LEXICOMETRIC AND INFORMATIONAL MEASURES IN POLITICAL AND LITERARY CORPORA

FLORENTINA ARMĂŞELU

Luxembourg Centre for Contemporary and Digital History (C^2DH), University of Luxembourg
florentina.armaselu@uni.lu

Cuvinte-cheie: *teoria informaţiei, lexicometrie, analiza corpusurilor.*
Keywords: *information theory, lexicometry, corpus analysis.*

1. INTRODUCTION

How can the frequency of word occurrence be interpreted in the context of an informational analysis of textual corpora? To what extent can the word frequency values and distribution serve as indicators of the degree of "surprise", "certainty" or "informativeness" conveyed by a text? What factors (e.g. language, genre, corpus size) influence these measures? Can digital tools support qualitative assumptions about a text or a collection of texts being more "surprising", more "predictable" or more "informative" than others? What else can be learned about the texts using this type of analysis? The paper addresses these questions by creating "informational profiles" for samples of multilingual, multi-genre corpora and by building a "digital test bed" to compare the informational measures computed for the selected data.

No unified definition of information exists; it is a "polymorphic phenomenon and a polysemantic concept" that can be "associated with several explanations, depending on the level of abstraction adopted and the cluster of requirements and desiderata orientating a theory" (Floridi 2017: 1). Several information measures have been devised so far (for overviews, see Arndt 2001; Soofi *et al*. 2010; Kowalski 2013). This paper will focus on three measures based on probabilistic models intended to capture the informational particularities of textual corpora.

Within a "mathematical theory of communication", Shannon (1948: 10, 11) defined the "entropy" of a set of possible events with different probabilities of occurrence as a measure of "how much 'choice' is involved in the selection of the event", "how uncertain we are of the outcome" or, more generally, as a measure of "information, choice and uncertainty". Thus, entropy was often related to as "surprisal and uncertainty, as a consequence of choice" (Bentz *et al.* 2017a: 1) or was referred to as "measuring information only in terms of the indeterminacy that

SCL, LXXI, 2020, nr. 2, Bucureşti, p. 163–178

BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

it removes" (Marcus 1970: 206).[1] On the other hand, the "informational energy" (Onicescu 1966) of a system with several possible states and corresponding probabilities was considered to provide information on the "degree of organization of a system or the mode of partitioning of its elements" (Sârbu 1999: 69). It was also supposed that it could "be used to quantify the degree of homogeneity of a system or structure" (Preda and Dedu 2015: 28).[2] Other studies have pointed out a negative correlation in relation to entropy, i.e. "the informational energy decreases when the informational entropy increases" (Marcus 1970: 193).[3] In response to the theory of communication, not taking into account "semantic aspects" deemed by Shannon (1948: 1) as "irrelevant to the engineering problem", a theory of "semantic information" was proposed (Carnap and Bar-Hillel 1952). Within this context, a "semantically related concept of information" (Dretske 1999: 52) was defined, often designated as "informativeness" or "information content" (Floridi 2017; Resnik 1995; Mintz *et al*. 2014), pertaining to the "inverse relationship principle", i.e. "an increase in available information" corresponds to a "decrease in possibilities, and vice versa" (Barwise 1998: 491). This correlation was also formulated in terms of probability –"as probability [of a concept] increases, informativeness decreases" (Resnik 1995: 449) – or of occurrence frequency – "the informativeness of a concept is inversely dependent on its occurrence frequency: the more frequent a concept, the less informative it is" (Mintz *et al*. 2014: 1).

Various research projects have applied such informational measures to study language-related phenomena. For instance, Shannon (1948, 1951) exemplified the use of n-gram entropy in assessing how well a letter can be predicted in a natural language, such as English, when the n preceding letters are known. Marcus (1970) evaluated first-order entropy and informational energy considering the occurrence frequency of letters to discuss prosodic aspects of a set of poems in Romanian. Bentz *et al*. (2017a) processed large parallel multilingual corpora to analyse word learnability and expressivity across languages based on unigram entropies calculated at word level. Kalimeri *et al*. (2014) compared unigram, bigram and trigram entropy values for word-length representations in Greek and English corpora to examine the sensitivity of these measures to language and text genre. Cisne *et al*. (2010) used lemma-based entropy estimations to illustrate how applications of information theory can prove the validity of editorial principles such as *difficilior lectio potior* (DLP), i.e. "the more difficult reading [is] preferable", in the reconstruction of ancient texts. Other researchers determined

[1] Ro. "[…] entropia lui Shannon măsoară informația doar sub aspectul nedeterminării pe care ea o elimină […]" (Marcus 1970: 206).

[2] Ro. "[…] energia informațională poate fi utilizată pentru a cuantifica gradul de omogenitate al unui sistem sau al unei structuri […]" (Preda and Dedu 2015: 28).

[3] Ro. "[..] energia informațională descreşte atunci când entropia informațională creşte" (Marcus 1970: 193).

BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

informativeness (or information content)[4] estimators to measure semantic similarity in IS-A taxonomies (Resnik 1995) or to assess students' reading comprehension through summarisation and predict summary scores for texts from different genres and topics (Mintz *et al*. 2014).

The present study investigates the application of informational measures to characterise textual corpora. Instead of starting from a particular language-related phenomenon to be examined, it focuses on methodological aspects in constructing a digital framework for testing and interpretation that can be used to: (a) discern and compare different factors that may influence this type of characterisation; (b) highlight peculiarities or nuances that may be less apparent in other forms of analysis. Given their low to medium computing complexity, the three measures described above were considered as basic elements in defining the "informational profile" of the corpora to be studied. The paper will discuss various facets of the construction and analysis process related to data selection and preparation, methods and tools applied in computing the measures, and the interpretation of results.

2. DATA SELECTION AND PRE-PROCESSING

To provide a comparative basis for the study, two types of multilingual corpora (in Romanian, French and English) were considered: (1) a selection of minutes of plenary sittings (2004 to 2012) from the Digital Corpus of the European Parliament (DCEP–PV); (2) a selection of poems by three authors (Eminescu, 2011; Hugo, 2009; Rossetti, 2005) from Project Gutenberg. The main selection criteria were related to genre, availability, size, format and language (three languages accessible to the author). The two genres – contemporary political history and literature – were chosen as potentially providing enough contrasting features for the intended research goals and because of the online availability of the texts. A temporal dimension was also taken into account, with texts not considered individually but grouped by year as a basic unit for analysis and comparison both within and across corpora.

The Digital Corpus of the European Parliament (DCEP) is a collection of documents published by the European Parliament[5] including press releases, minutes of plenary sittings, adopted texts, questions and answers, etc. The corpus[6] is available for download as full-text documents and as sentence-aligned data, in text-only (TXT) and structured (XML, SGML) formats, in more than 20 languages.

[4] The two terms are often used synonymously. In this paper, for clarity purposes, the first term will be used.

[5] http://www.europarl.europa.eu/portal/en.

[6] Number of documents: 1.5 million; number of words: 1.37 billion; 23 languages and 253 language pairs (according to https://ec.europa.eu/jrc/en/language-technologies/dcep, last updated: 10/03/2017).

BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

The DCEP–PV samples (TXT) comprised: 708 files, 2004 to 2012, 3,684,871 word tokens[7] in English; 702 files, 2004 to 2012, 3,866,870 word tokens in French; 504 files, 2007 to 2012, 2,633,585 word tokens in Romanian. The criteria determined various requirements for the selection of data: (1) available in the three languages in TXT or XML format (accepted by TXM[8], the software used in lexicometric processing); (2) indication of the date (year) in the file name (to create TXM year-based partitions); (3) not extremely large (given TXM data size limitations). Pre-processing was needed to automatically convert the texts to lower case[9] before they were imported into TXM for statistical analysis.

Project Gutenberg served as a source for the second set of samples. This online repository contains free literary eBooks in more than 50 languages[10] and multiple formats (HTML, EPUB, Kindle and plain text). The samples (TXT) comprised poems by three authors, covering a period of their creative life, with year indications[11]: 29 poems by Christina Georgina Rossetti from the volume *Goblin Market, The Prince's Progress, and Other Poems* (Miscellaneous Poems), 1848 to 1869, 15,623 word tokens in English; 70 poems by Victor Hugo from *Les contemplations* (books 4–6), 1834 to 1856, 65,496 word tokens in French; 87 poems by Mihai Eminescu from the volume *Poezii* (poems published during the poet's lifetime), 1866 to 1887, 80,368 word tokens in Romanian. The selection was guided by: (1) availability of download as plain text for each of the three languages and corresponding to a comparable period of time (mid- to late 19th century); (2) indication of the year of publication/creation for each poem; (3) text size not too large, to allow semi-automatic pre-processing. Before TXM import, preparation was necessary[12]: heading styles were added to the poems' titles; they were converted to lowercase; the downloaded files were automatically divided into individual files for each poem, with an indication of the year[13] added manually to the file name.

[7] Word type as a "unique string of unicode characters (lower case) delimited by non-alphanumeric characters (e.g. white space and punctuation marks)" and word token as "any recurring instance of a specific word type" (Bentz *et al.* 2017b: 3). Tokens or occurrences will also be used synonymously with this meaning. The figures per corpus for both samples (DCEP–PV and Project Gutenberg) are provided by TXM – *Properties* and include punctuation marks as tokens.

[8] http://textometrie.ens-lyon.fr/?lang=en.

[9] The conversion was performed in batch processing via a Linux virtual machine.

[10] Over 58,000 eBooks, 58 languages (according to https://www.gutenberg.org/, last updated: 19.01.2019).

[11] This is why poems with the year of creation/publication were selected for the study rather than prose. For the French sample, two poems without this temporal marker were excluded from the selection.

[12] Via Microsoft Word, heading styles, *Change Case*, View – *Outline* mode, saving as *Plain Text, UTF-8*.

[13] For English and French, the year was extracted from the beginning/end of the poems; for Romanian, it was taken from the *Tabel cronologic* (Chronological table) included in the volume.

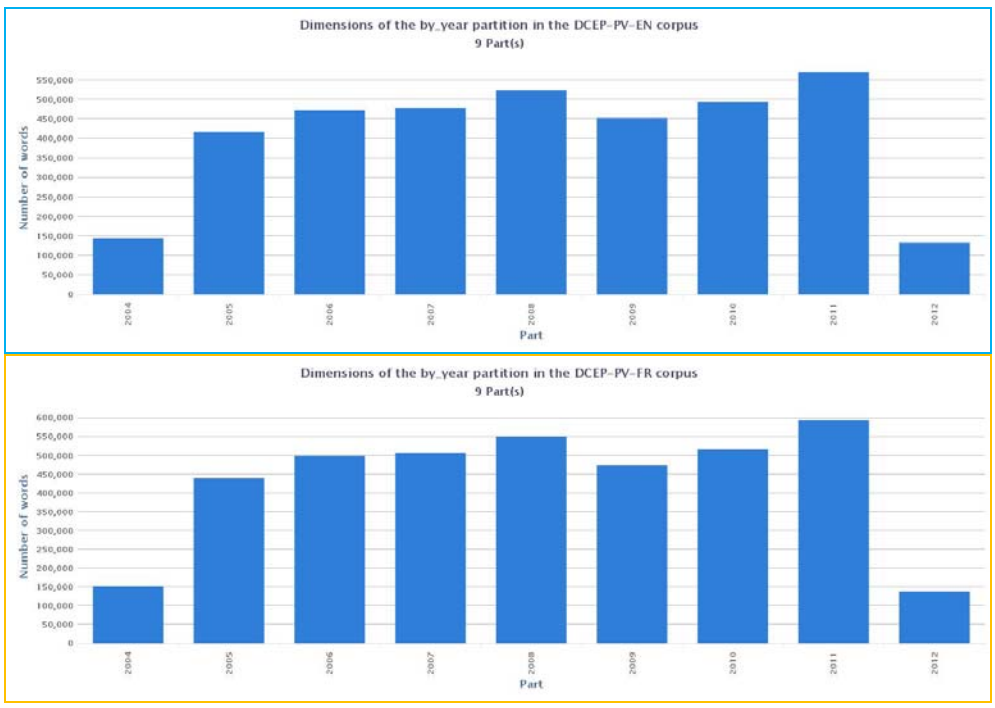
BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

3. LEXICOMETRIC AND INFORMATIONAL ANALYSIS

The six corpora, DCEP-PV and Project Gutenberg samples in English, French and Romanian, were then imported[14] into TXM, an open-source software platform (Heiden *et al*., 2010) for lexicometric analysis. The texts were part of speech tagged and lemmatised[15] during import. Partitions (TXM User Manual) for each corpus were manually created by year of publication/creation.[16]

Figure 1 (a and b) shows the dimension diagrams corresponding to these partitions, with years represented on the horizontal axis and number of word tokens on the vertical axis.

The three DCEP-PV samples (Fig. 1.a) display the minimum and maximum number of word tokens for 2012 (132,840 – EN, 137,663 – FR, 133,071 – RO) and 2011 (570,430 – EN, 593,750 – FR, 568,907 – RO) respectively.



[14] Option TXT + CSV.

[15] Via *Tree Tagger* configured for TXM and language models for English, French and Romanian.

[16] In TXM, *Partition – Assisted* mode. The files corresponding to a certain year were selected to form a group (part) in the partition – the parliamentary documents/poems produced in a year from the considered sample.

BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

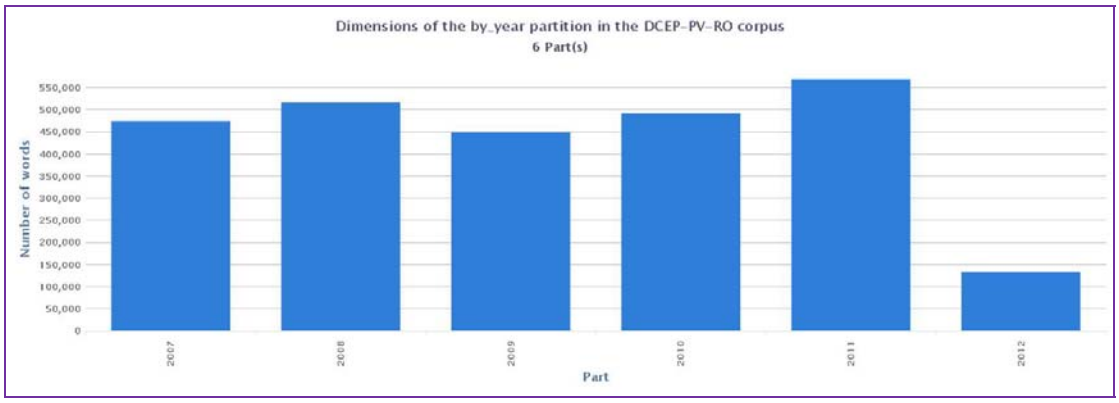
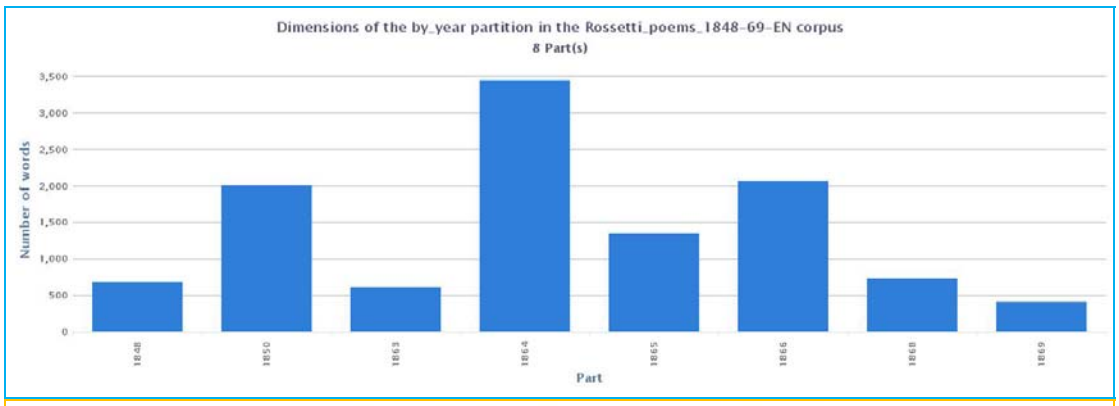
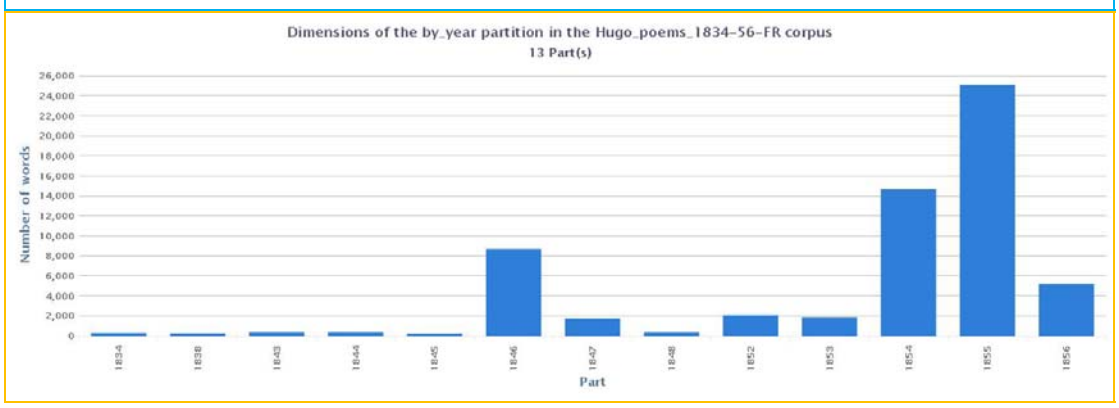


Figure 1.a. TXM year-based partitions by number of word tokens. DCEP-PV samples.

The selection of poems (Fig. 1.b) indicates more variation in size per year, with minimum and maximum values in 1869 (412 tokens) and 1864 (3,446 tokens) (Rossetti-EN), 1845 (229 tokens) and 1855 (25,094 tokens) (Hugo-FR), and 1880 (169 tokens) and 1881 (6,546 tokens) (Eminescu-RO).





BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

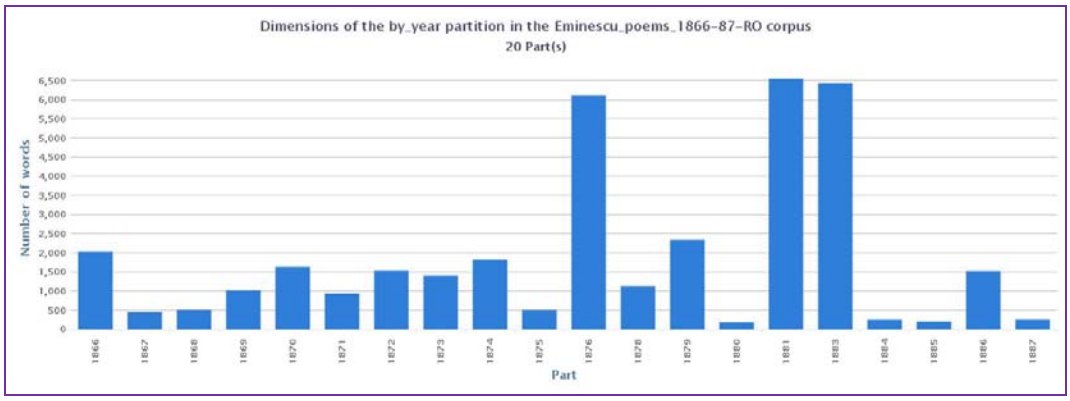


Figure 1.b.TXM year-based partitions by number of word tokens. Gutenberg samples.

The software allowed lexical tables (TXM User Manual) to be computed for each corpus and year partition. For a more general view, lemmas[17], instead of words, were considered for analysis.

Figure 2 presents an extract from the lexical tables for English DCEP-PV (left) and Romanian poems (right), sorted by decreasing order of frequency (second column). The first column in each table contains the lemmas. The other columns display at the top the total number of occurrences from the texts for a year (t), which sums up the individual values corresponding to each lemma. For example, *parliament* (rank 30)[18] occurs 10,913 times in total, 344 times in 2004 and 1,244 times in 2005; the total number of occurrences for 2004 is 91,422, for 2005 is 270,946; *vrea* (en. *want*) (rank 30) occurs 159 times, etc.

DCEP-PV-EN (rank 30–36)

enlemma	Frequency	2004 t=91422	2005 t=270946	2006 t=296733
parliament	10913	344	1244	1355
commission	10884	476	1310	1361
speak	10307	236	993	1088
following	10007	210	775	888
rapporteur	9990	210	1187	1471
debate	9360	310	1308	1376

Eminescu's_poems-RO (rank 30–36)

rolemma	Frequency	1866 t=1218	1867 t=248	1868 t=294
vrea	159	6	0	0
cel	146	4	2	0
ochi	142	2	0	4
lume	135	4	2	3
cum	134	7	2	2
dulce	114	19	8	2

Figure 2. Excerpts from TXM lexical tables by lemma and year.

The lexical tables exported from TXM were imported into Microsoft Excel and augmented with columns to compute the informational measures. Punctuation marks were discarded. The following formulae were used for entropy (1) (Shannon 1948: 11), energy (2) (Onicescu 1966; Marcus 1970: 192) and informativeness (3) (adaptation of Carnap and Bar-Hillel 1952, Dretske 1999: 52 and Floridi 2017: 27):

[17] Canonical forms as in dictionary entries (infinitive for verbs, singular and masculine for nouns, etc.).

[18] TXM also includes punctuation marks in the lexical tables.

BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

$$\mathrm{H} = -\sum_{i=1}^{N} p_i \log_2 p_i \quad (1)$$

$$\delta = \sum_{i=1}^{N} p_i^2 \quad (2)$$

$$\mathrm{INF} = -\sum_{i=1}^{N} \log_2 p_i \quad (3)$$

where N represents the number of unique lemmas for each part (year), and p_i the probability of lemma of rank *i* inside a part of the partition.

Table 1 shows the Excel informational measures, sorted by *Frequency* in descending order. The second row builds up the totals for the columns defined as follows: *Units* – lexical units (lemmas); *Frequency* – total occurrences per sample; *t_y_j* – total lemma occurrences per year (with *y_j* symbolising the year); *N_y_j* – number of unique lemmas per year; *Entropy_y_j, Energy_y_j, INF_y_j* – informational measures per year.

Table 1

Excel informational measures table (excerpts DCEP-PV-EN, 2004)

Units	Frequency	t_2004	N_2004	Entropy_2004	Energy_2004	INF_2004
Totals	**2,873,421**	**112,156**	**5,510**	**9.374208162**	**0.013188835**	**81,249.35965**
parliament	10,913	344	1	0.025607329	9.40745E-06	8.348882522
commission	10,884	476	1	0.033444817	1.80123E-05	7.880329513
speak	10,307	236	1	0.018711714	4.42771E-06	8.892504227
following	10,007	210	1	0.01696556	3.50585E-06	9.060901759
rapporteur	9,990	210	1	0.01696556	3.50585E-06	9.060901759
debate	9,360	310	1	0.023491361	7.63974E-06	8.499022871

The texts for a certain year may therefore be considered as "informational systems" with different possible states (lemmas) and different probabilities of occurrence. These probabilities were computed as relative frequencies (see also Marcus 1970: 199), i.e. the absolute frequency of a lemma divided by the total number of lemma occurrences in the texts corresponding to a year. For instance, the lemma of rank 20 in Table 1[19], *parliament*, has a probability (for 2004) calculated as 344 divided by 112,156. Some simplifications were applied. The informational measures were computed for unigrams, single lemmas as independent blocks whose influence on the overall values per sample was considered cumulatively. In reality, words in a text "exhibit short- and long-range correlations" (Bentz *et al.* 2017a: 5), which should involve more complex estimations taking into account more than one item. Each sample was also considered as being characterised by a "closed" vocabulary containing a certain

[19] Different rank from that in Figure 2, as punctuation marks were discarded.

BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

number of unique lemmas that may or may not appear in the different parts (years) of the partition. The lemmas absent from certain parts (frequency and probability 0) were ignored in the logarithmic calculus for the corresponding year.

4. DISCUSSION OF RESULTS

The Excel spreadsheets were also used to draw diagrams and analyse the different factors influencing the measures.

Table 2 presents a snapshot of the informational figures for the two samples. In their cross-linguistic comparison, Bentz *et al*. (2017a: 9) reported that unigram word entropies stabilised at a mean of 9.14 with a standard deviation SD of 1.12 for texts with more than 50K tokens. For the DCEP-PV sample, the unigram lemma entropy varied with average values slightly over that mean for groups of texts per year with more than 50K occurrences, but figures computed at the word level are not yet available for comparison. On the other hand, the Gutenberg selection showed higher variation between minimal/maximal values for the unigram lemma entropy and lower estimates as compared with the DCEP-PV samples.

Kalimeri *et al*. (2014) remarked that unigram, bigram and trigram entropies based on word length are sensitive to language and genre, with literature showing the lowest values and Greek exhibiting higher figures than English. In the current study, the results of the unigram lemma entropy per year indicated sensitivity to genre and language as well, with higher figures for the parliamentary minutes than for the poems, and the lowest values for French as compared with English and Romanian for both political and literary texts. However, it should be noted that the size of the Gutenberg samples was much smaller than the DCEP-PV samples, and testing with larger samples would be needed to allow more general assertions.

Table 2

Minimal, maximal and average informational values[20]

Sample/ Measure	**DCEP-PV**			**Project Gutenberg**		
	EN	**FR**	**RO**	**EN**	**FR**	**RO**
$\mathbf{H_{min}}$	9.37 (2004)	9.08 (2005)	9.74 (2008)	7.09 (1869)	6.41 (1845)	6.13 (1880)
$\mathbf{H_{max}}$	9.82 (2012)	9.60 (2012)	10.02 (2012)	8.53 (1864)	8.53 (1855)	9.32 (1881)
$\mathbf{H_{avg}}$	**9.56**	**9.30**	**9.88**	**7.76**	**7.34**	**8.02**
$\boldsymbol{\delta_{min}}$	0.0082 (2012)	0.0107 (2012)	0.0067 (2012)	0.0075 (1866)	0.0139 (1852)	0.0061 (1872)
$\boldsymbol{\delta_{max}}$	0.0139 (2005)	0.0177 (2005)	0.0116 (2008)	0.0118 (1848)	0.0242 (1856)	0.0196 (1880)
$\boldsymbol{\delta_{avg}}$	**0.0122**	**0.0154**	**0.0097**	**0.0099**	**0.0187**	**0.0090**
$\mathbf{INF_{min}}$	63,645 (2012)	65,063 (2012)	69,721 (2012)	1,548 (1869)	831 (1845)	608 (1880)
$\mathbf{INF_{max}}$	172,867 (2008)	174,181 (2008)	193,742 (2008)	9,668 (1864)	36,236 (1855)	24,384 (1881)
$\mathbf{INF_{avg}}$	**137,881**	**139,679**	**163,005**	**4,388**	**8,912**	**6,834**

[20] For visibility, the values have been truncated at a maximum of four decimals.

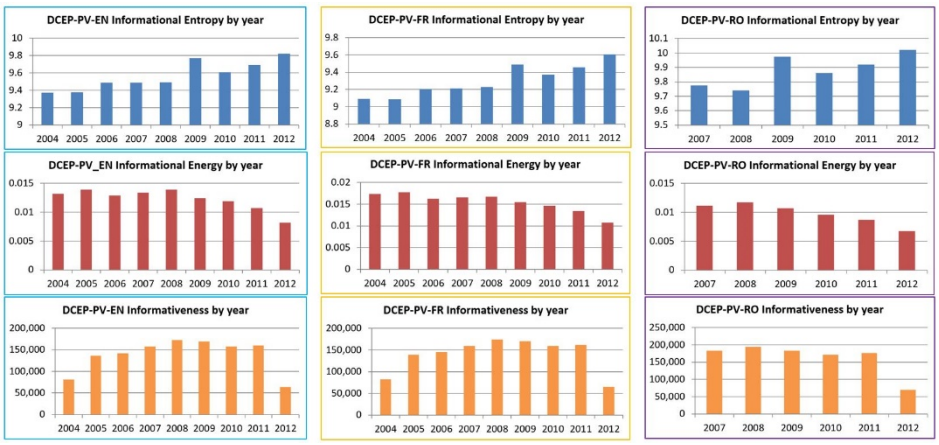
BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

The informational energy for lemmas displayed a lower range of values than entropy for both sets of samples. Marcus (1970: 200) reported similar differences between unigram entropy (around 4) and energy (around 0.06) computed for letters and three poems by Eminescu. However, the lemma values for Romanian in this study are higher for entropy and lower for energy than those corresponding to letters observed by Marcus. This can be interpreted in terms of "indeterminacy" related to choice that is higher for lemmas than for letters, which intuitively makes sense. The average energy values for DCEP-PV and Gutenberg exhibited an inverse sequence than that for entropy, i.e. the highest values for French, followed by English and Romanian for both sets of samples. The difference by genre was less clear-cut than for entropy, with lower values for literature in English and Romanian but slightly higher values for the French poems as compared with the parliamentary minutes, an issue that may again be related to the sample size, although further investigation and testing are needed to confirm this. Informativeness values[21] showed different patterns, with the highest averages for Romanian followed by French and English for DCEP-PV and for French followed by Romanian and English for Project Gutenberg.

Figure 3 illustrates the dependencies of the informational measures (vertical axis) for each sample by year (Fig. 3.a), number of unique lemmas (N) (Fig. 3.b) and total occurrences (t) (Fig. 3.c) (horizontal axis).

The rows 1, 2, 4 and 5 in Figure 3.a show a general tendency of inverse correlation between entropy and energy, as stated in previous studies (Marcus 1970), although not very strict and more discernible for the DCEP-PV samples, of a greater size.

The rows 3 and 6 indicate a similarity with the diagrams from Figure 1 (a, b), representing the corpus dimensions for each year, which suggests a relation between informativeness and the size of the considered samples, as further explained below.



[21] Not normalised by number of occurrences or unique lemmas.

BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

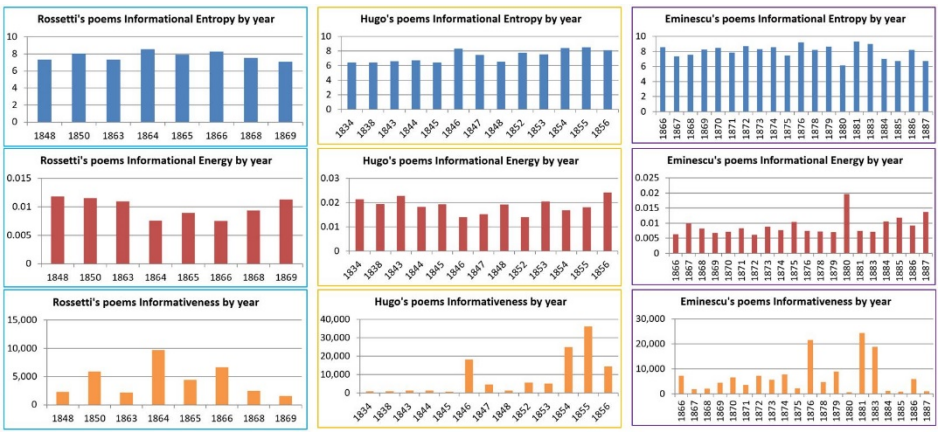


Figure 3.a. Informational measures by year.

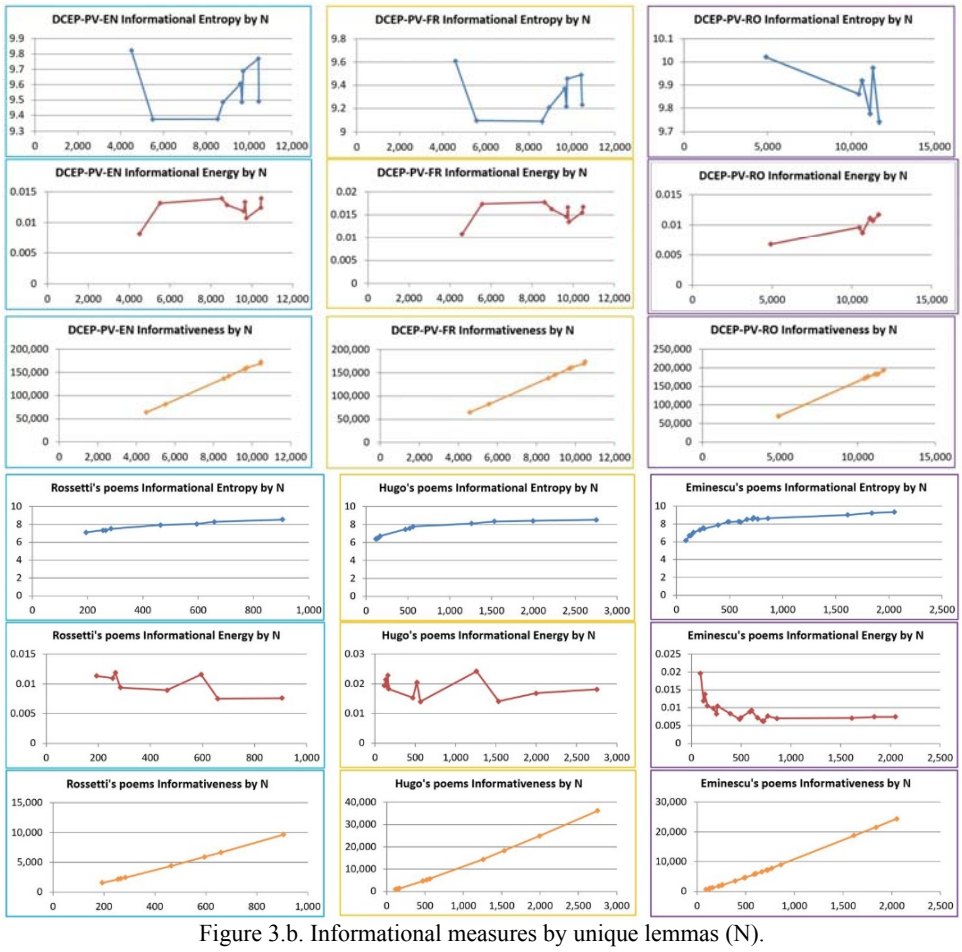


Figure 3.b. Informational measures by unique lemmas (N).

BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

In general, all the three measures displayed sensitiveness to sample size (t), as also observed by Bentz *et al.* (2017a) and Kalimeri *et al.* (2014), and to the number of unique lemmas (N).

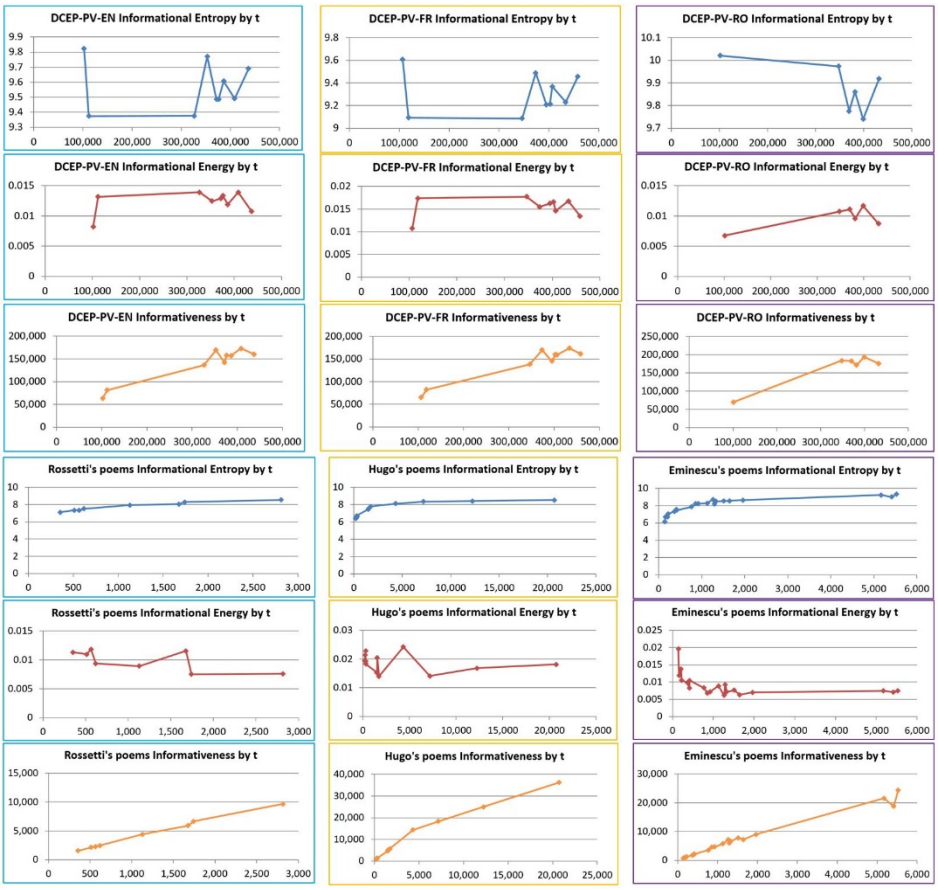


Figure 3.c. Informational measures by total occurrences (t).

Figure 3 (b, c) shows that for DCEP-PV entropy unexpectedly decreased with the increase of t and N, forming a plateau until a certain value (more perceptible for French and English), then starting to increase and alternate between higher and lower values. Variation in the opposite direction was observed for energy, as previously seen in other studies, although not strictly following an inverse shape as compared with entropy. The Gutenberg set displayed a more regular pattern, with entropy generally increasing with t and N (except in a few cases) and energy decreasing although in a less smooth way with some peak values towards the middle. Informativeness exhibited the steadiest tendency, increasing slightly irregularly with t and linearly with N. The latter behaviour might be interpreted as

BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

indicating that texts containing a higher number of unique lemmas or richer "vocabularies" are more informative.

Yet what is responsible for such a variation of entropy/energy (decreasing/ increasing) with t and N (increasing) as mentioned before? The fact that entropy and energy achieved maximal/minimal values for uniform probability or frequency distribution (the case with equal probabilities corresponds to the most uncertain situation) has been previously stated (Shannon 1948, Kalimeri *et al*. 2014, Marcus 1970, Sârbu 1999). A closer look at particular values from the two sets of samples might provide an explanation related to this observation. For instance, the steep decrease in entropy for DCEP-PV (EN) in Figure 3 (b, c, first cells top left) (H_2012 = 9.82; H_2004 = 9.37) corresponds to an increase in t and N (t_2012 = 102,181, N_2012 = 4,510; t_2004 = 112,156, N_2004 = 5,510). An examination of the parliamentary files for these years revealed duplicates (in separate files and in the minutes) of the "attendance register" listing the attendees' names for 2012. These duplicates seem to influence the mid-rank frequencies and eventually determine a higher entropy value, and more "surprisal" related to the choice of lemmas (see Fig. 4).

Kalimeri *et al.* (2014) and Bentz *et al*. (2017b) showed that different types of n-gram probability and word/lemma frequency distributions can be used to examine entropy differences and cross-linguistic phenomena across text corpora for various genres and languages. Figure 4 (left) presents a comparison of the lemma frequency distributions computed for 2004 and 2012, with $\log_{10}$ (frequency)[22] represented on the vertical axis and the frequency rank[23] on the horizontal axis. While mid-rank frequencies for 2012 are slightly higher, the diagram shows fewer values in the low frequency ranks than for 2004 (a shorter "tail") (a possible influence of size, as well), which may indicate less "informative" content for 2012 (INF_2012 = 63,645, INF_2004 = 81,249).

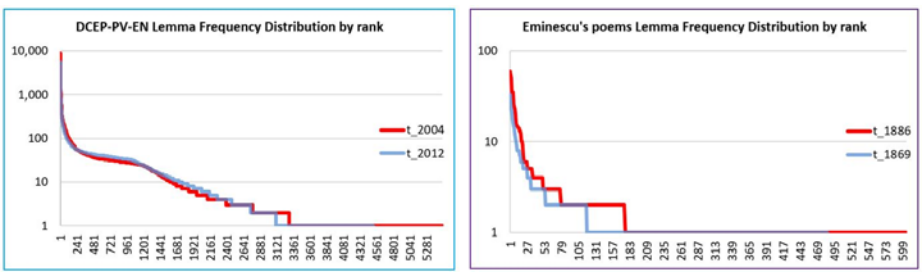


Figure 4. Frequency distribution comparison, $\log_{10}$ (frequency) per frequency rank.

[22] The logarithmic scale was used for a better visibility of the curves to be compared.

[23] With rank 1 representing the most frequent lemma in the sample, rank 2, the second most frequent, etc.

BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

A similar analysis for poems by Eminescu corresponding to 1869 (*Amicului F. I., Junii corupti, La moartea principelui Stirbey*)[24] and 1886 (*La steaua, Nu ma intelegi, Scrisoarea V*)[25] showed decreasing entropy (H_1869 = 8.25; H_1886 = 8.21) for increasing t and N (t_1869 = 857, N_1869 = 484; t_1886 = 1295, N_1886 = 604). The frequency distribution comparison for this case (Figure 4, right) indicates lower values in high frequency ranks for 1869 than for 1886, which can be interpreted as less "predictability" associated with lemma choices and thus higher entropy. However, the 1886 group appears to be more "informative", as it contains a longer "tail" of low rank frequency lemmas (possibly also due to size). A close reading of the 1869 and 1886 poems may confirm or disconfirm these hypotheses.

5. CONCLUSION AND FUTURE WORK

The paper presented a methodology for building "informational profiles" based on three measures, entropy, energy and informativeness, and a "test bed" to apply them to the analysis of multilingual, multi-genre corpora. It was shown that an open-source platform for lexicometric processing and a spreadsheet application may be used for such a purpose. The results highlighted various factors influencing the measures (e.g. corpus size, genre and language, and word frequency distribution) and some structural and stylistic particularities of the studied samples. Further testing is needed, e.g. for bigram and trigram analysis to take into account word correlations in context or to experiment with larger literary corpora and other genres so as to be able to make generalisations.

Acknowledgement. The author would like to thank Sarah Cooper, from the Language Centre of the University of Luxembourg, for English proofreading.

BIBLIOGRAPHIC REFERENCES

Arndt, C., *Information Measures: Information and its Description in Science and Engineering*, Berlin Heidelberg, Springer-Verlag, 2001.

Barwise, J., "Information and Impossibilities", *Notre Dame Journal of Formal Logic*, Volume 38, Number 4, Fall 1997, https://projecteuclid.org/download/pdf_1/euclid.ndjfl/1039540766.

Bentz, C., D. Alikaniotis, M. Cysouw, R. Ferrer-i-Cancho, "The Entropy of Words–Learnability and Expressivity across More than 1000 Languages", *Entropy*, 19, 275, 2017a, https://www.mdpi.com/1099-4300/19/6/275.

Bentz, C., D. Alikaniotis, T. Samardžić, P. Buttery, "Variation in Word Frequency Distributions: Definitions, Measures and Implications for a Corpus-Based Language Typology", Journal of

[24] En. *To my friend, F.I., Corrupt youths, On the death of Prince Stirbey.*
[25] En. *To the star, You don't understand me, Letter V.*

BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

Quantitative Linguistics 24, no. 2–3 (3 July, 2017): 128–62, 2017b, https://doi.org/10.1080/09296174.2016.1265792.

Carnap, R., Y. Bar-Hillel, "An Outline of a Theory of Semantic Information", *Technical Report No. 247*, Cambridge, Massachusetts, Research Laboratories of Electronics, Massachusetts Institute of Technology, 27 October 1952.

Cisne, J. L., R. M. Ziomkowski, S.J. Schwager, "Mathematical Philology: Entropy Information in Refining Classical Texts' Reconstruction, and Early Philologists' Anticipation of Information Theory", PLoS ONE, 5(1), 13 January 2010, https://doi.org/10.1371/journal.pone.0008661.

Digital Corpus of the European Parliament (DCEP), https://ec.europa.eu/jrc/en/language-technologies/dcep, downloaded from: https://wt-public.emm4u.eu/Resources/DCEP-2013/DCEP-Download-Page.html, document date: 11 March 2015.

Dretske, F. I., *Knowledge and the Flow of Information*, The David Hume Series, Philosophy and Cognitive Science Reissues, CSLI Publications, 1999.

Eminescu, M., The Project Gutenberg eBook, *Poezii*, Release Date: 18 February 2011, http://www.gutenberg.org/ebooks/35323.

Floridi, L., "Semantic Conceptions of Information", In *The Stanford Encyclopedia of Philosophy* Edward N. Zalta (ed.), Spring 2017 Edition, https://plato.stanford.edu/archives/spr2017/entries/information-semantic/.

Heiden, S., J-P. Magué, B. Pincemin, 2010, "TXM: Une plateforme logicielle open-source pour la textométrie – conception et développement". In Sergio Bolasco, Isabella Chiari, Luca Giuliano (eds.), *Proc. of 10th International Conference on the Statistical Analysis of Textual Data – JADT 2010*, Vol. 2, pp. 1021-1032, Rome, Edizioni Universitarie di Lettere Economia Diritto, https://halshs.archives-ouvertes.fr/halshs-00549779/fr/.

Hugo, V., The Project Gutenberg eBook of *Les contemplations*, v 2–2, Release Date: 29 August 2009, http://www.gutenberg.org/ebooks/29844.

Kalimeri, M., V. Constantoudis, C. Papadimitriou,K. Karamanos, F. Diakonos, H. Papageorgiou, "Entropy analysis of word-length series of natural language texts: Effects of text language and genre", *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, 22, 2014, https://www.researchgate.net/publication/259783179_Entropy_analysis_of_word-length_series_of_natural_language_texts_Effects_of_text_language_and_genre.

Kowalski, A. M. (ed.), *Concepts and Recent Advances in Generalized Information Measures and Statistics*, Bentham Science Publishers, 2013.

Marcus, S., *Poetica matematică*, Bucharest, Editura Academiei Republicii Socialiste România, 1970.

Mintz, L., D. Stefănescu, S. Feng, S. K. D'Mello, A. C. Graesser, "Automatic assessment of student reading comprehension from short summaries". In J. Stamper, Z. Pardos, M. Mavrikis and B. M. McLaren (eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (EDM 2014), pp. 333–334, 2014.

Onicescu, O., "Energie informationnelle", *Comptes Rendus Acad. Sci.* Paris, 263, 1966, 22, 841–842, cited in Marcus (1970).

Preda, V., S. Dedu, *Octav Onicescu – Omul şi opera. Restituiri: contribuţii la dezvoltarea cercetării economice Entropia informaţională în economie*, preliminary version, Bucharest, Academia română, Institutul Naţional de Cercetări Economice, SSN: 2285 – 7036 INCE – CIDE, 2015, http://www.studii-economice.ro/2015/seince151027.pdf.

Resnik, Ph., "Using information content to evaluate semantic similarity in a taxonomy", *JCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence*, Montreal, Quebec, Canada, 20–25 August 1995, Volume 1, San Francisco, Morgan Kaufmann Publishers Inc., pp. 448–453, 1995, https://arxiv.org/pdf/cmp-lg/9511007.pdf.

Rossetti, C. G., The Project Gutenberg eBook of *Goblin Market, The Prince's Progress, and Other Poems*, Release Date: October 26, 2005, http://www.gutenberg.org/ebooks/16950.

Sârbu, C., "Information Energy And Its Application", In Allen Kent, James G. Williams (eds.), *Encyclopedia of Computer Science and Technology*, Volume 41, Supplement 26, *Application*

BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)

of Bayesan Belief Networks to Highway Construction to Virtual Reality Software and Technology, pp. 67–81, CRC Press, 1999.
Shannon, C. E., "A Mathematical Theory of Communication", Reprinted with corrections from *The Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656, July, October 1948, http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.
Shannon, C. E., "Prediction and Entropy of Printed English", *The Bell System Technical Journal*, January 1951, https://www.princeton.edu/~wbialek/rome/refs/shannon_51.pdf.
Soofi, E. S., H. Zhao and D. L. Nazareth, "Information measures", In *Wires Computational Statistics*, Volume 2, Issue 1, January/February 2010, John Wiley & Sons, Inc., 2010, https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.62.
TXM User Manual, Version 0.7 A, LPHA, February 2018, http://textometrie.ens-lyon.fr/files/documentation/TXM%20Manual%200.7.pdf.

LEXICOMETRIC AND INFORMATIONAL MEASURES IN POLITICAL AND LITERARY CORPORA

Abstract

The paper presents a method for corpus-based informational analysis, using an open source platform for lexicometric processing and a spreadsheet application. This type of study may serve in illustrating the factors that influence informational measures such as entropy, energy and informativity, and in detecting certain structural or stylistic particularities of the analysed corpora.

MĂSURI LEXICOMETRICE ŞI INFORMAȚIONALE ÎN STUDIUL CORPUSURILOR POLITICE ŞI LITERARE

Rezumat

Articolul prezintă o metodă de analiză informațională a corpusurilor, folosind o platformă *open source* pentru procesarea lexicometrică şi un program de calcul tabelar. Rezultatele arată că un astfel de studiu poate fi util în ilustrarea factorilor care influențează măsurile informaționale de tip entropie, energie şi informativitate, şi în detectarea anumitor particularități de ordin structural sau stilistic ale corpusurilor examinate.

BDD-A32478 © 2020 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.25 (2026-11-07 19:48:46 UTC)