A Second Glance at a Stylometric Map of Polish Literature

Jan Rybicki

Introduction

A first glance at how stylometric statistics of frequently used words can lay out on one plain a map of Polish letters was put forth several years ago in a text with precisely that title.[1] In 2014, 500 Polish books seemed to be a substantial selection of texts. Today, we might revisit this subject, for the research presented below uses a sample over five times this size. Its scope includes Polish texts running from 14th-century sermons, *Kazania świętokrzyskie*, to a recent bestseller, *Tajemnica domu Helclów* by "Szymiczkowa", Polish novels, epic poetry and drama, as well as Polish translations of English-language texts (including numerous translations of Shakespeare) and from French, Russian, German, Czech, Italian, Portuguese, Spanish, Hungarian and Turkish texts.

One might venture to say that until now, distant reading has not yet been tried on such a vast collection of Polish texts. In the years since 2014, we have finally managed to usher this notorious term into Polish, mainly thanks to the Polish translation of the first book by the creator of this concept.[2] Surely this is not the case – though I am not at all convinced that we are truly dealing with *distant reading* here. As Matthew Jockers has noted, Franco Moretti, his former boss at Stanford University, studies texts "from the outside" and "from afar" by means of publication data, travel maps — actual and virtual — of authors and literary figures and, like a good Marxist, charts the Darwinian evolution of genres and literary forms in order to develop and broadcast his own literary genetics, even making explicit reference to DNA research. Of course, this is all rather interesting, but Jockers does not support expanding Moretti's term to computer stylistics undertaken alongside him and before him by scholars such as John Burrows, Hugh Craig, Karina van Dalen-Oskam, David Holmes, David Hoover Richard Forsyth and Fotis Jannidis (and in Poland, Adam Pawłowski, Maciej Eder and this article's author): this work entails "reading" many works at once with the help of statistics that

[1] J. Rybicki, *Pierwszy rzut oka na stylometryczną mapę literatury polskiej,* "Teksty drugie" 2014, issue 2, p. 106-128.

[2] F. Moretti, *Wykresy, mapy, drzewa. Abstrakcyjne modele na potrzeby literatury*, trans. T. Bilczewski and A. Kowalcze-Pawlik, Kraków 2016.

made their way into literary studies directly via authorship attribution and are oriented more towards the comparison of "linguistic" elements of the text (such as frequencies of words, of their the root forms, of parts of speech...). The discipline of knowledge most often appearing under the rubric of stylometry – a term we can attribute to Wincenty Lutosławski – Jockers has astutely called "macroanalysis"[3]... and it remains unclear if this name will stick. In any case, it is crucial to note this distinction, for different advantages– and disadvantages– emerge from these two related methods for observing literary texts.

Method

The major drawback of what we have agreed, at least provisionally, to call macroanalysis is its departure from traditional literary studies' emphasis on the meaning of the text, its "contents" and its "message". This comes about because in macroanalysis, the scholar commands an insentient machine to devour text after text, chop each one into individual words, and then count the decimated remains of sentences, paragraphs and chapters torn from all context in order to establish a list of the most frequently appearing words. These very words, however, repeat themselves *ad nauseum*, for any natural language – even that of dolphins – consists above all of the shortest words that have the least "semantic" value and are rarely assigned meaning (for this is how we might summarize Zipf's three laws of distribution). It follows that in each natural language, the set of the hundred most frequently used words hardly includes a single word of definite meaning (in the sample studied here, one is pressed to find words like "eyes" and "home", while drowning in a flood of conjunctions, prepositions and pronouns) – yet these words constitute roughly half of any given text. Long ago, John Burrows wrote that "in most discussions of English fiction, we proceed as if a third, two-fifths a half of our material were not really *there,*"[4] for we so often ignore the linguistic tissue that links them. Meanwhile, it unfortunately turns out to be precisely the statistics of these "unimportant" words – and not those winged "meaningful" ones – that best describe how one writes. This is because – to return briefly to the pulp metaphors of horror – having hacked literary masterpieces into pieces, stylometry extracts from the pieces only those words that best fit into its gloomy cells, creating a Frankenstein-like creature that is meant to replace a living being: one raw list of the same words for every massacred text.

At this point, it luckily turns out that although these lists are governed by Zipf distribution and all resemble one another, they build such a vast data set that the differences between them, though insignificant to the naked eye, become in fact rather meaningful when viewed through the lens of multidimensional analysis, whose dimensions are as numerous as the very words undergoing analysis. How many dimensions should that be? As a rule, the more the better, for elementary geometry has taught us that the distance between two points on a plane increases when we incorporate a third dimension into this distance, and although we might cease to "see" with our bare eyes even further dimensions, the distance will continue to increase with each subsequent dimension. It is a shame we cannot see this, but multidimensional analysis exists precisely to reduce many dimensions to two or three, usually from a sufficiently close perspective

[3] M. Jockers, *Macroanalysis. Digital Methods and Literary History*, Champaign 2013.

[4] J. Burrows, *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*, Oxford 1987, p. 1.

to "lend" them a distanced scale – or perhaps simply to differentiate them – by using the most frequently appearing words of individual authors, categorized by style, generation, gender, epoch, genre or literary type … In this way, new "graphs, maps, trees" begin to emerge, this time no longer resembling Moretti's, for they only reflect raw linguistic material. This suffices, however. These data visualizations often take on forms that surprise the traditional literary scholar in their similarity to traditional interpretations. It bears mention that although the rule "the more, the better" does apply here, in practice, the statistics become saturated, i.e. the results become more stable, at some 1000-2000 words, and tend not to change at higher wordlist lengths.[5,6]

One catch remains: although the existence of an authorial or chronological "fingerprint" has been empirically confirmed many times over, the very mechanism for establishing similarities and differences between texts has not been sufficiently justified in linguistics, and only its cognitive branch deigns to glance curiously at the conceptual framework of macroanalysis.[7] Of course, the fact that each writer uses – in part unconsciously, to be sure – these commonly frequent words according to their own individual proportions should not come as a surprise. Surely writers of a given epoch accommodate language forms at a shared developmental stage. It is worse (and somehow harder to reconcile) that the author's stylometric signal can even persist through the trauma of its translation into a foreign language. Although research on texts in their original language and in translation is undertaken using two compartmentalized frequency lists on which one might search in vain for exact correlations between, say, prepositions in two languages, graphs and maps made on their basis are quick to group texts by their original authors, disregarding the translator.[8] It becomes a bit easier to identify various translators when they consistently translate the same author or even a single text.[9]

Moreover, no other classification system has proven to be as effective as these boring lists of word frequency: neither keywords, nor n-grams (sequences) of adjacent words, nor even n-grams of words' grammatical values (a.k.a. part-of-speech tags) consistently provide such a clear picture of authorship or chronology.[10] Even those that occasionally turn up similar findings involve significantly more burdensome computer processing. Mere lemmatisation (converting all phrases to their root forms) lends no significant improvement to the results (which should come as no surprise, given that the author's grammatical choice, say, to narrate a story in the present tense, is stylistically meaningful). Stylometry has not, in fact, halted the attempts to disrupt this domination of the lexicon. In part, this is in order to produce research more "digestible" for traditional literary studies on the one hand, and linguistics on the other – yet for the time being, the results

[5] J. Rybicki, M. Eder, *Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?*, "Literary and Linguistic Computing" 2011, issue 26 (3), p. 315-32.

[6] M. Eder, *Does size matter? Authorship attribution, small samples, big problem*, "Literary and Linguistic Computing" 2015, issue 30 (2), p. 167-182.

[7] The Australian scholar Louisa Connors' doctoral thesis also bears mention here: *Computational stylistics, Cognitive Grammar, and the Tragedy of Mariam: Combining Formal and Contextual Approaches in a Computational Study of Early Modern Tragedy*, Newcastle 2013.

[8] J. Rybicki, *The Great Mystery of the (Almost) Invisible Translator: Stylometry in Translation,* [in:] *Quantitative Methods in Corpus-Based Translation Studies*, ed. M. Oakley, M. Ji, Amsterdam 2012, p. 231-248.

[9] J. Rybicki, M. Heydel, *The Stylistics and Stylometry of Collaborative Translation: Woolf's ‚Night and Day' in Polish*, "Literary and Linguistic Computing" 2013, issue 28 (4), p. 708-717.

[10] R. Górski, M. Eder, J. Rybicki. *Stylistic fingerprints, POS tags and inflected languages: a case study in Polish,* [in:] *Qualico 2014: Book of Abstracts*, Olomouc 2014, p. 51-53.

have been negligible. Only the gender signal seems to turn up more "meaningful" words, when one searches its results in national literatures of the 18th and 19th centuries.[11]

Since for now, we can offer nothing "better" or more "digestible", I will sketch the origins of the visualizations below. A more precise description of this procedure in full can be found in Maciej Eder's Polish text[12] and in the same author's significantly more "technical" article appearing in "Digital Scholarship in the Humanities".[13] For more information, a description from the programmer's perspective can be found in the prestigious "R Journal":[14] most computational procedures have been executed with the help of the stylometric package "stylo" described in that journal and written for R's statistical programming community.[15] This package brings together electronic versions of all texts, breaks them up into individual words, counts these words' frequencies throughout the entire corpus, and selects a number of the most frequently used words as specified by the researcher. Once the program has derived sequences of numbers in this manner for each text, it compares these sequences for each pair of texts. The comparison is based on an assigned metric of distance – or difference – between the texts. Among stylometry's various metrics, for this study, I chose one that has demonstrated the best aptitude for capturing the author's signal: the Burrows' Delta method,[16] which uses the cosine of the angle between vectors of word frequency for each pair of texts.[17] "Cosine Delta" ($\Delta\angle$) for two texts (T and T1) measures the angle α (the greater the angle, the greater the distance between the two texts):

$$\Delta\angle(T, T1) = \alpha,$$

is calculated according cosine similarity of Z-scores" between two vectors (x = z(T) i y = z(T1));

$$\cos \alpha = \frac{\sum_{i=1}^{n_s} x_i y_i}{\sqrt{(\sum_{i=1}^{n_s} x_y^z)}\sqrt{(\sum_{i=1}^{n_s} y_i^z)}},$$

where ns is the number of words analyzed in the study, and z(T) is the value of z-score of word frequency in text T, calculated according to the usual formula:

$$z\,(T) = \frac{f_s(T) - \mu_s}{\sigma_s},$$

[11] J. Rybicki, *Vive la différence: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies*, "Digital Scholarship in the Humanities" 2016, issue 31 (4), p. 746-761.

[12] M. Eder, *Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii*, "Teksty drugie" 2014, issue 2, p. 90-105. This study was a deliberate theoretical-practical companion piece to my text cited above and appeared in the same journal.

[13] M. Eder, *Visualization in Stylometry: Cluster Analysis Using Networks*, "Digital Scholarship in the Humanities" 2017, issue 32 (1), p. 50-64.

[14] M. Eder, J. Rybicki, M. Kestemont, *Stylometry with R: A Package for Computational Text Analysis*, "R Journal" 2016, issue 8 (1), p. 107-121.

[15] R Core Team. *R: A language and environment for statistical computing*, http://www.R-project.org/ 2014 [July 14 2017].

[16] J. Burrows, *Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship*, "Literary and Linguistic Computing" 2002, issue 17, p. 267-287.

[17] P. W. H. Smith, W. Aldridge, *Improving Authorship Attribution: Optimizing Burrows' Delta Method*, "Journal of Quantitative Linguistics" 2011, issue 18 (1), p. 63-88.

where fs(T), in turn, is the raw frequency of a given word s in text T, μs then the average frequency of word s in the set of texts to which T belongs, while σs is the standard deviation of the frequency of word s in that same set of texts.[18] By this method, we derive the value of the distance Δ∠ for each pair of texts. This produces a matrix of distances for the whole set of texts. On this basis, we can already reach some conclusions about which texts resemble one another, although two-dimensional visualizations organized according to select statistical methods can give us a much more legible picture of these relationships. All this amounts to a good attempt (and for the purposes of this study, we do not fear this term) at a polysystem of literature in Polish.

A matrix of distances can be studied by analyzing the concentrations that connect the most similar texts to one another in the context of the set as a whole, and for this study, I have done precisely this. Thus, for instance, the closest neighbor of *Ogniem i mieczem* is *Potop*, while the next closest neighbor of both texts is *Pan Wołodyjowski*. We might expect this cluster of the trilogy's three parts would then be linked in the following order with *Krzyżacy*, *Quo vadis*, and finally, *W pustyni i w puszczy*. We would be correct to expect that such a large cluster of Sienkiewicz's adventure novels might only be matched with *Bez dogmatu* and *Rodzina Połanieckich*, while the "full Sienkiewicz" is linked with the similarly constructed "full Prus" in the subsequent stage of linking texts on the basis of their stylometric resemblance. By these methods, the multidimensional space created by texts from the data set as a whole and all words included in the analysis is reduced to something we can present on a single plane.

Among the methods of data visualization, "network analysis" has made quite a name for itself. This method sorts data points according to two or three dimensions (in this case, individual texts) depending on their degree of resemblance: the greater the resemblance, the thicker and shorter the line between the two data points. Of course, for so many texts, this work requires elaborate mathematics. For the researcher, all of this labor is taken care of by the Gephi[19] program, aided by the force-directed algorithm Force Atlas 2. According to the algorithm's creators, Force Atlas 2 "simulates a physical system in order to spatialize a network. Nodes repulse each other like charged particles, while edges attract their nodes, like springs. These forces create a movement that converges to a balanced state".[20] In this study, Gephi collected the results of the "stylo" figures demonstrating how often the data of two respective texts is in close proximity; then the frequency of the "points of contact" becomes an indicator of the similarity between two texts, and the strength of their common spring, never allowing two texts to drift far apart from one another. As I mentioned above, sooner or later (depending on the size of the networks and the strength of the processor) the system reaches a state of equilibrium. A network emerges, and a "map" (in this case of Polish literature) is complete. On this map, we can isolate discrete clusters in two manners: either by drawing from traditional knowledge of literary history and ascribing individual texts to authors, epochs,

[18]S. Evert, T. Proisl, F. Jannidis, I. Reger, S. Pielström, C. Schöch, T. Vitt, *Understanding and explaining Delta measures for authorship attribution*, "Digital Scholarship Humanities" 2017 https://academic.oup.com/dsh/article-abstract/doi/10.1093/llc/fqx023/3865676/Understanding-and-explaining-Delta-measures-for [July 14 2017].

[19]M. Bastian, S. Heymann, M. Jacomy, *Gephi: an open source software for exploring and manipulating networks*. International AAAI Conference on Weblogs and Social Media, 2009.

[20]M. Jacomy, T. Venturini, S. Heymann, M. Bastian, *ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software*. PLoS ONE 2014, issue 9(6), e98679, doi:10.1371/journal.pone.0098679.

genres or time periods, or by sorting the network mathematically, using the function of modularity. For "weighted" networks (the kind that appear in this study), those for which linkages between individual nodes have various "weights", the network's modularity is calculated with the formula:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j),$$

where *Aij* is precisely the weight ("strength") of connections (similarity) between points (texts) *i* and *j*; *ki* = ΣjAij is the sum total of all connections coinciding at nodes *i*; *ci* is the cluster to which node *i* is assigned; and finally, the function *δ(u, v)* adopts the value 1 when *u* = *v* and the value 0 when u ≠ v, and m = ½ΣijAij.[21] One might say that the formula above serves, for the computer, as its substitute for human knowledge of authors, epochs, and literary genres...

Material

Before I move on to the results, it is worth elaborating on the body of texts included in the study. The largest group (1319 titles) consists of original Polish novels, epic poems and – especially in the case of the older texts – sermons, psalms and hagiographies. Of course, in this case, all genres of the prose novel appear in marked disproportion. This is driven by two interrelated factors: firstly, there is simply is more novels around than anything else, and secondly, they are also the genre that is mst readily available in electronic version. A second disproportion can also be observed – this time, chronological. The fourteenth century is represented by one text, while the fifteenth and sixteenth centuries contribute ten and nine texts respectively to the pool. The next century in Polish literature has been called the "century of manuscripts" for a reason – the number of available texts drops to eight; but the eighteenth century fares even worse: it is arguably the domination of non-epic genres that led, despite the exertions of a certain Bishop of Warmia, to only five texts. A boom in the production of novels – and in their later availability in electronic format– occurs in the nineteenth century, accounting for 426 titles within the set. This growth continues into the twentieth century (631 texts). Against this backdrop, the new millennium begins with a bang, for its first several years already boast 229 titles. This comes as no surprise, for it is precisely the twenty-first century that ushered literature into the electronic medium, often without printed matter as an intermediary.

So much for Polish prose and epics. A separate category includes the Polish drama, from Kochanowski to Mrożek (a span that includes 63 texts): apart from *Odprawa posłów greckich*, there is Fredro, of course, all three bards, Norwid, and many texts by Wyspiański, Zapolska, Przybyszewski and Witkacy. Aside from a number of individual texts by specific authors, Gombrowicz and Mrożek contribute many texts. All together, the texts indigenous to Poland amount to 1382. It is worth noting that to read through every text, even at the breakneck speed of one text every two days, would take a single person over seven and a half years.

And now we can move on to translations from foreign languages. Since the interwar period, Polish translations most often come from English texts.[22] The data set includes 408 translations from the

[21] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, *Fast Unfolding of Communities in Large Networks*, "Journal of Statistical Mechanics: Theory and Experiment" 2008, issue 10, p. 1000.

[22] See also: W. Krajewska, *Recepcja literatury angielskiej w Polsce w okresie modernizmu. (1887-1918). Informacje. Sądy. Przekłady*, Wrocław-Warszawa-Kraków-Gdańsk 1972.

language of Byron, and that not counting Shakespeare, who, in his own right, claims 135 translations, bringing us to a total of 543. There are significantly less representatives from the French – 242, from Russian – 103 and that same amount from German. Czech, Spanish, Hungarian, Italian, the Scandinavian languages, and Turkish altogether contribute 175 texts. The set therefore includes a proportion of foreign texts that is not significantly less than its Polish texts (1161). In total, the set consists of 2,548 titles. The scale of the entire data set amounts to 170,692,206 words.

How did I obtain all these texts in electronic format? Unfortunately, I did not record precise statistics. A significant portion of the texts – those in the public domain – were found in various free collections, from such noble and useful ventures as Free Readings (*Wolne lektury)*,[23] The Online Polish Literature Library *(Biblioteka literatury polskiej w Internecie)*[24] and Old Poland (*Staropolska)*.[25] These three archives proved to be the most useful. The oldest Polish texts come from the small but invaluable electronic "Library of the Treasures of Medieval Polish Letters" ("Biblioteki zabytków polskiego piśmiennictwa średniowiecznego") at the Polish Language Institute PAN, in Kraków.[26] The more recent texts were often simply sourced from online bookstores in the form of e-books– this of course sped up the process of obtaining texts, and additionally lowered costs, for electronic books are often (marginally) cheaper than their printed counterparts, though a large portion had to be transferred to electronic format by means of scanners or OCR.[27] The Institute of English Philology at Jagiellonian University's recent acquisition of sheetfeed scanners somewhat facilitated the unmediated digitalization of books, under the condition that each volume first had to be divided into individual pages.[28]

At this point, I will provide a short digression on the current accessibility of Polish-language literature – both original and translated – in electronic format. Since I managed to obtain over two thousand texts for the purposes of this study, one might get the impression that our national literature is already quite prevalent in digital form. From the "average" reader's perspective, this is even somewhat accurate: reading a book online or downloaded from the internet is, in fact, quite easy. It is harder, however, to prepare a text for quantitative analysis, for nearly every archive uses its own format, its own user interface and – understandably– rigorously defends its resources from being available in entirety or in large portions. It might not bother the conventional reader to read the first Polish translation of Hamlet (Wojciech Bogusławski's from 1797, based on Friedrich Ludwig Schröder's German adaptation) in DjVu format, which is difficult to process electronically. In fact, quite the contrary. *Le plaisir du texte* in which the eponymous hero thankfully survives, and

[23]http://wolnelektury.pl [July 14 2017].

[24]http://literat.ug.edu.pl/ [July 14 2017].

[25]http://www.staropolska.pl/ [July 14 2017].

[26]*Biblioteka zabytków polskiego piśmiennictwa średniowiecznego*, ed. W. Twardzik, Kraków 2006.

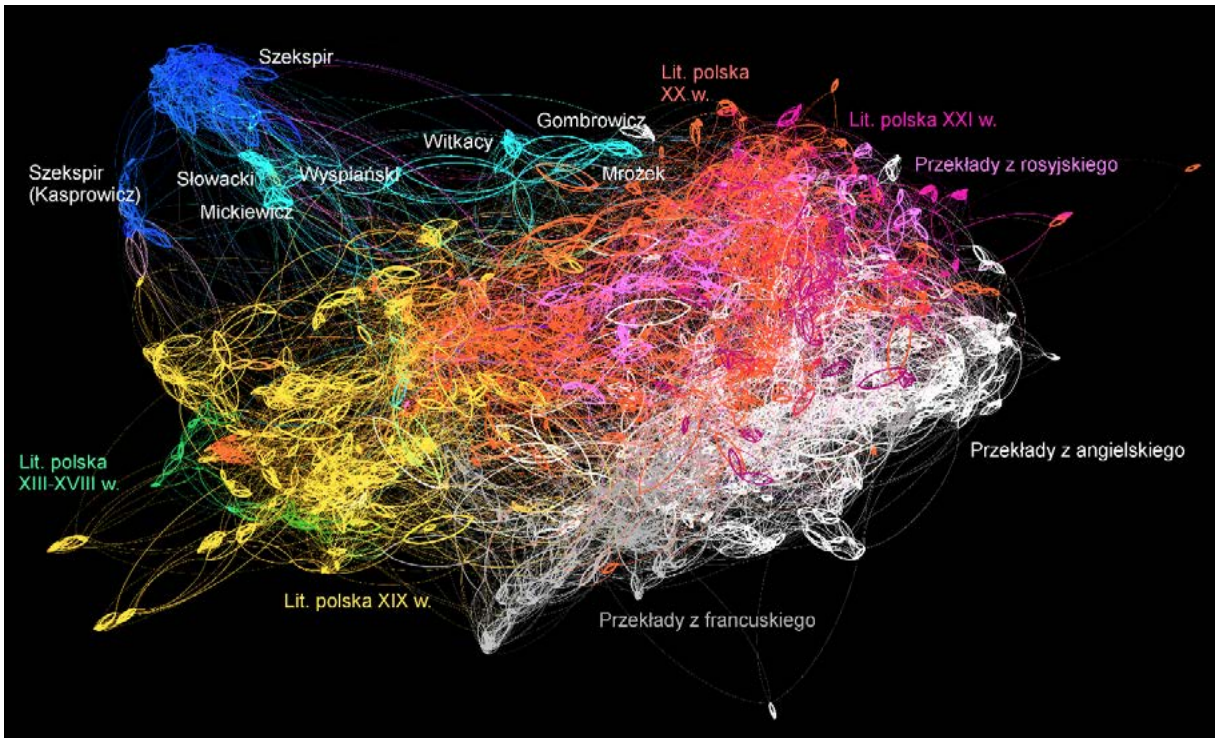
[27]It is impossible to fail to mention the heroic efforts of my two master's advisees, Anna Hołubiczko and Marta Kamuda, who compiled such an impressive set of Polish translations of Shakespeare, doggedly scanning print versions or with monk-like precision, correcting the difficult old scans of the Polona Library (https://polona.pl/ [July 14 2017]). The fruit of these labors – aside from the collection of Shakespeare's translations and a significant contribution to the subcategory of Polish drama – are two noteworthy masters theses: A. Hołubiczko, *"Porównania śmierdzą": porównanie równoległych tekstów polskich przekładów Szekspira* (master's thesis), Kraków 2017; M. Kamuda, *Stylometric Analysis of the Polish Translations of Shakespeare* (master's thesis), Kraków 2017.

[28]Here I must extend enormous thanks to the Volumin Bookbinding Workshop located at Św. Gertrudy 5 in Kraków that ruefully but willingly separated every volume in exchange for our vague promise to someday put them together again.

who is accompanied not by Horatio but by Gustav, is only enhanced when the computer screen shows the beautiful print format from 1823. The conventional reader somehow makes do even when some of the allegedly digitalized texts in Polish libraries are in fact images converted into PDFs, which only require that one be able to recognize the texts within… It's a small consolation that this issue does not only prevail in Poland, and even this level of availability exists in spite of the Text Encoding Initiative consortium's seemingly ironclad regulations regarding the digitalisation of text. Digital Humanities throughout the world, whose most acclaimed "discipline" is precisely the development of digital archives for all cultural artifacts, is in the position to create beautiful and valuable digital editions for the "non-specialized" reader; but apparently continues to fail its clients from its own environment– precisely the people engaged in the quantitative analysis of cultural data. And yet, according to Willard McCarty, one of the most esteemed authorities on the digital turn in the humanities, it is computer stylistics that have contributed most to this turn and that points the way forward.[29]

Findings

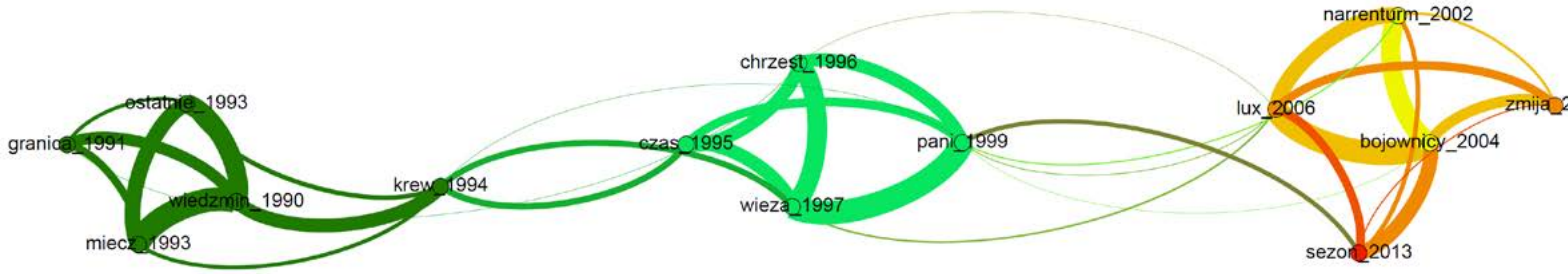
How does the representative sample of Polish-language literature look, then, through the lens of macroanalysis? Take Graph 1:



Graph 1. Network analysis of 2,548 texts on the basis of the frequency of the 2,000 most common words in the entire data set.

[29]W. McCarty, *Getting There from Here. Remembering the Future of Digital Humanities: Roberto Busa Award Lecture 2013*, "Literary and Linguistic Computing" 2014, issue 29 (3), p. 197.

Similarly to the macroanalysis cited above of the significantly smaller body of Polish literature,[30] this network reveals a marked chronological pattern on the part of original Polish texts – even if, in more distant orbits on the graphs, less disciplined satellites appear. The earlier texts in the data set, indicated by the color green, tend to group together in the lower left corner of the graph. Nineteenth century literature (in yellow) shifts slightly up and to the right, after which twentieth century texts (in red) gradually proceed. The dark purple clusters in the upper right corner of the graph represent writing from the twenty-first century. The most important observation for this kind of visualization, moreover, is not so much the existence of chronological clusters, but their progressive evolution in one consistent direction. Literature on the subject has long since recognized the staggered evolution we might describe with the phrase "tiptoeing towards the Infinite". Some have effectively argued that this is not only the product of linguistic shifts, but reveals the effects of the evolution of literary stylometry - not stylistics.[31] Perhaps the best argument for this kind of interpretation is the occurrence of directed and evolving trends in the scope of a single author's body of work, whereas the existence of marked changes in only the Polish language become more difficult to defend. A network analysis of the work of Andrzej Sapkowski makes a good example (Graph 2.). This analysis shows that word frequency falls into three explicit time segments of a few years.



Graph 2. Periodization of Andrzej Sapkowski's body of work on the basis of the frequency of the most frequent words: 1990-95 in dark green; 1995-2000 in light green; twenty-first century in yellow.

Let us return, however, to Graph 1, which reveals other interesting phenomena. Polish literature (or rather, its *mainstream* novels and epics) runs from the left to the right; attached to it from the bottom is a large, gray-white mass. The gray lines link together Polish translations of French literature, while white lines designate translations from English. If the chronological signal is a trend common to the entire graph, it is difficult not to connect the appearance of French translations "earlier", or further to the left with French writers' earlier influence on literary output in Poland. English and American texts arrive later, appearing further to the right, and manage to infiltrate the fields of indigenous Polish literature more effectively, though not entirely. It is no accident that the translations themselves of French literature were made earlier than translations from English.

[30] J. Rybicki, *Pierwszy rzut oka...*

[31] J. Burrows, *Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative,* [in:] ed. S. Hockey, N. Ide, *Research in Humanities Computing* issue 4, Oxford 1996, p. 1-33.

These, however, are not the only curious findings regarding translation produced by this visualization. Within the grey sea of Polish translations of French literature, we can discern several white islands: these are translations of Walter Scott. Moreover, on the grey-white border, we also find the early translations of Dickens. This case pertains to the two English-language novelists who were the first of their kind to be recognized on the banks of the Vistula. On this basis, one might guess that the system of translated novels is also impacted by chronological influence.

This is not, however, the only factor influencing the layout of clusters on this network graph. The first Polish translation of Charlotte Brontë's masterpiece *Jane Eyre* – translated into Polish in 1880 by Emilia Dobrzańska as *Janina* – clings closer to the grey French texts than the white English ones. This should come as no surprise, for the Polish version is not only abridged, but happens to use the French translation as an intermediary. Many French calques were observedin close readings and thus support this claim.[32] Several other older translations from English follow this trend for similar reasons, as we can reasonably suspect. In this manner "distant reading" can unearth surprising themes for the "close reader".

And finally: in this same French grey cluster, a red pattern emerges in a cluster of texts from Stendhal, Balzac and Proust, translated by Tadeusz Boy-Żeleński, also the original author of *Znaszli ten kraj* and *Marysieńka Sobieska*. With these texts, the Polish doctor found his way into a circle of translators whose stylometric fingerprint is not contingent on whether they write their own words or translate those of others. This is not the first time that quantitative research has revealed this exact feature of Boy's work[33], and it does not apply to him alone; there are several authors who translate in an entirely different style from how they write: when translating Juvenal from Latin into English, Samuel Johnson, just like Boy, "maintains his own tone", while John Dryden "finds a distinct not and holds it".[34] While these two great non-Slavic literatures have hardly come into contact with the bulk of the data set's original Polish texts, the light purple flicker of Russian translations, meanwhile, penetrates into the very center of the red zone of Polish twentieth-century literature. Russian science fiction, meanwhile, mingles with the dark purple cluster of contemporary Polish literature, within which there is no shortage of representatives of that same genre. The genre signal therefore appears in a rather characteristic manner;[35] while other behavioral patterns of translations from foreign Slavic languages suggest the presence of curious oscillations towards a certain *translationese*, which seems to grow in proportion with the differences between the original and target languages, especially given that the unfortunately sparse amount of translations from Czech (which are therefore unmarked on the graph) follow the behavioral trend of translations from Russian. However, the most curious effect associated with translation is the enormous distance between translations of Shakespeare

[32]D. Hadyna, *A controversial translation justified by the context: Janina, the first Polish version of Charlotte Brontë's Jane Eyre* (master's thesis), Kraków 2013.

[33]J. Rybicki, *Stylometric Translator Attribution: Do Translators Leave Lexical Traces?*, [in:] *The Translator and the Computer,* ed. T. Piotrowski, Ł. Grabowski, Wrocław 2013, p. 193-204.

[34]J. Burrows, *The Englishing of Juvenal: Computational Stylistics and Translated Texts*, "Style" 2002, nr 36, p. 677-699.

[35]See also: C. Schoech, *Fine-tuning our stylometric tools: Investigating authorship, genre, and form in French classical theater*, [in:] ed. K. Walter, K. Price, *Digital Humanities 2013 Conference Abstracts*, Lincoln 2013, p. 383-386.

– indicated in dark blue – and the white spot on the opposite edge of the network representing the rest of English literature in Polish translation. Of course one reason for such a distinct boundary is the typological distinction, for "white" texts are exclusively prose. This does not, however, fully explain the fact that Polish Shakespeare follows his own rules. Although the data set analyzed here represents the work of nineteen different translators of the English bard, Polish Shakespeare retains his own stylometric profile. Only Kasprowicz's translations and a few others from the turn of the century diverge from this pattern – and even then, only to a degree. Quite naturally, the light-blue trail of Polish dramas runs not so far away (its chronology following the same current as the rest of Polish literature, moving from left to right) the sphere surrounding Shakespeare most markedly prolongs those of its elements that use Shakespeare's influence as a standard and manifesto, as it were: the romantic dramas of Mickiewicz and Słowacki, and appearing adjacent, the neoromantic theater of the author of *The Tragicall Historie of Hamlet Prince of Denmark. According to the Polish text of Józef Paszkowski, read and reconceived by St. Wyspiański* (*The Tragicall Historie of Hamlet Prince of Denmark. Według tekstu polskiego Józefa Paszkowskiego, świeżo przeczytana i przemyślana przez St. Wyspiańskiego)*.[36]

All of these observations are united by one common law: as the machine is preoccupied with calculations and pure graphics, the work of arranging the data points on the graph and their classification still belongs to a rather "human" humanities. The human-interpreter knows, after all, which point designates which text (even if it might be difficult to single that point out from the thicket of dense networks), and that interpreter makes autonomous decisions to assign distinct colors to Shakespeare, twentieth-century Polish literature, translations from English, etc. The picture that begins to emerge must, by its very nature, rely heavily on a traditional history of literature, which remains the first point of departure in appraising the computer's visualizations; visualizations can reveal interesting inconsistencies and connections that might be counterintuitive from the perspective of traditional literary studies – and can likewise reveal a lack of connections. In light of this interpretation, man intrudes on his research, even when he only determines the numbers and categories by which he organizes his research: be it by author, epoch, genre, or language of origin…

A machine, however, might relieve man of one of these actions. A machine can be ordered to divide the analyzed texts into the desired number of groups. Man is still, in the end, choosing the amount of groups, but the divisions might (though not necessarily) run entirely counter to those arrived at by human knowledge of the texts in question. To achieve this end, the Gephi package mentioned above has enormous value.

Let us review what happens when a computer tries to autonomously indicate – on the basis, of course, of the smallest differences in the usage of frequently appearing words – how a set of works diverge if we allow for two or more main groups. Graph 3 is a set of visualizations using two, three, four and seventy groups.

[36]Kraków 1905.

Graph 3. Network analysis of the data set using a modular division into 2, 3, 4 and 70 groupings (starting from the upper-left).

With the possibility of dividing the texts into only two groups, the modular algorithm divides the data set into large clusters of prose in their original language and in translation (green) on the one hand, and Polish dramas and Shakespeare's dramas on the other (purple). A number of early Polish novels from the middle of the nineteenth century also belong to the second group (by Duchińska, Goszczyński, Niewiarowski, Michał Jezierski). The three-group graph is sorted into a group of early prose (green, dating up to the mid-twentieth century) and later prose together with the majority of translated works (purple); the Polish drama and Shakespeare (excluding Kasprowicz and his contemporaries) comprise a separate yellow cluster. After a further increase in the number of groups, author-based groupings begin to emerge. It is only when the computer can use seventy groups, however, that Polish romantic and neoromantic drama begin to diverge from Shakespeare. This is an interesting measure of the linguistic resemblance between these two categories that are literarily quite wedded.

Conclusions

The editor of this collection called Graph 1 an "unidentified Pollock"[37] – and truly, there is no point obscuring the fact that the description and commentary on visualizations of network analysis of over two-thousand texts begins to have ekphrastic connotations. One might even suggest that here, we are dealing with an interesting transformation rare in cultural studies:

[37]T. Mizerkiewicz, email from July 13 2017.

the aesthetics of the word, when passed through a linguistic-mathematical-statistical programming filter creates, in the end, a new aesthetics – the aesthetics of the image. If we are concerned, however, with scientific research and not with visual impressions, it is better not to continue down this path, for the very attempts at scholarly objectivity that once led to the foundation of quantitative analysis end here. Even in the last century, Edward Stachurski wrote that "using statistical methods in linguistic and stylistic textual research allows for a kind of confidence that the obtained results stand on objective foundations independent from the reader's subjective judgements".[38] David Hoover echoes him: "Quantitative approaches to literature represent elements or characteristics of literary texts numerically, applying the powerful, accurate, and widely accepted methods of mathematics to measurement, classification, and analysis."[39] Many a digital humanist has confessed to have been driven to the world of computers by the cognitive nihilism of postmodernism, whose one true claim should be that there is no one truth ...[40]

We should not overemphasize this objectivity – as Maciej Eder cautions in the text accompanying "a first glance at a map of Polish literature."[41] It is true that measurement and classification are undertaken is such a way that the researcher's subjective choices play a moderate role. This role is moderate, but still visible, for even the most self-aware and impartial scholar must make a number of decisions that weigh on his conscience. How large a data set? When does a data set become appropriately "representative"? Does "representative" imply: taking account for the differences in the number of works contributed by various authors – so that it is quite alright that Kraszewski contributes so many texts to the data set, for it is not his "fault" that Schulz managed to write so few texts? Or perhaps it would be best to treat the data "fairly", using egalitarian proportions? On the one hand, the disproportionate scales of the various authors' material interferes with the linguistic balance of the text – frequent words from Kraszewski's enormous body of work (as with those of Polish Shakespeare) leave a significantly deeper imprint on the list for the entire data set than, say, Schulz's little gems. At the same time, this does produce a complete portrait of Polish literature: Kraszewski, Jeż, Papi and Lem wrote vast amounts; others significantly less, and this is a fact we cannot modify after the death of an author – and often it is so for their lives.

The second moment in the process of non-objective choice is the stage of setting the parameters for quantitative analysis. To reiterate: it is true that stylometry has not yet reached a consensus on this point and continues to develop its methods to limit the influence of these and other programming settings on the obtained results –in fact, this is one of the main priorities of this academic community.[42] The doubts, however, linger: does averaging the results of many individual analyses truly render the most reliable findings? Should we instead determine a single but ideal set of parameters – most often, of course, being the number of words whose frequency we compare?

[38]E. Stachurski, *Słowa-klucze polskiej epiki romantycznej*, Kraków 1998, p. 11-12.

[39]D. Hoover, *Quantitative Analysis and Literary Studies*, [in:] ed. S. Schreibman, R. Siemens, *A Companion to Digital Literary Studies*. Oxford 2007, 518.

[40]W. McCarty, *Getting There*..., p. 190.

[41]M. Eder, *Metody ścisłe w literaturoznawstwie*...

[42]See also: eg. J. Rybicki, M. Eder, *Deeper Delta*....

And finally, a third moment occurs: the moment when everything is counted, the computer processing has mapped the data on one plain in all the colors of the rainbow – and the humanist arrives and observes. Does that humanist really see the objective truth in this tangle, or does he or she simply cater his own knowledge – and ignorance, for of course, he or she has not read all 2,500 texts, this much we can concede – to these colorful blobs?

I propose a slightly less ambitious scenario: given that we cannot say for sure how well the linguistic idiosyncrasies of the authors are conveyed in statistics of word frequency — those being "synsemantic" words, and since, for the time being, linguistics offers us no clear theories that might serve us as experimental humanists just as theoretical physics points the way for experimental research – let us make use of what we have. What we have is metaliterary and critical texts on the one hand, and on the other, a growing body of materials that support the claim that quantitative analysis often does reveal relationships as accurately as qualitative analysis. This being the case, every time stylometry turns up unexpected results incompatible with the "qualitative approach", perhaps it would be worthwhile to take this as a sign that new interpretations lie ahead. The simplest and least controversial usage of computer-based stylometry – authorial attribution – changes the landscape of the literary polysystem every time it uncovers or verifies who in fact wrote a given text. Perhaps it is worth extending our faith in quantitative analysis to apply to situations where quantitative analysis engages two texts or two authors whose similarities nobody thus far has considered? In the end, no doctor with any self respect would deliver an (often life-saving) diagnosis without reviewing the results of blood and urine tests. Stylometry offers the literary scholar precisely this type of laboratory method — perhaps it is worthwhile to make use of it?

KEYWORDS

MACROANALYSIS

Stylometry

DISTANT READING

Abstract:
This articles presents the results of a quantitative analysis of frequently appearing words in a data set of over 2,500 Polish texts: Polish literature from the fourteenth to twenty-first century, and Polish translations from English, French, Russian and (to a lesser degree) other languages. The data set reveals a visible signal by type and by original language. The results also point to a definite stylometric specificity of Polish translations of Shakespeare, and their stylometric resemblance to Polish romantic and neoromantic dramas.

multidimensional analysis

network analysis

Polish literature

Polish translations

Note on the Author:

Jan Rybicki was born in 1963 and is a graduate of the Institute of English Philology at Jagiellonian University (1987). He has lectured at Krakow's Pedagogical University (1991-2000, 2001-2011) and at Rice University in Houston (1996-1997, 2000-2001). Since 2011 he has been an adjunct at Jagiellonian University's IEP. His research interests focus mainly on computer-based stylometry of literary language in the original and in translation. Aside from academic articles he translates English-language literature – he already has around thirty translated novels under his belt, by authors such as Kingsley Amis, John le Carré, Douglas Coupland, William Golding, Nadine Gordimer, Francis Scott Fitzgerald, Kazuo Ishiguro, Kenzaburo Oe and Kurt Vonnegut.