ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY (ROMJIST)
Volume **25**, Numbers 3–4, 2022, 275–289

A Comparative Lexical Analysis of Three Romanian Works – The Etymological Metalepsis Role and Etymological Indices

Horia Nicolai TEODORESCU[1, 2, *] and Speranta Cecilia BOLEA[3]

[1]Gheorghe Asachi Technical University of Iaşi, Romania
[2]Romanian Academy – Iaşi Branch, Romania
[3]Institute of Computer Science, Romanian Academy – Iaşi Branch, Romania
Email: hteodor@etti.tuiasi.ro*,
cecilia.bolea@iit.academiaromana-is.ro
* Corresponding author

Abstract. We introduce an etymological perspective in computational linguistic and apply it for the analysis of literary works with self-biographic content. Our hypothesis is that etymological mixtures used by an author may be a powerful stylistic device. We propose and investigate in the frame of computational linguistics the notion of "etymological metalepsis"; we argue that this device is used by some writers for conveying a sense of the historical frame that their works depict, the specificities of the local context, and even of the sentiments assumed and conveyed by the writer. Three works are analyzed and contrasted from the standpoint of the proposed stylistic device reflecting the etymological mixture used by the author, based on a semi-automated lexicographic analysis and the etymology of the most frequent words. Stylometric indices are suggested in relation with the etymological metalepsis. We also use tools of computational linguistics to elucidate and validate literary critical elements previously published on these oeuvres.

Key-words: Computational linguistic, etymology, style, stylistic device, stylometric indices, Zipf law.

1. Introduction

The "etymological metalepsis" device, according to our suggestion, is the mixing of loanwords able to convey to the reader subtle information without expressing it, based on the knowledge of the (cultivated or native) reader on the period when loanwords entered the language, the

regions where they are most used, the period when they have been more used, and the specificities of the persons (characters) that are probable to use those loanwords. The proposed "etymological device" works much as an allusion that may point to an epoch or to specificities of the characters. It also works as metalepsis – 'pointing to something by means of something else', by means of association – here the use of a rarely used word with a specific etymology pointing to an epoch when words with the same origin were more frequently used, or regions of a country where those loanwords are more frequently used, or characters who, by the use of such words, fall into a specific category or have specific traits (implicit social stereotype). The limits of etymological allusions / metalepses include the variable degree of associativity different readers may have and the difficulty of translation, because languages have very different histories of borrowing words and thus may be incongruent at the level of etymological nuances. For the use of the foreign readers of the article, the loans in the Romanian language, their occurrence frequencies and some other etymological specificities of Romanian are briefly described in the Annexes 1 and 2.

We conjectured in previous studies [1–5], devoted mainly to works that fall in the memoirs and self-biographic genres, that the style non-uniformities determine the segmentation of the text in its main parts and used these non-uniformities for automatic segmentation. In this research we suggest other means for the same purpose and add a new tool for the general stylistic analysis of literary works. The stylistic and lexicographic analyses are tools that might support the literary critics, and we explore also this avenue. Among others, we try to bring a statistical perspective to some opinions expressed by articles of literary critics regarding one of the discussed novels. This study adds to the tools previously presented in [1–5]. We suggest that the indexes of etymology of the words, especially in languages that mix various etymologies (as many European languages do) are an indicator of style. These stylistic indices would represent the ratio of less used words today vs. neologisms, of words of etymologies from superposed languages to the words of the root language, and the ratios of words with recent or medium historical etymologies (which are somewhat archaisms, or 'old' age words) vs. neologisms with specific etymologies.

The remaining part of the article develops linearly: In Section 2 the three works comparatively analyzed are presented, with hints to their specificities; Section 3 proposes the etymological metalepsis notion; Section 4 introduces the etymologic vectors and of the etymological metalepsis indices; Section 5 presents the method and basic results; Section 6 presents detailed results; the final Section is conclusive.

2. Brief presentation of the three works

The three analyzed works were written during the same period, between the two World Wars; Memories of War (MoW), in Romanian "Notiţe zilnice din războiu" (War Diary) by Averescu, was published in 1935 [6] (see the note at [6] for some uncertainties regarding the year of the initial publication), Under Three Kings (UTK), by Nicolae Iorga, in 1932 [7], and "At Medeleni" ("La Medeleni", LM), by Ionel Teodoreanu, was published in 1925 (the first volume) [8]. Also, the three works roughly belong to similar genres, with MoW and UTK being memories and self-biographic works, while LM is considered by many as a self-biographic novel. Because the three works were published in a 12-year span, they are supposed to use the same form of Romanian language. No literary critics have considered the Memoirs of Averescu a literary work; the volume by Iorga was analyzed by critics from the literary viewpoint (Călinescu [9]), although its literary value was found minor; the volume by Ionel Teodoreanu was unanimously classified as a literary work of interest. Therefore, one may expect that only Teodoreanu's volume is written

in a style pertaining to the literature class and thus can be assumed to extensively use literary devices. For more contextual information about the three authors and the analyzed works, see Annex 3.

Because the work "La Medeleni" plays a critical part in our analysis, we subsequently present it with some details. "La Medeleni" is a three-volume novel, authored and published in different years by Ionel Teodoreanu. While the three volumes have continuity, they recount different epochs in the life of the main characters, under different settings, and have a degree of independence. "La Medeleni" has several peculiarities, some of them briefly discussed below. We analyze only the first volume, which is also the most referred by the critics; when not explicitly mentioned else, by "work" we refer to this first volume, 'Hotarul nestatornic' (The Fickle Border), of "La Medeleni". "La Medeleni" continues the literary debut of Ionel Teodoreanu in 1923 with the volume 'Uliţa copilăriei', also related with his childhood and reflecting the author's nostalgies for a world perceived as better. The title relates to a village on the border of the river Prut (Iasi County), Romania. The novel has been extensively analyzed by successive generations of literary critics, with quite different results. It is strange that a novelist who was derided as minor by some major Romanian critics (Călinescu [9], Manolescu [10]) nevertheless received a huge attention from critics. C. Olteanu [11] cites eight literary critics who analyzed Ionel Teodoreanu's works, one of them (A.E. Constantinescu [12]) writing a full volume on Teodoreanu.

The use of stylistic tools in the study of "La Medeleni" is well suited because this author is better known for his stylistic prowess than for other narrative qualities. In fact, critics have emphasized the style role in the works of I. Teodoreanu. M. Munteanu, in her article "How Style Makes Space" [13] notices about him and two other authors that "*The question of "style" in the literature of Moldavian authors such as ... Al. O. Teodoreanu is tricky and complicated since it has to be confronted with another important issue, that of spatial identity. The sirens of the past, ... the aristocratic, now decadent, families lure these authors ...*" [13]. As C. Olteanu says in [11], Teodoreanu "*creates images of a surprising and sometimes useless metaphorism, rebuked by critics*" ("creează imagini de un metaforism epatant, inutil pe alocuri, taxat de critici") and remarks his narrative discourse, with imagery excesses blamed by some critics ("*discursului narativ, cu excesele sale imagistice, blamate de unii critici literari*" – the narrative discourse, with its imagistic excesses, blamed by several literary critics). C. Olteanu [11] cites M. Bucur [14] who notes "the prose [of Teodoreanu is], lyrical by excellence". G. Ibrăileanu, also cited by Olteanu [11], recollects that, when talking, Teodoreanu always made comparisons and invoked images ("*când vorbia, făcea numai comparaţii şi imagini*" – when he was speaking, he continuously used comparisons and images). These style peculiarities could be revealed by our analysis. For further details of the literary context, see Ivănescu [15]. In addition to these contemplations by other authors, we believe that I. Teodoreanu contrasts the space as perceived in childhood with the one perceived, at the same geographical location, in adulthood.

3. Explanatory for the etymological metalepsis notion

During its evolution, every language has borrowed words from neighboring populations, from migrants and from other civilizations during cultural and commercial exchanges. The loanwords may remain in marginal use, or as regionalisms, or may become archaisms and even disappear. Either as neologisms, regionalisms, archaisms, or words that have a definitive and general status in the language, the use of mixtures of words with different etymologies may reflect the region to which the speakers belong, their cultural peculiarities, or their linguistic preferences

expressed in a natural, unconscious way.

As already said, our hypothesis is that speakers, and especially writers may use varying etymologic mixtures with a purpose, to convey specific traits to the characters, to invoke a place or an epoch, or to indirectly create a specific atmosphere. If so, they use the etymological mixture with the role of evoking, allusion; that is, the variation of etymological proportions becomes a referral to a frame (temporal, regional) by something (the etymological mixture) very loosely related to it. Therefore, we infer, this is a form of metalepsis.

4. The etymologic vectors and of the etymological metalepsis indices

We define the 'etymologic contrast' of a text (mainly, literary works) by the weights of words of various etymologies. Specifically, the degree of the etymological contrast is given by the ratio of the number of words of different origins, weighted by the average ratio of the words of various origins in the literature.

We hypothesize that the 'etymologic contrast' may be used by some authors as a literary device. On the other hand, the excessive, inappropriate, or wrong use of words with specific etymologies may confuse the reader, create the feeling of an anachronist description, and may produce a sense of artificiality of the characters, of distortion, exaggeration, and unnaturalness of the work.

A basic etymologic index is defined as

$$I_e(N,T) = \frac{1}{N}\sum_{h\in L} n_{h,T} \tag{1}$$

where N is the number of most frequent words considered, except stopwords (in our case, $N = 200, 250$, or 300), *L=sl.,mgh.,bg.,sb.,...* (sl. – Old Slavic language(s); mgh. – Magyar; bg. – Bulgarian; sb. – Serbian; ...) is the set of indices of the languages from where the words originated, T is a text (work, spoken discourse), and $n_{h,T}$ is the (total) number of words from the 'h' language (with the index h) from T text, among the first N words,

$$n_{h,T} = count(w), w \in L \cap T, rank(w) \leq N. \tag{2}$$

An alternative etymologic index is defined at the level of distinct words instead of occurrences, that is, the total number of distinct words (or lemmas), $\nu_{wh,T}$, with roots in the foreign language h, among the first N most frequent words in T,

$$I_{ew}(N,T) = \frac{\sum_{h\in L}\nu_{wh,T}(N)}{N}. \tag{3}$$

A weighted word-level etymologic index is

$$I_{e0}(N,T) = \frac{\sum_{h\in L}\nu_{h,T}(N)}{\nu_{0,T}(N)} = \frac{\sum_{h\in L}\nu_{h,T}(N)}{N - \sum_{h\in L}\nu_{h,T}(N)}, \tag{4}$$

where $\nu_{h,T}(N)$ are the numbers of occurrences of all the loanwords from language h that are among the most frequent N words, and $\nu_{0,T}(N)$ is the total number of occurrences of words from the root language (here, Latin) that are among the most frequent N words. All these "indices"

are functions of N; one needs to carefully verify if they are constant for a text, or at least if they tend to a limit value when $N \to \infty$; similar requirements apply to other indices used in computational linguistics. For computations, see Section Results. The etymologic vector is defined as $V_{e0}(N) = (n_{10}, \cdots, n_{l0})$. Both the etymologic vector and index depend on the number N of the most frequent words considered. A similar definition is obtained at occurrence level.

The departure of $I_{e0}(N,T)$ from the 'normal' etymological index (for a certain language) may be indicative for the use of the etymologic metalepsis. (It is also interesting to determine how many hapax legomena are from the set of languages considered, but we have not done it. However, it could add to the analysis power.)

A convenient and natural way to define a compressed index is the entropy of the etymological mixture, either at the level of occurrences, or at the level of words; in the second case it may be defined as

$$E_{w,T}(N) = -\frac{\nu_{w0,T}(N)}{N} \ln \frac{\nu_{w0,T}(N)}{N} - \sum_{h \in L} -\frac{\nu_{wh,T}(N)}{N} \ln \frac{\nu_{wh,T}(N)}{N}, \tag{5}$$

where $\nu_{w0,T}(N) = N - \sum_{h \in L} \nu_{wh,T}(N)$ is the number of words from the root language among the most frequent N words, $\nu_{wh,T}(N)$ are the numbers of words with roots in the language h; the ratios $\frac{\nu_{wh,T}(N)}{N}$ are estimations of the probabilities. We suggest that the last index and its extension for occurrence-level entropy could serve as etymological metalepsis indices because they contextualize the information in various families of words.

5. Analysis method and basic results

The processing method used is shown in Figure 1. The orthographic differences between the then rules and the current rules were removed for the computer analysis such that the parsers can correctly interpret the works. The TXT files resulted after corrections were parsed with TTL parser [16]. In the analysis, we removed the stopwords, whereby stopwords we considered: adverbs, adpositions, articles, conjunctions, numerals, residuals (foreign words or symbols), abbreviations, interjections, particles, auxiliary verbs. The definitions of these parts of speech were made according to the TTL parser. There are two reasons of including the adverbs in the stopwords list: first, most of them are of Latin origin and would contribute to masking the words of non-Latin etymology among the most frequent words; second, many adverbs have the same form as the corresponding adjective (e.g., "frumos"), increasing the count for the respective word and further masking the words of non-Latin etymology among the most frequent words. Yet, including the adverbs in the stopwords list remains a disputable decision.

After deriving the most frequent words in the three works, the etymology of the first 250 words for UTK and MoW and of the first 300 words for LM was determined. The summary of the etymological data is given in Table 1. Details on the analysis are given in Section 6.

6. A basic lexicographic analysis – "La Medeleni"

Partly, "*La Medeleni*" is a dream novel; it is enchanting and it is a sentimental indirect if not actual self-biography. The content is reverie more than action, although it seems full of dialogue

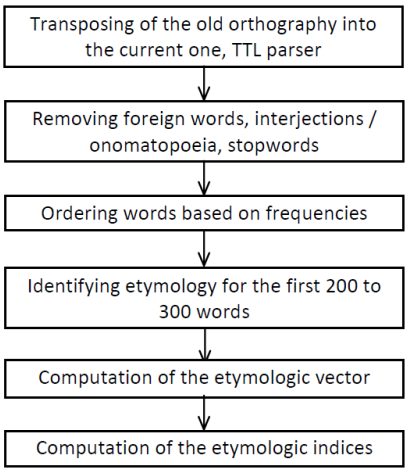

Fig. 1. Scheme of the operations in the analysis / study of the etymologic vectors and of the etymological metalepsis index.

Table 1. Etymological data

up to N = 250 or 300	alb. cognate	sl.	mghr.	bg.	srb.	ge.	tc.	gr./ngr.	fr.	unkn.	onomato-opoeia	ukr.	pol.
MoW, number of words among the first 250 most frequent	-	7	2	-	-	-	-	1	1	-	-	-	-
MoW, total number of occurrences among the first 250 most frequent	-	191	25	-	-	-	-	7	16	-	-	-	-
UTK, number of words among the first 250 most frequent	-	7	4	1	-	-	-	2	1	-	-	-	-
UTK, total number of occurrences among the first 250 most frequent	-	644	217	47	-	-	-	90	106	-	-	-	-
LM, number of words among the first 300	5	29	2	3	1	-	3	5	-	3	2	1	1
LM, total number of occurrences among the first 300	506	1,208	72	112	59	-	118	255	-	105	92	28	22
LM, number of words among the first 250	4	23	2	3	1	-	2	4	-	2	2	1	-
LM, total number of occurrences among the first 250	561	1,198	83	130	71	-	118	291	-	94	102	30	-

and action – actually, the dialogues seem to be heard in the mind and heart of the author, behind a veil, not in the actual world. The charm of this novel may come in part from the skillful mixing of neologisms of Latin provenience (mainly from French) with less used Romanian words of borrowed provenience, recalling past centuries when Romania was more cosmopolite; hence the complex etymology – Turkish, Ukrainian, Magyar, German etc. This makes this novel extremely difficult to translate, because languages have very different structures of strata of loanwords, which makes the use of the etymological device almost impossible to convert to an equivalent in another language.

This specific use of the vocabulary in LM is reflected in and proved by the statistics as subsequently discussed; the statistics also show the lexical differences between the three discussed works. The general statistic is shown in Tables 2 and 3. The number of paragraphs (including in dialogues) per page is larger in the first part of the volume 1 of "La Medeleni" than in the second part.

Table 2. Sentence and paragraph level statistics

Name of the work sections (in Romanian)			Avg. no. of para. per page	Avg. no. of the sentences	STDEV no. of para. per page	STDEV length of the sentences
LM (vol.1)	Part #1	I Potemkin şi Kami-Mura	34.64	6.48	6.47	5.84
		II Căsuţa Albă şi Rochiţa Roşie	39.21	5.94	10.01	4.72
		III Herr Direktor	35.93	6.51	6.75	5.84
		Part #1	**36.07**	**6.37**	**7.60**	**5.62**
	Part #2	Mediul Moldovenesc	32.13	7.01	10.06	6.60
		II Robinson Crusoe	30.20	7.74	11.21	7.08
		III Papuşa Monicăi	29.00	7.49	7.74	5.95
		IV Moş Gheorghe, Nu Tragi din Lulea	28.09	8.20	10.82	6.93
		Part #2	**30.44**	**7.43**	**9.92**	**6.63**
MoW	Part #1	21 VII 1916 – 4 XII 1916	20.79	15.87	4.19	11.28
	Part #2	6 XII 1916 – 22 XI 1917	23.14	14.64	4.48	10.54
	Part #3	26 XI 1917 – 5 III 1918	23.33	14.59	5.22	10.04
UTK	Part #1	I Factorii	5.43	38.03	1.93	26.45
	Part #2	II Conflictul	4.97	30.25	1.46	22.19
	Part #3	III Cele două lumi din nou faţă în faţă	5.36	30.77	2.13	22.41

* Avg. = average; STDEV = standard deviation; no. = number; para. = paragraphs

Table 3. Word general statistics

	Total number of words	Number of independent words	Average number of repetitions
MoW	11,138	2,028	5.49
UTK	75,642	9,223	8.20
LM (vol. 1)	45,726	6,017	7.60

When counting the number of sentences based on the TTL parser, the results were manually checked and the parser errors were removed. Two false sentences as determined by the parser in MoW and 106 false sentences in LM were found to consist only of punctuation signs; they have not been included in the counts in Table 2.

The number of independent words appears to linearly increase with the total number of words in the text, $y = 0.1116x + 824.95$, with a very good fit ($R^2 = 0.9996$), see Figure 2.

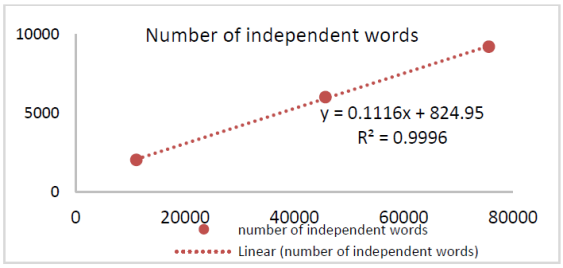

Fig. 2. The rank distributions of number of independent words (Table 3).

The variations of the number of paragraphs per section of 20 pages and of the number of paragraphs per page are shown in Figure 3.

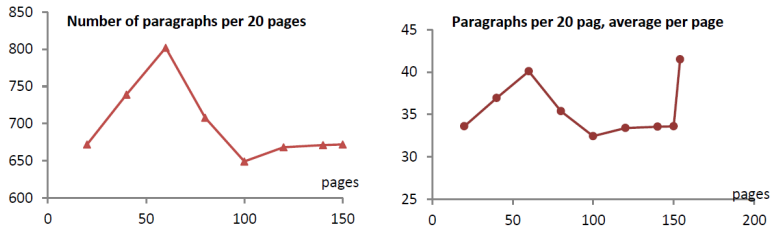

Fig. 3. Moving average over 20 pages in LM.

Remarkably, in "La Medeleni", there is a large number of rather rare loanwords; some have Turkish roots (e.g., "duduca", which has Persian origin and was borrowed in Turkish, from where in Romanian); others have Hungarian roots ("făgădui"– mgh. fogadni – 6 occurrences, used instead of the synonym with Latin root "promite", promis – used 4 times; "mohorâ" – 3 times), or Ukrainian origin ("bihuncă" – 14 times). Also, there is a significant number of words with unknown etymology, such as "cocon" (in forms more specific to the Easter region such as "conaş", "conul"), with Bulgarian or Serbian origin ("năduşi" – 5 times), and a few have Russian roots ("scripca", violin – once, instead of the usual word for violin, "vioara"). Notice that in some circumstances the use of lemma analysis instead of word analysis may hide fain details of the style. For example, "duduca" is used as a sign of respect for a young woman; the diminutive of duduca, "duduiţa", which is found 96 times in LM, was used to show a respectful attachment by an adult for a girl of higher social status. In contrast, the two forms of plural in use, "duduci/dudui", are rather pejorative. In LM, all three forms appear, with the respective meanings; we counted apart these forms.

The extensive use of words borrowed from other languages in a distant past instead of using neologisms much more frequent today is a device to indicate past times and enforce the sentiment of melancholy of timings passing and of childhood times. E.g., mixing "dulap" and "şifoniera" (armoire, wardrobe) in the same paragraph ("şi că-n dulapul rufelor . . . şi dacă ar mai fi deschis şi şifoniera. . . "), Teodoreanu skillfully jumps from the use of archaic or rarely used regionalisms to modern synonyms, depending on the speaker' (character's) sentiment. Also, in [8] the obsessive use of the Turkish word "geamandan" (24 occurrences) for baggage, without a single use of the more frequently employed neologism "valiza", underlines the sentiment of time past and, at the same time, underlines the perpetual and seemingly unavoidable move, spatially and through life, of the characters.

After this incursion hinting to the literary effects that may not be apparent from the statistics, we proceed to the analysis. We compared the etymologic mixture in the three works. The method is as follows: the most frequent words (not lemma) were selected, and their etymology determined. Then, the words with various etymologies are counted, excluding the words of Latin origin, as well and those (very frequent) of Italian or French origin (again, we concede that that may be a disputable decision). The results are shown in Table 1 and in Figures 4 and 5.

Notice in Figure 5 the much larger numbers of the loanwords, and the much larger numbers of their occurrences for LM ("La Medeleni"), compared with MoW and UTK.

Figure 5 is a representation of the sum in the index $I_{eW}(N,T) = \frac{\sum_{h\in L}\nu_{wh,T}(N)}{N}$. Because we used a fixed N, the index value for each language is the value on the graph divided by the

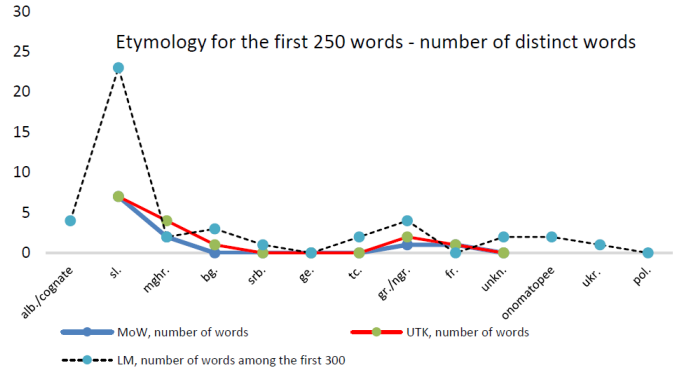

Fig. 4. Etymology for the most frequent 250 words (representing the numbers of distinct words), in the three works, except for LM where the first 300 words are considered.

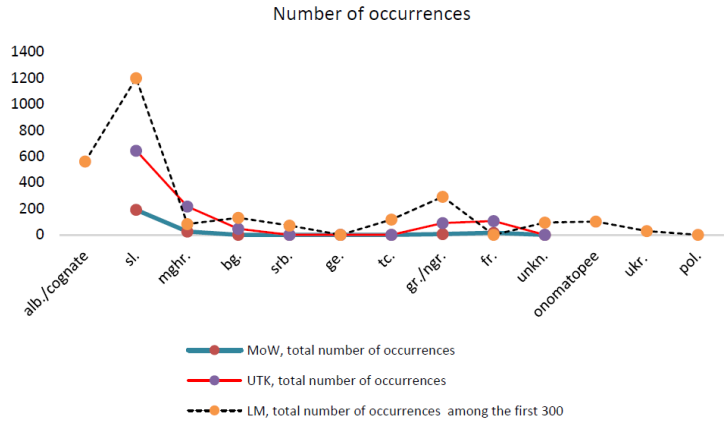

Fig. 5. Number of occurrences of words with various etymologies in the three analyzed works. For MoW and UTK, the first 250 words were considered; for LM the first 300.

respective N.

Interestingly, the number of occurrences of words with a specified etymology increases almost linearly ($R^2 > 0.8$) with the number of the most frequent words among which we look for, see Figure 6. This fact is surprising; we do not have an explanation for it. Indeed, assuming a uniform (totally random) mixing of words with different etymologies, Zipf's law should be valid for words of a specified etymology, as it is for the mixing. Then, the number of words $N[r]$ of rank r from one etymologic subgroup should be described by $N[r] = \frac{A}{r^\alpha}$, where A is a constant (A not depending on the etymology, in case of uniform mixing) and typically $\alpha = 1$, also not depending on the etymology. If that were true, by summing the number of occurrences of words with a specified etymology one should obtain the so-called harmonic numbers, $H_n = \sum_{r=1}^{n} \frac{A}{r} \approx \ln n + \gamma$, with $\gamma \approx 0.58$, (see for example [17]). If the etymology mixing were uniform, meaning that the ratio of the core etymology (Latin, for Romanian) and the specified etymology words is a constant $\lambda_l ang = \lim_{n\to\infty} \frac{H_n(lang)}{H_n(total)} \approx \lim_{n\to\infty} \frac{H_n(lang)}{H_n(core)} = \Psi$, then $H_n(lang) \approx \Psi H_n(total) = \Psi(\ln n + \gamma)$. This leads to $H_n(lang) \approx \ln n^\Psi + \Psi\gamma$, not

to a linear law. The results in Figure 6 contradict either the simplistic model of uniform mixing of etymologies, or the usual Zipf's rank statistic ($\frac{1}{r}$), indicating that the use of loanwords is not random, but a stylistic device. As the three works analyzed have very different manners of using loanwords, one can conclude that the use of this stylistic device is a literary or personal trait of the author – at least in a given work. On the other hand, as intuitively expected, the slope of the lines is larger for the groups of words of the etymology having the larger density in the language in the given text.

We may ask if the A/r^2 is not a better model. It is not, as $\sum_{r=1}^{\infty} \frac{1}{r^2} \to constant(1.645)$, see [17]. Because the linear model empirically determined is between these cases, we derive that there is some model different from $\lim_{n\to\infty} \frac{H_n(lang)}{H_n(core)} = \psi$ that explains this linguistic device. On the other hand, we can use a continuum representation $f(r) = r^{-u}$ and integrate $y = \int_1^x r^{-u}\, dr$, for $y(x)$ to be linear, one needs that $u = 0$. From the above, one derives that a small absolute value of the power u for the loanwords, when combined to a lower value of this slope compared with the absolute value of the slope for the entire vocabulary in the work, could provide a partial explanation the observed behavior. The departure of the specific sub-vocabularies from the standard Zipf's law is known; for example [18] evidenced such departures for patents. We find the slope of the rank distribution in the double logarithmic plane to be -0.58, which is sensibly lower than the slope for the entire vocabulary for LM, -0.89, see Figure 7. This may provide a partial explanation for the empirical results.

A further approach is to use more elaborate distributions instead of Zipf's power (law) distribution, for example the general Zipf–Mandelbrot distribution, used in biology for characterizing the abundance of species – a formally similar problem to the discussed one; here, the 'species' are etymological classes; this distribution is [19] $p(r) = \frac{a}{(1+br)^c}$, with a, b, c parameters; with suitable constants (e.g., $a = 6, b = 20, u = 0.5$), this distribution is able to approximate reasonably well a line segment between $r = 50$ and $r = 300$.

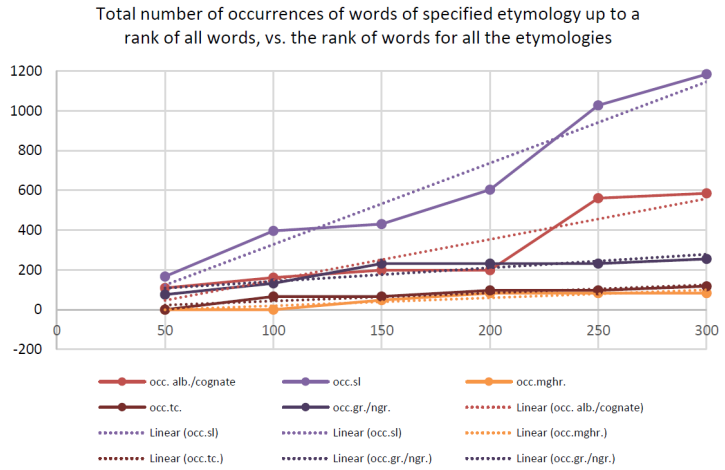

Fig. 6. Relationship between the sum of the number of occurrences of the loanwords with specified etymology and the lowest rank of words of that etymology considered in the total. $y = 4.09x - 80.73$, $R^2 = 0.94$ for sl.; $y = 2.05x - 55.67$, $R^2 = 0.81$ for alb./cognate; $y = 0.681x + 74.13$, $R^2 = 0.79$ for gr./ngr.; $y = 0.41x + 2.6$, $R^2 = 0.85$ for tc.; $y = 0.399x - 20.4$, $R^2 = 0.84$ for mgh.

Notice that Figure 6 is a representation of the sum $\sum_{h\in L} n_{h,T}$ in the basic etymologic index, $I_e(N,T) = \frac{1}{N}\sum_{h\in L} n_{h,T}$, where N is the number of most frequent words considered. Because $\sum_{h\in L} n_{h,T}$ seems to vary linearly in N, the slope is of interest; hence we used the normalization $\frac{1}{N}$ in the definition of the index.

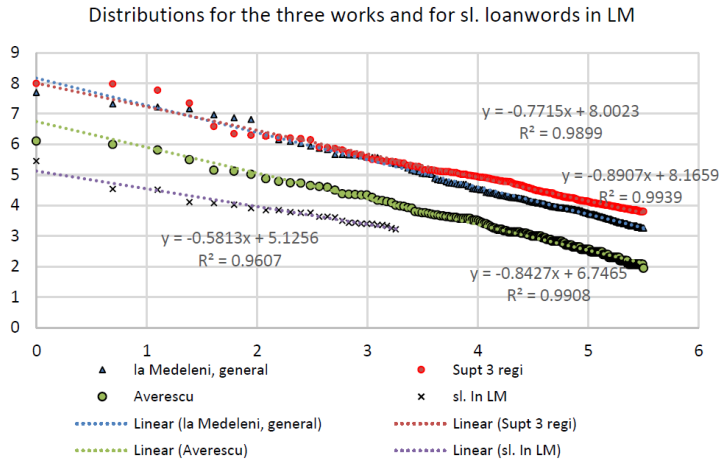

Fig. 7. Rank statistics for the three works and for the words with Slavic roots in LM (for the first 300 most frequent words).

One may consider that the results show that, in "La Medeleni", the ineffable is shared indirectly, by the abundance of the loanwords of various etymologies and by their contrast with neologisms. Also, the results, corroborated with the literary critical opinions in the literature, indicate that the skillful use of loanwords indeed operates as a stylistic device, for which we suggested the term *etymological metalepsis*.

The indices $I_{ew}(N,T) = \frac{\sum_{h\in L}\nu_{wh,T}(N)}{N}$, with $N = 250$, for the three works, as well as for $N = 300$ for LM, are given in Table 4. The values show that LM is much richer in frequently used loanwords than the other two works.

Table 4. The indices $I_{eW}(N,T)$ for MoW, UTK and LM

up to N = 250 or 300	alb. cognate	sl.	mghr.	bg.	srb.	ge.	tc.	gr./ngr.	fr.	unkn.	onomato-opoeia	ukr.	pol.
MoW, $N = 250$ $I_{ew}(N,T)$	-	7/250	2/250	-	-	-	-	1/250	1/250	-	-	-	-
UTK, $N = 250$ $I_{ew}(N,T)$	-	7/250	4/250	1/250	-	-	-	2/250	1/250	-	-	-	-
LM, $N = 300$ $I_{ew}(N,T)$	5/300	29/300	2/300	3/300	1/300	-	3/300	5/300	-	3/300	2/300	1/300	1/300
LM, $N = 250$ $I_{ew}(N,T)$	4/250	23/250	2/250	3/250	1/250	-	2/250	4/250	-	2/250	2/250	1/250	-

The entropy of the etymological mixture at the level of words is defined as

$$E_0(N,T) = \sum_{h,T} -\frac{\sum_{h\in L}\nu_{h,T}(N)}{\nu_{0,T}(N)} \ln \frac{\nu_{wh,T}(N)}{\nu_{0,T}(N)}. \tag{6}$$

For the three works, the entropy values are given in Table 5. The etymologic word entropy for LM is more than the double of the entropy for the other two works.

Table 5. Word-level entropies for the three works

Work, no. of most frequent words, N = 250	Entropy
MoW	0.204
UTK	0.285
LM	**0.763**

7. Conclusions

We hypothesized that the specific etymological mixture used by authors in their literary works may play a distinct stylistic role and named the hypothetical stylistic device "etymological metalepsis". For verifying our hypothesis, we compared three works at the stylistic level and found that the etymological mixture largely differs between these works, moreover, it varies along each work. We proposed a set of stylistic indexes based on the etymology mixture of the most frequent words and determined its value for the three oeuvres.

Each language, as used by a writer or speaker, has its own texture, as represented by the several layers of words included during the mixture of populations along the formation and evolution of the nation, or borrowed from neighbors or regional cultural radiating nations (such as France, Germany, and Austria in Central Europe). The use of the words from different layers is not uniform among speakers and writers. It is not uniform from work to work of the same writer either – the skilled, talented writers adapt the use of words to the topic, place, and epoch when the action develops. "La Medeleni" is highly specific in the use of the lexical texture to evoke bygone epochs, melancholy, sadness, and the change in generations of people, in making obvious the linguistic contrast between countryside and cities.

While we are convinced that some form of etymological metalepsis may have been used by authors in many languages, we are aware that the method of studying this stylistic feature should be very much dependent on the language. We argue that tools – such as etymological online dictionaries suitable for automatic use – have to be developed for the proposed analysis method to become easy to use.

Among various avenues to pursue, future work should detail the analysis in this paper; as a minimum, future work should comparatively deal with the comparison of the results in this paper and the results obtained excluding adverbs from the stopword list; moreover; one should extend the analysis to at least 500 of the most frequent words in the discussed works. Other etymological indices present themselves, but their usefulness should be investigated.

Acknowledgement. HNT thanks Prof. Mike Teodorescu for several ideas provided for this study, including the very basic idea of it. We thank Dr. Gabriela Haja for encouragements and advices and the referees for valuable corrections.

Authors' contributions: HNT proposed the indices, performed the etymological analysis and wrote most of the paper with feedback from SCB. SCB wrote the code for part of the analysis related to parsing, extracted the vocabulary lists, determined most of the descriptive statistics of paragraphs and sentences, performed all the operations related to parsing, and obtained the related software-based analysis of the stylometric indices, contributing Tables 2 and 3.

References

[1] H.-N. TEODORESCU and S. C. BOLEA, *Stylometric and topic analysis of a historical text*, Romanian Journal of Information Science and Technology (ROMJIST), **21**(2), pp. 99–113, 2018.

[2] H.-N. TEODORESCU and S. C. BOLEA, *Text sectioning based on stylometric distances*, The 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timisoara, Romania, 10-12 Oct., pp. 1–6, 2019.

[3] S. C. BOLEA and H.-N.TEODOPRESCU, *Automatic segmentation of texts based on stylistic features*, Proceedings of the Romanian Academy Series A, **21**(3), pp. 283–292, 2020.

[4] H.-N. TEODORESCU and S. C. BOLEA, *Automatic segmentation of texts based on stylistic features*, The 11th Conference on Speech Technology and Human-Computer Dialogue (SpeD), 13-15 Oct., Bucharest, Romania, pp. 1–6, 2021, [Online]. Available:doi: 10.1109/SpeD53181.2021.9587362.

[5] S. C. BOLEA, *Implementation of an algorithm for automatic segmentation of texts based on stylometric analysis*, 7th International Symposium on Electrical and Electronics Engineering (ISEEE), 28-30 Oct., Galati, Romania, 2021.

[6] A. AVERESCU, *Notiţe Zilnice din Războiu (1916-1918)*, Bucureşti: Cultura Naţională Publishers, 1935. Both volumes are said to be published in 1937. In fact, Averescu used his book as a political and self-promotion instrument, republishing it as "first edition" at several publishing houses. The Preface is dated 1935, July 1., [Online]. Available: http://enciclopediaromaniei.ro/wiki/Alexandru_Averescu.

[7] N. IORGA, *Supt trei regi, Istorie a unei lupte pentru un ideal moral şi naţional*, Ediţia a II-a,Bucureşti, 1932.

[8] I. TEODOREANU, *La Medeleni: Hotarul nestatornic (partea întâia)*, 1925.

[9] G. CĂLINESCU, *Istoria literaturii române: de la origini până în prezent: Ionel Teodoreanu*, Editura Paideia, Bucharest, Romania, pages 94, 2014.

[10] N. MANOLESCU, *Istoria critică a Limbii Române, 5 secole de literatură*, Editura Paralele 45, Bucharest, pages 1506, 2008.

[11] C. OLTEANU, *The historical and literary context of the emergence of Ionel Teodoreanu in The Romanian Literary Scene*, Journal of Romanian Literary Studies, 16 , pp. 1288–1296, 2019, [Online]. Available: http://old.upm.ro/jrls/JRLS-16/Rls

[12] A. E. CONSTANTINESCU, *Ionel Teodoreanu şi medelenismul*, Bucharest, Romania, pages 49, 2016.

[13] M. MUNTEANU, *How style makes space. Reflections on the forms of life in the Literature of Viaţa Românească circle*, Dacoromania Litteraria, IV, pp. 45–59, 2017.

[14] M. BUCUR, *Viaţa românească*,Anul XVI, nr. 6-7, iunie-iulie, pages 391, 1963.

[15] G. IVĂNESCU, *Istoria limbii române*, Iaşi, Junimea, Romania, pages 766, 1980.

[16] *TTL Parser – RACAI*, [Online]. Available: http://www.racai.ro/en/tools/text/.

[17] *Wolfram—Alpha*, Accessed April 4, 2022, [Online]. Available: wolframalpha.com; https://www.wolframalpha.com/input/?i=sum+1

[18] M. TEODORESCU, *Machine learning methods for strategy research*, Harvard Business School Research Paper Series No. 18-011, October 14, 2017, Accessed April 12, 2022. Available at SSRN: https://ssrn.com/abstract=3012524 or http://dx.doi.org/10.2139/ssrn.3012524.

[19] M. A. MONTEMURRO, *Machine learning methods for strategy research*, 9 Jul 2001, [Online]. Available: arXiv:cond-mat/0104066v2 [cond-mat.stat-mech].

Annex 1.

The Romanian language includes, beyond the words with a Latin root, which constitute the basis of the language, words that are (lexical) cognates with words from languages in the region, or borrowed form a variety of languages in the region, comprising old Slavic language(s), Albanese, Magyar, Bulgarian, Greek and Neo-Greek, Serbian, Ukrainian, Turkish, Russian, German, and French. The order of listing approximately reflects the order (period) of borrowings. Most of the words that were previously assumed to be of Albanese origin are now believed to be cognates, some of them of late-Latin origin. Magyar came since the arrival of the Magyar population in Pannonia in the 9^{th} century; most probably, the borrowings are from the 12^{th} to 16^{th} centuries. Old Slavic words were included in Romanian during the admixing of the original population and the Slaves, starting the early and middle of the VI century until the VIII century. Some Slavic words may have come through the Orthodox Church. German loanwords are related to the population arrived during the XIII-XIV centuries in Transylvania, close to the border with the other two Romanian provinces. Greek and neo-Greek loanwords came through the Orthodox Church but mostly through the reign of Phanariots during the XVII to XIX centuries. Bulgarian and Serbian loanwords may have penetrated since the XII century, when the Bulgarian and Serbian kingdoms were important powers in the Balkans, though the Church especially Serbian, and through continuous exchanges. Similar entry ports may have had Ukrainian and Russian words. French loanwords came mostly during the XIX and XX centuries. Romanian language has a considerable number of words of unknown or unsure provenience and a large number of onomatopoeias.

Annex 2. Comparison of the vocabulary from MoW, UTK, LM

Table A2. Comparison of the vocabulary of the three works: most frequent 31 words (except stop-words)*

Nr crt.	LM	MoW	UTK
1	da+nut+	se	care
2	olgut+a	care	se
3	se	este	era
4	-i	s-	el
5	monica	ce	o
6	eu	generalul	ce
7	ce	i-	s-
8	deleanu	mi-	lor
9	tu	spus	putea
10	o	fost	i
11	mos+	armatei	ei
12	gheorghe	mi	lui
13	care	fi	le
14	te	armata	-i
15	doamna	general	e
16	el	ma+	toate
17	herr	fa+cut	orice
18	-l	face	regele
19	direktor	situat+ia	bra+tianu
20	-s+i	sunt	-l
21	ma+	mine	mare
22	era	regele	-s+i
23	i	divizia	aceasta+
24	ochii	poate	fost
25	i+l	avut	erau
26	e	ra+spuns	poate
27	i+i	ne	sa
28	ei	noi	sale
29	-t+i	aceasta+	acest
30	i+s+i	mea	fi
31	papa	m-	face

* All words, including names, are written without capitals. The sign "+" denotes diacritics (for example "i" is represented as i+. Notice that except "se", the lists of the most frequent words differ; MoW and UTK are more similar.

Annex 3. Explanatory remarks on the three authors and their works

The remarks in this Annex may be useful for the readers not familiar with the Romanian literature.

Both Alexandru Averescu and Nicolae Iorga have been politicians, at least in their second part of the life; they established their own parties, with contested and disputable programs; both used to be ministers and prime ministers. Their politics was right-leaning, although in different ways. Averescu was initially populist, but he is known to have inclined to the extreme right for some part of his political career. He has made blunders in his external and internal politics and was involved in several political scandals. His sole work is MoW. Iorga was highly traditionalist, but he strongly opposed extreme right currents such as fascism; he was killed by them. Iorga was an internationally renowned historian of his time and a prolific writer of history volumes, with an inclination toward literary works. As a politician, he is known to have made disputable decisions. Both MoW and UTK may have been written as tools for the political career of their authors. Iorga's work explicitly states that in the subtitle: The full title of *Under Three Kings* (UTK), in Romanian, is "Supt trei regi: Istorie a unei lupte pentru un ideal moral şi naţional", translation: "Under Three Kings: A History of a Fight for a Moral and National Ideal".

Regarding Ionel Teodoreanu (1897-1954): He should not to be confused with his brother, Al. Teodoreanu, also a writer.

For the foreign readers: The word "dudui", the plural of "duduca", should not be confused with the verb "[a] dudui", which is onomatopoeic and means "making a low frequency, repeatable noise" (no equivalent in English).