Book Genre and Author's Gender Recognition Based on Titles: the Example of the Bibliographic Corpus of Microtexts*

Adam Pawłowski[1], Elżbieta Herden[1], Tomasz Walkowiak[2]

1) University of Wrocław, Institute of Information and Library Science
2) Wrocław University of Science and Technology, Department of Computer Engineering

Abstract:

The subject of this article was the application of automatic taxonomy methods to the corpus of microtexts, consisting of book titles. Two hypotheses have been tested. The first one claimed that on the basis of a book title only one can automatically recognize its genre (writing species). The second assumed the possibility of recognizing the author's gender on the basis of the book's title. FastText and word2vec methods were applied. The analyses gave a positive (and rather astonishing) result: it was shown that with properly chosen n-grams more than 70% of titles had a correctly assigned writing species, while the accuracy of the gender recognition of the author was almost 80%. Both values significantly exceed the levels of random recognition. The research was conducted on the corpus of titles derived from the Polish national bibliography.

Key Words:

corpus linguistics, automatic taxonomy, gender recognition, book genre, fastText, word2vec, bibliography, Polish

1. The Problem

Bibliographies have become in recent years the subject of quantitative and qualitative research conducted within the framework of digital humanities. Their volume, counted in millions of records available in digital form, is extensive enough to use NLP methods, statistics and text mining. NLP techniques allow the text to be processed at the morphosyntactic level (including, among others, lemmatization). Statistical tools enable the creation of full, quantitative descriptions of bibliographic corpora, whereas text mining methods are used to create advanced data representations and search tools, based on, among

* This research was supported by Polish National Science Center (NCN) under grant 2016/23/B/HS2/01323 "Methods and tools of corpus linguistics in the research of a bibliography of Polish book publications from 1997 to 2017".

others, machine learning techniques such as word2vec, topic modelling, statistical classifiers (e.g., linear soft-max classifier) or neural networks. The content of large bibliographies in the case of big data research represents a unique cognitive value from the point of view of cultural anthropology and also to some extent of scientometrics. Titles, although they belong to the class of microtexts, synthesize information important from the point of view of the author. When analyzed in large quantities, they reflect the general state of knowledge in society, its preferences, great civilizational trends, and intellectual fashions. These qualities can be fully utilized thanks to a specific feature of bibliographic structures, not found in literary or applied texts. This feature is the presence of meticulously and methodically prepared metadata, which allows the verification of the effects of empirical research, conducted on those fields of bibliographic records that contain information in text format (primarily titles).

The presence of metadata indicating, among other things, the date and place of publication, the genre of the text and the gender of the author, plays a special role in the broader context of quantitative research into language and corpus linguistics using automatic methods. The Achilles heel of these studies is the limited possibility of verifying the results obtained, combined with a multitude of text-mining methods. This is the result of several factors:

- instability of structural features of the text (e.g. difficulty of measuring unit length, questionable issue of text segmentation into smaller parts);
- the fluidity of semantic categories assigned to individual words or multi-word structures (meanings are constructed ad hoc in the process of sender-recipient interaction, the phenomenon of polysemia, homography, and word correlations occurs in mass);
- a practically unlimited number of potential results of certain text processing operations (this applies in particular to classification, but also to the generation of semantic clusters by topic modelling).

While the first two limitations are well known in empirical linguistics, the third limitation has only in recent years been recognized as a problem. The mass availability of electronic text and the development of NLP methods has led to a situation where clustering of texts (and microtexts) is relatively easy, but evaluation of the quality of the result obtained is often impossible as the number of potential divisions of classified objects into clusters (sets) is practically unlimited (it depends on the selection of relevant features of these objects and the metrics of similarity applied). In natural language text research, testing (evaluation) of results can be carried out only by people, or by application of formal measures (e.g. minimization of the system perplexity). The situation is completely different in bibliographic corpora, where the reference system for potential tests exists in the form of metadata. The above arguments (and counterarguments) indicate the possibility of effectively processing large bibliographies as text corpora with automatic methods.

2. Data and Research Hypotheses

The subject of our research is a corpus of approximately 1,850,000 bibliographic records extracted from the national bibliography of the Polish National Library via the API interface[1]. The records are stored in a strongly redundant and slightly archaic MARC format, which, however, is used in most of the bibliographic databases of the world. The fields containing the title of the work (only titles in Polish were included), the authors' data, key words and genology data (writing genre) were considered relevant for automatic analyses. The database includes literature published in the 20th and 21st centuries without taking into account the actual year that the text was written (however, twentieth century texts predominate).

Two hypotheses were formulated in preparation for the research. The first one concerns the efficiency of recognizing the literary genre of the text solely on the basis of the title and assumes that the method of machine learning can effectively attribute specific texts to the writing genre to which they actually belong. The problem of genology in this case was a complex one due to the number of genres and inconsistencies in the description of classification traits. Working solely on metadata, we distinguished as many as 7000 writing species (or genres). This number was then reduced to the set of 52 most salient genres because the smallest classes were eliminated (many of them were the outcome of human error). Further steps of text processing included the elimination of genres equivalent in terms of subject matter of publication but bearing different names. The final list obtained in this way consisted of 31 items and allowed for an effective verification of the automatic classification results (cf. 4.1).

The second hypothesis is more risky, but is worth consideration if only to eliminate possible assumptions about its validity. Recently, literature on the subject includes studies devoted to automatic recognition of the "cultural gender of the text". These studies focus primarily on belles-lettres (cf. Rybicki 2016, Walkowiak & Piasecki 2018), but also on microtexts available in social media, i.e., tweets and text messages (Mikros 2013, Mikros &

[1] http://data.bn.org.pl/

Perifanos 2013, Silessi et al. 2016). Having a corpus of titles at our disposal, we assumed that it would be possible to automatically recognize the gender of the author, especially in the case of genres that allowed the author relative freedom in composing the title (mainly in belles-lettres and biography). It was assumed that the hypothesis would be positively verified if the attribution obtained automatically would be significantly better than the purely statistical probability of gender attribution. As an additional working hypothesis, we posited that the effectiveness of recognizing "gender" would be directly proportional to the length of the title (for example, it is doubtful that a good result will be obtained on the basis of one-word titles). Both hypotheses can be verified based on the metadata contained in the records, which allows for virtually flawless validation.

3. Methodology

In our research, we used a recently developed deep learning package called *fastText* (Joulin et al. 2017). It consists of two different approaches: supervised and unsupervised. The first one, that we will call within this paper *supervised fastText*, is based on the representation of documents (doc2vec) as an average of word embedding (word2vec) and uses a linear soft-max classifier (Goodman 2001) to assign the doc2vec representation to one of a range of known classes. This hidden representation is used by a linear classifier for all classes (i.e., literary genres and the author's gender), allowing information about word embedding learned for one class to be used by others. *Supervised fastText* by default ignores word order, much like the classical bag of words (BoW) method (Harris 1954). The main idea behind *supervised fastText* is to perform word embedding and classifier learning in parallel (simultaneously). Since *supervised fastText* forms the linear model, it is very effective for training and achieves solutions faster by several orders of magnitude compared with other competing methods (Joulin et al. 2017). During classification, words that do not exist in the embedding model (because they do not exist in the training corpus) are omitted from the averaging. *FastText* allows us to build the embeddings not only for single words but also for word *n*-grams. It allows local word order to be taken into account.

The second approach, referred to in this paper as the *fastText language model,* is an extension of the word2vec (Le & Mikolov 2014) model built on Common Crawl and Wikipedia by *fastText* in unsupervised mode. The models were trained by CBOW (Continuous Bag of Words model) with position weights and subword information (Grave et al. 2018)[2]. The word representation is constructed as the sum of the character *n*-grams embeddings (for *n*-grams appearing in the word). It allows for generation of word embedding for words not seen in a training corpus and for working with inflected languages such as Polish. Since the *fastText language model* provides vector representations of individual words, we have represented documents as an average of these vectors. The vector representations were classified by the Multi-Layer Perceptron (*MLP*) based on a model trained by the back error propagation method on the learning set (Hastie et al. 2013).

4. Experiments and results

4.1 Recognizing the literary genre of the text

In order to verify the first hypothesis, the number of classes had to be limited based on the criterion of the number of elements. It was initially assumed that effective classification was possible if the class had no fewer than 5000 titles, while the less numerous classes were rejected (in such cases the training corpus would be too small to be effective). This allowed us to identify 52 writing genres, which included a total of about 570,000 titles. The data were randomly divided into training and testing set in the proportion of 2 to 1. This proportion was used throughout all experiments performed. The primary results of automatic classification using the *supervised fastText* method (with word unigrams) obtained accuracy of 54% on the test set. Detailed analysis showed that classes not recognized correctly by the algorithm had in fact a large similarity to certain classes recognized correctly (for example, sub-genres of different types of textbooks, novels, and stories were clustered together). This means that mistakes in attribution are actually due to the polysemy of the language, which means that “collections” and “anthologies”, for example, are treated as separate categories (although in reality they can describe the same objects).

As a result, the number of genre classes was reduced to 31, namely: academic textbooks (Pol. *podręczniki akademickie*), novels (Pol. *powieści*), anthologies (Pol. *antologie*), conference materials (Pol. *materiały konferencyjne*), popular publications (Pol. *wydawnictwa popularne*), guides (Pol. *poradniki*), biographies (Pol. *biografie*), albums (Pol. *albumy*), diaries (Pol. *pamiętniki i wspomnienia*), textbooks for vocational schools (Pol. *podręczniki dla szkół zawodowych*), stories (Pol. *opowiadania i nowele*), children’s literature (Pol. *literatura dla dzieci*), travel guides (Pol. *przewodniki turystyczne*), textbooks for primary schools (Pol. *podręczniki dla szkół podstawowych*), developing (Pol. *poradniki rozwoju osobistego*), Polish

[2] https://fasttext.cc/docs/en/crawl-vectors.html

journalism (Pol. *publicystyka polska*), support materials (Pol. *dokumenty towarzyszące*), comics (Pol. *komiksy*), children's poetry (Pol. *poezja dla dzieci*), youth novel (Pol. *powieść młodzieżowa*), religious considerations and meditations (Pol. *rozważania i rozmyślania religijne*), historical literature (Pol. *literatura historyczna*), publications for children (Pol. *wydawnictwa dla dzieci*), textbooks for high schools (Pol. *podręczniki dla szkół ponadgimnazjalnych*), statistical data (Pol. *dane statystyczne*), analyses and interpretations (Pol. *analizy i interpretacje*), commemorative books (Pol. *księgi pamiątkowe*), exercises and tasks (Pol. *ćwiczenia i zadania*), bibliography (Pol. *bibliografia*), encyclopaedias (Pol. *encyklopedie*), compendia and indexes (Pol. *kompendia i repertoria*).

Table 1. Accuracy of the attribution of titles to their literary genres

Classification method	Accuracy
supervised fastText, unigrams	64.9%
supervised fastText, bigrams	71.1%
supervised fastText, trigrams	**71.8%**
supervised fastText, fourgrams	71.7%
supervised fastText, fivegrams	71.5%
fastText language model+MLP	66.2%

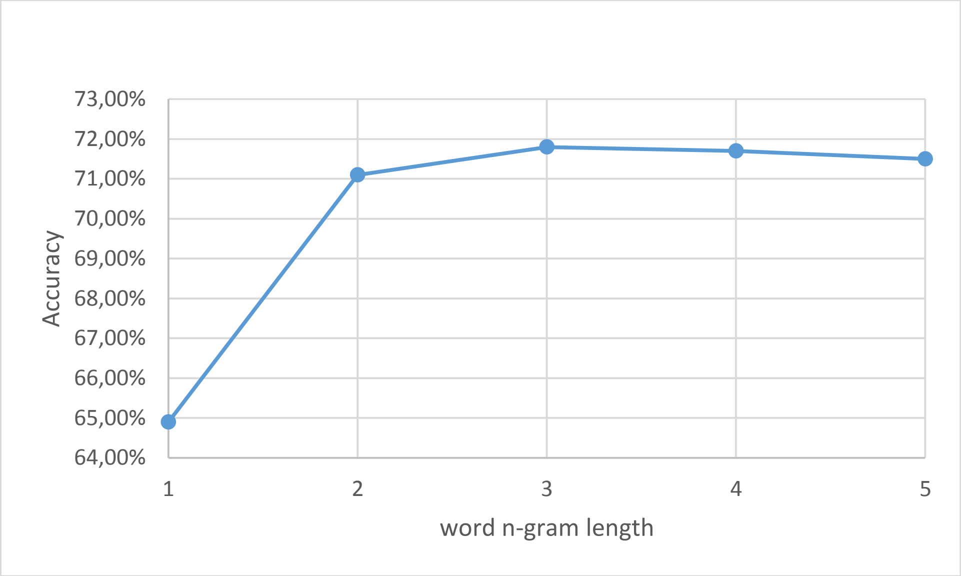

Figure 1. Accuracy of the literary genre attribution for the *supervised fastText* method as a function of word *n*-grams length.

The results of automatic classification using the *supervised fastText* method with word embeddings built on single words, word bigrams, trigrams etc., and verified on the basis of the metadata included in the database records, turned out to be surprisingly satisfactory. The relationship between the length of word *n*-grams and the accuracy of the attribution (Figure 1) prompted the observation that the most effective assignments can be carried out on tri-word titles. However, even single word titles allow one to achieve a nontrivial level of recognition of a literary genre. The accuracy in this case was at 64.9%, while random attribution accuracy for the most frequent class of academic textbooks in a test set would be at 19.6%. Although tri-word titles are the most effective in absolute numbers (71.8%), in fact, a caesura occurs between single- and multi-word titles. Interestingly, an increase in the length of the title (4-grams, 5-grams etc.) does not lead to an increase in the accuracy of assignment of titles to their proper classes of literary genre. This fact can be interpreted as an effect of significant limitations of the title naming system and its semantic saturation. Despite appearances, this system does not allow for free choice of the linguistic and/or stylistic measures. Therefore, the automatic method based on a well-trained algorithm allows for effective attributions of even very short microtexts, which would be less probable in the case of free speech samples or other text species.

We also conducted a detailed analysis of the attribution accuracy for selected literary genres (treated here as classification categories). First, we measured the number of correct decisions from all assignments made to a target specific class (precision), and the number of correct decisions from all assignments expected to a specific class (recall). Table 2 gives the precision and recall values for the first four genre classes (about 50% of all titles). In all cases, the quality of automatic matching of titles to the appropriate classes proved to be very high.

Table 2. Quality of writing genre attribution by automatic method (first 4 genres)

Writing genre	Precision	Recall	Support
Academic textbooks	88.2%	82.1%	19.6%
Novels	81.9%	67.8%	14.3%
Anthologies	68.5%	64.5%	11.1%
Conference materials	82.4%	80.6%	7.77%

The investigation of false decisions of the algorithm applied also proved interesting. Some text genres were intensely confused, which could suggest a method error. Table 3 shows that, for example, in the class indexed manually as ‘studies’ (Pol. *opracowania*) only 14.3% of titles were attributed correctly, while 20.6% were assigned to ‘academic textbooks’; 28.8% of the titles indexed manually as ‘children’s poetry’ were correctly recognized, but 21.1% were assigned to the ‘anthology’ class. Similarly, the algorithm assigned 24.7% of titles indexed manually as ‘youth novel’ to the ‘novel’ class, and part of ‘support materials’ was recognized as ‘academic textbooks’. However, on closer examination, it turns out that decisions based on the *supervised fastText* algorithm are correct, because at the stage of manual indexing of documents, due to human errors, some titles of works apparently similar in terms of content were placed in different classes. In fact, children’s poetry collections are also anthologies, a youth novel is a novel, and ‘studies’ or ‘support materials’ are also academic textbooks. And however paradoxical this may sound, it can be claimed that the *supervised fastText* algorithm, trained on a sufficiently large database, is capable of correcting errors or inconsistencies in human decisions.

Table 3. Genres most frequently misclassified (selected 4 genres)

Writing genre (A)	Classified correctly (precision)	Misclassified as (B)	Percentage of genre A classified wrongly as genre B
Studies (Pol. opracowania)	14.3%	Academic textbooks	20.6%
Children’s poetry	28.8%	Anthology	21.1%
Youth novel	61.3%	Novels	24.7%
Support materials	59.5%	Academic textbooks	19.1%

However, our research shows that the decision to choose an analysis algorithm is not obvious and there are several possible options here. Since the *supervised fastText* method is not able to take into consideration words not existing in the training set, we also tested the *fastText language model* (see Section 3) trained on a huge corpus of Polish texts capable of calculating vector representations of words for unseen words (due to a usage of character *n*-grams). The averaged word vectors were classified by the MLP (Multi-Layer Perceptron) classifier, achieving accuracy of 66.2%. The results (see Table 1) were better than for the *supervised fastText* method, in that usage of word unigrams had a lower success rate compared with usage of longer word *n*-grams.

4.2. Automatic recognition of the author’s gender

Another important aim of the research was to automatically recognize the author’s gender and evaluate the result obtained. This task was carried out in four stages. At the beginning all titles of multi-author works were eliminated, so that the author’s gender category was unambiguous. The second phase of the experiment included automatic recognition of the author’s gender and indexing database records with ‘M’ or ‘F’. This operation was necessary because metadata do not provide this information (apparently considered as being obvious).

This task turned out to be relatively easy because Polish as an inflected language is characterized by a very useful feature, which is the almost common occurrence of the female first name suffix *–a* ([cons.] *–a*). Male names have different endings, where consonants and sonants dominate. The gender of non-Polish authors of translated books, as well as some rare names, were recognized semi-automatically. The third phase of the experiment consisted in eliminating those literary genres where titles are constructed according to more or less strict rules preventing the free linguistic expression of gender. This applies primarily to texts related to science, education and administration. Genres giving more freedom of titling, defined broadly as belles-lettres, poetry, drama, biography, or some 'how-to' guides, were left in the corpus. We obtained in this way about 280,000 well profiled and annotated titles where 71% authors were male and 29% female. The last phase consisted in the application of the *fastText* algorithm to perform automatic attribution of author's gender using only titles. The task was carried out separately for titles composed of one word, two words, etc. The best result was obtained for four- and five-word titles (79%). However, as in the case of literary genre attribution (see above), a slightly lower value for unigrams can be observed, followed by a steep increase and stable values for longer titles (Table 4 and Figure 2).

Table 4. Accuracy of the author's gender attribution based on the title for selected genres

Classification method	Accuracy
supervised fastText, unigrams	75.7%
supervised fastText, bigrams	78.2%
supervised fastText, trigrams	78.9%
supervised fastText, fourgrams	79.0%
supervised fastText, fivegrams	79.0%
supervised fastText, sixgrams	78.9%

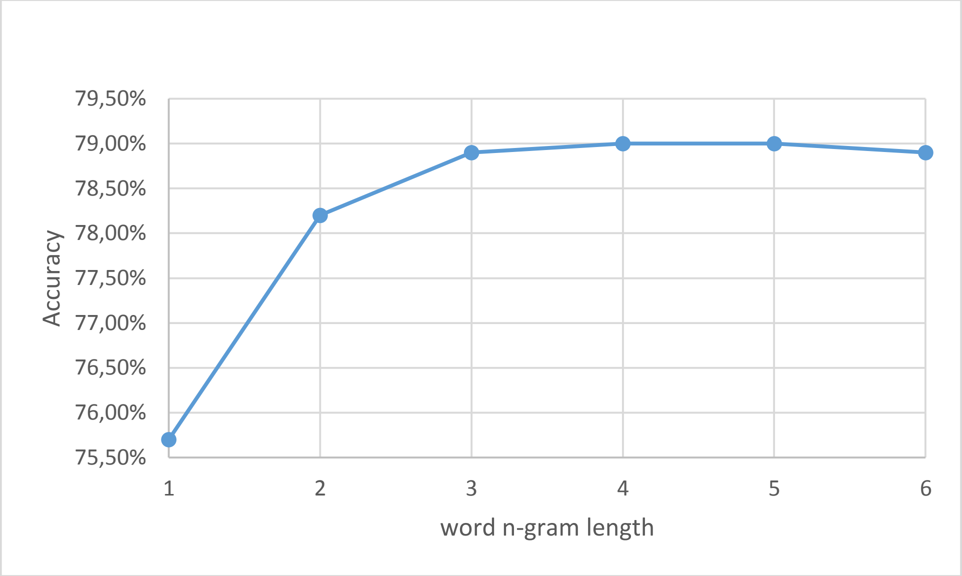

Figure 2. Accuracy of the gender attribution for *supervised fastText* method in a function of word *n*-grams length

Estimating this result is not easy. In the case of a proportional distribution of authors, it should be assumed that any value significantly higher than 50% will be considered as satisfactory (random attribution in such a corpus of texts should yield a 50% success rate). The 79% value is therefore significantly higher and could be considered satisfactory. However, in this case the distribution of authors is not balanced and it is much more advantageous to estimate the gender attribution separately for male and female authors. The result of this attribution (Table 5) turned out to be very good, as both precision and recall values were about 20 percentage points higher than the values of random attribution. Of course, the question remains open as to whether a 100% attribution is possible. Given the complexity of the human mind and the impact of culture on gender, this seems impossible – at least for very short and standardized microtexts, such as titles (it would be, however, more probable in the case of free text, e.g., in social media).

Table 5. Author's gender attribution based on four-word book titles for selected genres

Gender	Precision	Recall	Expected random attribution (support)
Male	90.8%	81.8%	71%
Female	49.6%	68.4%	29%

Using this result, a test of a subsidiary hypothesis was also carried out, which was taken for granted at the outset, namely that there are genres that allow for the free shaping of titles and those that impose the structure and vocabulary of the titles. Table 6 contains the results of the author's gender recognition accuracy test across the entire corpus of titles (over 855,000 items). The result turned out to be surprising, as the quality not only did not decrease, but even increased slightly. The results for different word *n*-grams in titles are presented in Table 6. Table 7 lists the precision and recall values obtained for word 4-grams.

Table 6. Accuracy of the author's gender attribution based on the title for all genres

Classification method	Accuracy
supervised fastText, unigrams	76.5%
supervised fastText, bigrams	80.3%
supervised fastText, trigrams	81.4%
supervised fastText, fourgrams	81.9%
supervised fastText, fivegrams	82.0%
supervised fastText, 6-grams	82.0%
supervised fastText, 7-grams	82.0%
supervised fastText, 8-grams	81.9%

Table 7. Author's gender attribution based on four-word book titles for all genres

Gender	Precision	Recall	Expected random attribution (support)
Male	91.4%	85.7%	76%
Female	50.9%	65.0%	24%

Does this result mean that titles of scientific or law books can be considered as an expression of the author's gender? Of course not. The high values of these parameters in the most standardized genres probably stem from the fact that certain subject areas are gender-specific. It is therefore not the case that the lexemes such as 'law', 'medicine', 'administration' or 'physics' connote masculinity or femininity. Rather, publications on certain topics are (or have been) more often written by women or by men and this principle was learned by the *supervised fastText* algorithm. In fiction and other 'free' genres there is no such dependence, which makes the previously obtained assignments attributes valuable from a cognitive point of view.

5. Conclusions

The research conducted proves that automatic taxonomy of microtexts as short as book titles is possible and gives positive results. Previous studies in microtext taxonomy and gender recognition were also successful but they were carried out on tweets which, despite being considered as short, seem extremely elaborate and rich in content compared to book titles (cf. Mikros & Perifanos 2013). Available research in book title automatic processing is rather scarce: it concerns only "book genres" and was conducted on smaller corpora derived from Amazon databases (Ozsarfati et al. 2019, Chiang et al. 2015). In our study an effective recognition of the book genre (called also writing species) as well as the author's gender was conducted on the basis of two-, three- and four-word sequences. We worked on an extensive bibliographic corpus in a flexional language (Polish). The result obtained is unexpected and intuitively not obvious, as artificial intelligence in application to language data derived from great bibliographies has apparently exceeded human capabilities. The study, however, seems difficult to challenge, as it is methodologically sound and based on solid empirical material.

Unfortunately, we do not have the results of similar tests carried out on human respondents. But our intuition and many years of scientific experience suggest that an artificial intelligence system should give "human-like" results in recognizing a genre of a book or of a scientific paper based just on its title. This is due to the fact that meanings of words or expressions in titles are more or less universally reproduced in the minds of persons speaking a given language and computer algorithms can only reproduce this competence. However, recognizing the author's gender solely from the titles would be difficult for human respondents as it requires something more than just semantic knowledge. Its condition is to be familiar with the entire knowledge base developed in the course of machine learning from the training set of data – in his case a bibliographical corpus consisting of hundreds of thousands of items. This explains, why it may seem that a computer has a linguistic competence that exceeds that of a human being. The source of this success, apart from the narrowly understood technology (memory capacity, processing speed etc.), is most likely the philosophy of artificial intelligence algorithms (here *fastText*), which assume extensive training on large and reliable data in order to build a knowledge base. This makes it possible to effectively process new items, such as words, clauses, sentences etc., created within the same communication system in a given language.

At the end of these reflections, it is worthwhile to ask one more general question: is AI the future of linguistics? The matter seems delicate, as humans have always felt a subconscious fear of "thinking" machines. This archetype is deep-rooted, as it dates back to biblical times, where the beginning of the story of artificial beings threatening humans is found (e.g. Golem). But today the facts are such that the number of texts artificially generated by algorithms is growing (it is hard to say, for example, who in the Amazon databases quoted above created the book metadescriptions – a human or a computer). In addition, we can observe a massive increase in machine-readable texts that people are no longer able to acquire or analyze. These reasons make automatic, complex and seemingly opaque algorithms of language analysis and creation the inevitable future of linguistics.

References

Chiang, Holly, Yifan Ge & Connie Wu. 2015. *Classification of book genres by cover and title*, http://cs229.stanford.edu/proj2015/127_report.pdf

Goodman, Joshua. 2001. Classes for fast maximum entropy training. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, https://arxiv.org/pdf/cs/0108006.pdf

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin & Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation* (LREC 2018), https://www.aclweb.org/anthology/L18-1550.pdf

Harris, Zellig S.1954. *Distributional structure*. *WORD* 10.2-3. 146-162. doi: 10.1080/00437956.1954.11659520

Hastie, Trevor, Robert Tibshirani & Jerome Friedman. 2013. *The elements of statistical learning: Data mining, inference and prediction* (Springer series in statistics), New York: Springer.

Joulin, Armand, Edouard Grave, Piotr Bojanowski & Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol. 2, Short Papers*, 427–431, https://www.aclweb.org/anthology/E17-2068.pdf

Le, Quoc & Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, http://proceedings.mlr.press/v32/le14.pdf

Mikolov, Tomas, Wen-tau Yih & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta: Association for Computational Linguistics, 746–751, https://www.aclweb.org/anthology/N13-1090.pdf

Mikros, George K. 2013. Systematic stylometric differences in men and women authors: a corpus-based study. In Reinhard Köhler & Gabriel Altmann (eds.), *Issues in Quantitative Linguistics* 3: *Dedicated to Karl-Heinz Best on the Occasion of His 70th Birthday*, 206–223, Lüdenscheid: RAM–Verlag, https://www.academia.edu/3429459/Systematic_stylometric_differences_in_men_and_women_authors_a_corpus-based_study

Mikros, George K. & Kostas Perifanos. 2013. Authorship attribution in Greek tweets using author's multilevel n-gram profiles. In *AAAI Spring Symposium*: *Analyzing Microtext*, https://www.aaai.org/ocs/index.php/SSS/SSS13/paper/viewFile/5714/5914

Ozsarfati, Eran, Egemen Sahin, Can J. Saul and Alper Yilmaz. 2019. Book genre classification based on titles with comparative machine learning algorithms. In *IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, 14-20. doi: 10.1109/CCOMS.2019.8821643

Rybicki, Jan. 2016. Vive la différence: Tracing the (authorial) gender signal by multivariate analysis of word frequencies. *Digital Scholarship in the Humanities* 31.4. 746–761, https://doi.org/10.1093/llc/fqv023

Schwartz, Roy, Oren Tsur, Ari Rappoport & Moshe Koppel. 2013. Authorship Attribution of Micro-Messages. In *Proceedings of EMNLP*, https://www.aclweb.org/anthology/D13-1193.pdf

Silessi, Shannon, Cihan Varol & Murat Karabatak. 2016. Identifying Gender from SMS Text Messages. In *5th IEEE International Conference on Machine Learning and Applications* (ICMLA), Anaheim, CA, 488–491. doi: 10.1109/ICMLA.2016.0086

Walkowiak, Tomasz & Maciej Piasecki. 2018. Stylometry analysis of literary texts in Polish. In Leszek Rutkowski, Rafał Scherer, Marcin Korytkowski, Witold Pedrycz, Ryszard Tadeusiewicz & Jacek M. Zadura (eds.) *Artificial Intelligence and Soft Computing* (Lecture Notes in Artificial Intelligence 10842), 777–787, Cham: Springer. doi: 10.1007/978-3-319-91262-2_68