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

Cracow, 11 April 2019

Texts like networks: How many words are sufficient to recognize the author?

We are more original than we think – this is what is being suggested by literary text analysis carried out by a new method of stylometry proposed by scientists from the Institute of Nuclear Physics Polish Academy of Sciences in Cracow. The author's individuality can already be seen in connections between no more than a dozen of words in English text. It turns out that in Slavic languages authorship identification requires even fewer words and, in addition, it is more certain.

Finding out the author of a text is usually not difficult: just read the signature. However, sometimes there is no signature since it has not been preserved or has been deliberately omitted by the author. Often, instead of a first and last name, we see a pseudonym. So, how can we verify who penned a historical text known only from fragments? How can we establish the true creator of an Internet lampoon? How can we really determine if the text of a thesis or doctoral dissertation is not plagiarized? In many cases, traditional stylometric methods fail or do not lead to sufficiently reliable conclusions. In *Information Sciences*, scientists from the Institute of Nuclear Physics of the Polish Academy of Sciences (IFJ PAN) in Cracow have presented their own statistical tool for stylometric analysis. Constructed with the use of graphs, it makes it possible look at the structure of texts in a qualitatively new way.

“The conclusions of our research are, on the one hand, encouraging. They indicate that the individuality of any person manifests itself clearly in the way they use a surprisingly small number of words. But there is also another, darker side of the coin. Since it turns out we are so original, it will be easier to identify us by our statements,” says Prof. Stanislaw Drozdz (IFJ PAN, Cracow University of Technology).

Stylometry – i.e. the science dealing with the determination of the statistical characteristics of the style of texts – is based on the observation that each of us uses even the same language in a slightly different way. Some have a broader vocabulary, others narrower, some like to use certain phrases and make mistakes, others avoid repetition and are linguistic purists. And when we write, we also differ in the way we use punctuation. In the typical stylometric approach, the basic features of a text are usually examined, e.g. the frequency of occurrence of individual words, whilst punctuation is ignored. Analyses are carried out for the studied text and for texts written by potentially well-known authors. The creator is deemed to be the person whose works have parameters with the values closest to those obtained for the material being identified.

“We suggested that the characteristic features of the style be sought in a network representation of the text, using graphs,” explains Tomasz Stanisz, PhD student at the IFJ PAN and the first author

of the publication, and he specifies: “The graph is a collection of points, or vertices of the graph, connected by lines, i.e. the edges of the graph. In the simplest case – in the so-called unweighted network – the vertices correspond to individual words and are connected by edges if and only if two given words have occurred adjacent to each other at least once in the text. For example, for the sentence ‘Jane is hungry’, the graph would have three vertices, one for each word, but there would only be two edges, one between ‘Jane’ and ‘is’, the other between ‘is’ and ‘hungry’.”

While constructing their stylometric tools, the IFJ PAN researchers tested different types of graphs. The best results were obtained for weighted graphs, that is, those in which each edge carries information about the number of occurrences of its corresponding connection between words. Two parameters turned out to be the most useful in such networks: the node degree and the clustering coefficient. The first describes the number of edges coming from a given node and is directly related to the number of occurrences of a given word in the text. In turn, the clustering coefficient describes the probability that two words connected by an edge with a given word are connected with an edge also between themselves.

Using statistical tools prepared in this way, the Cracow-based physicists looked at 96 books: six novels by eight well-known English authors (Austen, Conrad, Defoe, Dickens, Doyle, Eliot, Orwell and Twain) and eight Polish authors (Korczak, Kraszewski, Lam, Orzeszkowa, Prus, Reymont, Sienkiewicz and Zeromski). The authors included two winners of the Nobel Prize for Literature (Wladyslaw Reymont and Henryk Sienkiewicz). All the texts were downloaded from the internet libraries Project Gutenberg, Wikisource and Wolne Lektury. The group from the IFJ PAN then checked the reliability with which the authorship of 12 randomly selected works in one language could be determined, treating the rest of the pool of works as comparative material.

“In the case of English texts, we identified the authors correctly in almost 90% of cases. In addition, in order to achieve success, it was necessary to trace the connections between only 10-12 words of the examined text. Contrary to naive intuition, a further increase in the number of words studied did not significantly increase the effectiveness of the method,” says Tomasz Stanisz.

In Polish, the determination of authorship turned out to be even simpler: only 5-6 words needed to be traced. What is particularly interesting is that despite the fact that the pool of significant words was half as many as in English, the probability of correct identification was increased by up to 95%! Such high diagnostic accuracy, however, was only achieved when punctuation marks were also treated as separate words. In both languages, omitting punctuation resulted in a significant reduction in the number of correct guesses. The observed role of punctuation is another confirmation of the conclusions from the publication of the group of Prof. Drozdz of 2017, where it was shown that punctuation plays an equally important role in language as the words themselves.

“In comparison with English, Polish seems to give greater possibilities of revealing the style of the author. We think that the other Slavic languages are characterised by similar features. English is a positional language, which means that the order of the words in a sentence is important. This sort of language leaves less room for an individual style of expression than the Slavic languages, in which inflection, or variation, decides about the role of a word or phrase in a sentence. This allows for greater freedom to organize the order of words in a sentence, whilst its meaning remains unchanged,” sums up Prof. Drozdz.

CONTACTS:

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

SCIENTIFIC PAPERS:

1. "Linguistic data mining with complex networks: A stylometric-oriented approach"
T. Stanisz, J. Kwapień, S. Drożdż
Information Sciences 482 (2019) 301–320
DOI: https://doi.org/10.1016/j.ins.2019.01.040

2. "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 Polish Academy of Sciences.

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

IMAGES:

IFJ190411b_fot01s.jpg **HR:** http://press.ifj.edu.pl/news/2019/04/11/IFJ190411b_fot01.jpg
The author of an unsigned text can be identified by analysing the relationship between just a few words of the text, as shown by physicist-statisticians from the Institute of Nuclear Physics of the Polish Academy of Sciences in Cracow. (Source: IFJ PAN)