Romanian Reports in Physics, Vol. 60, No. 3, P. 417–422, 2008

ON THE SYMBIOSIS OF PHYSICISTS AND LINGUISTS

GABRIEL ALTMANN

Sprachwissenschaftliches Institut, Ruhr-Universität Bochum, Germany

(Received March 12, 2008)

Forty years ago, a physicist asked me what I am engaged in. I answered: "In quantitative linguistics". "Oh, yes," he said, "you count letters, don't you?" I was surprised at his excellent information about the 19th century and said: "No, I never did. I try to find laws in language, just as you try to do in the nature." Never in my life I saw a face more perplex than his. But since then, the situation changed drastically. Today, there are many physicists counting letters, hoping to find physical laws behind them. And once in a decade they discover that letters behave like mesons and create a wonderful theory. It is not valid, but it is wonderful.

But quite seriously, the engagement of physicists in linguistics was almost always associated with progress. It is not so much the mathematical apparatus they bring in, but rather the way of thinking which, due to their education, is quite "natural" to them but fully foreign to linguists who have a very modest mathematical and "non-linguistic" knowledge. But this way of thinking conceals the same danger if applied to language as that lying in wait for linguists who would try to transfer their fuzzy thinking in quantum theory. Physicists would politely smile with the left corner of their mouth concealing it behind a moustache, but linguists usually roar with malicious laughter because they finally found somebody who makes still greater errors than they themselves. In any case, linguists know that "to have a body" does not mean "to be physicist" but physicists mostly do not know that "to speak a language" does not mean "to be linguist".

As a matter of fact, the greatest problem is the undifferentiated identification of physical and linguistic entities. The "thinghoods" (as G. Klir called them) are different, but somewhere at the high level of general systems one may observe analogical behaviour. Nowadays no physicist tries to reduce linguistics to physics and even linguists know that if one finds an analogy, the physical entity is not basic and the linguistic entity is not just its special case under different boundary conditions but perhaps both of them display a behaviour which can be captured by the same very general stochastic process. We have the same rights. That's all. One should not try to transfer the gravitation theory to dialectology (once made by a linguist!) or the complete thermodynamics to a linguistic discipline only because the concept of entropy is applicable both in physics and linguistics. The differences

in thinghood are drastic. In linguistics, entropy is only an index of diversification which can be expressed in different ways, and Shannon's entropy can be transformed in different other measures having nothing in common with thermodynamics. Thus, if there is some analogy between physical and linguistic entities (which are, by the way, the most immaterial entities of our world, cf. *e.g.* the meaning; but even the lowest level, phonemics, contains only conceptual constructs), there must exist a way to find it without reductionist attempts and respecting the thinghood of physical and linguistic entities. The only possibility is to take into account the properties of models, *i.e.* to scrutinize the common abstract super-systems – if there are any.

The last way has been chosen by Ioan-Iovitz Popescu who began his late linguistic career with the study of the h-point on the rank-frequency curve of words. This first article appeared in 2007 but was known to linguists already in 2006. Instead of separating words in classes (a frustrating problem even for linguists but many times tried by physicists) and instead of modifying the famous Zipf's law (a never ending enterprise of physicists) he found a fixed-point on the empirical rank-frequency curve, showed several possibilities of its computation and asked what it could mean. Being a wise physicist aware of the above mentioned fact that "even to speak many languages" does not mean "to be linguist" he contacted linguists. And linguists found that the h-point in the word-frequency study is a kind of revolution, a kind of turning point in the history of word frequency study. Iovitzu himself proposed several interpretations, found some other remarkable points, brought new vistas concerning text and caused a paper-tsunami on my desk and a chaos in my computer. His e-mails full of new ideas came daily. For the sake of security I made a print-out of all. On the left side of my desk I erected a Mount Everest consisting of his new ideas, in the mid there was a shaky heap of his data, and his figures were placed at the right side. Fortunately, this right mountain could not break down because it soon reached the ceiling. I placed my computer in another room and performed my daily Iovitzu-Marathon between two rooms which is responsible for my present fitness. Every evening I looked under my first desk suspecting that Iovitzu is hidden there and enlarges the heaps himself.

Just in order to reduce the heaps we began to write down his ideas and each article passed about 20 versions. This was the usual norm to cook me ready. We began with "Some aspects of word frequencies" (2006a), continued with his geometrical ideas "Some geometric properties of word frequency distributions" (2006b), used his h-point to define thematic concentration and autosemantic compactness in "Writer's view of text generation" (2007a), "On the dynamics of word classes in texts" (2007b) and "Autosemantic compactness of texts" (2008a). Silly as I am I hoped that the Himalaya on my desk will get smaller. On the contrary, the Iovitzu-idea-heap converged to the ceiling and the figure-heap got a twin. The rest of the desk was occupied by data. Sometimes I saw Dracula on my monitor but it was surely a hallucination.

In this hopeless situation I employed Kant's categorical imperative and said: "Let us write a book!" This was the only remedy – except for fire brand in my house. We asked several colleagues to send us data in 20 languages and wrote "Word frequency studies" (2008b), a book with a short title but with eleven authors from 6 states. Even a mathematician (J. Maèutek) and the founder of synergetic linguistics (R. Köhler) took part in writing it. In order to show the possibility of testing we wrote a separate article "Confidence intervals and tests for the h-point and related text characteristics" (2007c) which appeared separately before we finished the book. Peculiar enough, it was not Iovitzu who insisted on the normality of deviations – an assumption never holding true in linguistics (!) but quite usual in physics – but our mathematician whose fate would be sealed in his community if he admitted anything else. In some sciences normality is a necessary assumption, in linguistics it is an absolutely irrational one.

If one analyzes one unique language and uses a known mathematical model, there are two possibilities: the model is valid or it is not valid, even if any empirical corroboration is only a matter of degree. One should determine the rejection criteria a priori. If one applies the model to another language, its validity will usually be disturbed because languages – even related ones – can behave very differently. In that case one "modifies" the model. But if one takes 20 languages at once, the models are exposed to enormous risk of falsification and in case of validation they get a high degree of corroboration. This is why "modelling" linguists should never set up their models for one language only – a children's ailment not own to physicists. If we strive for finding laws, they must hold for all languages (even dead ones) and, perhaps, for other communication systems, too (whales, bees, etc.). But using 20 languages simultaneously can show even the differences between languages. And this was Iovitzu's next discovery concerning the importance of texts. Though linguistically compelling and logical, it could not easily be shown up to now because linguists usually analyze frequencies only in one language. But it is quite logical that highly synthetic languages have more forms, and if we count word forms, the number of hapax legomena will be greater than in highly analytic languages in which the words are repeated more frequently because there are fewer forms. Thus word-counts and morphological typology of languages are closely connected. And since morphological properties are closely connected with all other properties of language, the frequency study of texts gets a quite new importance. In order to show the significance of this fact, we wrote "Hapax legomena and language typology" (2008c), "Zipf's mean and language typology" (2008d) and "On the diversity of word frequencies and language typology" (2008i) opening a number of possibilities for further research.

The first ten articles and a book within 18 months was the result of Iovitzu's first engagement in linguistics. I ordered the lorry of a paper mill and we loaded the Himalaya through my window filling the lorry to its full capacity. I put ten CDs from my earlier desk to my book-case, transported the computer to its old place

and checked my daily e-mails. There were about 50 spams and a short e-mail from Iovitzu: "Dear Gabriel, I stated that everything holds also for music!" In that very moment I heard the beginning of Beethoven's Fifth Symphony from the radio – but that was definitely by chance. I checked the data he sent me in the attachment and answered: "Dear Iovitzu, you are right, as usually."

Since we are people knowing that knowing to play the piano does not mean to be music theoretician, we sought a victim and found Zuzana Martináková, Slovak music theoretician engaged in quantitative musicology. She supplied us with containers full of data and her theoretical and historical knowledge and Iovitzu had again his fling. We were sure that other musicologists will not understand us because they look at the restricted matter and not at a possible super-system, but we all have problems with our nearest relatives. Taking also J. Mačutek in our TA (= Transsylvanian Alliance) we wrote three articles: "Some problems of musical texts" (2008e), "Ord's criterion for musical texts" (2008f) and "On stratification in music" (2008j) in which we described some new properties of compositions and their development in European music. Just as in linguistics, in musicology, too, one will need some time to understand Iovitzu's different view, even if he never tried to hurt the thinghood of language and music and never operated with physical theories. He sees structures and their similarities and sets up common models starting from real data. Hence he drastically differs from those theoreticians who set up an abstract model and seek its realizations. In mathematics we ourselves can determine the assumptions under which a theorem holds, but in empirical sciences like linguistics the assumptions are given.

After this musicological intermezzo we turned back to our original destination, to linguistics. While writing the next article "A new text indicator" in which Iovitzu defined a very simple index associated with morphological properties of language and tested it, as usually, on several languages, he made perhaps the most courageous step in the domain of Zipf's law. It can be shown that Zipf's power law (or zeta distribution) holds in all cases in 20 languages, and its numerous modifications (made by linguists, mathematicians and physicists) are not quite necessary though they bring mathematicians a deep aesthetic satisfaction. It is the dominant paradigm in word frequency studies since 70 years, it has been introduced to different scientific disciplines – even to chaos theory, fractals, sand-piles, etc. (cf. http://www.nslij-genetis.ord/wli/zipf), it can be derived in different ways and everything in its domain is only a variation of a given melody. One goes from the zeta to the polylogarithmic and the Lerch functions, adds different components to the exponent, etc. But in linguistics it has a weak interpretation and this was the cause of many (vehement) discussions in the 20th century. Zipf's discovery was a discovery of a regularity, and its first (lucky) approximation – performed axiomatically – survived for decades. However, linguistic entities behave in texts in a very special way. The words build many different strata and each stratum has its own ranked course. This course can be Zipfian (zeta) or even

exponential, there is no reason to prefer one of them! Now, Iovitzu proposed to use the exponential function (this is perhaps the only symptom of his knowledge of physics) and to pool all strata resulting in a superposition of exponentials. Peculiarly, two components of the superposition are sufficient to capture the rank-frequency distribution as a sequence in all texts he tested in 20 languages. The dependent variable is the proportion at the given rank. Needless to say, in almost all cases the fitting is better than using the zeta function and need not be rejected even in one case. Now, since we have a model which has a good linguistic interpretation and can be better fitted to data than the original Zipf, we can take leave of the beloved good old paradigm. Nevertheless, R. Köhler and myself – the coauthors of Popescu's "Zipf's law – another view" (2008f) – remain in all other domains persuaded Zipfians. If one wants to set up theories in linguistics, one cannot avoid Zipf, on the contrary, his ideas must be taken into account at every occasion.

At last, Iovitzu put his thumb at the peculiar problem of diversification initiated by G. K. Zipf. Diversification is a process ruling over the entire speech activity of Man, it is one of the causes of development and variation. Iovitzu has shown thwt is a very regular process having the same form at all levels of language but each level can be identified by means of the value of a coefficient which is very stable in all languages analyzed. Fengxiang Fan, a Chinese specialist for English, helped us to process English data (Popescu, Mačutek, Altmann 2008k, Fan, Popescu, Altmann 2008l, and Popescu, Altmann 2008m).

This time, the intervention of a physicist in linguistics can be considered a full success. Quantitative linguists already accept Iovitzu's ideas and try to catch his train taking a running jump. As can be seen in the references, their number increases. Under Iovitzu's baton, text analysis develops to a quite new science. It must be remarked that no linguist – except for Zipf himself - has ever made as deep changes in our thinking about texts as Iovitzu. We know that in human sciences paradigms come and go, especially because they are made "qualitatively", based on one language only (in linguistics) and using ad hoc concepts. Their main aim is description and classification. Since texts, *i.e.* the use of language means (parole, performance) decides at long sight about the image of language we have in our heads – and not only the other way round – the study of texts is the door which must be passed if we want to learn something about "the" language. It does not consist only of words and grammatical rules, it abides by mechanisms which are not even subject of curiosity in qualitative linguistics. They are hidden somewhere, not conscious and not learnable, but effective in the same way as natural laws. Needless to say, they are all stochastic. There is nothing deterministic in language, even if grammatical rules lying on the surface of language may sometimes evoke this impression. If something changes in language or in text, then something else changes too. However, not functionally but stochastically, with time delay, etc. There is a strict self-regulation in language discovered by G. K. Zipf, and this is the domain where physicists and linguists can meet. Language is not a special case of

matter, words are not acoustic oscillations, but physics and linguistics can meet on the way taken by Iovitzu.

We all hope that this way will continue but nobody can look in Iovitzu's head and predict all his surprising discoveries. He himself probably at least. I personally would be glad if they could continue further 75 years, because in that case he would live in health further 75 years.

PS. The idea that all linguists would be forced to study 4 semesters mathematics and physics fills my soul with malicious joy.

REFERENCES

1. Popescu, I.-I. (2007), Text ranking by the weight of highly frequent words, in: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 555–565.* Berlin, New York.
2. Popescu, I.-I., Altmann, G. (2006a), Some aspects of word frequencies, *Glottometrics,* **13**, 23–46.
3. Popescu, I.-I., Altmann, G. (2006b), Some geometric properties of word frequency distributions, *Göttinger Beiträge zur Sprachwissenschaft,* **13**, 87–98.
4. Popescu, I.-I., Altmann, G. (2007a), Writer's view of text generation, *Glottometrics* **15**, 71–81.
5. Popescu, I.-I., Best, K.-H., Altmann, G. (2007b), On the dynamics of word classes in texts, *Glottometrics,* **14**, 58–71.
6. Popescu, I.-I., Altmann, G. (2008a), *Autosemantic compactness of texts*, in Altmann, G., Zadorozhna, I., Matskulyak, Y. (eds.), *Problems of General, Germanic and Slavic Linguistics*. Papers for the 70th Anniversary of Professor V. Levickij, 2008, Chernivtsi: Books – XXI.
7. Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mehler, A., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B. D., Grzybek, P., Altmann, G. (2008b), *Word frequency studies,* Berlin, New York (in print).
8. Mačutek, J., Popescu, I.-I., Altmann, G. (2007c), Confidence intervals and tests for the h-point and related text characteristics, *Glottometrics,* **15**, 42–52.
9. Popescu, I.-I., Altmann, G. (2008c), Hapax legomena and language typology, *Journal of Quantitative Linguistics* (in print).
10. Popescu, I.-I., Altmann, G. (2008d), Zipf's mean and language typology, *Glottometrics,* **16**, 31–37.
11. Martináková, Z., Mačutek, J., Popescu, I.-I., Altmann, G. (2008e), Some problems of musical texts, *Glottometrics*, **16**, 80–110.
12. Martináková, Z., Mačutek, J., Popescu, I.-I., Altmann, G. (2008f), Ord's criterion for musical texts (submitted).
13. Popescu, I.-I., Altmann, G., Köhler, R. (2008g), Zipf's law – another view (submitted).
14. Popescu, I.-I., Altmann, G. (2008h), A new text indicator (msc).
15. Popescu, I.-I., Altmann, G. (2008i), On the diversity of word frequencies and language typology, *Göttinger Beitrage zur Sprachurssenschaft*, **14**, 2008, 83–91.
16. Popescu, I.-I., Martináková, Z., Altmann, G. (2008j), On stratification in music (submitted).
17. Popescu, I.-I., Mačutek, J., Altmann, G. (2008k), Word frequency and arc length, *Glottometrics*, **17**, 18–44.
18. Fan, F., Popescu, I.-I., Altmann, G. (2008l), Arc length and meaning diversification in English, *Glottometrics*, **17**, 82–89.
19. Popescu, I.-I., Altmann, G. (2008m), On the regularity of the diversification in language, *Glottometrics*, **17**, 97–111.