STATISTICAL METHODS APPLIED IN A CROSS-DISCIPLINARY MANNER IN ENUMERATIVE OR NON-NUMERICAL ATTRIBUTIVE SERIES. STATISTICAL FREQUENCY ANALYSES IN PHILOLOGICAL (MINI)CORPORA

Profesor Gheorghe Săvoiu, PhD., D.H.,
University of Piteşti
Associate professor Constantin Manea, PhD., D.H.,
University of Piteşti

Abstract

One of the simplest series of statistical data, emerged as a result of the grouping manner of processing linguistic or philological reality, and scientifically in-depth searched, by the name of either non-numerical attributive series, or enumerative series, is subjected, in this paper, to the specific investigation of early statistical thinking, respectively the analysis of the frequencies of occurrence of specific words or expressions. The method of frequency analysis in philological (mini)corpora can thus become, through its cross-disciplinary applicability, a useful validation method in modern linguistics, standing the chance to offer rigorous criteria selection solutions or pertinent arguments, extracted from the use of language in situations of ambiguity, and even of selective uncertainty. The series type investigated in this article, a dominant one in the universe of philology or linguistics, acquires a special utility that can be ensured by rapid processing using the method or frequency analysis. The frequency method allows for prompt decisions in the context of philological or linguistic uncertainty, by providing various statistical indicators capable of giving truth value to the evidence-based argumentation focused on language use, by statically, dynamically, spatially and structurally developing quantification in philological (mini)corpora of emergence frequencies, by means of confronting options, evaluating concentrations or diversifications, and finally even by capitalizing on the statistical profile in modern linguistics. The authors offer a number of pertinent examples of the usefulness of the cross-disciplinary method, which, in this case, aims at simplifying the decision, both in traditional philology and in modern linguistics, by capitalizing on past statistical information, in order to infer the actual use and ussage trends in classical or modern science language, in a more accurate way. For the sake of applicability, the modern philological

Internet-type (mini)corpus, accessed with search engines, becomes the statistical population observed and processed, and the tabular presentation and the adequate graphic representation draw the quantitative image of the evidence. Finally, the authors anticipate the evolution of modern language towards a significant liquidity, or even towards an excessive volatility, and by analyzing the data flows of a number of enumerative series, the paper treats and statistically identifies solutions emphasizing the major impact of cross-disciplinary methods in any philological or linguistic scientific research.

Keywords: cross-disciplinarity, statistical method, enumerative series, homograde or attribute series, frequency analysis, philological or linguistic (mini)corpus, concentration–diversification indicators, statistical profile method, linguistic usage.

JEL codes: C46, C49.

1. **Introduction**

Theoretical statistical methods are constantly expanding their practical applicability and thus become increasingly useful in various fields of scientific research, outlining an obvious cross-discipline character. As conceived and described in this paper, cross-disciplinarity *"is an approach that selects, combines, associates, aggregates, applies single methods to various scientific realities, or puts into practice well-defined disciplinary methods to one science within the methodological body of others, thus becoming a generic concept in the creativity of the investigative or methodological approach, and is, finally, a first step towards the emergence and delimitation of new disciplines or sciences*" (Săvoiu, et al., 2020, p. 8).

Modern cross-disciplinarity offers complex solutions through the originality of its specific approaches, or through the creativity of their transposition into other sciences, starting from the simple finding that it simplifies a classic approach, which has become much too usual and sometimes increasingly inefficient, compared to evolving reality. In the investigations and arguments of the researchers facing increasingly varied and complicated phenomena, from sociology to philology, biology, demography, etc., the only option is to transform the isolating vision, long confirmed unidisciplinarily, into a team investigation of a cross-disciplinary type, by resorting to various methods, including, the statistical, mathematical and physical ones, which appear as a priority. Thus, a lot of the linguistic issues that seem rather difficult to sort out can be solved by resorting to the simplest statistical methods. Text analysis based on statistical recurrences sometimes becomes a matter of life and death, as any student taking his/her first English lessons could learn today from an amazing linguist, who is both a researcher and writer, David Crystal, who exemplifies the expression of *a matter of life and death*, in his famous book *The English Language: A Guided Tour of the Language*, where he describes how the same student attracted by the correct learning of English could escape from a serious charge of plagiarism or punishment by hanging, offering as life-saving evidence a profound analysis – simultaneously linguistic and statistical – of a letter belonging to an actual criminal, which was falsely attributed to the said student, following only the frequency in use of a few expressions, structures and phrases, a certain style or the distinctive colouring of language… (Crystal, 2002).

This article is the result of a common search by the authors, arising from the desire to simplify the decisions related to the correct philological or linguistic use of language, and to validate or invalidate various hypotheses of frequency analysis, or single out, as useful indicators, data on usage frequency for several terms that are sensitive in terms of linguistic standardization or normalization. The passion for cross-discipline investigation of some recurrences and linguistic convergences, the team spirit and the desire to apply statistical methods in different fields, as investigation processes influenced by space, time and structure (Manea, Săvoiu, 2014), generated two periods of investigations by accessing the net, in 2013-2014 and in 2019-2020, respectively.

2. Review of the specialized literature

In Romanian philological research, be it older or more recent, starting from, and based on, the statistical method of frequency analyses, the first studies set out either from dictionaries of the Romanian language, or from individual texts, chosen in such a way as to have a certain, requisite degree of representativeness for the Romanian language. The first philologist concerned with the application of statistical quantifications, a genuine pioneer in the field of ensuring a method or evidence-based probative analysis in the study of the etymological structure of the Romanian lexicon, was A. de Cihac. Thus, this evaluator made a false statistic by consulting dictionaries and glossaries (in an inaccurate methodological manner), as those books had been specially selected to obtain certain results, without a scientific justification regarding the complete list from which the unrepresentative sampling was performed for the period 1870-1879, as well as the very selection method; the latter was obviously guided or directed in a dedicated manner, and implicitly subjective, as literally and self-confessedly admitted in the preface to the second volume of his etymological dictionary, entitled *Dictionnaire d'*étymologie daco-romane. Éléments slaves, magyars, turcs, grecs-modernes *et albanais*, Francfort s/M, which was published in 1879, and virtually represents the first scientific dictionary of the Romanian language. A. de Cihac finally made an approximate assessment of some relative frequencies, taking into account the words defined as "not derived" in the dictionary he compiled, without dwelling on the lexical units of the Romanian language itself. Starting from the "*approximately 500 Latin words, 1,000 Slavic words, 300 Turkish words, 280 modern Greek words and 20 to 25 Hungarian or Albanian words*" (Dimitriu, 1973, p. XIII), A. de Cihac erroneously appreciated that the Latin element "*which undoubtedly constitutes the substance of the Romanian language [...] not only remained almost stationary, after being received, as far as the basic vocabulary stock is concerned, but the latter must even have lost many words as a result of so many troubles for which these unfortunate territories have been the scene for many centuries*" (Dimitriu, 1973, p. VIII). The conclusions of A. de Cihac's approximate and guided or directed statistical evaluations diminish the place and role of the Latin element inherited by the Romanian language vocabulary, judging it subjectively. Unfortunately, the reality of language facts was seriously marred by the data provided, through such quantifications, by A. de Cihac, as well as the one in which Sextil Puşcariu justified the former one, using percentages that were rather similar, and revalidated it in 1920. The relative frequencies calculated, listed and presented there seemed to indicate that the Romanian language had a *predominantly Slavic* etymological structure of the lexicon (with a value going up to just over two-fifths), and the *Latin element*, i.e. the words inherited, was in fact only one of the other constitutive *fifths*, as *the other two fifths* reunited Turkish lexical elements and elements of a heterogeneous origin: Hungarian, Neo-Greek, Albanian, etc.)

To begin with, the statistical quantifications mentioned bove were not based on a coherent methodology or an adequate statistical-mathematical apparatus, because, on the one hand, they do not correspond to the reality of the figures given by the "*index*" at the end of the second volume, and on the other hand, the word count did not take into account the relationship between *words* and *variants*. In connection with the first aspect, which also calls into question the statistical quantification made by Sextil Puşcariu according to Cihac's word "*index*", Mircea Seche shows that the count does not correspond to reality for two clear reasons (Seche, 1966, p. 107): (i) first, the words contained in A. de Cihac's "index" are more than 8,900 (rather than 5,765, as Sextil Puşcariu had falsely estimated); (ii) the total sum of the words recorded by the dictionary (that is, not only the index at the end of the second volume) is 17,645, as not only the toponyms were excluded, as was but natural, but also the variants (either phonetic or lexical). Both of the above arguments are of paramount importance in order to emphasize the complete lack of accuracy in statistical quantification and eventually the lack of a reasoned, well-grounded and scientifically proven assessment. Furthermore, the lack of a unitary methodological treatment and the absence of an ensured statistical comparability are equally obvious, in terms of Latin origin words, since A. de Cihac considered only the bases/roots, and for the others – the derivatives as well. Mircea Seche shows that all the words in A. de Cihac's dictionary are distributed, in accordance with the various etymological layers, as follows: "*elements of Latin origin (and their derivatives) – 6,141; elements of Slavic origin – 4,691; elements of Turkish origin (and their derivatives) – 1,250; elements of Neo-Greek origin (and their derivatives) – 1,100; elements of Hungarian origin (and their derivatives) – 1,026; Romanian items common with Albanian (and their derivatives) – 90*" (Seche, 1966, p.108). The percentages corresponding to these new data sets are: the Latin lexical element (and its derivatives) represent 45.6% of the total, the Slavic element (and its derivatives) represent 34.8%, the Turkish element (and its derivatives) represent 7.1%, the neo-Greek element (and its derivatives) represent 6.2%, the Hungarian element (and its derivatives) represent 5.8% of the total, and the lexical element shared with Albanian (and its derivatives) represents only 0.5% of the total words searched. It can therefore be observed that the difference is sensitive, if compared both to the relative frequencies indicated by A. de Cihac, and to those indicated by Sextil Puşcariu.

Secondly, A. De Cihac's statistics put on the same plane words unequal in terms of *circulating* power and *semantic* volume or content. Cihac's count (like Puşcariu's, later on) does not take into account the essential fact that the lexical units of a language can by no means be on the same plane in terms of their *relative importance*.

The *weight* of words of various origins marks the identity of a vocabulary, as well as their degree of representation in the basic vocabulary of a language unde study. This *statistical weight*, measured as relative frequency, must be studied from the standpoint of the dynamics of the vocabulary of the respective language.

However, Sextil Puşcariu demonstrated, in several works, for example *Locul limbii române între limbile romanice*, *Limba română*, vol. I (*The place of the Romanian language among Romance languages. The Romanian language*), that the Latin(ate) or Romance nature of the Romanian language, which is visible from its entire structure, can also be deduced from the "construction material" of the vocabulary, yet not by merely counting.

"*Any etymological dictionary is unilateral, because it takes into account only the origin, and not the circulation of words in the language, as well. In an etymological dictionary, words known and used on a daily basis by every Romanian, coming from any region of the country, occupy a locational* *unit, just like the words used only in one region – and very rarely even there – and unknown to all other areas*" (Puşcariu, 1976, p. 181). In the same paper (Seche, 1966), a possible reference of approximations, and/or a probable source of A. de Cihac's false estimates, namely a similar "statistical count" dated 1840, by the Russian linguist I. Hinkulov, from which it seems that the controversial A. de Cihac could have been inspired as concerns the percentage of the various etymological elements in the lexicon.

The linguist who first noticed the fundamental falsity of A. de Cihac's statistical observation was the Romanian scientist Bogdan Petriceicu Haşdeu, who developed the theory of circulation as an adaptation of the theory of circulation in economics, applying it to the words in the Romanian lexicon. Bogdan Petriceicu Haşdeu clearly and unmistakeably showed that the frequency or circulation of words is decisively important in establishing the lexical *physiognomy* of a language. riticizing the etymological classification made by A. de Cihac – for whom "*L'*élément latin de la langue roumaine ne représente guère aujourd*'hui qu'un cinquième de son vocabulaire, tandis que l'*élément slave y entre pour le double ou pour 2/5 à peu près" ("*The Latin element of the Romanian language hardly represents today one fifth of its vocabulary, while the Slavic element is comprised twice as much, i.e. for about 2/5*" (Haşdeu, 1984, p. 73) – Bogdan Petriceicu Haşdeu revealed the flawed nature of some false, inaccurate "statistics" that put on the same plane words that are not equal in point of semantic volume or power of circulation. Bogdan Petriceicu Haşdeu revealed the flawed nature of some false, inaccurate "statistics" that put on the same plane words that are not equal in point of semantic volume or power of circulation: "*The dictionary does not give us, since it cannot give us, the circulation in language; and this is the key point*". Dimitrie Macrea compiled another statistic in 1942, based on the words contained in CADE (*Dicţionarul enciclopedic ilustrat „Cartea Românească"*. Partea I: *Dicţionarul limbii române din trecut şi de astăzi* de I.-A. Candrea. Partea II: *Dicţionarul istoric şi geografic universal* de G. Adamescu, Bucureşti, 1926-1931 – *The Illustrated Encyclopedic Dictionary "Cartea Românească".* Part I: *The Dictionary of the Romanian language in the past and today*, by I.-A. Candrea. Part II: *The Universal Historical and Geographic Dictionar* by G. Adamescu, Bucharest, 1926-1931). Its final conclusion (considered at least astonishing, in the opinion of the late academician Alexandru Graur) was that the Latin elements represent 20.58% of the total, the Slavic ones 16.41%, the French items 29.69%, and the remaining 33.32% was allegedly made up of words originating in languages that did not matter too much in terms of percentage, or of words whose origin was unknown. The gist of the matter is therefore the crucial significance, unanimously recognized in general linguistics, of the fundamental lexicon of a natural language, or rather its most resilient and important section, the very *core*, indeed, of the lexicon of a language; the concept is the direct opposite of the *mass of the vocabulary* or *secondary vocabulary*, and several names are used to name it, such as: *basic vocabulary*, *fundamental vocabulary*, *essential vocabulary*, *main* or *principal lexical stock*, and rather infrequently *the usual lexical stock*. (Hristea et al., 1984, p. 14)

The role and the relative value that the different elements hold within the lexical make-up of the Romanian language can be better specified if we consider the data provided by several subsequent statistics, performed on the basis of works of greater depth, which were philological and statistical to practically the same extent. In such a statistical quantification compiled by Sever Pop, in 1948, starting from the *The Dictionary of the Romanian language in the past and today*, by I.-A. Candrea (published by *Cartea Românească* in Bucharest, 1931), one can find that the number of words of Latin origin amounts to 8,800, to which must be added 14,000 neologisms received from Romance languages, which gives a total of 22,800 terms (a comparison would be welcome with the situation in French, where, according to the *Dictionnaire de l'Académie*, 1878, out of 32,000 words, 20,000 were of learned or foreign origin and only 12,000 were French words of native origin). The number of Slavic origin words (taken over from common Slavic, as well as modern Slavic languages such as Bulgarian, Serbian, Ruthenian, Russian and Polish) was also high, i.e. nearly 7,800, but many of those words are obsolete, or else are terms of a regional, rather restricted, use. An interesting statistical research by Mihaela Bîrlădeanu, published in the paper titled *Structura etimologică a două vocabulare reprezentative: român şi francez*, in *Studii şi cercetări lingvistice*, 6/1983 (*The Etymological Structure of Two Representative Vocabularies: Romanian and French*, in *Linguistic Studies and Researches*, no. 6/1983), the author compared the representative vocabularies of Romanian and French, and found the following etymological structure for the representative vocabulary of our language: Latin: 1. inherited: 30.45%, 2. learned terms 1.77%; internal forms 24.81%; substratum 0.96%; Old Slavic superstratum 8.91%; Neo-Greek 1.11%; loans from modern Slavic languages 1.80%, Romance loans from: French 7.64%, Italian 0.54%; Turkish 0.73%; Hungarian 1.27%; germane 0.27%; English; onomatopoeic 0.23%; multiple etymology 17.36%; unknown and/or uncertain origin 2.08%. Among the relatively recent papers dealing with quantification and statistical analysis in the philological field, a few came from the authors of this article (Manea, 2004; 2009; Manea, Săvoiu, 2014). The method of frequency statistical analysis was the major method applied in one of the main sections of the book *Structura* *etimologică a vocabularului neologic (cu specială referire la anglicismele din limba română) – The Etymological Structure of the Neological Vocabulary (with Special Reference to the English Loans in Romanian)* – published in 2004, and also in Încercare statistică asupra ortografierii cu â *şi* î (*A Statistical Essay on Spelling with â and î*), published in 2009.

3. Methodology

The enumerative series is the simplest form of statistical presentation of a population grouped in keeping with the most commonplace criterion-based solutions in the non-quantitative universe, namely in the world of words, specific to the philologist or linguist (Săvoiu, 2012). A series of that type can be represented by the simplest list of first and last names of people, grouped according to a certain organizational, administrative, structural, spatial, temporal, etc. criterion. Following the method of capitalizing on the method of statistical grouping of this population in relation to the first name or last name, there results a series of distribution or frequency distribution of first names or first names identified as distinct, a series known as a *homograde* or *attributive non-numerical series*, the first string being enumerative or qualitative, and the second – quantitative or numerical (Săvoiu, 2003, p.118). In a simplified way of rendering it, any enumerative statistical series processed by grouping becomes a non-numerical attributive series – with reference to the first string, which is essential as it is the discriminating one) – and the second string becomes numerical due to the fact that frequencies of occurrence were assigned to words, phrases and expressions, qualifiers, hierarchies, etc. (Săvoiu, et al, 2006). Figure 1 describes, in a general model, a few usual non-numerical or enumerative attributive series:

Forms of existence of the non-numerical or enumerative attributive statistical series processed by grouping

Figure no. 1

Variant (x_i)	word	phrase	qualificative	hierarchy	correct/incorrect	yes/no
Frequency (n_i)	n_1	n_2	n_3	n_4	n_5	n_6

Source: Made by the authors

The statistical series of this kind includes two parallel strings, the former including variants of the non-numerical or enumerative attributive variable discriminated according to a linguistic, philological, hierarchical, evaluation, knowledge, etc., criterion, and the latter including the occurrence frequencies of the variants in the first string. Variant (x_i) can represent a word (noun, first name, verb, adjective, adverb, etc.), a noun (proper names, usual or not, name of a common object), a plural form, a verb form, a form of writing or pronunciation in use, a spelling or writing solution in local or general use, an input source from a foreign language of a neologism, a correct or incorrect form of pronunciation or wording, a qualifier, a nuance or a hierarchical qualifier, a level of coverage, a level of acceptance or rejection, a monosyllabic or multisyllabic answer (yes, no or yes, no, I don't know / I don't answer), a way of accepting or not accepting, etc. Frequency (n_i) is a numerical information resulting from the quantitative aggregation of identical variants, the sum of all frequencies bringing together all the cases in the series.

A handy example of the most simple and relevant enumerative series by its simple constitution, an example rediscovered in the last three decades, considered famous for the elitism of those recorded, but also for their association with anti-religious militancy or communist atheism in Romania, is that of the series of the personalities incinerated (cremated) rather than buried in the Greek-Orthodox tradition (without observing the Orthodox rituals of the vigil of the deceased, the funeral and the memorials, which over time became alms, whose content was in the range of the most inexplicable things possible)… ot only legally, but also Biblically, cremation is a type of funeral very similar to burial in point of finality, in which all those who die return to the dust, the former describing an accelerated process of decomposing a corpse by burning and thus turning a human body into ashes, and the latter a much longer process as measured in years, a natural process of metamorphosis of human body into the dust, as a natural "return to the land from which you are taken; for you are earth, and you will return to earth" (*The Bible or Holy Scripture*, 2001, version translated and annotated by Bartolomeu Valeriu Anania, *Genesis* 3:19, p. 22). This simple enumerative series includes, as non-numerical attributive variables, both the name and surname of 2,153 personalities incinerated, and their professions, as well as the desciption provided by the implicit gesture of incineration. The series denotes the absence of the Orthodox religious faith in the case of all those recorded in this list of individuals in a Romanian society that is permanently declared, in practically all censuses, as dominantly Greek-Orthodox. The list of the Romanians considered different from most of Romania's inhabitants apparently hides more interesting information, which it can promptly reveal if it is statistically processed in an elementary manner, or subjected to a frequency analysis and transformed into a non-numerical attributive series. In terms of classical statistics, a simple list of cremated people turns into a *homograde series* of professions or occupations, which are unfortunately amalgamated within its body. The resulting homograde series ptovides not only simple statistics, but true paradoxes contrasting with the common opinion about the process and purpose of incineration, as can be seen from the analysis of those who used this solution more frequently… Cremation historically became a practice officially recognized by Romanian laws, with legal status just like burial or interment, an increasingly current status in the conditions of the Covid-19 pandemic, currently experienced by the population of our country with great intensity. In relation to burial, incineration offers some advantages, which are not only sanitary, but more extensive, ensuring a "more dignified exit from the scene of life", in a context disrupted by the impact of increasingly aggressive infectious diseases, communicable or contagious diseases capable of generating epidemics and even pandemics, accidents that disfigure corpses, human remains severely affected by wars, explosions, floods, etc.: (i) it involves much lower costs than burial, and, over a longer period of time, eliminates the costs of traditional burial traditions; (ii) it is a much more aesthetic and ecological practice; (iii) it induces a sense of equality between people in the final act of their lives; (iv) it expands the cult of the dead, dilating the emphasis on the soul compared to the emphasis on the body; (v) the urn with human ashes may be kept at home or buried in the graves of any cemetery; (vi) cremation involves observing the will freely expressed by the deceased not to subsequently repesent more than a memory rather than an obligation for the family, etc. (http://www.incinerareamurg.ro/romani-celebri-care-au-fost-incinerati).

Without intending in any way to promote incineration to the detriment of classical burial or to advertise for companies whose object of activity is incineration, not burial, a statistical analysis of the impact of the former in Europe shows that over 2/3, and as much as 70% of people choose the incineration solution in the United Kingdom, Sweden, Denmark, the Czech Republic, Hungary, reaching even Pareto optimum proportions in Switzerland (85%). Numrous surprising or apparently unexpected frequency hierarchies are identified in the enumerative series of people incinerated in Romania, according to the data available online at http://www.incinerareamurg.ro/romani-celebri-care-au-fost-incinerati nd processed by the authors. The resulting homograde series is presented in part by segmentation in relation to the profession or occupation declared by the families of those cremated into two limiting subgroups, which provide simple statistical information focused on maximum and minimum frequencies (Table 1).

The lowest frequencies, or the scarcest (left), and maximum frequencies, or most frequent occurrences (right), calculated from the enumerative series of incinerations in Romania

Table 1

Profession/job	**Minimal frequency (n_i)**		**Profession/job**	**Maximal frequency (n_i)**
Businessmen	2		Teachers	333
Political scientists	2		Generals	164
Race drivers	2		Writers	158
Printers	2		Actors	125
Orthodox priest, MP	1		Engineers	111
Banker	1		Researchers	109
Statistician	1 (Mircea Biji)		Communist militants	108

Source: Made by the authors by processing the occurrence frequencies from: http://www.incinerareamurg.ro/romani-celebri-care-au-fost-incinerati

The system of statistical indicators used for frequency analysis in a series of homograde distribution consists of a variety of components: "*absolute frequencies (n_i), relative frequencies ($n*_i$), cumulative frequencies in increasing order (n_i↑ or $n*_i$↑) or in decreasing order (n_i↓ or $n*_i$↓), frequency distribution densities (n_i/h_i sau n_i*/h_i). The ascending or descending cumulative frequencies allow the identification of the quantiles (Cv) to the numerous family of which the following belong: the median (Me), the quartiles (Q_1, Q_2, Q_3), the deciles (D_1, ..., D_9) and percentiles (C_1, ..., C_{99}), which divide the statistical population into two, four, ten and one hundred equal parts*" (Săvoiu, et al., 2020, p. 130).

A wide variety of statistical concepts and tools are applicable and, as has often been shown, very useful in virtually all areas of research and in all types of scientific approaches. This truth is difficult, or even impossible to challenge in philology or linguistics, starting from the capitalization of the system of frequency indicators, passing through the concentration and diversification coefficients Herfindah –Hirschman and Gini–Struck, and finally innovating the statistical profile in the fields of philology or linguistics, in the particular case of a linguistic (mini)corpus, by relieving it of the ambiguities and uncertainties related to standardization, usage, accuracy or correctness, etc. The concepts that are subject to observation, quantification and frequency statistical analysis, in what follows, are small-sized, and their role is rather methodological, and amplifying the role of cross-disciplinarity (their size ranging between 25 and 50 items, nearly all being terms belonging to the scientific-technical vocabulary of contemporary English (as shown in the detailed tables of the results). The search engines used were *Google* and *Ask*, and the corpora of words (texts) accessed were made up of academic material (texts, articles, etc.), found on the Internet.

There is a major methodological problem, namely that of subjectivism or counterfeiting the results of a *direct* investigation in a philological (mini)corpus, in a certain sense desired by the person who makes the construction of the philological (mini)corpus and the sampling, observed and processed later; it cannot be eliminated in the absence of the researcher's honesty, regardless of the degree of correctness of the frequency analysis or sampling, observing the random extractions of each word or phrase / expression starting from mathematically known practical probabilities. To all these attitudes related to the incorrect application of the survey theory, premeditated actions or not by the coordinators of direct research that thus become unscientific or mere philological or linguistic opinions, is added the existence of permanent premises for *the Hawthorne effect*. In a direct or field research, the Hawthorne effect is a classic effect known as a form of psychological reactivity of respondents, through which the subjects of a partial or experimental research modify certain aspects of their behavior, in the use of language with philological or linguistic impact, as a result of the fact that they are bing studied, without this attitude being a response to manipulations with subjective purposes (Sesardić, 2018, p. 21).

The most important categories of serious methodological errors, related to the appearance in *indirect* or *documentary* research, focusing on frequency analysis, can be considered: i) errors caused by the technique of generating philological (mini)corpora, namely due to the inadequacy of the form or technique of random, directed or mixed sampling; ii) methodological or systematic errors, in strict connection with the difficulties of sampling and quantification from the Internet, techniques with erroneous results as long as they lack tools and software to control or verify specific records, from plural forms to verbal forms, from phrases or expressions, to full expressions or specific abbreviations, etc.; iii) errors caused by non-fulfillment of the constraints of temporary spatial and structural programming of the searches; iv) errors generated by the differentiated limitations of search engines; v) errors caused by search engine optimization algorithms; vi) errors due to the inadequate pre-processing used to reduce the size of the solution space, and finally the research results to those validated, credible, verified, etc.

4. Statistical frequency analyses in philological (mini)corpora

Static (absolute) frequency analysis is practically the simplest way to apply a cross-disciplinary and investigative method in order to solve a philological (linguistic) problem, such as, for example, identifying the correct grammatical plural in keeping with the dominant usage in a philological (mini) corpus or in a distinctive database (Internet). Table 2 presents such an analysis with its specific relevance and irrelevance, starting from the idea of additional confrontation, based on statistical principles, of at least two search engines within the same (mini)corpora.

Static frequency statistical analysis of a correct grammatical plural according to the majority use in a philological (mini)corpus (Internet)

Table 2

English nouns with multiple plural forms	Results obtained by search engines		Observations
	Google (search)	Ask (search)	
apsides	96,300	8,330	
apses	427,000	76,300	Relevant
apsises	12,200	-	
octopuses	651,000	225,000	Relevant
octopodes	124,000	12,800	
octopi	523,000	133,000	
addenda	5,700,000	430,000	Relevant
addendums	468,000	105,000	
addendas	115,000	-	
criteria	445,000,000	42,800,000	Relevant
criterions	511,000	131,000	
criterias	679,000	244,000	
antennae	7,570,000	832,000	***Irrelevant****
antennas	2,840,000	4,770,000	
apexes	416,000	67,600	
apices	607,000	193,000	Relevant
apparatus	169,000,000	12,700,000	Relevant
apparatuses	7,670,000	515,000	
appendixes	2,580,000	264,000	
appendices	17,400,000	2,150,000	Relevant
aquariums	29,300,000	3,210,000	Relevant
aquaria	9,790,000	1,260,000	
automatons	528,000	1,260,000	
automata	29,200,000	1,450,000	Relevant
bureaux	73,700,000	1,150,000	***Irrelevant*****
bureaus	30,600,000	3,210,000	
cerebellums	44,800	6,510	
cerebella	393,000	50,200	Relevant
curricula	17,700,000	2,440,000	Relevant
curriculums	6,860,000	694,000	
formulas	57,000,000	8,200,000	Relevant
formulae	13,100,000	1,940,000	
genera	96,300,000	5,450,000	Relevant
genuses	117,000	15,900	
hiatuses	279,000	42,500	
hiatus	34,900,000	4,140,000	Relevant
maximums	2,860,000	407,000	
maxima	131,000,000	6,640,000	Relevant
minimums	8,030,000	1,080,000	
minima	63,500,000	1,340,000	Relevant
nuclei	22,200,000	4,050,000	Relevant
nucleuses	73,200	8,270	
phenomena	68,900,000	10,800,000	Relevant
phenomenons	462,000	99,900	

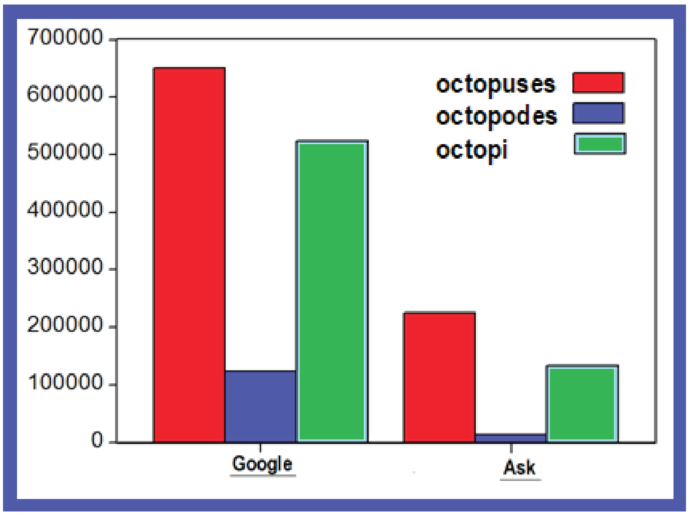
syllabuses	532,000	144,000	
syllabi	4,930,000	799,000	Relevant
strata	65,600,000	4,060,000	Relevant
stratums	336,000	24,200	
vortexes	486,000	113,000	
vortices	2,610,000	448,000	Relevant

Souce: Taken over by the authors from an earlier research paper (Manea, Săvoiu, 2014, pp.13-17).

Statically or absolutely evaluating the frequency distributions of the appropriate grammatical plural in accordanc with the specific use, or usage, of most subjects in a philological (mini)corpus (Internet) for words like *apse*, *addendum*, *antenna*, *apex*, *apparatus*, *appendix*, *automaton*, *bureau*, *criterion*, *cerebellum*, *curriculum*, *formula*, *genus*, *hiatus*, *maximum*, *nucleus*, *octopus*, *phenomenon*, *syllabus*, *stratum*, *vortex*, relevant data were identified for the absolute majority (with only two exceptions: *antenna* and *bureau*), we could draw valuable practical conclusions concerning the usefulness of the frequency analysis method, approached by at least two search engines. Graphical representations of comparative frequency values can contribute, by better visibility, to the assessment of the relevant – and probably implicitly correct plural, where quantitative information is relevant, by detailing the absolute frequencies by several search engines (for example: *Google* and *Ask*).

Comparative graphical statistical analysis of a grammatical plural, using two search engines in a linguistic (mini)corpus (Internet)

Fig. no. 2



Source: Taken over by authors after an earlier own paper (Manea, Săvoiu, 2014, pp.13-17)

Figure 2 points out that a graph comparing the distributions of the plural forms studied with the help of two search engines, too, may be more useful, through its higher visibility, in assessing the hypotheses regarding the grammatical-phonetic adequacy of one of the three plural forms, analyzed statistically in terms of frequency. Relevant quantitative information ensures that the frequency of the occurrences will become relevant, and if there practically occur situations of ambiguity, one can naturally use several search engines (*Google* and *Ask*).

In complex analyzes one can move on to a subdivision of the extensive list of specialized or technical terms (some of which were scientific only initially), in as far as: a) thy illustrate an issue related to morpho-phonematics; b) they illustrate a mere issue of spelling; c) they illustrate a category of terms that can hardly be called typically technical or scientific terms, although they are undoubtedly *learned* terms.

"Here are some examples of terms from the subcategories deduced from the extended list: a) *antenna*, *apex*, *apparatus*, *appendix*, *automaton*, *cactus*, *calyx*, *cerebellum*, *cerebrum*, *cicada / cicala*, *colloquium*, *cranium*, *criterion*, *curriculum*, *dilettante*, *discus*, *fauna*, *flora*, *formula*, *fungus*, *genus*, *hiatus*, *iambus*, *larynx*, *libretto*, *memorandum*, *novella*, *nucleus*, *palazzo*, *phenomenon*, *radius*, *radix*, *retina*, *rhombus*, *stratum*, *syllabus*, *tableau*, *tempo*, *trapezium*, *vacuum*, *vertebra*, *vertex*, *vortex*; b) *bureau*, *flamingo*, *fresco*, *grotto*, *halo*, *manifesto*, *memento*, *motto*; c) *aquarium*, *candelabrum*, *cicerone*, *colossus*, *focus*, *grotto*, *gymnasium*, *hippopotamus*, *maximum*, *millennium*, *minimum*, *narcissus*, *persona grata*, *referendum*, *sanatorium*, *symposium*, *terminus*, *ultimatum*" (Manea, Săvoiu, 2014).

The detailed observations that the authors were able to make in support of the idea of factual particularization and concrete grounding of some hypotheses regarding the recurrence of words in data (mini)corpora as early as 2014, revealed that there are, for example, very few terms that have three provable plural forms (viz. *octopus* "(Rom.) caracatiță" – pl. *octopuses*, *octopodes*, *octopi*, although the last form is actually inappropriate / not recommended).

Equally instructive may prove further investigations focusing on a similar search for the distribution of the nouns *apsis* (pl. *apsides, apses, apsises*), *addendum* (pl. *addenda, addendums, addendas*) and *criterion* „*criteriu*" (pl. *criteria, criterions, criterias*), possibly also *frustrum*.

In quite a few cases, marking linguistic usage, as understood and illustrated by some top British and American dictionaries, was so important that the lexicographers in question (e.g. the authors of the *MacMillan* dictionary) found it appropriate to gloss irregular plural forms as *legitimate* *lemmas* in their own right (e.g. *bacteria*, *criteria*, *algae*, *data*). However, when the issue of meaning is also considered, the procedure itself seems to prove its total impotence: how could we verify the meaning and use of each occurrence that appears in the texts or (mini)corpora on which searches were made? (Examples: *genius* (pl. *geniuses* vs. pl. *genii*), *domino* – pl. *dominoes* („game”) / *dominos* („article of dress”), *index* – pl. *indexes* / *indices*, *stamen* – pl. *stamens* / *stamina*, *milieu* – pl. *milieus* / franc. *milieux* [‘mi:ljø], *calculus* – pl. *calculuses*; med. *calculi* [‘kælkjulai], *polypus*– pl. *polypi*; *data* (pl., though usually considered an uncountable noun) → sg. *datum*, *agenda* (pl. form of *agendum*; currently considered a singular form, in spite of its Latin derivation).

Dynamic and instrumentally compared statistical frequency analysis describes a more advanced way of cross-disciplinary investigation having the same role and purpose, namely identifying the correct grammatical plural according to the dominant usage approached in two different moments of the evolution of a philological (mini)corpus or a distinctive database (Internet). Table 3 presents this analysis against the same thematic background of the correct plural (2014), yet in a repetitive manner (2020), with an additional calculation of the *evolution index of the relevance of the assessment*:

Dynamic frequency statistical analysis of the correct grammatical plural according to the majority usage in a philological (mini)corpus (Internet)

Table 3

English nouns having multiple plural forms	**Results obtained using search engines in different years**				**Observations**
	2014		**2020**		
	Google(search)	**Ask (search)**	**Google (search)**	**Evolution index - %-**	
apsides	96300	8330	127000	131,88	
apses	427000	76300	**637000**	**149,18**	Relevant
apsises	12200	-	*5680*	*46,56*	
octopuses	651000	225000	**4450000**	**683,56**	Relevant
octopodes	124000	12800	162000	130,65	
octopi	523000	133000	2130000	407,27	
addenda	5700000	430000	**8560000**	**150,18**	Relevant
addendums	468000	105000	1170000	250,0	
addendas	115000	-	200000	173,91	
criteria	445000000	42800000	**698000000**	**154,83**	Relevant
criterions	511000	131000	1060000	207,43	
criterias	679000	244000	2060000	303,39	
antennae	7570000	832000	8890000	113,21	
antennas	2840000	4770000	**74100000**	**26091,55**	**Relevant
apexes	416000	67600	597000	143,51	
apices	607000	193000	**2100000**	**349,42**	Relevant
apparatus	169000000	12700000	**190000000**	**112,43**	Relevant
apparatuses	7670000	515000	5320000	69,36	

appendixes	2580000	264000	3110000	120,54	
appendices	17400000	2150000	**22800000**	**131,03**	Relevant
aquariums	29300000	3210000	**59100000**	**201,71**	Relevant
aquaria	9790000	1260000	10600000	108,27	
automatons	528000	1260000	1370000	259,47	
automata	29200000	1450000	**52100000**	**178,42**	Relevant
bureaux	73700000	1150000	72800000	98,78	
bureaus	30600000	3210000	**118000000**	**385,62**	*Relevant
cerebellums	44800	6510	65500	146,21	
cerebella	393000	50200	**460000**	**117,05**	Relevant
curricula	17700000	2440000	**28000000**	**158,19**	Relevant
curriculums	6860000	694000	8100000	118,08	
formulas	57000000	8200000	**155000000**	**271,93**	Relevant
formulae	13100000	1940000	52000000	396,94	
genera	96300000	5450000	**156000000**	**161,99**	Relevant
genuses	117000	15900	238000	203,42	
hiatuses	279000	42500	396000	141,94	
Hiatus	34900000	4140000	**59400000**	**170,20**	Relevant
maximums	2860000	407000	3930000	137,41	
maxima	131000000	6640000	**548000000**	**418,32**	Relevant
minimums	8030000	1080000	9210000	114,70	
minima	63500000	1340000	**276000000**	**434,65**	Relevant
nuclei	22200000	4050000	**34800000**	**156,76**	Relevant
nucleuses	73200	8270	96600	131,97	
phenomena	68900000	10800000	**112000000**	**162,55**	Relevant
phenomenons	462000	99900	1210000	261,90	
syllabuses	532000	144000	1500000	281,95	
syllabi	4930000	799000	**6080000**	**123.33**	Relevant
strata	65600000	4060000	**66500000**	**101,37**	Relevant
stratums	336000	24200	928000	276,19	
vortexes	486000	113000	1020000	209,88	
vortices	2610000	448000	**4260000**	**163,22**	Relevant

Source: Made by authors for the columns referring to 2020 and after (Manea, Săvoiu, 2014).

Note*: The aggregation of frequency data over time turns irrelevant situations into relevant ones in a way that is at least unexpected, if not actually prompt.

Note**: The search for **antennae** returned 8890000 results, being surpassed by the search for **antennas**, which returned 74100000 results, giving an extension in use multiplied maximally in 6 years, i.e. 260 times, that is **26091.55 - 100.00 = 25991.55 %** more, if compared to the past.

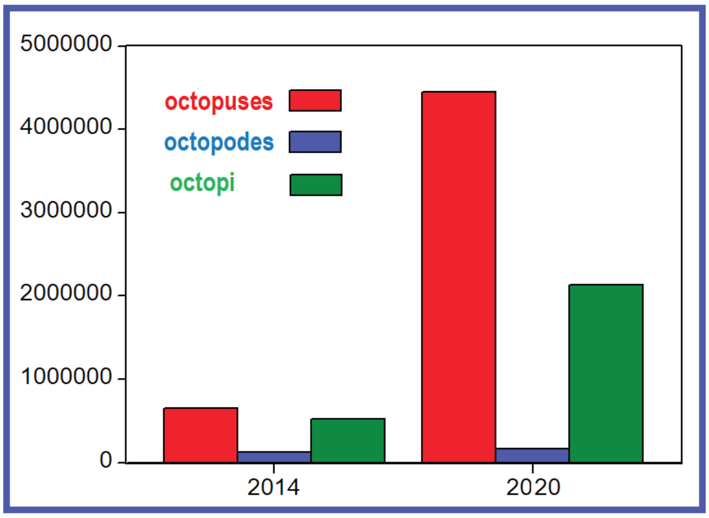
Time kills the initial irrelevance and, as can be seen, solutions are obtained for all examples, and, what is more, different dynamics can be identified according to the frequency statistical analysis. Thus, some words or plural forms exemplified in both Table 2 and Table 3 are of much greater relevance as a result of an explosion in their use in the (mini)corpus investigated (Internet).

Resuming the frequency analysis in the case of the potential triple plural: *octopuses*, *octopodes* and *octopi*, three completely different dynamics are identified from the final data, which reconfirms the relevance of the form *octopuses*, while also showing an acceleration of polarization between the forms *octopuses* and *octopi*, which offers linguistic evidence of the continuity of their increasing linguistic use.

The graphical statistical representations of the frequency values compared in time in Figure 2 contribute to the appreciation of the relevant (correct) plural, through better visibility, where the quantitative information is relevant, by detailing the absolute frequencies, after several investigations performed at different points in time (viz. 2014 and 2020).

Graphic statistical analysis of a comparative type of the variants of a grammatical plural in a linguistic (mini)corpus (Internet), conducted at two different moments in time

Fig. no. 3



Source: Made by the authors in keeping with the data in Table 3

The calculation of the concentration-diversification coefficients, of the HH type (Herfindahl–Hirschman) or C_{G-S} (Gini–Struck), describes – by the ascending values (for HH), or by values that tend more and more clearly towards 1 (for C_{G-S}) – a concentration in the analyzed linguistic corpus (Internet), or in a specific linguistic (mini)corpus (a distinctive database). The limits of an excess concentration are given by the values of 0.677 (HH) and 0.409 (C_{G-S}) for the situation where the linguistic offer is reduced to 3 solutions (n = 3), as can be inferred from the calculations made in the 2010 paper authored by Săvoiu, Crăciuneanu, Ţaicu.

Table 4 shows that the plural form *octopuses* is apt, by quantified linguistic use by means of the frequency statistical analysis, to represent a single plural solution, generating values above the limits of excessive concentration in 2020 (in keeping with both the actual value of the HH coefficient and that of C_{G-S}).

Dynamic frequency statistical analysis of the correct grammatical plural using statistical concentration-diversification coefficients

Table 4

Variant	2014	(gi)	$(gi)^2$		2020	(gi)	$(gi)^2$	
octopuses	651,000	0,501	0,251	**HH = 0,650**	4,450,000	0,660	0,436	**HH = 0,732**
octopodes	124,000	0,096	0,009		162,000	0,024	0,001	
octopi	523,000	0,403	0,163	**C_{G-S} = 0,367**	2,130,000	0,316	0,099	**C_{G-S} =0,551**
Total	1,298,000	1,000	0,423		6,742,000	1,000	0,536	

Source: Made by the authors

The previous statistical method frequently motivates rigorously the choice of the correct octopus form, according to the dominant linguistic use (with structural values and located over 0.5). At the same time, this method validating a process of excess concentration, guaranteed by the quantifications of the Gini-Struck coefficient (CG-S) that exceed values of 0.41 (Săvoiu, Crăciuneanu, Ţaicu, 2010, pp. 15-27) finally leads to the idea that frequency analysis complemented by concentration-diversification analysis ensures a much greater credibility than the exclusive application of simple frequency analysis in philology in investigating the variation of language use in time and space.

Another manner of better using frequency statistical analyses can be conducted in order to identify the source of a particular neologism, with reference to the scientific literature that has established it, through extensive or maximum use. To exemplify this aspect, Table 5 summarizes the occurrences of a neologism in Romanian in parallel with the synonyms (or near homonyms) in French and English:

Frequency statistical analysis of the source of frequent Romanian neologisms, apparently inspired from French or English, within a linguistic (mini)corpus (Internet)

Table 5

Romanian	French	English	Source (abbr.)
fenomen 30100000	*phénomène* 62300000	*phenomenon* 647000000	Eng.
teoremă 12800000	*théorème* 3370000	*theorem* 60100000	Eng.
dilemă 29600000	*dilemme* 3480000	*dilemma* 101000000	Eng.

Source: Made by the authors

It goes without saying that the Greek derivation or etymology (or rather the primary Greek origin of the three words exemplified) cannot be questioned, yet the frequency analysis points out that these scientific terms, although having entered our lexicon via French, according to linguistic use certainly (or paradoxically) *belong* to the English scientific literature. Obviously, the etymology or historical origin of words cannot be underestimated or abandoned, because in this way one can become misinformed instead of being well-informed. In the case described in Table 5, frequency statistical analysis cannot be accused of misinformation and even less of ill-information, which can quantitatively prove the manner of capitalizing scientific language in one language or another.

To what extent are incorrect expressions being used in accordance with the norms and standards that are (apparently) in force, without however forgetting that human language and speech have an extraordinary vitality, and when do their dynamics begin to exceed that of the forms considered correct? Such a question can be answered by a similar statistical quantitative analysis of a frequency type, starting from the 20/80 Pareto balance, and we believe that as soon as the incorrect form (F_{INC}) exceeds 20% of the total aggregate uses, and has an *evolutionary index of the relevance of the assessment* that is higher than the form considerd or declared correct (F_{COR}), an evident competition is established in their parallel use. Table 6 illustrates two possibilities of putting to better use the method of frequency statistical analysis in identifying language competitions related to alternative use (correct-incorrect), and also a not yet declared competition between a well-established form and two other forms considered together as correct:

Frequency statistical analysis of the Paretian or non-Paretian ratio between correct and incorrect forms of the same word in a philological (mini)corpus (Internet)

Table 6

Frequency F_{COR}	Frequency F_{INC}	Aggregated use	Ratio F_{INC}/F_{COR}
preşedinţie 819000	*preşedenţie*78300	897300	8,7% vs 91,3%
să aibă 25300000	*să aivă + să aibe* 2780000	28080000	9,9% vs 90,1%
găluşti 45700	*găluşte* 387000	*432700	10,6% vs 89,4%

Source: Made by the authors.

*Note: Both forms are considered correct, but even so the limit of Paretian optimum that could have opened a real linguistic competition related to their alternative use was not exceeded.

The range of arguments that justifies the use of the optimal Paretian type of ratio starts from the concept of variation, and implicitly from the derived one, i.e. homogeneity, in the case of alternative or binary variables, where the limit of the homogeneity coefficient is reached in the 20% versus 80% evaluations (YES vs. NO). For values smaller than 80% compared to the Paretian limit, certain possibilities of heterogeneity are proved within any population, including a philological (mini)corpus.

The method of the statistical profile, especially in the classic variant of confronting a set of statistical profiles, has a rather high potential linguistic or philological applicability. Possessing a great power of synthesis, the method of confrontation through statistical profiles can synthesize the philological dialogue and argumentation in an essential way, as can be seen in Table no. 7.

Philological confrontation of statistical profiles conducted in the same philological (mini)corpus

Table 7

The fake statistical profile of the etymological structure of Romanian made by A. de Cihac and resumed by Sextil Puşcariu, in accordance with the Russian linguist I. Hinkulov	**The statistical profile of the etymological structure of Romanian, made in an objective and balanced manner by Mircea Seche**
Words of Latin origin: 20,2%	Words of Latin origin: 45,6%
Words of Slavic origin: 41,0%	Words of Slavic origin: 34,8%
Words of Turkish origin: 16,7%	Words of Turkish origin: 7,1%
Words of Neo-Greek origin: 11,0%	Words of Neo-Greek origin: 6,2%
Words of Hungarian origin:10,2%	Words of Hungarian origin: 5,8%
Words of other origins: 0,9%	Words of other origins: 0,5%

Source: Made by authors, in the guise of a synthesis and final confrontation.

There are obviously many more statistical methods applicable in philology or linguistics, which gives a much wider space for action to cross-disciplinarity, and even to multidisciplinarity, which truth current and future studies and research will certainly prove.

5. Conclusions

This paper is the result of applying statistical methods in a cross-disciplinary manner, and the authors were guided by the desire to identify useful solutions to several controversial problems (even though those solutions were sometimes exclusively partial), with objective and reliable results, likely to confirm the logical and analogical systematics of any natural language. Any frequency analysis of the statistical type, if appropriately applied to philological corpora and minicorpora, leads eventually support the common user of the language, whose multiple capacities of anticipated association, of simple comparison for the purpose of ranking and confrontation with the purpose of elimination, can be permanently optimized, offering proven linguistic solutions and proving their practical use through statistical arguments. According to the authors, this direction in cross-disciplinary research can bring about additional clarifications as to the adequacy of statistical methods to philological investigation, by making better direct use of frequency tools, evolutionary indices of relevance in language usage, concentration-diversification coefficients meant to identify the dominant plural of several Anglicisms, the sources from which some neologisms have been taken that already have an internationalized character, or the intensity of the competition between the correct and incorrect forms of the same word in linguistic use (Săvoiu, et al., 2020, p. 146).

The authors aim to continue the investigations carried out by attempting similar ones through cross-disciplinarity, and even to initiate, within a wider team tending towards multidisciplinarity, more extensive research by multiplying the methods, simultaneously with the extension of the degree of coverage through ever larger corpora. In the short term, it is worth analyzing, in terms of use and frequency of use, the hybrid semantic variants (or "barbarisms") that occur quite frequently in Romanian, more often than not borrowed directly from English, or at least modeled after Anglo-American models, such as *onerous*, *intrepid*, *vocal*, *versatile*. Unfortunately, even in the medium term, it hardly possible to identify optimal solutions for researching the distinct ways of pronouncing such terms searched for, since their pronunciation / phonetics cannot be recorded in the texts from large corpora with universal access – of the Internet type. (Săvoiu, et al., 2020, pp. 147)

The only viable option for the near future is therefore the methodical study of philological (mini)corpora.

What other useful things could be included in the focus of frequency statistical analyses? Most probably, one can identify a set of multidisciplinary-team projects in the field of linguistic standardization or normalization, and comparative-contrastive didactics that have become more and more useful or relevant. Analogously, the mere construction of data corpora, mainly through searches conducted on the immense multitude of texts hosted by the Internet, could promptly and concretely prove what the real situation is of some words that exhibit uage difficulties, e.g. how to make the verb/predicate – subject agreement, or the actual use of certain prepositions or conjunctions, as well as the prevalence or non-prevalence, in the real use of the language, of certain plural forms declared incorrect, yet provably evolving or dynamic, etc. Starting from the model of the dictionary authored by J. C. Wells, linguistic or philological comparisons can be made, in terms of internationalization, of several Romanian words with those similar in meaning that are used (in this particular case) in English, and also by frequency statistical evaluation, as well as structuring or weighting of words that start with a certain dictionary letter, or what percentage of nouns or verbs have irregular forms or uses, or are considered "aberrant", starting from a rigorous statistical quantification with increasingly efficient software.

A book published quite recently in the area of cross-disciplinary, and sometimes even multidisciplinary research (Ross, Greenhill, Atkinson, 2013), where genetics was used simultaneously with the history of folk tales, revealed an interesting fact with consequences related to the territorial dynamics of language usge, finding that tales and stories are grouped geographically in a way very similar, or even identical, to genes. This is an aspect that could allow the determination of language in terms of usage area, and its ethnic-linguistic borders, starting from the variants of folk tales and popular stories, in parallel with the evolution of DNA. The impact of language and stories proves greater than that of genes, or, in the synthetic expression used by Australian journalist Christine Kenneally, in 2019, "*a human couple can more easily mix their genes without sharing the same language, much easier than a story can cross a language barrier*" (Kenneally, 2019, p. 385).

One could finally conclude by saying this: "*Culture and language are, and remain, solid in their essence, while genes are liquid (Khan, 2013), and a statistical method can apparently study a stable or momentry phenomenon more easily (...), but it shuld not forget to do so in a dynamic [methodological] manner (at periodic intervals), and thus with intentions generating or providing vitality and philosophical flow (...)*" (Săvoiu, et al., 2020, p. 150).

References

1. Graur, A., 1989. *Iarăşi â şi î*, in *Puţină gramatică*, vol. II, Bucureşti: Editura Academiei R.S.R.
2. Haşdeu, B. P. 1984. *Cuvente den bătrâni*, Bucureşti: Editura Didactică şi Pedagogică,
3. Hristea, T. (coord.) 1984. *Sinteze de limba română*, Bucureşti: Editura Albatros
4. Iordan, I., 1978. *Istoria lingvisticii româneşti* (Coordonator: acad. Iorgu Iordan), Bucureşti: Editura Ştiinţifică şi Enciclopedică.
5. Iordan, I., Robu, V. 1978. *Limba română contemporană*, Bucureşti: Editura Didactică şi Pedagogică, Bucureşti.
6. Kenneally, C., 2019. *Povestea secretă a speciei umane, Cum ne sunt modelate identitatea şi viitorul de ADN şi de istorie*, Bucureşti: Editura Humanitas

7. Khan, R. 2013. Why Culture is Chunky and genes are Creamy, *Gene Expression*.
8. Lombard, A., 1992. Despre folosirea literelor â şi î, *Limba română*, nr. 10. pp. 531-540.
9. *MacMillan Dictionary*, MacMillan, 2012
10. Macrea, D., 1943. Fizionomia lexicală a limbii romăne, *Dacoromania*, vol 10/2. pp. 362-373.
11. Manea, C. 1993. *Consideraţii statistice asupra ortografiei lui â din a şi î din i*, comunicare prezentată la Sesiunea de comunicări ştiinţifice „Rolul presei şi învăţământului în relansarea economico-socială a României", de la Universitatea din Sibiu, 21-22 mai 1993.
12. Manea, C. 1997. Structura etimologică a limbii române literare contemporane în lumina frecvenţei cuvintelor, *Studii şi cercetări lingvistice*, anul XVIII, nr. 2.
13. Manea, C. 2009. ***Încercare statistică asupra ortografierii cu â şi î***, în vol. *Distorsionări în comunicarea lingvistică, literară şi etnofolclorică românească şi contextul european*, Academia Română, Institutul de Filologie Română „*A. Philippide*", Iaşi: Editura Alfa, pp. 209-218
14. Manea, C., 2004. S*tructura etimologică a vocabularului neologic (cu specială referire la anglicismele din limba română)*, Piteşti: Editura Universităţii din Piteşti.
15. Manea, C., 2009. ***Încercare statistică asupra ortografierii cu â şi î***, în vol. *Distorsionări în comunicarea lingvistică, literară şi etnofolclorică românească şi contextul european*, Academia Română, Institutul de Filologie Română „A. Philippide", Ed. *Alfa*, Iaşi, pp. 209-218.
16. Manea, C., Săvoiu, G. 2014. *Frequency Analysis of the Use of a Number of Morphological Variants in Scientific Language*, in ESMSJ, vol. IV, no. 2 (special issue), pp. 13-17.
17. Marcu, F. 1997. *Noul dicţionar de neologisme*, Bucureşti: Editura Academiei Române.
18. Marcu, F. 2000. *Dicţionar uzual de neologisme*, Bucureşti: Ed. Saeculum I.O
19. O'Keeffe, A., M., McCarthy, J., and Carter, R. A. 2007. *From Corpus to Classroom*, Cambridge: Cambridge University Press.
20. Onu, L., 1992. Oportunitatea reformei ortografice, *Limba română*, nr. 4, pp. 199-207, Bucureşti.
21. Săvoiu, G., 2003. *Statistică generală. Argumente în favoarea formării gândirii statistice*, Piteşti: Editura Independenţa economică.
22. Săvoiu, G. (coord.), Asandei, M., Grigorescu, R., Manole, S. 2005. *Cercetări şi modelări de marketing. Metode cantitative în cercetarea pieţei*, Bucureşti: Editura Universitară.
23. Săvoiu, G., 2012. *Statistică generală cu aplicaţii în contabilitate,* Bucureşti: Editura Universitară.
24. Săvoiu, G., Crăciuneanu, V. Ţaicu, M. 2010. A New Method of Statistical Analysis of Markets' Concentration or Diversification. *Romanian Statistical Review*, vol. 58(2), pp. 15-27.
25. Săvoiu, G., et al. 2020. *Metode statistice şi crosdisciplinaritate*, Bucureşti: Editura Universitară.
26. Sesardić, N., 2018. *Când raţiunea pleacă în vacanţă. Filozofii în vacanţă***,** Bucureşti: Editura Humanitas.
27. *** *Biblia sau Sfânta Scriptură*. 2001. Ediţie jubiliară a Sfântului Sinod, versiune redactată și adnotată de Bartolomeu Valeriu Anania (ed.), Bucureşti: Editura Institutului Biblic şi de Misiune al Bisericii Ortodoxe Române, p. 22. Available on - line la: http://www.diacronia.ro/en/indexing/details/B347/pdf. Accessed on 30.04.2020.