The Henryk Niewodniczański
Institute of Nuclear Physics
Polish Academy of Sciences

Radzikowskiego 152
31-342 Kraków, Poland
switchboard: +48 12 6628000
director: +48 12 6628200
fax: +48 12 6628458
www.ifj.edu.pl
dyrektor@ifj.edu.pl

Krakow, 30 November 2016

Overlooked elements of language and literature play a key role

Everything is pointing towards success in unravelling the mysteries inherent in every human language, which for nearly 100 years have been an object of intrigue for mathematicians and linguists working on studies into statistics of literature. New analysis of the frequencies of word occurrence in the most famous works of literature, undertaken at the Institute of Nuclear Physics of the Polish Academy of Sciences in Krakow, have shown that our languages are structurally more complex and more exhaustive than they ever before seemed.

It's been said that 80% of a person's success is achieved from only 20% of their efforts. That famous ratio holds up over a surprising number of domains. For example, it is apparent that in every language, whether spoken or written, that 80% of all statements are made up of merely 20% of the most common words. One possible reason is that when we talk to each other we want to convey as much content as possible with the least effort (among other factors). This phenomenon of dependency was one of the earliest of the series of power laws to be discovered, and is known as Zipf's law. It has turned out that it is not as trivial as it might seem at first glance. Scientists from the Institute of Nuclear Physics of the Polish Academy of Sciences (IFJ PAN) in Krakow have established that certain puzzling features of Zipf's law, for decades a source of intrigue for those involved in the statistical analysis of literary texts, are a consequence of neglecting one of the basic components of language.

"A year ago, with the help of detailed multiscale analysis we showed that the length of sentences in literature – that is, the distance between the sentence-ending punctuation – shows a very complex dependency of a multifractal nature, especially evident in the works of the genre known as stream of consciousness. It was an intriguing result that prompted us to look with greater attention to the role of other punctuation marks, especially in the context of Zipf's law. The results have provided us with a new way to look at not only the role of punctuation among languages, but even within the same language," says Prof. Stanislaw Drozdz (IFJ PAN, Cracow University of Technology).

Charts showing Zipf's law for literary texts are created with the use of an uncomplicated procedure. Each word is counted per how often it occurs in the text. Those which occur most often are assigned to rank 1, the next are placed in rank 2, etc. (in richer texts the ranks can even exceed ten thousand, and exotic words usually appear far above rank 1,000). Zipf's law states that the probability of a word is inversely proportional to its rank: the larger the rank, the lower the probability. Graphs showing the relationship are (on a logarithmic scale) on a straight line.

Ever since the American linguist George Zipf popularized his law, it never ceases to amaze. How can something as complex as a structure created using language be described by such a straightforward law? There were more puzzles. Quite early on it was noticed that the graphs

relating to the frequency of words for ranks closer to unity curve slightly downward from a straight line. That deviation particularly intrigued Benoit Mandelbrot, the great French mathematician of Polish origin, who worked on this issue for many years. He even suggested his own amendment to Zipf's original law, to better map deviation (it is worth mentioning at this point, that his work on Zipf's law, among other interests, helped guide him toward the concept of fractals).

In this latest study, the physicists at IFJ PAN analyzed texts written in six Indo-European languages, two of each belonging to these groups: Germanic (English and German), Romance (French and Italian) and Slavic (Polish and Russian). The selected literary works come from the archives of Project Gutenberg (www.gutenberg.org), and each is at least five thousand sentences long. For each of the languages at least five different texts were chosen, and merged to form one text totaling about a million words. All words unrelated to the transmitted content were removed, such as 'chapter', 'part' and 'epilogue', and for all language-specific abbreviations like 'Mr.' and 'Dr.' the dots were removed and they were treated as separate words. Also deleted were annotations and footnotes, page numbers, and punctuation marks of a more typographical nature: quotation marks and parentheses.

"The eventual punctuation marks considered for analysis were commas, colons, and semicolons, and those which end sentences: periods, exclamation marks, question marks, and ellipses," specifies Prof. Jaroslaw Kwapien (IFJ PAN), one of the co-authors of the scientific paper published in the renowned journal Information Sciences.

Among the studied works of literature are: *1984* by George Orwell, *Moby Dick* by Herman Melville, Ulysses by James Joyce, *Gulliver's Travels* by Jonathan Swift, *Gone with the Wind* by Margaret Mitchell, *Thus Spake Zarathustra* by Friedrich Nietzsche, *The Trial* by Franz Kafka, *The Magic Mountain* by Thomas Mann, *Madame Bovary* by Gustave Flaubert, *The Phantom of the Opera* by Gaston Leroux, *Foucault's Pendulum* by Umberto Eco, *Giacinta* by Luigi Capuana, *The Spring to Come* by Stefan Zeromski, *The Promised Land* by Wladyslaw Reymont, *The Doll* by Boleslaw Prus, *Anna Karenina* and *War and Peace* by Leo Tolstoy, and *The Brothers Karamazov* by Fyodor Dostoevsky.

Including punctuation marks led a to highly significant result: the downward bend seen on the original Zipf's graph for ranks close to unity practically disappears. The 'new words' (the punctuation marks) fall into place almost exactly so that together with the 'real words' the rank-frequency distribution now fits into the straight line for all ranks, thus extending the original form of Zipf's law to all the scales. Mandelbrot's amendment turned out to be generally redundant.

"When we begin to treat punctuation marks like they are words they start occupying ranks closer to unity and with frequency relative to the ordinary words, so that the original Zipf departure from the straight line for small ranks basically disappears. Thus, upon considering punctuation, our language emerges as a more consistent composition! That's why it seems well-founded to say that punctuation is just as important to a language as its words, and language without it is basically incomplete," says Prof. Drozdz.

New graphs reveal several novel and significant features. For example, considering punctuation in Slavic languages the thus generalized Zipf rank-frequency distribution falls almost perfectly along one line for all ranks. Some trace of the original Zipf's deviation remains for Romance and Germanic languages, and this is especially apparent with the English language.

"What if while analyzing non-Slavic languages we didn't consider their additional specific features?" wonders Prof. Drozdz, being mindful of other interesting interpretations: "Might it also be that the cause of incomplete reduction of curvature is rooted in the language itself? For instance, in English there might be a source of easily discernable tendencies of authors to limit the number of punctuation marks. If this last cause holds true, it might be worth asking the question: can we be sure that excessive reduction of punctuation is a beneficial action that doesn't harm the integrity of the language?"

The latest discovery from the IFJ PAN could potentially have implications beyond linguistic research. Zipfian deviation for ranks closer to unity is being observed in many areas and has diverse origins, which are often not fully understood. In the graphs prepared based on literary works the deviation disappears after accounting for the common factor, but so far this has been considered negligible. Perhaps in other cases it could also be eliminated by including elements which have thus far been deprived of a greater role.

The Henryk Niewodniczański Institute of Nuclear Physics (IFJ PAN) is currently the largest research institute of the Polish Academy of Sciences. The broad range of studies and activities of IFJ PAN includes basic and applied research, ranging from particle physics and astrophysics, through hadron physics, high-, medium-, and low-energy nuclear physics, condensed matter physics (including materials engineering), to various applications of methods of nuclear physics in interdisciplinary research, covering medical physics, dosimetry, radiation and environmental biology, environmental protection, and other related disciplines. The average yearly yield of the IFJ PAN encompasses more than 450 scientific papers in the Journal Citation Reports published by the Thomson Reuters. The part of the Institute is the Cyclotron Centre Bronowice (CCB) which is an infrastructure, unique in Central Europe, to serve as a clinical and research centre in the area of medical and nuclear physics. IFJ PAN is a member of the Marian Smoluchowski Kraków Research Consortium: "Matter-Energy-Future" which possesses the status of a Leading National Research Centre (KNOW) in physics for the years 2012-2017. The Institute is of A+ Category (leading level in Poland) in the field of sciences and engineering.

CONTACTS:

Prof. **Stanisław Drożdż**
The Institute of Nuclear Physics of the Polish Academy of Sciences
tel. +48 12 6628220
email: stanislaw.drozdz@ifj.edu.pl

SCIENTIFIC PAPERS:

“In narrative texts punctuation marks obey the same statistics as words”; A. Kulig, J. Kwapień, T. Stanisz, S. Drożdż; Information Sciences 375 (2017) 98–113; DOI: http://dx.doi.org/10.1016/j.ins.2016.09.051

LINKS:

http://www.ifj.edu.pl/
The website of the Institute of Nuclear Physics of the Polish Academy of Sciences.

http://press.ifj.edu.pl/en/
Press releases of the Institute of Nuclear Physics of the Polish Academy of Sciences.

IMAGES:

IFJ161130b_fot01s.jpg **HR:** http://press.ifj.edu.pl/news/2016/11/30/IFJ161130b_fot01.jpg
Recent research conducted at the Institute of Nuclear Physics of the Polish Academy of Sciences in Krakow reveals that in narrative texts punctuation plays as important role as words. (Source: IFJ PAN)

IFJ161130b_fot02s.jpg **HR:** http://press.ifj.edu.pl/news/2016/11/30/IFJ161130b_fot02.jpg
Probability of occurrence of words (vertical axis) versus their rank (horizontal axis) for corpora representing different European languages. The original puzzling downward departure from the straight line for ranks close to unity, observed for the ordinary words (brighter colors), disappears (corresponding darker colors) when the punctuation marks are also taken into account. (Source: IFJ PAN)