphasizing the particular position of `be` and `have`

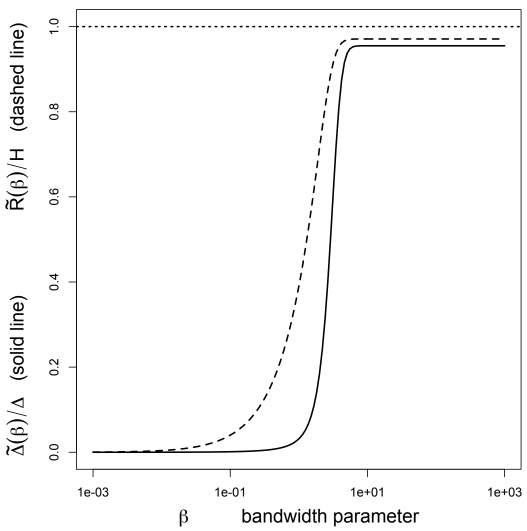

Figure 7: The larger the bandwidth parameter β, the less similar are the terms, and hence the greater are the reduced inertia $\tilde{\Delta}(\beta)$ as well as the reduced entropy $\tilde{R}(\beta)$ (3)

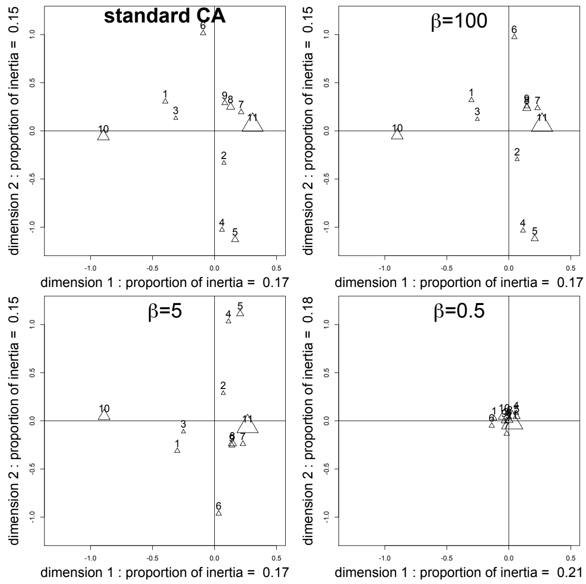

Figure 8: In the limit $\beta \to 0$, both diagonal *and off-diagonal* similarities $\mathbb{s}_{ij}(\beta)$ tend to one, making all terms semantically identical, thus provoking the collapse of the cloud of document coordinates.

where $\hat{\mathbb{d}}_{ij}$ is the semantic dissimilarity (7) and $\hat{\mathbb{\Delta}}$ the associated semantic inertia (8).

As a matter of fact, it can be shown that a binary $\mathbb{S}$ makes the similarity-reduced document dissimilarity $\tilde{D}_{kl}$ (6) identical to the chi2 dissimilarity (5), with the exception that the sum now runs on *cliques of synonyms* rather than terms. Also, the limit $\beta \to 0$ in (9) makes $\tilde{D}_{kl} \to 0$ with a reduced inertia $\tilde{\Delta}(\beta) = \frac{1}{2}\sum_{kl} \rho_k \rho_l \tilde{D}_{kl}$ tending to zero. In the opposite direction, $\beta \to \infty$ makes $\tilde{D}_{kl} \to D^{\chi}_{kl}$ provided $\hat{\mathbb{d}}_{ij} > 0$ for $i \neq j$, a circumstance violated in the case study, where the $n = 234$ verbs display, accordingly to their first sense in WordNet, 15 cliques of size 2 (among which `do-make` and `appear-seem`, already encountered in figure 5) and 3 cliques of size 3 (namely, `employ-apply-use`, `set-lay-put` and `supply-furnish-provide`). In any case, the *relative reduced inertia* $\tilde{\Delta}(\beta)/\Delta$ is increasing in β (figure 7).

Performing the similarity-reduced correspondence analysis on the reduced dissimilarities (6) between the 11 document, with similarity matrices $\mathbb{S}(\beta)$ (instead of $\hat{\mathbb{S}}$ as in figure 4) demonstrates the *collapse*of the cloud of document coordinates (figure 8). As a matter of fact, the bandwidth parameter β controls the *paradigmatic sensitivity* of the linguistic subject: the larger β, the larger the semantic distances between the documents, and the larger the spread of the factorial cloud as measured by reduced inertia $\tilde{\Delta}(\beta)$ (figure 7). On the other direction, a low β can model an illiterate person, sadly unable to discriminate between documents, which look all alike.

Conclusion and further issues Despite the technicality of its exposition, the idea of this contribution is straightforward, namely to propose a way to take semantic similarity explicitly into account, within the classical distributional similarity framework provided by correspondence analysis. Alternative approaches and variants are obvious: further analysis on non-verbs should be investigated; other definitions of $\tilde{D}$ are worth investigating; other choices of $\mathbb{S}$ are possible (in particular the original $\hat{\mathbb{S}}$ extracted form Wordnet). Also, alternatives to WordNet path similarities (e.g., for languages in which WordNet is not defined) are required.

On the document side, and despite its numerous achievements, the term-document matrix still relies on a rudimentary approach to textual context, modelled as p documents consisting of *bag of words*. Much finer *syntagmatic* descriptions are possible, captured by the general concept of *exchange matrix* E, giving the joint probability to select *a pair of textual positions* through textual navigation (by reading, hyperlinks or bibliographic zapping, etc.). E defines a weighted network whose nodes are the textual positions occupied by terms (Bavaud et al., 2015).

The parallel with *spatial issues* (quantitative geography, image analysis), where E defines the *"where"*, and the features dissimilarities between positions $\mathbb{D}$ defines the *"what"*, is immediate (see e.g. Egloff and Ceré, 2017). In all likelihood, developing both axes, that is taking into account semantic similarities on generalized textual networks, could provide a fruitful extension and renewal of the venerable term-document matrix paradigm, and provide a new approach to the distributional hypothesis, which can be reframed as a spatial autocorrelation hypothesis.

References

François Bavaud. 2011. On the Schoenberg transformations in data analysis: Theory and illustrations. *Journal of Classification* 28(3):297–314. https://doi.org/10.1007/s00357-011-9092-x.

François Bavaud, Christelle Cocco, and Aris Xanthos. 2015. Textual navigation and autocorrelation. In G. Mirkros and J. Macutek, editors, *Sequences in Language and Text*. De Gruyter Mouton, pages 35–56.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):77–84.

Jinho D Choi. 2016. Dynamic feature induction: The last gist to the state-of-the-art. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 271–281.

Frank Critchley and Bernard Fichet. 1994. The partial order by inclusion of the principal classes of dissimilarity on a finite set, and some of their basic properties. In *Classification and dissimilarity analysis*, Springer, pages 5–65.

Michel Deza and Monique Laurent. 2009. *Geometry of cuts and metrics*, volume 15. Springer.

Mattia Egloff and Raphaël Ceré. 2017. Soft textual cartography based on topic modeling and clustering of irregular, multivariate marked networks. In *International Workshop on Complex Networks and their Applications*. Springer, pages 731–743.

Zellig S Harris. 1954. Distributional structure. *Word* 10(2-3):146–162.

Jay J Jiang and David W Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008* .

C. Leacock and M. Chodorow. 1998. *Combining local context and WordNet similarity for word sense identification*, In C. Fellbaum (Ed.), MIT Press, pages 305–332.

Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788.

Tom Leinster and Christina A Cobbold. 2012. Measuring diversity: the importance of species similarity. *Ecology* 93(3):477–489.

Dekang Lin et al. 1998. An information-theoretic definition of similarity. In *Icml*. Citeseer, volume 98, pages 296–304.

Eric Marcon. 2016. *Mesurer la Biodiversité et la Structuration Spatiale*. Thèse d'habilitation, Université de Guyane. https://hal-agroparistech.archives-ouvertes.fr/tel-01502970.

Barbara McGillivray, Christer Johansson, and Daniel Apollon. 2008. Semantic structure from correspondence analysis. In *Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing*. Association for Computational Linguistics, pages 49–52.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Philip Resnik. 1993a. Semantic classes and syntactic ambiguity. In *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, pages 278–283.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007* .

Philip Stuart Resnik. 1993b. Selection and information: a class-based approach to lexical relationships. *IRCS Technical Reports Series* page 200.

Carlo Ricotta and Laszlo Szeidl. 2006. Towards a unifying approach to diversity measures: bridging the gap between the shannon entropy and rao's quadratic index. *Theoretical population biology* 70(3):237–243.

Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Disability Studies* 20:33–53.

Adam Smith. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations; Book I*. Project Gutenberg, Urbana, Illinois. Also known as: Wealth of Nations. http://www.gutenberg.org/ebooks/3300.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 133–138.

Appendix: *proof of the squared Euclidean nature of $\mathbb{D}$ in (7).*

The number ℓ_{ij} of edges is the shortest path (in the WordNet hierarchical tree) linking the concepts associated to i and j is a a *tree dissimilarity*[2], and hence a *squared Euclidean dissimilarity* (see e.g. Critchley and Fichet, 1994). Hence, (1) and (7) entail

$$\hat{\mathbb{d}}_{ij} = 1 - \hat{\mathbb{s}}_{ij} = 1 - \frac{1}{1+\ell_{ij}} = \frac{\ell_{ij}}{1+\ell_{ij}}$$

that is $\hat{\mathbb{d}}_{ij} = \varphi(\ell_{ij})$, where $\varphi(x) = x/(1+x)$. The function $\varphi(x)$ is non-negative, increasing, concave, with $\varphi(0) = 0$. For $r \geq 1$, its even derivatives $\varphi^{(2r)}(x)$ are non-positive, and its odd derivatives $\varphi^{(2r-1)}(x)$ are non-negative. That, is, $\varphi(x)$ is a *Schoenberg transformation*, transforming a squared Euclidean dissimilarity into a squared Euclidean dissimilarity (see e.g. Bavaud, 2011), thus establishing the squared Euclidean nature of $\mathbb{D}$ in (7) (and, by related arguments, the p.s.d. nature of $\mathbb{S}$).

[2]provided no terms posses two direct hypernyms, which seems to be verified for the verbs considered here