Challenges in Creating a Representative Corpus of Romanian Micro-Blogging Text

Vasile Păiș, Maria Mitrofan, Elena Irimia, Verginica Barbu Mititelu, Roxana Micu, Carol Luca Gasan
Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy
Bucharest, Romania
{vasile,maria,elena,vergi}@racai.ro

Abstract

Short text messages used in micro-blogging platforms have specific characteristics making them difficult to process with existing natural language tools, trained on regular text. This paper presents challenges encountered while creating a representative corpus of Romanian micro-blogging text; in this phase we focus on Twitter messages. Once completed, this would become an extension of the Representative Corpus of the Contemporary Romanian Language (CoRoLa) and will be made available to the research community using similar interfaces.

Keywords: large corpus, micro-blogging text, natural language processing, Romanian language

1. Introduction

Following the successful creation of a national representative corpus of the contemporary Romanian language (Tufiș et al., 2019), we turned our attention to the social media texts, as present in micro-blogging platforms. These platforms are characterized by brevity, thus the high number of contractions, abbreviations and emoticons to convey messages. They are also informal manifestation of communication, sometimes even colloquial. Using snippets of text in a foreign language is sometimes a way of making the message shorter and faster to deliver, but such strings also carry pragmatic information (Vogh, 2022). Code-switching can occur frequently in such messages, making them difficult to process and even to detect automatically the languages employed by the user (Das and Gambäck, 2013). Furthermore, specific features such as hashtags, user references and links are also present. All these characteristics of the language used in micro-blogging make models trained on regular texts to be less effective on micro-blogging texts. This has lead to the creation of specialized language models, such as the contextual model BERTweet (Nguyen et al., 2020) or the Spanish COVID-19 Twitter embeddings (Miranda-Escalada et al., 2021a; Miranda-Escalada et al., 2021b). By using such dedicated contextual models, it becomes possible to outperform other general models on downstream tasks applied to micro-blogging text. However, such dedicated models are not available for all languages, and more specifically they are not available for the Romanian language.

Twitter is one of the most popular micro-blogging platforms. In recent years it has been used for studying the propagation of different news, including COVID-19 information (Lopez and Gallemore, 2021; Larson, 2020). It offers a high-level API, allowing searching for tweets based on different criteria, including specific queries and language. We started the creation of a Romanian micro-blogging corpus by employing this API to gather a large collection of text[1]. However, in spite of the easiness to build the raw text collection, we are faced with different issues regarding corpus annotation and management. Even though additional platforms will be considered for inclusion in a later stage (such as Reddit, Tumblr or Gab), we consider that tackling the problems related to Twitter messages will be a relevant step for all the micro-blogging content. Hence, currently, we are focusing on the Twitter platform.

Our main goal is to make the micro-blogging corpus part of the representative corpus of the Romanian language, allowing it to be exploited in the same way. This means a similar processing pipeline must be applicable. Finally, the resulting data needs to be indexed consistently. The national corpus is indexed by the KorAP Corpus Analysis Platform (Bański et al., 2012) and is queried by users familiar with its query languages and web interface. For centralization purposes and for facilitating the user experience, it would make sense to use the same platform for the new micro-blogging corpus.

In this paper, we present the current activities as well as the challenges faced when trying to apply existing tools (for both annotation and indexing) to a Romanian language micro-blogging corpus. These challenges are encountered at all annotation levels, including tokenization, and at the indexing stage. We consider that existing tools for Romanian language processing must be adapted to recognize features such as emoticons, emojis, hashtags, unusual abbreviations, elongated words (commonly used for emphasis in micro-blogging), multiple words joined together (within or outside hashtags), and code-mixed text: see the adaptations to social media of processing tools such as Stan-

[1] The gathering process is still in progress

Proceedings of the CMLC-10 Workshop (@LREC 2022), pages 1–7
Marseille, 20 June 2022

ford part of speech tagger (Derczynski et al., 2013), OpenNLP (Ritter et al., 2011), or GATE (Bontcheva et al., 2013). We analyse these features with emphasis on the Romanian language.

The paper is organized as follows: Section 2 presents related work, Section 3 describes the corpus collection process, Section 4 provides challenges related to corpus indexing, Section 5 introduces a manually annotated sub-corpus, and finally conclusions are given in Section 6.

2. Related work

Among different types of corpora, a new one has emerged in the last decades: computer-mediated communication (CMC) corpus. It includes collections of blog posts, forums posts, comments on news websites, social media, mobile phone applications, e-mails and chat rooms exchanges. Corpora of texts from social media platforms of the type micro-blogging have been collected for various languages: German and Danish (Bick, 2020), English (Sharma et al., 2020), Turkish (Çöltekin, 2020), Chinese (Wang et al., 2012), Romanian (Manolescu and Çöltekin, 2021), Arabic (Zaatari et al., 2016), Italian (Sanguinetti et al., 2018), French (Mazoyer et al., 2020) and others[2]

The interest in working with texts collected from such sources manifest in connection to various tasks, such as sentiment analysis applications development (Sharma et al., 2020; Cieliebak et al., 2017), the need to improve NLP tasks such as word segmentation (Wang et al., 2012), annotation of emotions (Roberts et al., 2012), credibility analysis (Zaatari et al., 2016), event detection (Mazoyer et al., 2020), linguistic phenomena manifested on micro-blogging platforms (Coats, 2019) and others. However, detection of hate speech is the interest preoccupying most of those focusing their research on micro-blogging platforms (Bick, 2020; Çöltekin, 2020; Manolescu and Çöltekin, 2021; Sanguinetti et al., 2018).

Developers and maintainers of large, usually national corpora have manifested interest in reflecting the language from social media sites, including micro-blogging platforms, in their data (Kren, 2020).

3. Corpus collection

For the purposes of gathering the Twitter corpus, we constructed a crawler employing the Twitter API for Academic Research. Since we are interested both in Romanian-only tweets and in code-mixed texts (employing at least a few Romanian words), the crawler can use either the Twitter language detection or queries based on lists of expressions constructed by hand. The queries are periodically executed retrieving newly posted messages. Furthermore, even though it currently integrates only the Twitter API, the crawler is built in a modular way, allowing the use of other APIs in the future.

Listing 1: Example retweeted message

```
RT @AnonymousUser1: A long retweeted
message that gets trunc...
```

Messages are retrieved in the API specific JSON format. Following the retrieval, a second process transforms the messages into text documents. At this step a filtering operation is applied in order to remove duplicated messages (employing the message identifier) and to apply a primary anonymization function by removing usernames and URLs. We further remove messages that contain less than 3 words.

Specific to social networks is the sharing of messages with a user's friends or followers. In Twitter this mechanism is called retweeting. The same message is redistributed by another user with only small changes: possibly adding "RT" in front of the message, and sometimes adding the user that initially posted the message. In case of long messages the retweeted message could get truncated to obey the API size restrictions. An example of a possible retweet associated with the message "A long retweeted message that gets truncated." is given in Listing 1. It is worth noting that technically there is no rule about the way a retweeted message should look like. The actual format is dependent on the application used to generate the message. Some retweets do not start with "RT", do not contain a user being mentioned or even contain a list of users.

From a linguistic perspective, the presence of retweets does not provide any useful information. Truncation of messages further complicates their processing. Therefore, in the final version of the corpus, such messages will be removed to avoid unnecessary text duplication. Preliminary statistics on the collected data indicate a number of 759,719 raw JSON files. After applying the process of removing tweets with less than 3 words and converting to text, we are currently left with 741,940 text files. These files will need to undergo a final operation of removing retweets. The already removed files contain mostly user mentions, URLs, emojis or emoticons. However, a closer look at the removed files show the presence of messages such as "Felicitări, @AnonymousUser! <url>" ("Congratulations, @AnonymousUser! <url>"). Even though such messages may be deemed uninteresting, it is still debatable whether or not to keep a small number of files for indexing or for training language models.

4. Corpus indexing

Krill[3], the search module in KorAP, indexes and provides search opportunities on textual data (the Twitter content in our case), various layers of annotation data

[2] Some CMC corpora are available for browsing or download in a CLARIN repository at https://www.clarin.eu/resource-families/cmc-corpora.

[3] https://github.com/KorAP/Krill

and the documents metadata (Diewald and Margaretha, 2016). Micro-blogging posts have specific characteristics in terms of metadata, that are not in the lines of the metadata used to describe and index CoRoLa: for example, instead of an author (as documents in CoRoLa and other KorAP indexed corpora have), a tweet has a username associated to it. However, due to anonymization requirements, this username may not be used in a publicly available interface. Furthermore, to reduce de-anonymization attacks it is not feasible to replace a username with the same identifier in multiple instances. Document classification metadata fields usually used to index corpora may be difficult to provide: the domain for each post is not easy to identify, even if the corpus gathering process is based on a curated term list, while a literary genre specific to social media is yet to be theorised. A Twitter post has no title, publisher or other regular metadata fields. Nevertheless, other characteristics may be present, such as if the message is part of a conversation or if it is a retweet.

For indexing the corpus in KorAP Corpus Analysis Platform, a conversion chain has to be executed to convert the local data and metadata files first to the I5 format (Lüngen and Sperberg-McQueen, 2012) (which is a TEI customization used in the German Reference Corpus DeReKo (Kupietz et al., 2010)), then to a proprietary KorAP-XML [4] format and finally to a format compatible with the Krill indexer. At the moment, there is a simple solution to deal with the Twitter metadata that has already been used for Twitter-Sample Corpus in DeReKo: I5 metadata format was extended to support external links with arbitrary titles to reference the Twitter posts, since the title field is a mandatory field in KorAP indexing process. For dealing with metadata information about retweets, replies, hashtags and other Twitter specific metadata information, a special class could be written in the future in korapxml2krilll[5]. The KorAP platform distributes data annotation on different layers, with a base layer containing the form of the word and subsequent layers dealing with e.g. lemma information, morpho-syntactic information (POS tagging), syntactic information, etc. The Twitter corpus comes with a supplementary layer for the specific named entity (NE) annotation, which will require further adaptations of the indexing process.

For releasing the corpus, we need to provide sufficient anonymization, as demanded by different regulations, such as the GDPR, and also comply with Twitter's requirements. We examine the suitability of existing anonymization solutions for the Romanian language (Păiș et al., 2021a), and find they also need to be made aware of micro-blogging specific features, such as user specification and people names appearing in hashtags or in other unusual formats (lowercase letters, elongated names, first name and last name joined together without spaces).

A micro-blogging corpus comes with the additional challenge of being composed of a large number of files. Each file contains only a small number of sentences (usually one sentence). This may impose additional restrictions on the storage sub-system, for both processing and querying, requiring the ability to handle such a large number of files. However, it also offers an opportunity to exploit parallel processing pipelines, where the text can be distributed across a large number of processes, hosted on multiple servers. Prior to indexing it, the corpus must be tokenized and enhanced with token-level annotations. For this purpose, the available parallelization features in our RELATE platform (Păiș et al., 2020) are exploited in order to process a large volume of text in a manageable amount of time.

5. Manually annotated sub-corpus

In order to properly evaluate existing Romanian text processing pipelines and potentially train new ones specific to micro-blogging text, a small sub-corpus will be manually annotated. In a first phase, this annotation process will include named entity identification and classification of code-mixed messages (identifying also messages mostly written in foreign languages). Named entities will be marked at text span level, thus enabling us to check existing tokenization tools. As previously mentioned, we expect to encounter issues with named entities embedded in hashtags or other specific text structures.

In order to annotate the corpus with named entities, nine classes of entities were chosen. These classes will allow for evaluating recently created Romanian language NER systems (Păiş et al., 2021; Păiș, 2019; Mitrofan and Păiş, 2022; Mitrofan, 2019) and will account for the social messaging activities in the context of the COVID-19 pandemic. Each class of entities is briefly described below:

- Organization (ORG) entities are limited to corporations, agencies, and other groups of people defined by an established organizational structure. The annotation process will mark text spans clearly indicating the name of an organization. Examples: *Facebook*, *Guvernul* ("*the government*"), *PSD*, *#ConsConRo*.

- Person (PER) entities are regularly limited to humans. A person may be a single individual or a group. By extension, the same label is attached to fictional characters or references to religious figures. Examples: *Adela*, *Moș Crăciun* ("*Santa Claus*"), *Niculina Stoican*.

- Location (LOC) entities are limited to geographical entities such as geographical areas and landmasses, bodies of water, and geological formations, denoted by a proper name. The annotation process will identify the name associated with a

[4] https://github.com/KorAP/KorAP-XML-Krill#about-korap-xml

[5] https://github.com/KorAP/KorAP-XML-Krill

location entity, without additional words, unless these words are part of the official entity name. Examples: *România*, *Parcul Tineretului*, *Lacul Sfânta Ana* ("*Lake Saint Ana*").

- Time (TIME) expressions tell us when something happened, how long something lasted, or how often something occurs. Sometimes the precise date cannot be determined, allowing for expressions indicating periods of time. Examples: *astăzi* ("*today*"), *15 septembrie* ("*September 15*"), *Crăciun* ("Christmas").

- Legal references (LEGAL) are designations (the title of a legal document) or expressions pointing to another legal document. Examples: *legea 13/2021* ("*law 13/2021*"), *constituția* ("*the Constitution*").

- Anatomical parts (ANAT) class contains mentions of anatomical parts, parts of the human body, organs, components of organs, tissues, cells, cellular components. Examples: *cap* ("*head*"), *mâini* ("*hands*"), *ficat* ("*liver*").

- Chemical and drugs (CHEM) class contains mentions of amino acids, peptides, proteins, antibiotics, active substances, drugs, enzymes, hormones, receptors. Examples: *sodiu* ("*sodium*"), *vaccin* ("*vaccine*").

- Disorders (DISO) class contains mentions of anatomical abnormalities, congenital anomalies, diseases, syndromes, lesions, symptoms. Examples: *diabet* ("*diabetes*"), *COVID*.

- Medical devices (MED_DEVICE) class contains mentions of any device intended to be used for medical purposes. Examples: *stetoscop* ("*stethoscope*").

The annotators followed specific guidelines, inspired in part by the Linguistic Data Consortium (LDC) guidelines [6] for annotation of named entities. More specifically, regarding the annotations with the ORG, PER, LOC and LEGAL classes, the guidelines presented in both previous works (Păiș, 2019; Păiş et al., 2021) were followed. In order to annotate the corpus with named entities specific to the medical domain (CHEM, DISO, MED_DEVICE), the annotators followed the specific guidelines described in (Mitrofan, 2017; Mitrofan et al., 2019). However, these guidelines had to be adapted to include elements specific to micro-blogging texts, such as NEs present in hashtags, unusual abbreviations or spelling, and words linked together.

Similar to other NE gold corpora creation activities, we had to clearly define each type of entity. During the annotation process some issues were identified and required further clarifications. Nevertheless, since we also wanted to be able to use the newly annotated corpus to evaluate and adapt existing tools to the social-media domain, we were constrained by already existing annotation guidelines, such as the one[7] used for annotating the LegalNERo corpus (Păiș et al., 2021b). Some interesting NE annotation instances that we encountered and needed to be deliberated were:

- metonymies of the type places for organizations are also annotated as LOC: in "Thailand will organize the voting process" the word "Thailand" is annotated as LOC, though it refers to the government of the country;

- imbricated entities are not annotated: only the wider string is annotated: e.g., in the string *primăria din Tecuci* one could identify two entities: the LOC *Tecuci* and the ORG *primăria din Tecuci*; however, only the latter is annotated. An exception is made for the LEGAL entity class which has sub-entities annotated (this is due to the LegalNERo guidelines);

- only sequences that unambiguously identify a named entity are annotated: e.g., *un frate al lui Mbape* "one of Mbape's brothers" may refer to any of Mbape's brothers, thus not being annotated, while *Mbape* is a clearly identified person. However "podul peste Dunăre de la Brăila" ("Brăila bridge over the Danube") is a location entity since it is clearly defined, even though it lacks an actual name.

Classification of the tweet files is done according to 4 different axes, which will be encoded in the corpus as attributes at metadata level:

- Language = *Romanian* or *Mixed RO+English* or *Mixed RO+Other* or *Other*. For this attribute, we based our classification on the distinction between linguistic borrowing and code-switching phenomena: the borrowing occurs at lexical level - mostly when the concept to be expressed is not lexicalised in the spoken language or when the speaker has a momentary lapse - and it involves using a single (simple or compound) word from another language; the code-switching occurs at the syntactic level - for pragmatic reasons like communicating emotions or the need to be understood only by some listeners and not others - and it involves the alternative use of (most often) two languages by combining longer sequences of words.

- Sentiment = *Neutral* or *Positive* or *Negative*. The Sentiment is Positive or Negative if it is directly

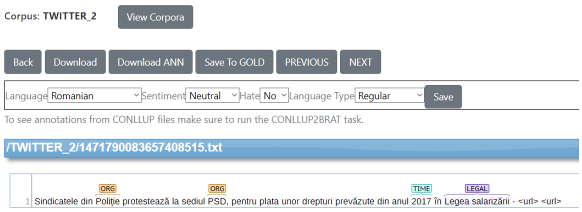
[6] `https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-edt-v4.2.6.pdf`

[7] `https://relate.racai.ro/resources/legalnero/legalnero_annotation_guide.pdf`

expressed by the tweet author but the text is classified as Neutral if it speaks in objective/journalistic manner about an unhappy event.

- Hate = *No* or *Yes*. This attribute encodes the presence in the tweet text of hate speech elements, expressed by harmful and offensive statements against specific categories of persons or even certain persons.
- Language Type = *Regular* or *Social Media Slang*. This attribute is meant to spot messages using micro-blogging specific language (the so-called social media slang), clearly different from regular text (for example ”LOL!!! :) :D”).

Both annotation and classification are handled within the RELATE[8] platform (Păiș et al., 2020; Păiș, 2020). For NER annotations, we defined a custom profile for the integrated BRAT[9] (Stenetorp et al., 2012) component. Classification is handled through a custom component available in the RELATE platform. The annotator’s interface is shown in Figure 1.



Corpus: TWITTER_2

Language Romanian Sentiment Neutral Hate No Language Type Regular

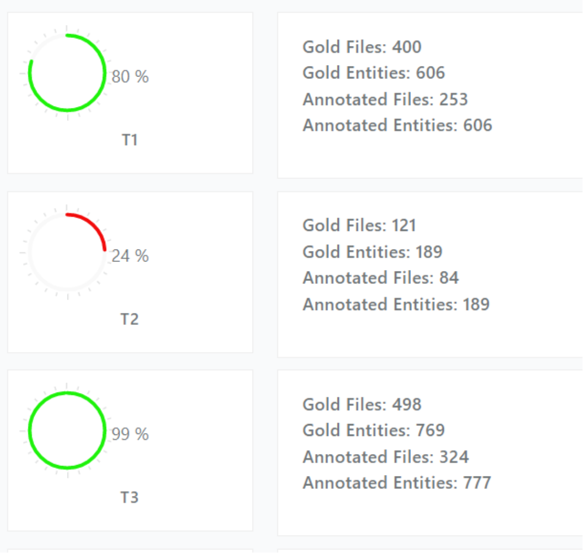
To see annotations from CONLLUP files make sure to run the CONLLUP2BRAT task.

/TWITTER_2/1471790083657408515.txt

1 Sindicatele din Poliție protestează la sediul PSD, pentru plata unor drepturi prevăzute din anul 2017 în Legea salarizării - <url> <url>

Figure 1: Annotation interface within the RELATE platform.

The tweets were split into multiple batches (with 500 messages in each batch) which were distributed amongst annotators (currently 7 annotators are involved). A number of files are common between at least three annotators, which will allow us to compute inter-annotator agreement metrics at the end of the annotation process. Periodic meetings are being held in order to identify and document potential issues early in the process, discuss and decide upon the right solutions. In order to encourage a certain level of competition between annotators, a simple dashboard was developed within the RELATE platform. This presents basic information, such as the number of files each annotator worked on and a graphical display (with changing colors) indicating the remaining work to be done. A snapshot of the dashboard is given in Figure 2.
The current annotation effort aims at annotating 21 batches of 500 messages. Each batch is split into 300 files unique to the batch and 200 files found in two other batches for agreement calculation. This leads to a number of 7,800 total distinct messages. Computing the average number of entities annotated in 7 batches, we notice the presence of 733 NEs in each batch (corresponding to 1.47 NEs in each file). Therefore we estimate a final value of approximately 11,466 total entities.



Gold Files: 400
Gold Entities: 606
Annotated Files: 253
Annotated Entities: 606

Gold Files: 121
Gold Entities: 189
Annotated Files: 84
Annotated Entities: 189

Gold Files: 498
Gold Entities: 769
Annotated Files: 324
Annotated Entities: 777

Figure 2: Dashboard for the annotation process.

6. Conclusion

This paper introduced the first steps taken towards extending the representative corpus of the contemporary Romanian language (CoRoLa) with a micro-blogging corpus. At this stage we focused only on Twitter, while developing the mechanisms which will allow us to extend the endeavour to other social media platforms as well. We presented the challenges encountered while working on this new Romanian corpus and we are actively working on solving the remaining issues. Furthermore, we are currently creating a manually annotated gold sub-corpus which will allow us to evaluate existing tools for micro-blogging text and train dedicated models. In turn, this will allow us to extend existing Romanian anonymization tools (Păiș et al., 2021a) to properly anonymize micro-blogging text.
We aim to make the final corpus available through the same indexing platform (KorAP) used for CoRoLa, thus enabling existing users to take advantage of the new resource in a similar way. Properly anonymized sub-corpora, such as the manually annotated gold corpus introduced in this paper, will also be made available for download in different formats, enabling other researchers to train and evaluate their own language models.

7. Bibliographical References

Bański, P., Fischer, P. M., Frick, E., Ketzan, E., Kupietz, M., Schnober, C., Schonefeld, O., and Witt, A. (2012). The new IDS corpus analysis platform:

[8] https://relate.racai.ro
[9] https://brat.nlplab.org/

Challenges and prospects. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2905–2911, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Bick, E. (2020). An annotated social media corpus for German. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6127–6135, Marseille, France, May. European Language Resources Association.

Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. A., Maynard, D., and Aswani, N. (2013). Twitie: An open-source information extraction pipeline for microblog text. In *Proceedings of Recent Advances in Natural Language Processing*, pages 83–90, Hissar, Bulgaria.

Cieliebak, M., Deriu, J. M., Egger, D., and Uzdilli, F. (2017). A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain, April. Association for Computational Linguistics.

Coats, S. (2019). Lexicon geupdated: New German anglicisms in a social media corpus. *European Journal of Applied Linguistics*, 7(2):255–280.

Çöltekin, Ç. (2020). A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France, May. European Language Resources Association.

Das, A. and Gambäck, B. (2013). Code-mixing in social media text. the last language identification frontier? *Trait. Autom. des Langues*, 54:41–64.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of Recent Advances in Natural Language Processing*, page 198–206, Hissar, Bulgaria.

Diewald, N. and Margaretha, E. (2016). Krill: Korap search and analysis engine. *Journal for language technology and computational linguistics (JLCL)*, 31(1):73–90.

Kren, M. (2020). Czech national corpus in 2020: Recent developments and future outlook. In *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*, pages 52–57, Marseille, France, May. European Language Ressources Association.

Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). The German reference corpus DeReKo: A primordial sample for linguistic research. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Larson, H. J. (2020). A call to arms: helping family, friends and communities navigate the covid-19 infodemic. *Nature Reviews Immunology*, 20(8):449–450, Aug.

Lopez, C. E. and Gallemore, C. (2021). An augmented multilingual Twitter dataset for studying the COVID-19 infodemic. *Social Network Analysis and Mining*, 11(1):102, Oct.

Lüngen, H. and Sperberg-McQueen, C. M. (2012). A TEI P5 document grammar for the IDS text model. *Journal of the Text Encoding Initiative*, (3).

Manolescu, M. and Çöltekin, Ç. (2021). ROFF - a Romanian Twitter dataset for offensive language. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 895–900, Held Online, September. INCOMA Ltd.

Mazoyer, B., Cagé, J., Hervé, N., and Hudelot, C. (2020). A French corpus for event detection on Twitter. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6220–6227, Marseille, France, May. European Language Resources Association.

Miranda-Escalada, A., Aguero, M., and Krallinger, M. (2021a). Spanish COVID-19 Twitter embeddings in FastText, January. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Miranda-Escalada, A., Farré-Maduell, E., Lima-López, S., Gascó, L., Briva-Iglesias, V., Agüero-Torales, M., and Krallinger, M. (2021b). The ProfNER shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 13–20.

Mitrofan, M. and Păiş, V. (2022). Improving Romanian BioNER using a biologically inspired system. In *Proceedings of the 21st BioNLP workshop (paper accepted)*. Association for Computational Linguistics.

Mitrofan, M., Mititelu, V. B., and Mitrofan, G. (2019). Monero: a biomedical gold standard corpus for the romanian language. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 71–79.

Mitrofan, M. (2017). Bootstrapping a Romanian corpus for medical named entity recognition. In *RANLP*, pages 501–509.

Mitrofan, M. (2019). *Extragere de cunoștințe din texte în limba română și date structurate cu aplicații în domeniul medical*. Ph.D. thesis, Romanian Academy.

Nguyen, D. Q., Vu, T., and Tuan Nguyen, A. (2020). BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online, October. Association for Computational Linguistics.

Păiş, V., Mitrofan, M., Gasan, C. L., Coneschi, V., and Ianov, A. (2021). Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Păiș, V., Ion, R., and Tufiș, D. (2020). A processing platform relating data and tools for Romanian language. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 81–88, Marseille, France. European Language Resources Association.

Păiș, V., Irimia, E., Ion, R., Tufiș, D., Mitrofan, M., Barbu Mititelu, V., Avram, A.-M., and Curea, E. (2021a). Romanian text anonymization experiments from the CURLICAT project. In *The 16th International Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 165–178.

Păiș, V., Mitrofan, M., Gasan, C. L., Ianov, A., Ghiță, C., Coneschi, V. S., and Onuț, A. (2021b). Romanian Named Entity Recognition in the Legal domain (LegalNERo), May.

Păiș, V. (2019). *Contributions to semantic processing of texts; Identification of entities and relations between textual units; Case study on Romanian language*. Ph.D. thesis, School of Advanced Studies of the Romanian Academy (SCOSAAR), Bucharest, Romania, November.

Păiș, V. (2020). Multiple annotation pipelines inside the RELATE platform. In *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 65–75.

Ritter, A., Clark, S., and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK.

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., and Harabagiu, S. M. (2012). EmpaTweet: Annotating and detecting emotions on Twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3806–3813, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., and Stranisci, M. (2018). An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Sharma, R., Verma, A., Grover, R., Pandey, D., Pandey, B., and A, L. (2020). Microbloging as a corpus for sentiment analysis structure and feeling mining. *Journal of Xi'an Shiyou University, Natural Science Edition*, pages 229–234, 11.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April. Association for Computational Linguistics.

Tufiș, D., Barbu Mititelu, V., Irimia, E., Păiș, V., Ion, R., Diewald, N., Mitrofan, M., and Mihaela, O. (2019). Little strokes fell great oaks. creating CoRoLa, the reference corpus of contemporary romanian. *Revue Roumaine de Linguistique*, 64(3):227–240.

Vogh, K. (2022). Code-mixing and semantico-pragmatic. In *Points of Convergence in Romance Linguistics: Papers selected from the 48th Linguistic Symposium on Romance Languages (LSRL 48), Toronto, 25-28 April 2018*, volume 360, page 243. John Benjamins Publishing Company.

Wang, L., Wong, D. F., Chao, L. S., and Xing, J. (2012). CRFs-based Chinese word segmentation for micro-blog with small-scale data. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 51–57, Tianjin, China, December. Association for Computational Linguistics.

Zaatari, A. A., Ballouli, R. E., ELbassouni, S., El-Hajj, W., Hajj, H., Shaban, K., Habash, N., and Yahya, E. (2016). Arabic corpora for credibility analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4396–4401, Portorož, Slovenia, May. European Language Resources Association (ELRA).