Linguistic Frontiers • 3(1) • 23–27 • 2020
DOI: 10.2478/lf-2019-0002

sciendo

Linguistic **Frontiers**

Once again about the hapax grammar: Epigenetic Linguistics

Original Study

Dan Faltýnek, Ľudmila Lacková, Hana Owsianková*
Palacký University Olomouc, Department of General Linguistics

Received: June 2019; Accepted: January 2020

Abstract: In this article, we deal with the similarity between epigenetic marks in DNA and hapax legomena in language; based on the so-called hapaxes, a grammar description is designed. We reflect hapax analysis of Czech language provided by Novotná (2013) and avoid random selection of the corpus. For this reason, we analyze a corpus of 12 authentic books from 12 authors who elaborated the theme "What's new in…" concerning their field of science, assigned by Nová beseda publishing. By analyzing a middle-sized corpus, we expected results similar to those of large-scale national corpus (see Novotná 2013). We chose to classify hapaxes into different categories in comparison to Novotná, yet the results show similar language productive categories. This kind of language potentiality seems to be analogical to epigenetic processes in biology, which is briefly introduced.

Keywords: norm, epigenetics, linguistics, hapax legomena, corpus

NORM AND INDIVIDUALITY

From their very beginning, the tendency of humanities is to generalize specific human behavior under the assumption that the interpersonal contact, communication, and the structure of the society are based on norms affecting and enabling every individual act of the subject in the social structure. Description of the social norms—including social structure, literary forms, arts, etc.—seems to be the best way to understand the subject of interest of humanities. Individual behavior, communication, and work of art are regarded as a usage of the norm, and their specifics are excluded out of study as random influences of the usage conditions. Study of these influences seems to be uninteresting, unless they have an impact on the norm. Exclusion of individuality from the study interest of the humanities has happened in order to improve their methodology. Linguistics works mainly with the system units (phoneme, morpheme, sentence). Even though a terminology to define actual occurrences of given system units (phone, morph, utterance) exist as well, in fact, the actual occurrences of system units are a marginal theme within linguistics. They are but isolated occurrences anchored at a given time, space, and speech. Semiotics also has a similar distinction between semiotic (system) qualities and physical qualities (Pattee 2001; Pattee 2008; Pattee, Kull 2009). Eco (2007) defined a distinction between molar qualities, significant for semiotics, and molecular-physical or chemical qualities, insignificant for semiotics (2007). As Peirce (1906) pointed out, physical object (dynamical object in his terminology) is untouchable by the means of semiotics and signification or communication. We can talk just about the immediate object—interpretation of a physical one. Again, the focus of the discipline is a model, not an individual thing. Theory of literature does the same by the death of the author (Barthes 1974); description of the literary work from the point of view of the author—author's intention—is by some literary scholars considered to be naive theoretical approach and lethal mistake in the practice of literary criticism.

On the other hand, individual qualities of the text are the matter of forensic linguistics, whose audience sits

* Corresponding author: Hana Owsianková, e-mail: h.owsiankova@gmail.com

in private hearing at the court. This is just one of few examples of subdisciplines that are actually interested in individual qualities of the study object. The loss of individuality in the humanities was caused by the attempt to have a controlled way of description and some kind of formalization—an attempt to be a science. In similar position as humanities stays the science of life—biology. The matter of interest of biology is life, which is the eponym for individuality. As phonemes and sentences, biology describes species and families. The death of the author speaks about the model organisms and their genomes. Neo-Darwinian revolution promotes life as controlled by genes and tries to find norm in their unique occurrences. Such balancing of biology between individual and model quite resembles humanities. But there is epigenetics in the contradiction to neo-Darwinism in the biological science.

WHAT IS EPIGENETICS

Epigenetics is a field representing a huge amount of biological processes that are important for organisms' development or even for the evolution but are not anchored at the level of the genetic script. In other words, they are not encoded by the classical four-letter genetic alphabet (Lacková 2018; Markoš, Švorcová 2019). Epigenetics accounts for features that, if considering only and exclusively four letters of the genetic dictionary (adenine, guanine, thymine, and cytosine), would remain unexplained. It has to be noted here that epigenetics cannot exist without genetics, because epigenetics is a kind of modification on the genetic script. Both epigenetics and genetics are essential parts of genome. Therefore, other "letters" or, better, "modified versions" of the four letters are studied by epigenetics. Epigenetic modifications can be chemically expressed by the addition of a special mark to a letter (nucleobases), for instance, adding a methyl group to the DNA molecule so that one of the four letters of the genetic alphabet becomes a modified letter (e.g., cytosine becomes methylated cytosine); this modification serves many cell processes such as aging or inactivation of the X chromosome. DNA methylation is only one example of the entire scale of adding letters to the four-letter genetic alphabet. Markoš speaks about "diacritic marks" of the DNA. Often the biological changes connected to the changes in the DNA script at the epigenetic level are inheritable—this particular feature is in contradiction with the classical neo-Darwinian understanding of heredity. For more information, see, for example, Jablonka and Raz (2009), who defined epigenetics as an inheritance of developmental variations that are not connected to differences in the sequence of DNA. In addition, epigenetic mutations can also be reversible, which means that the changes might but not must be inheritable. All these new discoveries at the epigenetic level in some way put into question the very idea of genes as blueprints. In fact, rather than simple automatic copying of a genetic script, a term context-depending reading is proposed by some biologists (Markoš 2002). With his approach, Markoš explains epigenetic processes metaphorically as subjective reading of a given text, where every single reader (organism) or group of readers (organisms) add a special interpretation (or meaning) of the text. In other words, epigenetic marks are an enlarged version of the genetic alphabet that permits to increase the variety of phenotypes. Here we come back to our introduction and to the idea of getting individuality back to science, in this case to biology. Epigenetic modifications are subjective features within a genetically normalized text; by consequence, we can say that they are examples of individual occurrences external to a norm. Although these individual occurrences might become a norm as well, one and the same epigenetic script might be further personalized. Even though they might be inherited and transported to future generations, thus becoming a certain kind of "norm" initially, they were created as a response to given environmental conditions, a context-depending reading.

HOW MANY LETTERS IN THE ALPHABET?

So far, we introduced the field of epigenetics as a biological example of incorporation of individual intra-generational occurrences in a scientific discourse (differently from genetics, which concerns trans-generational changes). Biology is one of the disciplines that proposed to pay attention not only to schematized norms (genetic script) but also to concrete usages of this norm and possible changes in it.[1] We will demonstrate that this scientific approach might be very fruitful also in the field of linguistics, but, first, let us make some more notes on the very analogy between epigenetics and linguistics. We already mentioned Markoš's metaphor of epigenetic chemical marks as diacritic marks in a linguistic text.

We would like to question this metaphor, which works as a perfect illustration to understand how epigenetics is formally manifested, but stops being explanative once we try to understand the function of epigenetic processes. Let us analyze Markoš's diacritic metaphor deeper. If we limit ourselves to understand epigenetics as mere diacritic signs, the finite number of "genetic universals" consequently comes to question. What can be accomplished at this moment, having an inventory of epigenetic "diacritic" marks, is simply to enlarge the genetic alphabet by extending it with new marked/modified letters. However, this proposal would not resolve the problem, because epigenetics accounts for the influence of environmental stress and interaction with other organisms

1 Ji (1991, 52—56) proposed the so-called Principle of Slow and Fast Processes (PSFP), where "slow" processes occur on an evolutionary time scale and "fast" processes take place on an individual life span time scale.

as actively participating in creating new marked letters.[2] Thus, given that changes in the environment are unpredictable, a simple extension of the alphabet will never encompass exhaustively all possible new formed epimutations.

And here we come to a problem: diacritics in language are a limited set of symbols (and also normalized); epigenetic modifications, on the other hand, are not countable, because they are unpredictable and context-dependent. To sum up, it seems that the "epigenetic diacritics" are not a mere addition of new letters to already established "genetic alphabet". As a matter of fact, the nature of epigenetically marked letters is different from the nature of the basic four letters. The difference can be seen as manifold, but the most striking point is the instability or unpredictability of their presence. The "diacritic" metaphor is very clever, but, in the same way, it does not exhaust the very characteristic of the epigenetic marks and could be misleading. Epigenetic marks on the DNA script are not mere "diacritics". A linguistic text can work perfectly even without the use of diacritics, one can normally understand a written text even without diacritic marks. Epigenetic marks on the DNA are, on the contrary, very important for the final interpretation of the message. Thus, we are asking the question: Is there a more suitable analogy between epigenetics and language?

CAN WE SPEAK ABOUT "EPIGENETICS OF LANGUAGE"?

Diacritics are a matter of orthographical norm and, as mentioned above, they are not necessary for understanding the statement (perfectly functional for writing messages or conversations on social networks). But we have to admit that it may change meaning at the word level—for example, in Danish én (one) versus en (the), in French là (there) versus la (the), and in Czech paní (lady) versus páni (lords). Nevertheless, the pragmatic context always clarifies the meaning of the word whether it does not have diacritical marks at all or whether they are used badly. Epigenetic marks do not have this character; to achieve a specific change, the correct mark has to be present. When dealing with epigenetics, one has to consider a large number of possible factors, such as environment, culture, and diet. To find a suitable linguistic analogy, it is advisable to focus on the phenomena affected by external factors. Which language level is thus most affected—phonology, morphology, syntax, or lexicology?

We already refused the phonological level (diacritic marks mostly represent different phonemes, e.g., Czech "a" is different from "á"), and we think that neither of the mentioned language levels offers a valid analogy for epigenetic linguistic features. As a matter of fact, we think that epigenetics is not a matter of linguistic levels, but yet is a matter of linguistic usage (in contradiction to the norm). As a consequence, we prefer Markoš's metaphor of reading rather than diacritics.

Now we can ask the question: How could a linguistic usage with individual occurrences be approached scientifically, or even how could it be useful for a language description (grammar)? Similarly to neo-Darwinism, in modern linguistics (since F. de Saussure), a "dogma" of understanding language as a normalized set of internal rules has also had a major influence on the research in the field. Notwithstanding, in the recent years, studies analyzing individual occurrences also appear. A small shift in the linguistic paradigm can be illustrated by analyses of the so-called hapaxes.

The hapax grammar of Czech was described by Novotná (2013), who analyzed corpus SYN (Czech national corpus)—the range of the corpus is 1.2 billion word forms. She carried out 20 random probes in the range of 3000 words. Before the selection, corpus was reduced by omitting numerals and non-alphabetical characters—after the selection, the numerals, errors, and so on were omitted. Hapax grammar was based on the remaining 30,000 word forms. The word forms were categorized (compounds words, derivation types, etc.). By this procedure, Novotná identified productive types of Czech grammar. In our analysis, we would like to avoid random selection of hapaxes, which is necessary for an analysis of a large-scale corpus. We would like to show that hapax analysis of much smaller corpus which contained full and authorized texts could lead to similar results. In this way, relations of the hapaxes to the rest of vocabulary can be preserved and the behavior of hapaxes in the text can be observed.

HAPAX GRAMMAR

Our analysis is based on a corpus of 12 popular science books from Nová beseda publishing company. The publications have the similar length—around 70 standard pages. Whole of the corpus encompasses 198,215 word forms (types). Authors had the same requirements for the text form—the explanation in the book should be accessible to readers from a discipline other than that presented by the book. This way of interpretation assumes a free way of handling the language, especially the terminology. For this reason—although the corpus is relatively small—a sufficient amount of the hapaxes was expected. Having the corpus prepared, the question was: Is it reasonable to analyze hapaxes? Is the epigenetic analysis of the text just a list of individual occurrences? We excerpted 8,150 single occurrences from the corpus. The sample contained proper names, numerals, abbreviations, terminology, and a few errors. The following were excluded from the analysis:

typing errors: mooc (too much), zacni (begin), zhromáždit (assemble), etc.

2 To specify, both genetics and epigenetics are affected by environmental influences, only at different time scales. For the purposes of this article, we only pay attention to fast changes influencing epigenetics.

foreign words: vicieux (brutal, cruel), bandcamp, nadah (advantages), millaise (what type, which kind), ereignis (incident), etc.
terms: kruhoústí (Cyclostomata), branchiální (branchial), nukleozomy (nucleosomes), neoliberalismus (neoliberalism), protoplazma (protoplasm), etc.

personal adjectives: Zarathustry (Zarathustra's), Deleuzovu (Deleuze's), Saussurovské (Saussurean), etc.

The rest of the vocabulary was compared with Czech National Corpus SYNV7 (4,255 billion words). Then we identified specific word forms for the Nová beseda series. We applied this procedure on the whole of the sample. In the reduced list of hapaxes, we found several types of words (some of the items belong to several categories):

compound words: psychospisovatel (psychowriter), sebedokumentování (self-documenting), trojprstost (threefingerness), sociáldarwinista (socialdarwinist), sebezmocnění (self-take possession), hyperčistý (hyperclean), mnohoalgoritmický (multialgorithmic), filmozofie (filmosophy = film + philosophy), sebekontrolovat (self-control), kvazivysvětlení (quasiexplanation), sebebližší (no matter how close), psychozábava (psychofun)

lexical negations: nediskurzivní (non-discursive), neusmívání (non-smiling), netestovatelný (non-testable), neobjektivovatelný (non-objectivable)

conversions: následkový (resulting), montážnický (installation), přetržitost (non-stopness)

aktionsart verb variants: zkritičněla (to become critical), zúzkostňovat (to make anxious), zahledání (search), odstigmatizovat (to unstigmatize)

affixation: prakinematografie (pracinematography), antikurátorská (anticurator)

suffixations by -ový: 35milimetrový (35 millimeter), multialbumový (multialbum)

derivations by -ost: stejnotvarost (homomorphic), nereprezentovatelnost (unrepresentativeness)

adapted loans: insistuje (to insist), leaknout (to leak), auteuři (from fr. auteurs = authors), neurologizace (neurologization), mashupový (mushup), reuniová (reunion), mockumentarista (mockumentarist = mock + documentarist)

occasional nouns: nezpřítomnitelnost (unpresentness), odracionalizovaný (derationalized), plnočíselný (fullnumerical)

neologisms: splynutec (mergenous one), zmocňovač (possessioner).

Words listed above are hapaxes from Nová beseda corpus whose occurrence in Czech National Corpus SYN V7 was 3 at most. Here are some words whose incidence was higher than 3 but still lower than 10: relevanční (relevant), důkazově (evidently), vrstevnatěji (more layered), nevěc (non-thing), nářečové (vernacular), rozhraničená (with determined boundaries).

As we can see, the analysis of the hapaxes shows the potentiality of the language (i.e., language needs to be and must be considered a norm because the norm is the possibility to be realized, see Mathesius 1911). With hapaxes, we cannot describe language norm—the word forms are not common—but the appearance of the norm is very much reflected in the text. Thus, we can highlight the epigenetic analogy: epigenetic processes must be based on a given genetic text, as individual word forms reflect common vocabulary of a given language.

After the hapax legomena analysis, we can consider text to be its own designer—in the same way that epigeneticist Anton Markoš considers life itself to be its own designer (Markoš et al. 2009). One note at the end: By repeating the hapaxes from Anton Markoš and Co je nového (What's new in…), we approximate these word forms to the language norm. By including Co je nového publications or this article into the Czech National Corpus, they become non-hapaxes. At the beginning of every norm stands a deviation which in time can become a usage and, consequently, a rule. The border of transformation from one state (hapax) to the other (norm) is really thin and usually is just a matter of time. As we mentioned earlier, time is the essence in the same way when it comes to epigenetic marks and their way towards becoming a norm. Thus, at the origin, everything starts with something epigenetic.

BORDERLINES OF BIOLOGY

Now more than ever, humanities and natural sciences are interdisciplinary. There are overlaps between disciplines, borrowing methods, using metaphors—the boundaries between disciplines are often thin and unclear. One issue is usually examined by multiple disciplines from different angles that interfere in some places. No field is the only one which has „know-how" on a particular subject anymore. Norms are being replaced by usages. (Epi)Genetics, and biology in general, are no exceptions. We are becoming more and more aware that somatic issues are not just a matter of the body. They are the result of many factors—from inborn attributes, through the influence of the environment, to the individual's specificities. This range of factors is exactly the reason and the area where biology establishes relations with other fields.

In order to explore the relations of biology with other sciences, we analyzed Anton Markoš's Co je nového v biologii (What's new in biology) and other 11 texts from Co je nového (What's new in…) about different sciences from Nová beseda publishing company. According to Anton Markoš's text, biology is close to linguistics and artificial intelligence, and we can see the rest of the relations between contemporary sciences (see Figure 1).

CONCLUSION

As we pointed out, beside the tendency of humanities, as well as natural sciences, to generalize, create ideal models, and search proofs for existence of norms, there is space for individual occurrences of research.
In case of biology, we discussed epigenetics and hapax legomena within linguistics, which led us to exploring an analogy between them. Based on a hapax analysis in

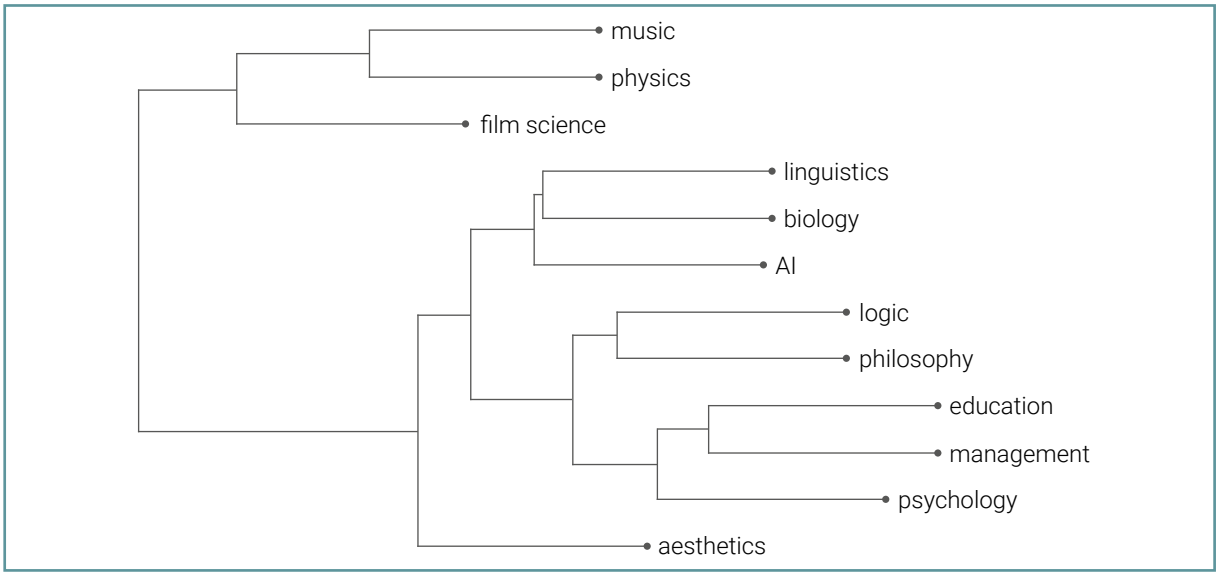

Figure 1: Presence discourse proximity of selected sciences based on edition Co je nového (What's new in…). The sample contains What's new in music, physics, biology, film science, aesthetics, psychology, education, management, logic, philosophy, linguistics, and artificial intelligence: bag-of-words analysis of bigrams of lemmatized words.

a corpus of 12 authentic books and creating their new classification, we proved the potentiality of Czech language to produce new categories by stepping out outside the norm.

Acknowledgment: The work was supported by Technology Agency of the Czech Republic, Project: Interactive technical books—Redefinition of electronic publishing in the field of non-fiction TL02000530.

REFERENCES

Barthes, R., 1974. *The Death of the Author*. New York: Hill and Wang.

Jablonka, E., Raz, G. D., 2009. Transgenerational epigenetic inheritance: Prevalence, mechanisms, and implications for the study of heredity and evolution. Q. *Rev. Biol*., 84(2), 131-176.

Eco, U., 2007. *Dall'albero al labirinto: Studi storici sul segno e l'interpretazione, Delfini Pocket*, La nave di Teseo.

Ji, S., 1991. *Molecular Theories of Cell Lie and Death* (Ji, S. ed.). New Brunswick, NJ: Rutgers University Press, pp. 52–56.

Markoš, A., 2002. *Readers of the Book of Life: Contextualizing Developmental Evolutionary Biology*, Oxford University Press.

Markoš. A., Grygar, F., Hajnal, L. et al., 2009. *Life as its own designer: Darwin's Origin and Western thought*. Springer.

Mathesius, V., 1911. *O potenciálnosti jevů jazykových*. Praha: Věstník Královské české společnosti nauk.

Novotná, R., 2013. Jazyková potencialita: studium na bázi hapaxů legomenon. *KGA* 8, 47–58.

Pattee, H. H., 2001. The Physics of Symbols: Bridging the Epistemic Cut. *Biosystems*, 60(1-3), 5–21.

Pattee, H. H., 2008. Physical and Functional Conditions for Symbols, Codes, and Languages. *Biosemiotics*, 1(2), 147–168.

Pattee, H. H., Kull, K., 2009: A biosemiotic conversation: Between physics and semiotics. *Sign Systems Studies*, 37(1/2), 311–331.

Peirce, Ch. S., 1906. *On the System of Existential Graphs Considered as an Instrument for the Investigation of Logic*. MS [R].

Švorcová, J., Markoš, A., 2019. E*pigenetic Processes and Evolution of Life*. Boca Raton. CRC Press, https://doi.org/10.1201/9781351009966

Vachek, J., 2003. *Dictionary of the Prague school of Linguistic*. John Benjamins: Amsterdam/Philadelphia.

Return to front page ↑