Using complex networks to quantify consistency in the use of words

D R Amancio, O N Oliveira Jr and L da F Costa

Institute of Physics of São Carlos, University of São Paulo, PO Box 369, Postal Code 13560-970, São Carlos, São Paulo, Brazil
E-mail: diegoraphael@gmail.com, chu@ifsc.usp.br and ldfcosta@gmail.com

Received 28 August 2011
Accepted 7 December 2011
Published 10 January 2012

Online at stacks.iop.org/JSTAT/2012/P01004
doi:10.1088/1742-5468/2012/01/P01004

Abstract. In this paper we have quantified the consistency of word usage in written texts represented by complex networks, where words were taken as nodes, by measuring the degree of preservation of the node neighborhood. Words were considered highly consistent if the authors used them with the same neighborhood. When ranked according to the consistency of use, the words obeyed a log-normal distribution, in contrast to Zipf’s law that applies to the frequency of use. Consistency correlated positively with the familiarity and frequency of use, and negatively with ambiguity and age of acquisition. An inspection of some highly consistent words confirmed that they are used in very limited semantic contexts. A comparison of consistency indices for eight authors indicated that these indices may be employed for author recognition. Indeed, as expected, authors of novels could be distinguished from those who wrote scientific texts. Our analysis demonstrated the suitability of the consistency indices, which can now be applied in other tasks, such as emotion recognition.

Keywords: data mining (experiment), pattern formation (experiment), random graphs, networks, communication, supply and information networks
Using complex networks to quantify consistency in the use of words

Contents

1. Introduction 2
2. Methodology 3
   2.1. Dataset 3
   2.2. Modeling texts as complex networks 3
   2.3. Consistency indices 6
      2.3.1. $C$ based on the histogram of co-occurrence of neighbors 6
      2.3.2. $C$ based on the cosine similarity 7
      2.3.3. $C$ based on the Pearson correlation coefficient 8
      2.3.4. $C$ based on the Leicht–Holme–Newman index [36] 8
      2.3.5. $C$ based on the Sorensen index [37] 9
      2.3.6. $C$ based on the frequency of the shared neighbors 9
3. Results and discussion 10
   3.1. Analysis of distribution and inter-correlation of consistency indices 10
   3.2. Analysis of correlation between consistency and linguistic features 12
   3.3. Using the consistency index to recognize authorship 16
4. Conclusion and final remarks 16

Acknowledgments 18
References 18

1. Introduction

Since the dawn of humanity, the ability to communicate has been proven to be an essential factor for preservation of life and for the maintenance of social relations. Writing, a major manifestation of communication, whose invention dates back to 3200 BC, also established itself as one of the key skills developed by mankind. Among its main advantages over the spoken language are the joint capacity of portability and permanence, which ensure that thoughts, ideas, facts and stories are preserved. Despite this ubiquity, the writing skill cannot be considered a trivial or ordinary task. Even after language acquisition, the construction of a concise, precise and well concatenated text requires organized thought, the ability to use expressive linguistic resources and the analytical interpretation of reality.

In addition to the difficulties imposed by the grammar mechanisms in written language [1], there is a factor related to the semantic level of detail to recreate and interpret the author’s original idea. Since it is not possible to specify all the details in a finite piece of text, the author must always focus on the desired level of generality. Thus, if little detail is provided the reader must fill in the blanks using his/her own semantic knowledge and experience about the world. On the other hand, in an excessively detailed text there is no room for inferences, which makes reading more objective. This dichotomy between objectivity and subjectivity has been explored in various ways in different genres of writing. While scientific texts, newspapers, magazines and reports tend to use a more
objective approach, literary and artistic texts tend to present themselves subjectively. In both cases, the degree of objectivity (or subjectivity) varies depending on both the text length (long texts tend to be more detailed) and number of descriptive words (a text with many adjectives for instance tends to be very detailed). Equally important seems to be the context induced by words [2,3], since words inducing restricted contexts somehow limit the ability to extrapolate ideas, and this makes the text more objective.

With this correlation between induction and objectivity as motivation [4], in this paper we address the problem of quantifying the level of consistency inherent in words, i.e. the degree of preservation of their neighbors, to understand the reasons why a word is used in a more or less consistent way. We use the term consistency because words whose neighbors are preserved will tend to be used in the same, consistent way by different authors in distinct types of text. Using the concepts and methodologies from complex networks [5]–[7] to analyze the relationship between concepts, we developed a series of indices to measure consistency. These indices are based on the idea that if a word induces a limited set of contexts, then the neighborhood of that word tends to be maintained even in texts written by different writers. In fact, this seems a reasonable assumption, since it is known that syntactically related words also tend to be semantically related. With the indices created using this methodology, we shall show that the distribution of consistency does not follow a power law [8,9], unlike the case of the frequency of words (Zipf law) [10]. Instead, the distribution seems to follow a log-normal distribution. Furthermore, the greater the familiarity, the number of distinct neighbors and the frequency in the language, the more consistent the word is. As for the semantic factors, we showed that consistent words tend to preserve not only the lexical neighborhood, but also the semantic context. Finally, we show how the quantification of consistency can be useful in tasks such as those related to authorship recognition.

2. Methodology

2.1. Dataset

The distinct contexts in which words are used were investigated with a database comprising several books from the Gutenberg project repository¹, whose list appears in table 1. Although the number of books differs among authors, the size of the corpus for each author has a fixed size (180 500 tokens). Thus, the difference of the corpus size has little interference in the analysis of consistency of words.

2.2. Modeling texts as complex networks

Language issues have attracted the interest of many researchers in recent years [11]–[18]. For instance, physicists have used dynamical systems [19,20] and complex networks [21]–[25], [7,26,27] to study various aspects of language. Many of these studies have used the co-occurrence model [26]–[28] to link adjacent words in the text. The idea behind this modeling comes from the use of co-occurrence statistics at various scales, from bigram statistics to discourse scale windows, which has been widespread in document analysis and retrieval for at least two decades [29]–[32]. Because we are interested in the neighborhood

¹ http://www.gutenberg.org/.

doi:10.1088/1742-5468/2012/01/P01004
Table 1. Database employed in the experiments.

| Author                      | Book                                                                 |
|-----------------------------|----------------------------------------------------------------------|
| Arthur Conan Doyle          | Uncle Bernac—A Memory of the Empire                                  |
|                             | The Tragedy of the Korosko                                            |
|                             | The Valley of Fear                                                    |
|                             | The War in South Africa                                               |
|                             | The White Company                                                    |
|                             | Through the Magic Door                                                |
|                             | The Adventures of Sherlock Holmes                                    |
| Bram Stoker                 | Dracula’s Guest                                                      |
|                             | The Jewel of Seven Stars                                              |
|                             | The Lady of the Shroud                                               |
|                             | Lair of the White Worm                                               |
|                             | The Man                                                               |
| Charles Darwin              | Coral Reefs                                                          |
|                             | On the Origin of Species by Means of Natural Selection               |
|                             | The Voyage of the Beagle                                             |
|                             | The Different Forms of Flowers on Plants of the Same Species          |
| Charles Dickens             | American Notes                                                       |
|                             | A Tale of Two Cities                                                 |
|                             | Hard Times                                                           |
|                             | The Old Curiosity Shop                                               |
| Thomas Hardy                | A Changed Man; and Other Tales                                       |
|                             | Desperate Remedies                                                   |
|                             | Far from the Madding Crowd                                           |
|                             | The Hand of Ethelberta                                              |
| Pelham Grenville Wodehouse  | My Man Jeeves                                                        |
|                             | Tales of St. Austin’s                                                |
|                             | The Adventures of Sally                                              |
|                             | The Clicking of Cuthbert                                             |
|                             | The Gem Collector                                                    |
|                             | The Man with Two Left Feet                                           |
|                             | The Pothunters                                                       |
|                             | The Swoop!                                                           |
|                             | The White Feather                                                    |
| Virginia Woolf              | Jacob’s Room                                                         |
|                             | Monday or Tuesday                                                    |
|                             | Night and Day                                                        |
|                             | The Voyage Out                                                       |
| William Wordsworth          | Lyrical Ballads, with Other Poems—Volume 1                           |
|                             | The Poetical Works of William                                        |
|                             | Wordsworth—Volumes 1–3                                               |

doi:10.1088/1742-5468/2012/01/P01004
Using complex networks to quantify consistency in the use of words

Figure 1. Example of the network obtained for the extract: ‘It was not that he felt any emotion akin to love for Irene Adler. All emotions, and that one particularly, were abhorrent to his cold, precise but admirably balanced mind.’ from the book *The Adventures of Sherlock Holmes*, by Arthur Conan Doyle.

Table 2. The pre-processing step involves two procedures: (i) removal of stopwords and (ii) lemmatization of the remaining words.

| Original                                      | Without stopwords | Lemmatized             |
|----------------------------------------------|-------------------|------------------------|
| It was not that he felt any emotion          | Felt emotion      | Feel emotion           |
| akin love for Irene Adler                    | akin love Irene Adler | akin love Irene Adler |
| All emotions, and that one                  | Emotions          | Emotion                |
| particularly, were abhorrent to his         | Abhorrent         | Abhorrent              |
| cold, precise but admirably balanced mind.   | Cold precise      | Cold precise           |
|                                              | Balanced mind     | Balanced mind          |

properties of words to examine the preservation of induced contexts, we chose to use this model, which is described as follows.

The modeling procedure started with a pre-processing step, where stopwords (i.e., words with little semantic meaning), such as articles and prepositions, were removed from the text. Although the frequency of such words may be useful in distinguishing writers’ personal characteristics [33] we decided to ignore them because we are only interested in the contextual semantic preservation. The remaining words were then lemmatized so that conjugated verbs and nouns in the plural form were converted respectively to their infinitive and singular forms. Thus words related to the same concept but with distinct inflections were taken as a single node in the network. This lemmatization was performed with the MXPOST part-of-speech tagger [34], based on a maximum entropy model [35]. This was done to resolve ambiguities during the conversion to the canonical form (infinitive and plural). After this pre-processing, each distinct word became a node and the neighborhood relationship between words defined the set of edges. To illustrate the procedure, we created a small network for the following text extract: ‘It was not that he felt any emotion akin to love for Irene Adler. All emotions, and that one particularly, were abhorrent to his cold, precise but admirably balanced mind.’ obtained from the book *The Adventures of Sherlock Holmes*, by Arthur Conan Doyle. Table 2 summarizes the pre-processing steps and figure 1 displays the resulting network.
The networks for each of the books of an author were then joined to obtain the so-called author’s network, reflecting the association of words generated by that author. That is to say, if a given node appears in one of the networks of books, then it will also appear in the author’s network. Similarly, two vertices were connected in the author’s network if both appeared connected at least in one network of the books. The derivation process is illustrated in figure 2.

2.3. Consistency indices

In this section we describe the indices proposed to measure the consistency $\mathcal{C}$ with which words are used. Because obtaining $\mathcal{C}$ demands that the neighborhood of each word for each author is known, it was calculated only for the 2880 words that appeared in the networks for all authors.

2.3.1. $\mathcal{C}$ based on the histogram of co-occurrence of neighbors. To derive this index for a given word $w$, consider the vector $\mathring{v}_w^i$ storing the frequency of each of the $k(w)$ distinct neighbors of $w$ over all authors’ networks. Hence, $v_i^w$ stores the number of authors for which a given neighbor $i$ appears connected to the word $w$. The element $v_i^w$ of $\mathring{v}_w^i$ ranges from 1 to 8, since $n = 8$, the number of authors studied. The consistency based on $\mathring{v}_w^i$ increases with
the sum of the vector components, since high \( v_i^w \) means that the corresponding association of words is repeated by several authors. The consistency index is calculated as

\[
C_{\text{hist}}(w) = \frac{s(w) - k(w)}{(n-1)k(w)} = \frac{s(w) - k(w)}{7k(w)},
\]

where \( s \) is given by

\[
s(w) = \sum_{i=1}^{k(w)} v_i^w.
\]

Note that while the numerator in equation (1) represents the sum of \( v_i^w \) derived from the minimum possible sum (which occurs when each of the \( k \) neighbors occurs in only one of the authors’ networks), the denominator is the range of the sum (the largest sum occurs when every neighbor appears in all \( n \) networks). Consequently, \( C_{\text{hist}} \) ranges between 0 and 1.

2.3.2. Consistency based on the cosine similarity. Consistency may be calculated for words considering two distinct authors’ networks \( A^i \) and \( A^j \), comprising the sets of nodes denoted by \( V_i \) and \( V_j \) respectively. Initially, \( V_i \) and \( V_j \) were expanded (keeping the original links) so that the new set of vertices \( V_{ij} = V_i \cup V_j \). Let \( w \) be the word whose consistency is being calculated. The number of neighbors of \( w \) appearing in both \( A^i \) and \( A^j \) is

\[
N_{ij}(w) = \sum_k A^i_{wk} A^j_{wk}.
\]

In spite of capturing the number of neighbors in common, it is difficult to interpret whether the value of \( N_{ij} \) is high or low, since it is not normalized. We have therefore normalized equation (3) dividing it by the geometric mean of the degrees of the word \( w \) in both networks. Thus, if \( k^i_w \) and \( k^j_w \), given by

\[
k_w = \sum_x A_{wx},
\]

represent the degrees of the word \( w \) respectively in \( A^i \) and \( A^j \), then the normalized number of shared neighbors is

\[
N'_{ij}(w) = \frac{\sum_k A^i_{wk} A^j_{wk}}{\sqrt{k^i_w k^j_w}}.
\]

In order to take into account all pairs of authors in computing the consistency, we defined consistency as the average of \( N'_{ij} \) over the \( n(n-1)/2 \) pairs of distinct networks:

\[
C_{\text{cos}}(w) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} N'_{ij}(w).
\]

It is also worth mentioning that \( C_{\text{cos}} \) ranges between 0 and 1, analogously to \( N'_{ij} \). Actually, \( C_{\text{cos}} \) can be interpreted as the cosine of the angle between the vectors whose elements indicate the presence or absence of a particular word as neighbor. Thus, if these vectors are similar, the angle between them is small and the cosine is high, indicating high consistency.
2.3.3. \( C \) based on the Pearson correlation coefficient. One can estimate whether \( N_{ij} \) in equation (3) is high by comparing it with the expected number of neighbors assuming a random choice of neighborhood. The expected numbers of common neighbors for word \( w \) in the networks \( A^i \) and \( A^j \) are given by \( k^i_w \) and \( k^j_w \), as defined in equation (4). If \( w \) in the network \( A^i \) randomly chooses one of its \( k^i_w \) neighbors, then the probability of choosing a node which is also a neighbor of \( w \) in \( A^j \) would be \( k^j_w / N \). Applying this reasoning for the \( k^i_w \) neighbors of \( w \) in \( A^i \), the expected number of common neighbors is \( k^i_w k^j_w / N \), where \( N \) represents the number of nodes in the network. Therefore, if \( N_{ij} > k^i_w k^j_w / N \), the consistency is higher than expected. Thus, defining consistency as the difference between the actual and expected numbers of neighbors in common, we obtain \( C_r \):

\[
C_r(w) = N_{ij} - \frac{k^i_w k^j_w}{N} = \sum_k A^i_{wk} A^j_{wk} - \frac{k^i_w k^j_w}{N}
\]

\[= \sum_k A^i_{wk} A^j_{wk} - N \bar{k}^i_w \bar{k}^j_w
\]

\[= \sum_k (A^i_{wk} A^j_{wk} - \bar{k}^i_w \bar{k}^j_w)
\]

\[= \sum_k (A^i_{wk} - \bar{k}^i_w) (A^j_{wk} - \bar{k}^j_w), \tag{7}
\]

where the notation \( \bar{k} \) represents the degree normalized by the number of nodes in the network:

\[
\bar{k}_w = \frac{1}{N} \sum_x A_{wx} \tag{8}
\]

Equation (7) can be interpreted as a covariance, i.e., a non-normalized correlation. To transform \( C_r \) into a normalized measurement, \( C_r \) is divided by the corresponding standard deviations of the vectors \( A^i_{wk} \) and \( A^j_{wk} \), \( k = 1 \cdots n \). The covariance then becomes the Pearson correlation coefficient \( r_{ij} \):

\[
C'_r(w) = r_{ij} = \frac{\sum_k (A^i_{wk} - \bar{k}^i_w)(A^j_{wk} - \bar{k}^j_w)}{\sqrt{\sum_k (A^i_{wk} - \bar{k}^i_w)^2} \sqrt{\sum_k (A^j_{wk} - \bar{k}^j_w)^2}} \tag{9}
\]

With consistency defined as in equation (9), \( C'_r \) ranges between -1 and 1. In order to restrict the range between zero and 1, the following linear transformation is performed in \( C'_r \), deriving \( C''_r \):

\[
C''_r = \frac{C'_r + 1}{2}. \tag{10}
\]

Now, if \( C''_r > 0.5 \), then the quantity of shared neighbors is greater than that expected by chance.

2.3.4. \( C \) based on the Leicht–Holme–Newman index [36]. In the consistency index based on the Pearson correlation coefficient, \( C_r \) is derived from the difference between the number of common neighbors found and the number of common neighbors expected by chance.
For the Leicht–Holme–Newman index \([36]\), the new measure of consistency is obtained from the ratio between these numbers, as follows:

\[
C_{\text{LHN}} = \frac{N_{ij}(w)}{k_i^{w}k_j^{w}} = N \frac{\sum_k A_{i}^{w}k_{i}^{w}A_{j}^{w}k_{j}^{w}}{\sum_k A_{i}^{w}k_{i}^{w}k_{j}^{w}}. \tag{11}
\]

The important threshold now is the number 1, since for \(C_{\text{LHN}}\) above 1 the consistency is higher than expected.

2.3.5. \(C\) based on the Sorensen index \([37]\). Unlike the metrics described above, this measure is based on the dissimilarity of the set of neighbors of a given word in two networks. The vectors \(A_{i}^{w}\) and \(A_{j}^{w}\), which store the neighborhoods of the word \(w\), are used again, with the dissimilarity of neighbors being computed as the corresponding squared normalized Euclidean distance:

\[
C_{\text{euc}} = \sum_k \left(\frac{A_{i}^{w}k_{i}^{w} - A_{j}^{w}k_{j}^{w}}{k_i^{w} + k_j^{w}}\right)^2 = 1 - \frac{2N_{ij}(w)}{k_i^{w} + k_j^{w}}. \tag{12}
\]

Since the distance between the vectors was divided by the maximum possible distance (given by \(k_i^{w} + k_j^{w}\)), \(C_{\text{euc}}\) ranges between 0 and 1. To make this measure consistent with other metrics, we converted it from a dissimilarity to a similarity measure with the following transformation:

\[
C_s = 1 - C_{\text{euc}} = 2\frac{N_{ij}(w)}{k_i^{w} + k_j^{w}}. \tag{13}
\]

2.3.6. \(C\) based on the frequency of the shared neighbors. In this index, we also consider the frequency with which a given neighbor was employed by each author. This new measure is justified because neighbors may appear with quite different frequencies. For example, if a given association of words occurs 100 times for one author and only once for another author (considering texts of the same size), the consistency index would still be maximum according to the indices described so far. However, this combination of words is clearly not consistent, since the frequencies are quite different.

The disparity in frequency is considered as follows. Suppose that word \(w\) has \(N\) distinct neighbors in the corpus of \(n\) authors. Let \(v_{k}\) be one of the neighbors of \(w\). If \(v_{k}\) appears associated to \(w^{f_k^{i}}\) times in \(A^{i}\) and \(w^{f_k^{j}}\) times in \(A^{j}\), then the consistency of \(w\) regarding the association \(w \leftrightarrow v_{k}\) is

\[
C_{ij}^{\text{freq}}(w \leftrightarrow v_{k}) = 1 - \frac{|f_k^{i} - f_k^{j}|}{f_k^{i} + f_k^{j}}. \tag{14}
\]

To consider all the neighbors, \(C_{ij}^{\text{freq}}(w)\) is computed as an average over \(C_{ij}^{\text{freq}}(w \leftrightarrow v_{k})\), provided that \(f_k^{i} + f_k^{j} > 0\). Assuming that this condition occurs \(t\) times, \(C_{ij}^{\text{freq}}(w)\) is

\[
C_{ij}^{\text{freq}}(w) = \frac{1}{t} \sum_{f_k^{i} + f_k^{j} > 0} C_{ij}^{\text{freq}}(w \leftrightarrow v_{k}). \tag{15}
\]
Using complex networks to quantify consistency in the use of words

Figure 3. Distribution of consistency for 2880 words. Regardless of the definition employed to quantify consistency, a similar distribution was obtained.

Table 3. Parameters obtained from fitting the consistency probability density functions. All distributions follow a log-normal distribution, since \( R^2 \sim 1 \) and \( \chi^2 \ll 1 \).

\[
\begin{array}{cccccc}
\text{C} & x_c & \sigma & A & R^2 & \chi^2 \\
\hline
\text{Frequency} & 0.02416 \pm 0.0011 & 0.97641 \pm 0.03091 & 0.00542 \pm 0.00014 & 0.991 & 1.53 \times 10^{-5} \\
\text{Histogram} & 0.02389 \pm 0.0018 & 0.89291 \pm 0.04733 & 0.00342 \pm 0.00019 & 0.972 & 2.68 \times 10^{-5} \\
\text{Sorensen} & 0.05076 \pm 0.0020 & 1.04088 \pm 0.02737 & 0.01055 \pm 0.00022 & 0.994 & 8.25 \times 10^{-6} \\
\text{Pearson} & 0.04134 \pm 0.0037 & 1.11074 \pm 0.04754 & 0.00683 \pm 0.00036 & 0.986 & 1.49 \times 10^{-6} \\
\text{Cosine} & 0.06411 \pm 0.0052 & 1.08812 \pm 0.04677 & 0.01232 \pm 0.00058 & 0.989 & 1.52 \times 10^{-5} \\
\text{LNH} & 3.97112 \pm 0.0536 & 0.65551 \pm 0.01073 & 0.90414 \pm 0.01281 & 0.994 & 1.17 \times 10^{-5} \\
\end{array}
\]

3. Results and discussion

3.1. Analysis of distribution and inter-correlation of consistency indices

The consistency indices defined in section 2.3 were first used to examine the distribution of consistency for the 2880 words under analysis. The distributions display a peak and two asymmetric tails for all the indices used, as illustrated in figure 3 for the Sorensen index and the indices based on the frequency of words in common and on histograms. One infers that it is rare for a word to take extreme consistency values (low or high consistency), but it is far more rare for a word to take very high consistency values. Formally, this finding is confirmed by the log-normal distribution that is obeyed by the indices. In all cases, the data could be explained by a log-normal function

\[
f(x; x_c, \sigma, A) = \frac{A}{\sqrt{2\pi}\sigma x} e^{-\left(\ln x / x_c \right)^2 / 2\sigma^2},
\]

where \( x_c, \sigma \) and \( A \) correspond to the free parameters of the distribution. Table 3 summarizes the parameters for each case, and confirms the suitability of the fitting with Pearson-squared \( R^2 \sim 1 \) and chi-square \( \chi^2 \ll 1 \).
Log-normal distributions are largely found in non-linguistic contexts (see, e.g., [38]–[43]), but are less common in linguistics, for which statistical distributions such as Zipf’s law [10] prevail, as in the case of frequency and ranking of words [8]. Nevertheless, log-normal distributions have been reported for random variables in natural language issues. For example, Williams [44] showed that the lengths of sentences seem to follow a log-normal distribution. Similarly, Herdan [45] found that the lengths of spoken words in phone conversation also follow this distribution. Log-normal distributions are usually generated by processes following proportionality laws [46]–[48]. For this reason, the consistency can be thought of as a result of a growth process governed by the $\alpha_k$ constant, which also follows a log-normal distribution. To verify why this statement is true, suppose that a given word is initially used by only two authors with consistency $C_0$. For each new writer who uses this word, the current consistency increases or decreases according to the factor $\alpha_k$:

$$C_k = \alpha_k C_0.$$  

After $n + 2$ authors have used the word, the final consistency $C_n$ will be given by

$$C_n = C_0 \prod_{i=1}^{n} \alpha_i.$$  

Actually, the consistency of the word is quantified as a percentage of the current consistency with each new use and this percentage is independent of the consistency currently observed. Since we assume that $\alpha_k$ is log-normally distributed, then $C_k$ will also follow a log-normal distribution, because the product of log-normal distributions also generates a log-normal distribution [48]. The requirement that $\alpha_k$ follows a log-normal distribution can be disregarded if we consider that many authors use the word. Because $\ln C_n$ is given by

$$\ln C_n = \sum_{i=1}^{n} \ln \alpha_i + \ln C_0,$$

and since according to the central limit theorem $\sum_{i=1}^{n} \alpha_i$ follows a normal distribution, it is possible to state that $\ln C_n$ is normally distributed and then $C_n$ follows a log-normal distribution.

The similar behavior for the consistency indices in figure 3 may mean that the indices are correlated. Indeed, the Pearson correlation coefficients ($r$) [49] between pairs of indices were all close to 1 (results not shown), with the exception of the LHN index. The correlation ranged from 0.902 for $C_r$ and $C_{\text{hist}}$ to 0.996 for $C_{\text{cos}}$ and $C_s$. Hence, five of the indices are equivalent and can be used interchangeably. As for LHN, it is weakly correlated with the other five indices, as shown in the first column of table 4. Even when the correlation of ranks was used, low values were observed, as shown in the second column of table 4, which summarizes the values of the Spearman’s rank correlation coefficient ($\rho$) [50]. Therefore, for quantifying consistency in texts in future works, it suffices to consider the measures based on the difference and on the ratio between the number of common neighbors and the expected number of neighbors by chance. For the sake of completeness, however, we chose to analyze all the indices in the following sections.
Using complex networks to quantify consistency in the use of words

Table 4. Pearson ($r$) and Spearman ($\rho$) correlations between LHN and other consistency indices. In both cases the correlations were low.

| C        | $r$  | $\rho$ |
|----------|------|--------|
| Histogram| 0.044| 0.166  |
| Pearson  | 0.318| 0.347  |
| Cosine   | 0.301| 0.315  |
| Sorensen | 0.293| 0.295  |

Table 5. Number of words $F$ in each group and range of the number of distinct neighbors $\eta$. While class A comprises words with few distinct neighbors (up to 150), class D comprises words with many distinct neighbors.

| Class | $\eta$       | $F$  |
|-------|--------------|------|
| $C_A$ | $0 \leq \eta \leq 150$ | 1070 |
| $C_B$ | $151 \leq \eta \leq 300$ | 791  |
| $C_C$ | $301 \leq \eta \leq 500$ | 463  |
| $C_D$ | $501 \leq \eta \leq 800$ | 256  |

3.2. Analysis of correlation between consistency and linguistic features

To understand how consistency measured according to the indices proposed is related to other linguistic factors, we examined the relationship between consistency and the number of distinct neighbors of the word. The scatter plots for four consistency indices are given in figure 4, indicating a strong positive correlation between the number of distinct neighbors and the consistency for all the indices (the indices not shown in figure 4 exhibit the same behavior). Words with larger numbers of neighbors in the networks tend to be more consistent. It is possible that words with many neighbors (probably frequently used) are more consistent because they are used in a more uniform fashion as they are widely employed by different writers. Conversely, words with fewer neighbors are probably less frequent words that are more prone to the influence of the writer’s personal experiences.

Since the distribution of number of neighbors for the 2880 words analyzed is very broad, we divided them into four classes ($C_A$, $C_B$, $C_C$ and $C_D$) according to the number of neighbors, as depicted in table 5. The correlations between consistency and the following features were calculated: (i) age of acquisition (i.e., the age at which a child begins to use the word as part of his/her spoken vocabulary), (ii) familiarity, (iii) imaginability, (iv) frequency of use in the English language; and (v) ambiguity (i.e., the number of distinct meanings extracted from the WordNet [52]). The correlations for each class and the corresponding consistency that provided the strongest correlations are shown in table 6.

The only case where the correlation was high occurred for familiarity in classes $C_C$ and $C_D$, with the other correlations being below 0.3. Notwithstanding, interesting trends can be inferred from the signs of the correlations. For example, the negative correlation

---

2 The quantities (i)–(iv) were obtained from the MRC Psycholinguistic Database [51].

doi:10.1088/1742-5468/2012/01/P01004
with the age of acquisition suggests that less consistent words take longer to learn. This should be expected since a heterogeneous use of weakly consistent words makes them more difficult to acquire. An analogous reasoning applies to the familiarity and imaginability (ability to visualize a concept), as familiar words tend to induce the same concepts (i.e., well-known words usually bring to mind the same concepts). As for the frequency in the language, the positive correlation indicates that widely used words are more consistent. Indeed, the widespread use of a word probably causes it to be used in a more homogeneous way. Finally, the ambiguity of the word correlated negatively with consistency. This result was also expected, since if a word is ambiguous then it can appear in multiple contexts, with its neighborhood tending to be more heterogeneous.

Figure 4. Correlation between the number of distinct neighbors and the corresponding consistency. Regardless of the consistency index, there is strong correlation between the consistency and the number of neighbors.
Table 6. Correlation obtained by comparing linguistic features and consistency of words for classes $C_A$, $C_B$, $C_C$ and $C_D$.

| Linguistic measure | Correlation for $C_A$ | Consistency index |
|--------------------|-----------------------|-------------------|
| Age of acquisition | $-0.09$               | Histogram         |
| Familiarity        | $+0.18$               | LHN               |
| Imaginability      | $+0.07$               | Pearson           |
| Frequency          | $+0.12$               | Sorensen          |
| Ambiguity          | $-0.13$               | LHN               |

| Linguistic Measure | Correlation for $C_B$ | Consistency index |
|--------------------|-----------------------|-------------------|
| Age of acquisition | $-0.15$               | Histogram         |
| Familiarity        | $+0.27$               | Histogram         |
| Imaginability      | $+0.13$               | LHN               |
| Frequency          | $+0.13$               | Sorensen          |
| Ambiguity          | $-0.13$               | LHN               |

| Linguistic Measure | Correlation for $C_C$ | Consistency index |
|--------------------|-----------------------|-------------------|
| Age of acquisition | $-0.24$               | Pearson           |
| Familiarity        | $+0.37$               | Histogram         |
| Imaginability      | $+0.14$               | Sorensen          |
| Frequency          | $+0.03$               | Cosine            |
| Ambiguity          | $-0.14$               | LHN               |

| Linguistic Measure | Correlation for $C_D$ | Consistency index |
|--------------------|-----------------------|-------------------|
| Age of acquisition | $-0.20$               | Sorensen          |
| Familiarity        | $+0.43$               | Sorensen          |
| Imaginability      | $+0.07$               | Histogram         |
| Frequency          | $+0.20$               | Sorensen          |
| Ambiguity          | $-0.20$               | LHN               |

In an attempt to understand why some words are more consistent than others, we examined the neighborhood of words possessing the highest and the lowest consistency values. By way of illustration, let us analyze the following strongly consistent words in $C_D$: `sleep`, `silence`, `happy`, `hair`, `God` and `cold`. For each neighbor of each one of these words, we counted the number of authors that associated them. The histograms with the frequencies in figure 5 indicate that these highly consistent words tend to have semantically related neighbors. For instance, all the eight authors associated `sleep` to `night`, `wink` and `dream`. To confirm this hypothesis of semantic proximity, we compared the neighbors for the texts written by the eight authors with the neighbors from the semantic network derived from the Edinburgh Associative Thesaurus (EAT) [53], which connects semantically related concepts. For each of the selected words with high consistency values, we computed its distance from the neighbor of the authors’ network in the EAT network. Then, we calculated the ratio between the number of neighbors that are at a specific distance $d$ and the expected number of neighbors that are at the same distance, assuming that the nodes were randomly chosen in the Edinburgh Associative Thesaurus network. The results in table 7 show that the number of neighbors immediately related in the EAT network is much higher than expected (which would give a ratio equal to 1), thus confirming that consistency is related to the limitation of semantic context.
Figure 5. Frequency of association of neighbors for the words *sleep*, *silence*, *happy*, *hair*, *God* and *cold*. Note that for the words shown in this figure (with high values of consistency), their neighborhood seems to be restricted to a single context.

Table 7. Ratio between the number of neighbors of the authors’ networks which are at a distance $d$ from the central concept in the EAT network and the expected number of neighbors at a distance $d$ assuming random associations. Interestingly, the immediate neighborhoods of the most consistent words are also concentrated in the immediate neighborhoods ($d = 1$) of the EAT networks. It turns out that consistent words induce specific contexts.

| Word  | $d$   |   |   |   |
|-------|-------|---|---|---|
|       | 1     | 2 | 3 | 4 |
| Sleep | 15.74 | 0.66 | 0.00 | 0.18 |
| Silence | 20.95 | 2.82 | 0.30 | 0.36 |
| Happy | 19.92 | 1.75 | 0.05 | 0.00 |
| Hair | 26.14 | 1.31 | 0.15 | 0.00 |
| God | 32.79 | 1.18 | 0.06 | 1.21 |
| Die | 18.20 | 2.28 | 0.30 | 0.86 |
| Cold | 9.19 | 0.69 | 0.00 | 0.60 |
| Chair | 18.31 | 1.93 | 0.40 | 0.00 |
3.3. Using the consistency index to recognize authorship

The as-expected correlations with linguistic features confirm the suitability of the consistency indices to quantify semantic aspects in text. We now check whether these indices can be used to identify authorship as authors may use words in more or less consistent ways. In the extreme case in which one compares authors of distinct genres (such as storytellers and authors writing scientific books), a very good distinction should be expected. This hypothesis was investigated by defining for each author an average consistency as

$$\bar{C} = \frac{1}{n_f} \sum_{k=1}^{n_p} C_k f_k,$$  \hspace{1cm} (20)

where $n_p$ and $n_f$ represent respectively the number of distinct words and the number of tokens in the corpus of a given author, $C_k$ is the consistency index for the $k$th word and $f_k$ is the frequency of the $k$th word. Using $C_{\text{hist}}$ to compute $\bar{C}$ for each author in equation (20), we compared the average consistencies of all 28 pairs of distinct authors. Figure 6 illustrates with grayscales the $p$-values from the comparison of $\bar{C}$. While some pairs of authors are easily distinguishable (see, e.g., Stocker and Darwin in $C_A$, Woolf and Hardy in $C_B$, Darwin and Doyle in $C_C$ and Wodehouse and Dickens in $C_D$), others are quite similar (Stocker and Woolf in $C_D$). In addition, counting the number of darkish grayscales one notes that the words in $C_A$ provide a better ability of distinction, while those in class $C_D$ provide the worst.

Finally, to examine how the authors are clustered in terms of consistency, we used the hierarchical clustering algorithm UPGMA [55] based on the cosine similarity\(^3\) to generate a hierarchy of authors. Figure 7 shows that four clusters can be identified upon choosing an appropriate distance threshold. Significantly, authors of novels (such as Wodehouse), are separated from those of scientific works (such as Darwin), which indicates that the difference in style may be reflected in the use of words of distinct consistency indices.

In summary, consistency indices are useful for detecting authorship, and now can be combined with other conventional methods [33], [56]–[59] to enhance accuracy rates in distinguishing authors.

4. Conclusion and final remarks

In this paper we have studied the problem of quantifying the complexity of writing considering the consistency of words. Assuming that consistent words induce grammatically/semantically limited contexts, we defined several indices to measure the tendency of a word to be used homogeneously (i.e., preserving the context). With the various indices proposed, we found that the consistency of words is well fitted by a log-normal distribution, in contrast to Zipf’s law for ranking words according to frequency. Interestingly, we found that consistency can be seen as a multiplicative growth process. Each new writer using the word modifies the current consistency by adding or removing a fraction of the current value which is independent of the current consistency. Furthermore,\(^3\)

\(^3\) Each author is characterized as a vector so that each element of the vector stores the frequency of the corresponding word in the author’s book. This model is widely used in text mining research and is known as ‘bag of words’ [54].
Using complex networks to quantify consistency in the use of words

Figure 6. Comparison of the average consistency indices for classes $C_A$, $C_B$, $C_C$ and $C_D$. The brighter the grayscale the smaller the $p$-value and the higher the difference in the average consistency.

more frequent and familiar words tend to be used more consistently, while ambiguous words and words which take a long time to learn tend to be less consistent. Finally, we confirmed that the quantification of complexity in terms of the characterization of the consistency of words is able to distinguish authors, especially those with different genres of writing.

As future work we propose the use of new metrics for consistency in order to ascertain whether the results are preserved. As a starting point, we intend to make use of the so-called Katz similarity [60] to quantify consistency using the distance between concepts. Similarly, we intend to further study the behavior of consistency extending the measures developed here considering not only the immediate neighborhood of the word, but also more distant neighborhoods. Since the consistency of words is related to the semantics, we wish to combine it with the features related to the writing style (for instance using CN topological measurements [6, 59]) to verify whether the distinguishability is enhanced. Finally, we suggest that the consistency indices may also be useful in applications to quantify subjectivity (in written texts or in transcribed speech), as subjective words...
Figure 7. Hierarchical grouping using the average consistency. Note that the texts by Darwin can be considered as outliers, since they are predominantly scientific, where words should be used in a more consistent way.

might have lower consistency values because distinct persons probably associate distinct neighbors to subjective concepts.

Acknowledgments

This work was supported by FAPESP and CNPq (Brazil).

References

[1] Arnold J, Wasow T, Losongco A and Ginstrom R, Heaviness versus newness: the effects of structural complexity and discourse status on constituent ordering, 2000 Language 76 1
[2] Kovacs L, Repasi T, Baksza-Varga E and Barabas P, Clustering Based on Context Similarity, 2008 First Int. Conf. on Complexity and Intelligence of the Artificial and Natural Complex Systems, Medical Applications of the Complex Systems, Biomedical Computing pp 157–65
[3] Rapp R, Discovering the senses of an ambiguous word by clustering its local contexts, 2005 Studies in Classification, Data Analysis and Knowledge Organization (Berlin: Springer) (New York: Wiley)
[4] Wiebe J and Mihalcea R, Word Sense and Subjectivity, 2006 Proc. 21st Int. Conf. on Computational Linguistics/44th Annual Mtg of the Association for Computational Linguistics
[5] Costa L F, Oliveira O N Jr, Travieso G, Rodrigues F A, Boas P R V, Antiqueira L, Viana M P and Rocha L E C, Analyzing and modeling real-world phenomena with complex networks: a survey of applications, 2011 Adv. Phys. 60 329–412
[6] Costa L F, Rodrigues F A, Travieso G and Villas Boas P R, Characterization of complex networks: a survey of measurements, 2007 Adv. Phys. 56 167–242
[7] Newman M E J, The structure and function of complex networks, 2003 SIAM Rev. 45 167–256
[8] Clauset A, Shalizi C R and Newman M E J, Power-law distributions in empirical data, 2009 SIAM Rev. 51 323–51
[9] Newman M E J, Power laws, Pareto distributions and Zipf’s law, 2005 Contemp. Phys. 46 661–703
[10] Zipf G K, 1949 Human Behavior and the Principle of Least Effort (Reading, MA: Addison-Wesley)
[11] Serva M, Petroni F, Volchenkov D and Wichmann S, Malagasy dialects and the peopling of Madagascar, 2012 J. R. Soc. Interface 9 (66) 54–67
[12] Petroni F and Serva M, Lexical evolution rates derived from automated stability measures, 2010 J. Stat. Mech. P03015
[13] Montemurro M A and Zanette D H, Universal entropy of word ordering across linguistic families, 2011 PLoS One 6 e19875

doi:10.1088/1742-5468/2012/01/P01004
Using complex networks to quantify consistency in the use of words

[48] Mitzenmacher M, A brief history of generative models for power law and lognormal distributions, 2004
Internet Math. 1 226–51
[49] Neter J, Kutner M H, Nachtsheim C J and Wasserman W, 1996 Applied Linear Statistical Models
[50] Spearman C, The proof and measurement of association between two things, 1904 Amer. J. Psychol. 15 72–101
[51] Coltheart M, The MRC Psycholinguistic Database, 1981 Q. J. Exp. Psychol. A 33 497–505
[52] Miller G A, WordNet: a lexical database for English, 1995 Commun. ACM 38 11
[53] Kiss G R, Armstrong C, Milroy R and Piper J, An associative thesaurus of English and its computer analysis, 1973 The Computer and Literary Studies (Edinburgh: University Press)
[54] Lewis D, Naive (Bayes) at forty: the independence assumption in information retrieval, 1998 Proc. ECML-98, 10th European Conf. on Machine Learning
[55] Jain A K and Dubes R C, 1988 Algorithms for Clustering Data (Englewood Cliffs, NJ: Prentice-Hall)
[56] Juola P, Authorship attribution, 2008 Found. Trends Inform. Retrieval 1 233–324
[57] Oakes M, Ant colony optimisation for stylometry: the federalist papers, 2004 Proc. 5th Int. Conf. on Recent Advances in Soft Computing
[58] Brennan M and Greenstadt R, Practical Attacks Against Authorship Recognition Techniques, 2009 IAAI: Proc. Twenty-First Conf. on Innovative Applications of Artificial Intelligence
[59] Amancio D R, Altman E G, Oliveira O N Jr and Costa L F, Comparing intermittency and network measurements of words and their dependency on authorship, 2011 New J. Phys. 13 123024
[60] Katz L, A new status index derived from sociometric analysis, 1953 Psychometrika 18 39–43

doi:10.1088/1742-5468/2012/01/P01004