STATISTICS OF A LARGE-SCALE ROMANIAN CORPUS FOR LANGUAGE MODELLING

T.A. DIAC[1], A.F. NEAGOE[2], M.C. RAPORTARU[3], A. OPREA[4,†], R.-M. DRĂGAN[3,‡], A. NICOLIN-ŻACZEK[3,*]

[1]Lund University, Faculty of Science, Box 118, 22100 Lund, Sweden
[2]Faculty of Physics, University of Bucharest, Atomiştilor 405, Măgurele, Romania
[3]Institute of Space Science – INFLPR Subsidiary, Atomiştilor 409, Măgurele, Romania
[4]T2, Ceauş Firică, 145100, Roşiori de Vede, Teleorman, Romania
[†]Email: anatol.oprea@t-2.srl, [‡]External collaborator
Corresponding author[*]: alexandru.nicolin.zaczek@spacescience.ro

Received June 12, 2025

Abstract. We report here a series of detailed statistical analyses on a novel large-scale multi-domain Romanian corpus that we use to train a small-language model. We identify the core vocabularies pertaining to six different domain-specific subcorpora and show that they follow the so-called Zipf's law independent on how we count words. Moreover, we introduce two novel frequency-word maps for a domain-specific subcorpus, one showcasing word ranks and one measuring the deviation of the word structure from a perfect vowel-consonant or consonant-vowel repeating pattern. Finally, we show a few examples of prompt/response instances.

Key words: Large-scale multi-domain Romanian corpus, Zipf's law, Language modelling.

DOI: https://doi.org/10.59277/RomJPhys.2025.70.111

1. INTRODUCTION

The large-scale availability of AI tools, particularly language models, has given a huge boost to AI-generated synthetic texts pertaining to task-specific user-defined subjects [1]. Despite this unprecedented surge of task-specific synthetic texts, there is little actual transparency on how the training of these models is done and the impact the training corpora have on the properties of the generated synthetic texts. Understanding these details is essential for *i*) ensuring that calibrated models comply with ethical, IPR and security constraints and *ii*) assessing output quality. This, in turn, enables proper governance of generative AI solutions. To this end, we present here a series of statistical results for a novel large-scale multi-domain Romanian corpus that allows us to easily quantify the properties of a large collection of texts using both well-known indicators like word frequency, word length and word ranking, and a newer one, which originates in statistical physics. This new indicator describes the structure of a given word by quantifying the differences between its structure and a perfect vowel-consonant or consonant-vowel repeating pattern. We refer to this indicator – originally introduced in the context of phonetically balanced utterances for biometrics – as the m-function and use it to distinguish between two subcorpora of seemingly similar structure.

The rest of the article is structured as follows: in Section 2 we present all technical details – from initial aggregation to tokenization – pertaining to the novel large-scale Romanian corpus, while in Section 3 we present our statistical results and briefly discuss the output of a small-language model (SLM) trained on the distinct subcorpora previously analyzed. Finally, Section 4 gathers our conclusions and some considerations for future extensions.

2. ROMANIAN CORPUS

Over the past two decades, Romanian language resources have grown from the first manually annotated treebanks – UD–RRT, containing about `59 000` tokens and `3 900` sentences [2] – to web-scale collections exceeding 100 million tokens. This three-orders-of-magnitude expansion has been driven by the availability of scalable web-scraping pipelines, automatic quality filters for language identification and deduplication, and the public availability of large open-data platforms such as Wikimedia dumps, the *OSCAR* initiative [3], and, more recently, the *LiRo* benchmark [4]. This steady increase in data volume has been accompanied by a larger multi-domain coverage, though there are gaps with respect to regional dialects, contemporary spoken registers, and recent literary texts. Moreover, full transparency of cleaning and filtering procedures is still required to guarantee reproducibility and to assess corpus-induced biases. This shows that training directly on native Romanian data is no longer a technical luxury, but a prerequisite for robust, scalable, and transparently-evaluated natural-language applications.

Using a series of custom built, automated web-scraping and extraction tools, we collected a large-scale, multi-domain Romanian language corpus of around 800 million tokens. Emphasis was placed on content preprocessing and curation to ensure that the resulting dataset is suitable for a wide-range of statistical research and AI training objectives. The data was acquired from publicly available sources spanning multiple domains of interest, including encyclopedic content, literature, news articles, legal documents, technical articles, and general web content such as blogs and forums. Encyclopedic content was sourced from the Romanian `Wikipedia` [5], while literary texts were extracted from Romanian `Wikisource` [6]. For both of them, we used the official April 2024 Wikimedia XML dump files, publicly available at `https://dumps.wikimedia.org`. The dumps were processed using the open-source tool `WikiExtractor` [7], which allowed us to extract and clean the raw MediaWiki markup content into structured `.jsonl` files, each containing one article per line, formatted as a JSON object with the full text as content. The subcorpora pertaining to the other domains of the corpus were also structured into `.jsonl` files, a format chosen for its advantages in data streaming and efficient access during processing. Table 1 summarizes the size and structure of each domain-specific subcorpus, including total token counts, vocabulary sizes, and the number of extracted objects, everything prior to filtering.

Table 1

Statistics of each domain-specific subcorpus, raw form (before filtering)

Domain	**Tokens**	**Unique Tokens**	**No. of Objects**	**Size (GB)**
Legal	251,749,350	2,864,995	609,881	1.81
Websites	230,440,267	3,146,952	242,512	1.57
News	194,062,691	2,533,599	689,745	1.28
Wikipedia	96,351,135	2,819,344	602,312	0.65
Wikisource	19,905,775	981,681	17,740	0.12
IT	4,899,784	188,778	14,011	0.03
Total	**797,408,002**	—	**1,434,607**	**5.46**

In the data acquisition process a significant number of artifacts such as `HTML` tags, markup residues, and formatting inconsistencies were encountered, which made a robust preprocessing stage essential. Sentence boundaries were detected and split using punctuation markers such as `.`, `,`, `;`, `?`, `!`, while all non-textual elements including emojis, symbols, and extra punctuation were discarded. Capitalized words appearing mid-sentence, often indicative of named entities or formatting noise, were excluded from token counts to maintain linguistic consistency. Tokens containing digits, hyphens, or consisting of a single character were filtered out, as these were generally found to induce noise or lack statistical value for our purpose. Before initiating the preprocessing stage, we first extracted lemma families for as many Romanian words as possible. This was accomplished using the publicly available DEX Online Romanian Dictionary dumps, accessible through their official GitHub repository [8]. By parsing the `SQL` database structure, we extracted around `313 000` lemmas and their associated inflected forms, grouping them into a `.txt` file with morphological families. This structured representation allowed us to treat inflectional forms as part of the same lexical unit. The lexical families `.txt` file is organized as a two-column dataset, where the first column lists the main forms, and the second column contains all lexical variants associated with a given main form, a small example being presented below:

```
tastare     tastare,tastarea,tastări,tastării,tastările,tastărilor
anulare     anulare,anularea,anulări,anulării,anulările,anulărilor
minge       minge,mingea,mingi,mingii,mingile,mingilor.
```

After completing the preprocessing and lemmatization steps for all six subcorpora, we obtained six distinct `.jsonl` files, each containing the cleaned and normalized version of its respective subcorpus, see Table 2. We have scrutinized the vocabularies of the six subcorpora and did not find words of significant occurrence frequency pertaining to obscene language, racial narratives, or criminal activities. Finally, let us mention that as the materials were gathered from public sources, our subcorpora can be used for the calibration of language models without the risk of IPR infringements or the accidental disclosure of non-public information.

Table 2
Statistics of each domain-specific subcorpus, lemmatized version (after filtering)

Domain	**Tokens**	**Unique Tokens**
Legal	206,324,988	340,066
Websites	201,830,501	554,980
News	164,738,095	513,256
Wikipedia	89,615,083	546,033
Wikisource	19,881,521	203,809
IT	4,526,419	44,442
Total	**686,916,607**	—

3. RESULTS

3.1. STATISTICAL RESULTS

We present here the results of our statistical analyses on the six subcorpora which make up the main corpus. In Fig. 1 we depict in log-log plots the frequency-rank distributions of words for all six domains, showcasing three distinct situations: in the first case we consider all alpha-numeric words in the corpus (see blue curves), regardless of their presence in the dictionary, in the second case we consider the main form of all words in the dictionary plus all other alpha-numeric words (see red curves), while in the third case we consider only the main forms (see yellow curves). Also shown in dashed line is the ideal Zipf's law which states that the frequency of a word is inversely proportional to its rank, see Ref. [9] for a general presentation and Ref. [10] for previous results on Romanian, and one can easily notice that the law holds over five to six orders of magnitude, depending on the subcorpus under scrutiny. In Fig. 2 we show the distribution of word frequencies considering all alpha-numeric words (right panel) and main forms only (left panel). In both cases we see that the distributions are very close to one another and show a clear scale-free nature, which, however, is more pronounced in the case of main forms, for which we also show (in dashed line) the scale-free approximation for comparison. To the best of

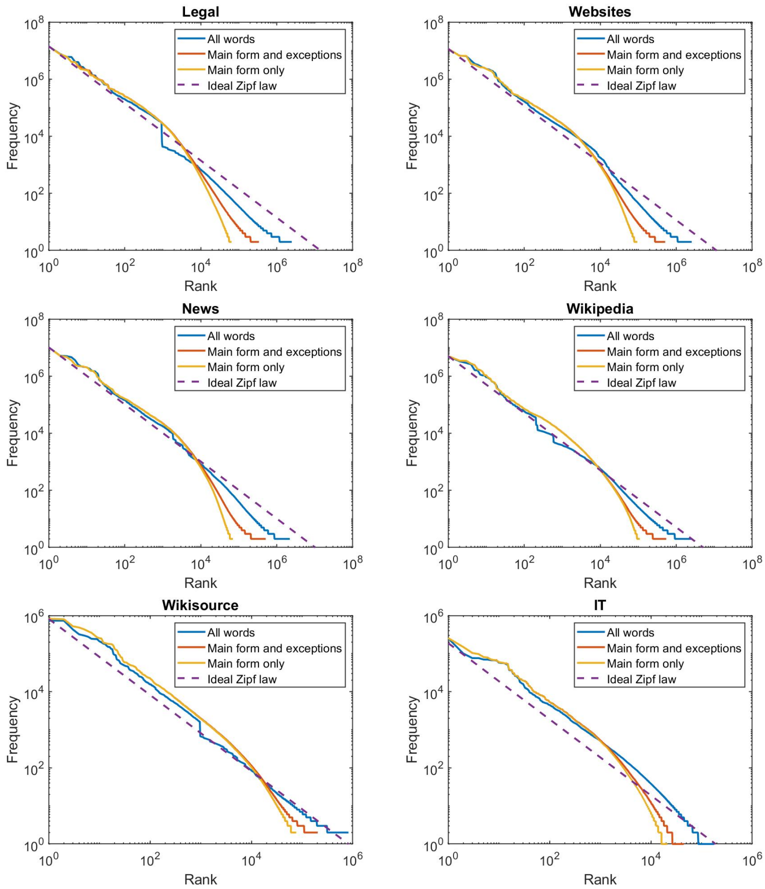

Fig. 1 – (Color online) Verification of Zipf's law for the six domains covered by the corpus using all alpha-numeric words in the text (blue upper curve), the main form of the words found in the dictionary and other alpha-numeric words (red middle curve), and only the main forms (lower yellow curve). The dashed line corresponds to the ideal Zipf law.

our knowledge this is the most thorough verification of Zipf's law for the Romanian language in terms of number of domains and overall number of words in the corpus. Our results are comparable in terms of complexity with those published in Ref. [11].

Next, in Fig. 3, we show the word-length–frequency isoline plots pertaining to the News and Wikipedia domains for the number of words, the sum of word ranks in the local subcorpus, and the sum of word ranks in the global corpus. Let us note that plots similar to the first row in Fig. 3, *i.e.*, the ones depicting the isolines of the number of words, have been published by Corral and Serra in Ref. [12]. Also note that the global corpus is obtained through the aggregation of the aforementioned domain-specific subcorpora and it is used to check the extent to which the word ranking has

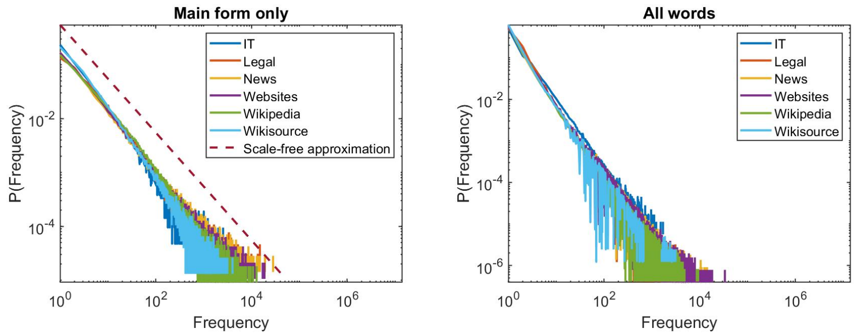

Fig. 2 – (Color online) Distribution of word frequencies for the six subcorpora considering only the main forms (left figure) and all alpha-numeric words (right figure). Both distributions are akin to a scale-free distribution, but the one for main forms provides a better fit as can be seen from the depicted scale-free approximation.

domain-specific features. Similar plots (not shown here) have been observed for the other four subcorpora, the differences between plots being merely quantitative. This common behavior observed in the word-length–frequency maps for the isolines depicted in Fig. 3 is important from a theoretical point of view, as it shows that the number of words and the sum of ranks have very similar properties with respect to word frequencies and word lengths. This similarity also shows that our subcorpora are large enough for finite-size effects to be negligible, a result which is mirrored in the virtually identical distribution of letters across the six subcorpora, see Fig. 4. The aforementioned result conveyed by Fig. 3 has, however, limited practical utility as these maps cannot be effectively used as *fingerprints* of these subcorpora due to their intrinsic similarity, thus a different indicator should be used.

The m-function – originally introduced in Ref. [13] for the Romanian language and later used in Ref. [14] for the Greek language – quantifies the differences between the structure of a given word and a perfect vowel-consonant or consonant-vowel repeating pattern, with m being equal to -1 for perfect series of vowel-consonant or consonant-vowel clusters and $+1$ in the case of a series of vowels or consonants. The m-function is akin to the standard Ising model used in statistical physics, where letters are replaced by spins and the values associated to the letters represent the so-called up and down states of the spins. The function is defined as:

$$m = \frac{1}{n-1}\sum_{j=1}^{n-1} l(j)\cdot l(j+1) \tag{1}$$

where $l(j)$ stands for the numerical value of the j-th letter in a word of length n and is equal to $+1$ if the letter is a vowel and -1 if the letter is a consonant [13]. As the m-function accounted for the fine (*i.e.*, vowel-consonant) structure of words it allowed for a clearer distinction between the six subcorpora than that offered by the

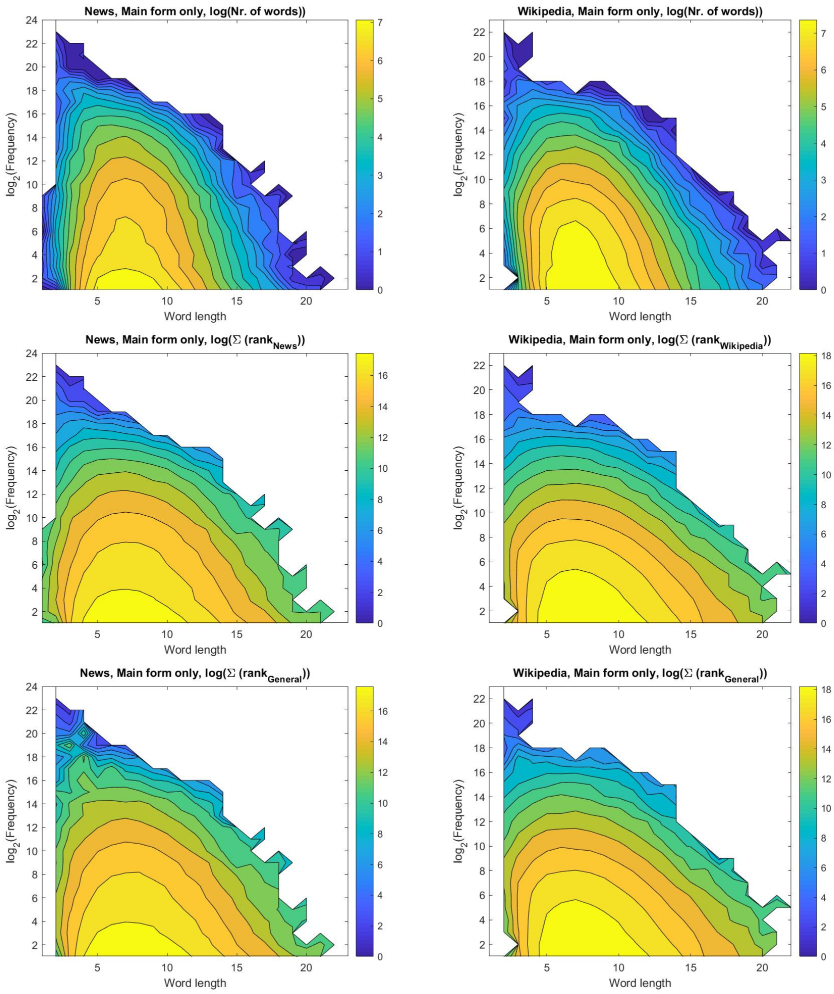

Fig. 3 – (Color online) Word-length–frequency isoline plots pertaining to the News (left column) and Wikipedia (right column) domains for number of words (upper row), sum of word ranks in the local subcorpora (middle row), and sum of word ranks in the global corpus (lower row).

indicators in Fig. 3. We noticed, however, that the differences between two distinct subcorpora are more visible on a word-length–frequency isoline plot if we include in Eq. (1) a term which accounts for the number of three-letter clusters of either vowels or consonants, namely

$$m = \frac{1}{n-1}\sum_{j=1}^{n-1} l(j)\cdot l(j+1) + \tilde{n}\cdot p, \tag{2}$$

where $\tilde{n}$ is the number of clusters in a word and p is a numerical value. In Fig. 5,

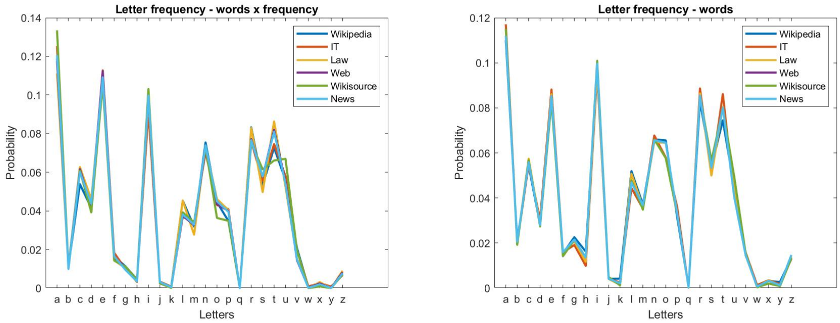

Fig. 4 – (Color online) Letter occurrence probability computed with and without accounting for the occurrence frequency of each word, see left and right plot respectively.

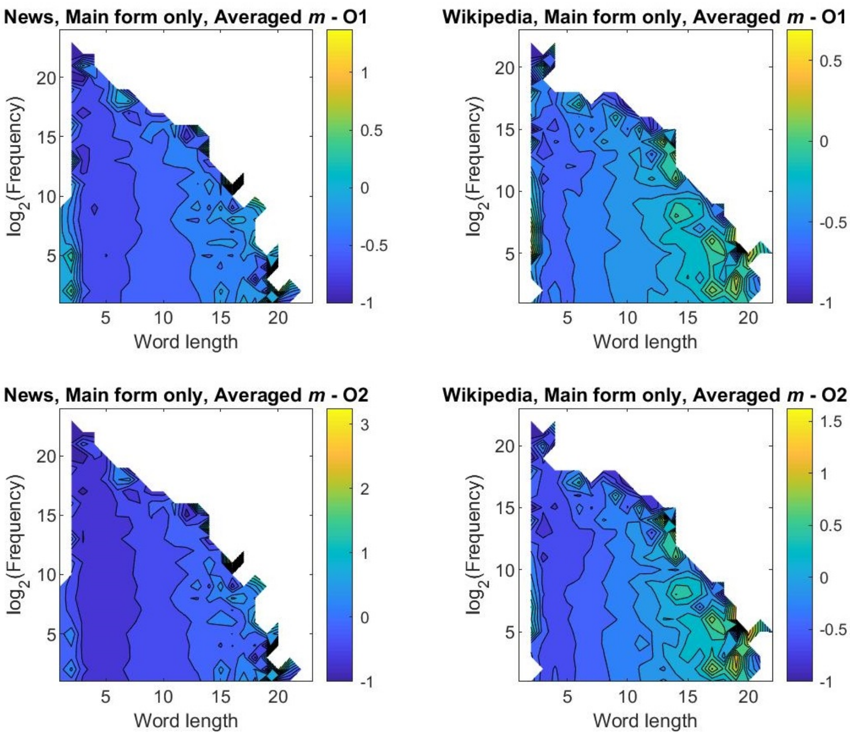

Fig. 5 – (Color online) Word-length–frequency isoline plots pertaining to the News (left column) and Wikipedia (right column) domains for two versions of the averaged m function, upper row corresponding to $p = 2$ and lower row corresponding to $p = 3$.

considering the News and Wikipedia domains, similar to Fig. 3, we show the new isolines for the average value of m computed according to Eq. (2) with $p = 2$ (upper row) and 3 (lower row). The differences observed between the plots in Fig. 5 suggest that the m-function can be used as an effective stylometric indicator [15], as well as a complementary indicator to those already used in the quantitative analysis of texts [16, 17].

3.2. LANGUAGE MODEL OUTPUTS

To show the actual usage of the previously discussed subcorpora we present here a few prompt/response instances obtained from a custom-built small language model (to be discussed elsewhere). For the scope of this article we mention that for each subcorpus we have trained a tokenizer considering vocabularies ranging from `8 000` tokens for IT to `50 000` for Law. For the actual model architecture we have considered the following parameters: Hidden Size equal to `384`, Number of Layers equal to `12`, and Attention Heads equal to `6`. Following training, the model is able to generate texts that respect all grammatical rules, though often its responses are semantically problematic, like Chomsky's famous *Colorless green ideas sleep furiously* example, more so when the prompts are single words. As the model is now being fine-tuned through different reinforcement learning strategies, we will not report here a detailed analysis of prompts and their associated responses and showcase just a few examples that illustrate the current accuracy of the model, usually limiting the responses to the first sentence.

Using the Wikisource trained model and the prompt *Ţăranii sunt* (in English: *Peasants are*) we get the answer *Ţăranii sunt, nu s-a schimbat nimic în noi* (in English: *Peasants are, nothing changed in us*) followed by grammatically correct sentences of similar semantic incoherence. For the News subcorpus, however, in the case of the prompt *Furtul a avut loc* (in English: *The theft took place*) the model provides a more logical output, namely *Furtul a avut loc sâmbătă seara, în jurul orei 20:30* (in English: *The theft took place Saturday evening around 20:30*) and continues with more details, which largely fit the context and are semantically coherent. In the case of Websites we mention here the prompt *Artistul este* (in English: *The artist is*) which is completed by the model as *Artistul este un film despre viaţa ta, iar această poveste te va transforma într-o comedie romantică, în care personajul tău este deja pierdut* (in English: *The artist is a movie about your life and this story will transform yourself into a romantic comedy in which your character is already lost*). For the language model trained on the Legal subcorpus the prompt *Constituţia este* (in English: *The Constitution is*) generates *Constituţia este organizată şi funcţionează pe bază de convenţie între părţi* (in English: *The Constitution is organized and functions based on an agreement between parties*) and continues with legal technicalities which are not entirely coherent semantically. The model trained on the Wikipedia subcorpus provides similar results and for the prompt *Românii sunt* (in English: *Romanians are*) it outputs *Românii sunt împărţiţi în funcţie de populaţia şi familiile acestora* (in

English: *Romanians are divided based on population and their families*), which is a followed by a text with good local (*i.e.*, four to six words) semantic coherence, but overall disconnected word sequences. Lastly, the model trained on the IT subcorpus resulted in a calibrated model that is able to generate grammatically correct texts, but the words rarely form meaningful sentences, which was to be expected given the small size of the subcorpus. For example, the prompt *Calculatorul este* (in English: *The computer is*) generates *Calculatorul este deja OS încât vom avea parte de un sistem de protecţie adiţional special conceput pentru Windows 10, folosind unelte specializate de genul ventilatoarelor grafice sau în sistemul de răcire* (in English: *The computer is already an OS so we will benefit from an aditional protection system especially designed for Windows 10 using specialized tools like graphical fans or in the cooling system*).

4. CONCLUSION

Using a series of automated data aggregation and text-processing tools, we have put together a multi-domain Romanian corpus of around 800 million tokens covering law at large, websites, news outlets, Wikipedia, Wikisources, and IT. We have reported a series of statistical analyses which allow for a unique description of each domain-specific subcorpus using both well-known indicators such as the word rank, word frequency, and word length, as well as an Ising-derived indicator, the so-called m-function, which allows us to quantify the deviation of a given word structure from a perfect vowel-consonant or consonant-vowel repeating pattern. Lastly, we showcase the usability of these six domain-specific subcorpora through a brief discussion tailored around a series of prompt/response instances using a custom-built small language model. The quality and the domain-specificity of the aforementioned subcorpora will allow for a more transparent calibration of language models at large, with preliminary results on the m-function suggesting we can investigate correlations between the texts used for calibration and those generated by the models. We expect to employ these results in cybersecurity applications, particularly the deviations from Zipf's law, and we foresee further use cases in the governance of generative AI tools involving natural language processing.

Acknowledgements. This work was supported by the Romanian Ministry of Education and Research under Romanian National Core Program LAPLAS VII – contract no. 30N/2023. The numerical simulations reported here were performed in the computing center of the Institute of Space Science – Subsidiary of INFLPR, Măgurele, România. The authors acknowledge fruitful discussions with T. Ivănoaica and V. Gheluţu.

REFERENCES

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems*, 5998–6008 (2017).
2. V. Barbu Mititelu, D. Tufiş, R. Ion, R. Simionescu, and E. Irimia, The Romanian Reference Treebank in Universal Dependencies Format, *Proceedings of LREC 2016* (2016).
3. P. Ortiz Suárez, B. Sagot, and L. Romary, OSCAR – A Large-Scale Open Multilingual Corpus, *Proceedings of the 1st Workshop on Building and Using Comparable Corpora* (2019).
4. S. D. Dumitrescu, A. M. Avram, D.-C. Cercel, F. Pop, LiRo: Benchmark and Leaderboard for Romanian Language Tasks, *Romanian Journal of Information Science and Technology* **25**, 203–219 (2021).
5. Wikipedia, *Romanian Wikipedia*, Wikimedia Foundation, `https://ro.wikipedia.org`
6. Wikisource, *Romanian Wikisource*, Wikimedia Foundation, `https://ro.wikisource.org`
7. Giuseppe Attardi, *WikiExtractor*, 2015. GitHub repository. Available at: `https://github.com/attardi/wikiextractor`
8. DEX online, *DEX Online Romanian Dictionary Database*, `https://github.com/dexonline/dexonline`
9. S. Piantadosi, Zipf's word frequency law in natural language: A critical review and future directions, *Psychonomic Bulletin and Review* **21**, 1112–1130 (2014).
10. A. Cocioceanu, C.M. Raportaru, A.I. Nicolin, and D. Jakamovski, Rank-frequency distributions in Romanian words, *AIP Conference Proceedings* **1916**, 040006 (2017).
11. Isabel Moreno-Sánchez, F. Font-Clos, and Á. Corral, Large-scale analysis of Zipf's law in English texts, *PLoS ONE* **11**, e0147073 (2016).
12. Á. Corral and I. Serra, The brevity law as a scaling law, and a possible origin of Zipf's law for word frequencies, *Entropy* **22**, 224 (2020).
13. A. Cocioceanu, T. Ivanoaica, A.I. Nicolin, and M.C. Raportaru, Computer-based statistical description of phonetical balance for Romanian utterances, *Advances in Intelligent Systems and Computing* **665**, 59–67 (2017).
14. A.N. Cocioceanu, S. Micluţa-Câmpeanu, M.C. Raportaru, E.G. Spanakis, and A.I. Nicolin, Statistical description of phonetic balance for Greek utterances, *AIP Conference Proceedings* **2071**, 040016 (2019).
15. J. Savoy, *Machine learning methods for stylometry. Authorship attribution and author profiling*, Springer, Cham, 2020.
16. I.-I. Popescu *et al.*, *Word frequency studies*, Mouton de Gruyter, Berlin – New York, 2009.
17. I.-I. Popescu *et al.*, *Quantitative analysis in poetic texts*, Mouton de Gruyter, Berlin/Boston, 2015.