From morpho-syntactic features to digital metrics in StyloMetrix. Automatic classification of advertising and non-advertising posts by detecting the dominance of either perlocution or illocution.

In the proposed paper, we aim to present an approach to distinguish advertisement from narrative blog posts. We claim that the style of the texts, including the occurrence of perlocutionary acts should play significant role here. The analysis is performed with StyloMetrix, which is a tool that creates a fully interpretable vector representation of text. Each linguistic feature is calculated using appropriate digital metrics. The tool is based on the study of morpho-syntactic features (grammatical categories, vocabulary models, syntactic models and e.g. nominal phrases), token types (units of language, e.g. autosemantic words), token frequency distribution, as well as on selected psycholinguistic aspects of words, disregarding the semantics of the text; here we refer to well-known studies in linguistics showing the relationship between style/genre and grammatical properties of a text. StyloMetrix vectors makes it possible to identify texts and classify particular documents into specific (e.g. human tagged by researchers) classes of texts, which contributes to the notion that stylistic properties can be considered a vehicle for semantics.

In this paper, we decided to subject specific speech acts (according to Austin's theory), or rather the blog posts containing them - to automatic analysis. We have focused on texts in which illocution (on the one hand) and perlocution (on the other) are dominant. Thus, referring to Searle's findings, we subjected to a contrastive analysis of directive (persuasive) acts, especially those of an indirect nature, meaning texts that should be read contextually, and expressive acts - mainly direct ones. We gathered a corpus of approx. 1000 texts with three classes: a) direct directive texts (at which our tool was 'targeted'), b) direct expressive texts, and c) indirect expressive texts. All of them coming from blog amateur entries on the Polish Internet, containing information and emotion - that is, direct expressive texts or indirect expressive texts, and texts containing any elements of advertising (surreptitious advertising, custom texts, product placement, etc.) - that is, indirect directive texts. The data set was split into a training set (where we taught our tool to recognize texts of the direct directive type), and a test set, where the already 'trained' tool is predicting the right class of a text. The initial set of linguistic metrics of StyloMetrix have been extended by those addressing specifically what we thought of as the morpho-syntactic characteristics of perlocution and illocution.

Since StyloMetrix is used to offer not only effective model, but also a fully interpretable vector representation, we can elaborate on linguistic properties of stylistic features of analyzed perlocutionary texts on top of having successfully fulfilled the main exercise of automatic differentiation of the group of texts. Therefore, our aim is to prove that strictly linguistic analysis, based on morpho-syntactic properties, allows to identify the most important features and map them onto the linguistic elements of style. Which in turn, can denote certain semantics.

Bibliografia

Austin J. L., *How to Do Things with Words*, 1962.

Banasiak-Mrozek, M., (2020), „Wykorzystanie stylometrii i uczenia maszynowego w informatyce śledczej", *Payload. Magazyn o ofensywnym bezpieczeństwie IT*, https://payload.pl/stylometria/

Bańko M., (2002) *Wykłady z polskiej fleksji*, Warszawa: Wydawnictwo Naukowe PWN.

Bąba S., (1978), *Stylistyczne funkcje potoku składniowego we współczesnej prozie polskiej (na wybranych przykładach)*, [w:] *Studia nad składnią polszczyzny mówionej. Księga referatów konferencji poświęconej składni i metodologii badań języka mówionego (Lublin 6-9 X 1975)*, pod red. S. Grabiasa, J. Mazura, K. Pisarkowej i T. Skubalanki, Wrocław: Zakład Narodowy imienia Ossolińskich, s. 271-283.

Choiński M., Rybicki J., "Is God Really Angry at Sinners? A Stylometric Study of Jonahan Edwards's Representations of God" , in: Rhys Bezzant, Eugene (2017), *The Global Edwards*, Wipf & Stock, pp. 349-359

Eder, M. (2017), "Visualization in stylometry: cluster analysis using networks", *Digital Scholarship in the Humanities*, 32(1), pp. 50-64.

Eder, M., Rybicki J., Kestemont M., (2015), "Stylometry with R: A Package for Computational Text Analysis", *The R Journal* [on-line].

Goswami, S., Sarkar, S., & Rustagi, M. (2009, March), „Stylometric analysis of bloggers' age and gender", in: *Third international AAAI conference on weblogs and social media*.

Gramatyka współczesnego języka polskiego. Morfologia, (1999), tom 1-2, pod red. R. Grzegorczykowej, R. Laskowskiego, H. Wróbla, Warszawa: Wydawnictwo Naukowe PWN.

Gramatyka współczesnego języka polskiego: składnia, morfologia, fonologia, (1984), [T. 1], *Składnia*, pod red., Z. Topolińskiej, Warszawa: Wydawnictwo Naukowe PWN.
Grochowski M., (1982) *Zarys leksykologii i leksykografii : zagadnienia synchroniczne*, Toruń: UMK.
Grzegorczykowa R., (1996), *Wykłady z polskiej składni*, Warszawa: Wydawnictwo Naukowe PWN.
Grzegorczykowa R., *Zarys słowotwórstwa polskiego. Słowotwórstwo opisowe*, Warszawa 1979
Jakus-Borkowa E., (1987), *Nazewnictwo polskie*, Opole: Wyższa Szkoła Pedagogiczna im. Powstańców Śląskich w Opolu.
Kania S., Tokarski J., (1984), *Zarys leksykologii i leksykografii polskiej*, Warszawa: Wydawnictwa Szkolne i Pedagogiczne.
Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010, „Boruta–a system for feature selection", *Fundamenta Informaticae*, 101(4), pp. 271-285.
Lewandowska-Tomaszczyk B., (2008) *Corpus linguistics, computer tools, and applications-state of the art : PALC 2007*, w: International Conference on Practical Applications in Language Corpora (2007), pod red., B. Lewandowskiej-Tomaszczyk i in., Frankfurt am Main: Wydawnictwo P. Lang.
lsemann, H. (2019), Forensic stylometry, in: *Digital Scholarship in the Humanities*, 34(2), pp. 335-349.
Łaziński M., (2003), *Grammar and Gender in Polish Corpora*, w: PALC 2001: „Practical Applications in Language Corpora", s. 309-327
Łaziński M., (2003), *Uwarunkowania stylistyczne użycia aspektu czasownika*, w: „Z polskich studiów slawistycznych. Seria 10. Językoznawstwo", *Prace na XIII Międzynarodowy Kongres Slawistów w Lublanie*, Warszawa 2003, s. 127-136.
Matuszczyk B., (2004), *O języku i stylu współczesnych kazań*, w: *Kościół w życiu publicznym. Teologia polska i europejska wobec nowych wyzwań*, pod red. Góźdź K. in., t. 2, Lublin.
Matuszczyk B., Wojtak M., (2004), *O wyznacznikach stylu współczesnych kazań*, „Studia Językoznawcze", t. III, s. 73-78.
McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014), *Automated evaluation of text and discourse with Coh-Metrix*, Cambridge University Press.
Mikołajczak S., (1970), *Składnia „Quo vadis" Henryka Sienkiewicza w ujęciu statystycznym (dla potrzeb stylistyki)*, „Język Polski", z. 2, s. 85–97.
Mikołajczak S., (1983), *Składnia wybranych utworów Bolesława Prusa i Stefana Żeromskiego*, Poznań, Wydawnictwo Naukowe UAM
Mikołajczak S., Pachowicz B., (1967), *Próba ustalenia typowych konstrukcji syntaktyczno-stylistycznych (głównie na podstawie polskiej literatury współczesnej)*, „Językoznawca", nr 16–17, s. 73–92
owak T., (2017) *Transformacje morfosyntaktyczne w badaniach eksperymentalnych, czyli lingwistyka między matematyką a psychologią*, „Białostockie Archiwum Językowe", nr 17
Piasecki, M., Walkowiak, T., & Eder, M. (2018), "Open stylometric system WebSty: integrated language processing, analysis and visualization", *Computational Methods in Science and Technology*, 24(1), pp. 43-58.
Podstawy językoznawstwa korpusowego, (2005), pod. red. B. Lewandowskiej-Tomaszczyk, Łódź: Wydawnictwo Uniwersytetu Łódzkiego.
Przyczyna W., Skowronek K., (2006), *Język listów pasterskich episkopatu Polski w latach 2000– 2004*, w: *Człowiek i sacrum. O pojęciach religijnych w języku i kulturze*, pod red., D., Sarzyńskiej i R. Tokarskiego, Sandomierz: Towarzystwo Naukowe Sandomierskie, s. 13-32.
Rybka M., Graf M., (2004), *Kilka uwag językowo-stylistycznych o tekstach zamieszczonych na kościelnych stronach internetowych*, w: *Język religijny dawniej i dziś*, pod. red. S. Mikołajczaka, T. Węcławskiego, Poznań: Wydawnictwo „Poznańskie Studia Polonistyczne", s. 233-240.
Sambor J. (1974), *Słownictwo bardzo częste w pięciu stylach polszczyzny pisanej*, „Poradnik Językowy", s. 466-475 oraz 533-537.
Sambor J. (1975), *O słownictwie statystycznie rzadkim*, Warszawa: Państwowe Wydawnictwo Naukowe.
Savoy, J. (2020), Machine Learning Methods for Stylometry, Springer International Publishing.
Searl J. R., *Speech Acts: An essay in the Philosophy of language*, 1969.
Skorupka S., Kurkowska H., (1959), *Stylistyka polska. Zarys*, Warszawa: Państwowe Wydawnictwo Naukowe.
Tuora R., Kobyliński, Ł. (2019), Integrating Polish language tools and resources in Spacy. In *Proceedings of PP-RAI 2019 Conference*, pages 210–214, Wrocław, Department of Systems and Computer Networks, Faculty of Electronics, Wroclaw University of Science and Technology.
Wilkoń, A. (2000), *Typologia odmian językowych współczesnej polszczyzny* (in Polish) (2 ed.), Katowice: Wydawnictwo Uniwersytetu Śląskiego. pp. 87–88.