MARCELL – A project to remember: hard work of a friendly consortium under wise coordination

Dan Tufiș[1], Vasile Păiș[1], Verginica Barbu Mititelu[1], Radu Ion[1], Elena Irimia[1], Andrei Avram[1], Eric Curea[1]

[1] Institutul de Cercetări pentru Inteligență Artificială "Mihai Drăgănescu"
{tufis, vasile, vergi, radu, elena, eric}@racai.ro
avram.andreimarius@gmail.com

1. Collection and Annotation of the Romanian Legal Corpus

In this section we review the results of the Romanian team in the first part of the MARCELL project (https://marcell-project.eu/) whose ultimate goal was to enable the enhancement of automatic translation in CEF.AT[1] on the body of national legislation in seven EU official languages. For this task, all the seven teams from Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia, and Slovenia cooperated in order to produce a comparable corpus heavily annotated (part-of-speech, syntactically parsed and semantically labelled by EUROVOC[2] identifiers and IATE recognized terms[3] appropriately marked-up).

EuroVoc is a multilingual thesaurus which was originally built up specifically for processing the documentary information of EU institutions. The covered fields encompass both European Union and national points of view, with a certain emphasis on parliamentary activities. The current release of EuroVoc (4.4) was published in December 2012. The new edition was the result of a thorough revision, among others, according to the concepts introduced by the Lisbon Treaty. It includes 6,883 unique IDs for thesaurus concepts (corresponding to the preferred terms), classified into 21 domains (top-level domains), further refined into 127 subdomains. Preferred terms and different variations of the preferred terms are assigned the same ID, subdomains and top-level domains.

IATE ('**I**nter**a**ctive **T**erminology for **E**urope') is the EU's terminology database. It has been used in the EU institutions and agencies since summer 2004 for the collection, dissemination and management of EU-spe-

[1] https://ec.europa.eu/inea/sites/inea/
[2] https://eur-lex.europa.eu/browse/eurovoc.html
[3] https://iate.europa.eu/home

DOI: https://doi.org/10.18135/VT70.16

cific terminology. It contains over 8 million terms in 24 official languages of the EU. The IATE database contains about 55,000 terms in Romanian.

The language-specific corpora were cross-lingually aligned at the top-level domains identified by EUROVOC descriptors. A general view of the project activities is given in another article (Váradi et al., 2020). As for the Romanian language, the current legal database includes more than 144k processed legislative documents. There are five main types of Romanian legal documents: governmental decisions (25%), ministerial orders (18%), decisions (16%), decrees (16%) and laws (6%). After the statistics were calculated, we found that there were six main issuers of the documents: Government (28%), Ministers (19%), President (14%), Constitutional Court (12%), Parliament (6%) and National Authorities (4%). Concerning the timestamp, most of the published documents were issued after 2000. Almost 4,000 documents were issued before 1990, and around 21,000 legal documents were published between 1990 and 2000. Following 2000, the number of issued documents increased. On average, more than 6,000 documents were issued every year, reaching a total of 120,000 until 2018, in 19 years. In terms of document length, there are around 6,000 short documents (less than 100 words per document, most of them being updates to other previously published legal documents), 70,000 documents contain between 100 and 500 words per document, more than 18,000 documents have around 1000 words per document and 52,000 contain more than 1000 words.

2. Linguistic Annotation

The corpus is annotated in batches as new documents are collected. All partners produced processing flows that were dockerized and stored as "ready-to-use" on the RELATE portal (Păis et al., 2019). The processing flows include language specific text normalization, sentence splitting, tokenization, POS tagging, lemmatization, dependency parsing, named entity recognition and classification, chunking, IATE term annotation and top level EUROVOC labeling.

The Romanian preprocessing pipeline, excluding IATE and EUROVOC annotations, is performed using the TEPROLIN text preprocessing platform (Ion, 2018). TEPROLIN offers the user various choices for each processing step and can be easily configured to different specific algorithms. It only needs a list of desired text annotations to infer and construct the pipeline getting these annotations out. TEPROLIN includes

mostly tools developed by our institute (e.g. TTL, MLPLA, NER, BIOner, Diac, TextNorm), however, not exclusively: we incorporated some other open-source algorithms (such as UDpipe, NLP-cube, Korap) into the preprocessing platform and will continue to add new better algorithms as they become freely available.

Dependency parsing is produced by NLP-Cube (Boros et al., 2018) which, according to the evaluations done in the CoNLL 2018 shared task "Multilingual Parsing from Raw Text to Universal Dependencies", has a labelled attachment score of around 85% for Romanian.

3. Automatic Identification of Legal Terms in Romanian Law Texts

As specified in the Grant Agreement of the project, each language-specific corpus was enriched with IATE and EUROVOC labels, then classified and multilingually clustered based on these annotations.

For term identification in both IATE and EuroVoc, the Romanian team used an algorithm similar to the Aho-Corasick algorithm (Aho and Corasick, 1975), using a language specific calibrated compressing function largely described in (Coman et al., 2019). This method only implies linear-time transformations of the IATE dictionary (through the compression function) and a single pass through the Aho-Corasick structure, the overall complexity of the proposed algorithm is linear and has a term matching rate of approximately 98%. Besides the compression function, the Aho-Corasick structures were not language-specific and were created during runtime (for the Romanian IATE terms, consisting of about 55,000 terms, the computation time was approximately 10 seconds).

Thus, the algorithm would be available for any language, provided that a specific compression function relevant to that respective language was accessible.

The processing allowed the annotation of the corpus through the introduction of EuroVoc and IATE labels. Thus, every occurrence of the IATE term inside the corpus is now annotated by its respective position and is accompanied by the corresponding EuroVoc categories. In the analysed testing sample, we were able to detect no false positive matches and a nearly perfect precision for detecting true matches. The overall matching rate over the used testing samples was approximately 98%-99%, however, as mentioned earlier in the paper, we expect the real matching rate to have a slightly lower value due to unexpected collisions

which may have occurred during the term identification process. Moreover, the matching rate, as defined in the paper, has a rather simple definition and it presents the accuracy of our work only to some extent.

Secondly, by using the aforementioned annotation with EuroVoc categories, we were able to create a statistical database documenting the occurrence frequency of all the categories in each legal document. For each file, we determined the number of terms falling in each of the 21 EuroVoc categories. After multiplying each frequency with a predefined weight, the mentioned file was placed in the category corresponding to the maximum number of terms.

The predefined weights were roughly determined over a testing sample of medium size (approx. 100 documents), due to the lack of pre-processed legal data. As most of the EuroVoc categories present in the description of the IATE terms yielded correct classifications, we only needed to slightly modify some of them (such as Geography) to obtain a better classification over the testing sample. This simple classification method was replaced later with a more sophisticated one taking advantage of available word embeddings (see further).

The computation time for simply identifying the matches with the hybrid algorithm was approximately 4,250 seconds, which yields a rate of almost 35 documents per second. This experiment was performed using a server with two Xeon 4210 CPUs at 2.2 GHz with 20 annotation threads, yielding over 1 document per second considering a single annotation thread. Thus, as the XML-formatted corpus had a size of approximately 31.2 Gb, the processing rate was 7.5 Mb of text per second. Because of using the Aho-Corasick structure, the memory usage was also linear, which made the computation possible on almost any machine. The number of matches was significantly increased by working with the lemmatised corpus instead of the unlemmatised one. By using the described algorithm, we have identified a total of 51,517,877 matches (IATE terms), out of which 29,162,667 were short terms (single word terms) and 22,355,210 were long terms (multiple word terms). The term identification step is based on the encoded list of IATE terms, which brings our approach closer to a gazetteer-based processing. This is why the estimated precision is so high.

4. Document classification and evaluation

In order to obtain an objective evaluation of our document classification algorithm it was necessary to have reference data, classified and validated by human experts. These requirements were not met for the MARCELL Romanian corpus, but instead we resorted to a well known multilingual corpus, JRC-Acquis (Steinberger et al., 2006), which was processed by a publicly available program called JEX.

JEX (Steinberger et al., 2012) is a multi-label classification software developed by JRC, trained to assign EuroVoc descriptors to documents. Its primary concern was to cover the activities of the EU. Written using Java, it provides scripts for pre-processing a collection of documents, training a new model, post-processing the results and evaluating a new model. Each script employs a configuration file for the required parameters. The toolkit also comes with a graphical interface (GUI) for users to label new text, XML, HTML documents or to interface with training scripts for obtaining a classifier on their own documents. However, the usage of the GUI interface is optional and the toolkit allows for simple command line execution over collections of documents.

Based on (Pouliquen et al., 2003), JEX classification algorithm relies on a list of lemma frequencies obtained from normalized text together with associated weights, statistically related to each descriptor. These are called associates or topic signatures. At runtime, given the new document's list of lemma frequencies the algorithm picks the descriptors of the associates that are the most similar to it. The JEX package offers pre-trained classifiers for 22 official EU languages, including Romanian (trained on over 25,000 documents, consisting of manually annotated ACQUIS and OPOCE corpora). (Steinberger et al., 2012) reports a F1 score of 47.84% (derived from P=45.55% and R=50.43%, computed for predicting 6 EuroVoc identifiers).

More recently, researchers tried to further improve the perfomance of JEX regarding different languages. For instance, the Italian language, Boella et al. (2012), mono-label transformations (Tsoumakas and Katakis, 2007), and employing Support Vector Machines (SVM) (Joachims, 1998) for classification, achieves an F1 score of 58.32%. While applying the Croatian CroVoc (an extended EuroVoc terminology) on the NN13205[4] corpus (different from the JEX training and testing sets), Šarić et al. (2014)

[4] http://takelab.fer.hr/data/nn13205

reports an F1=68.60%. We are unaware of other studies regarding EuroVoc classification for Romanian, therefore JEX is the only available tool/algorithm.

5. Dataset used for the evaluation exercise, the word-embeddings and the method

After downloading and extracting the JEX extended package,[5] we are presented with the corpora on which it was trained, consisting of the ACQUIS and OPOCE corpora. JEX used a regular cross-validation approach for evaluation, consisting of creating multiple splits out of the training data and evaluating each one followed by averaging the results. However, the individual splits are not provided, therefore we had to create our own splits. Before doing this, we first annotated the corpora using our RELATE platform with a pipeline similar to the one used for the Marcell project. Furthermore, the platform was applied to extract statistics on the corpora. Finally, a script took care of creating 10 split folders, each consisting of 80% training data and 20% test data with gold annotations from the original corpora. Furthermore, the 80-20 split rule was applied on both corpora, thus producing balanced splits.

Word representations learned using artificial neural network approaches (Mikolov et al., 2013) have previously been used successfully in a number of natural language processing tasks, including classification (Joulin et al., 2017). However, this had not been applied to EuroVoc classification. Facebook Research introduced the FastText[6] tool initially intended for training neural embeddings together with sub-word information (Bojanowski et al., 2017). Using this tool, we previously created and evaluated word representations on the Reference corpus for Romanian language CoRoLa (Barbu Mititelu et al., 2018). These results were reported by Păiș and Tufiș (2018) and can be freely downloaded from the website of our Institute.[7] An advantage of word embeddings representation is that once trained and evaluated, these representations can be used directly for converting words into numeric (floating point) vectors, suitable as input to other algorithms. This ensures a starting point given by the accuracy of the word representation and reduces the time needed for training more advanced algorithms.

[5] https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer#Download%20JEX
[6] https://fasttext.cc/
[7] http://corolaws.racai.ro/word_embeddings/

The FastText tool was further enhanced, enabling training a linear classifier based on word embeddings and encoding of input documents. Therefore it seemed like an obvious choice for using the previously generated representations to try and classify texts using the EuroVoc terminology. The tool allows for adapting the model parameters to a specific language by considering the minimum and maximum lengths of character and word n-grams. Additionally, other parameters such as learning rate can be further fine-tuned.

For each of the previously created splits, a Romanian language classifier was trained. Then it was evaluated on each of the test corpora and the results were finally averaged to produce the final data (similar to the JEX evaluation approach). This allowed us to obtain an average F1=53.53% (compared to the JEX reported F1 of 47.84%, this gives us an increase of 5.7%). Similarly, we noted increased performance for both precision (50.93% our result compared to 45.17% from JEX) and recall (56.41% our result compared to 50.19 from JEX).

For the purposes of the Marcell project, we further converted the EuroVoc identifiers into MT labels and finally top-level domains. This is possible given that the mapping is present in the EuroVoc. There is a direct mapping from an identifier to an MT label and further to a top-level domain, represented by the first two digits of the MT label. Reverse mapping is not possible directly, since multiple identifiers are associated to a MT label.

In the context of the Marcell project, documents are classified using only EuroVoc top-level domains. In order to give an estimate of our classifier at this level we converted both the gold corpora annotations and the classifier automatic annotations to top-level annotations (considering both our approach and the JEX approach). We then applied again the scoring algorithm on all the splits and computed a final average over all the splits. This produced a top-level F1 score of 70.80% for our approach, compared to a score of 64.88% for JEX, thus providing an increase of almost 6% (comparable, yet slightly better than the identifier based evaluation). Similarly, we noted increases in both precision (64.90% our method vs 59.34% JEX) and recall (77.89% our method vs 71.56% JEX).

After performing the evaluations, a classifier was trained on the entire training corpora, providing a model which should have similar performances to the reported averaged ones (however in this case no further evaluation can be performed since there is no additional data to compare against). This final model was used to classify the Romanian Marcell

legislative corpus. This produced a rather unbalanced distribution of top-level domains with more than 80,000 documents assigned to the domains geography (72) and European Union (10). At the other end, less than 1,000 documents were assigned to the domains international organisations (76), science (36) and industry (68).

Apart from the EuroVoc classification, the Marcell corpora annotations include term identification. There are three options for this purpose: identifiers, MicroThesaurus (MT) labels or top-level domains. We consider MTs to be more useful for tasks like multilingual clustering, which was one of the goals of the project. This happens because MT labels include semantic information as opposed to the identifiers, which are used only as record ID in terminology. Furthermore, the large number of identifiers (6883), compared to 127 MT labels makes it more challenging for cross-lingual clusterization to decide identifier similarity (possibly requiring additional processing), while the MTs already utilize the hierarchical nature of the EuroVoc terminology. Finally, the MT architecture is more stable to changes in terminology, contrary to the identifiers which are growing in number or get removed (as certain terms may become obsolete).

The new EuroVoc classification method for Romanian legal documents presented in this section was integrated in the RELATE platform. First, it is possible to annotate single documents.[8] In this mode, the user can enter the text document, the number of identifiers to be predicted and a threshold for the identifier association probability. By default, the number of predicted identifiers is set to 6 and the threshold to 0. This corresponds to the same values used during the JEX comparison. The platform page is presented in the following image.

[8] https://relate.racai.ro/index.php?path=eurovoc/classify

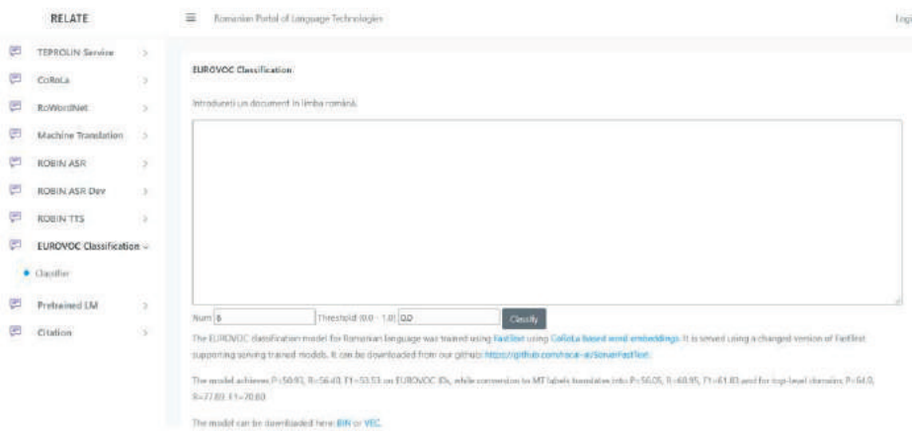

Figure 1. Single document EuroVoc classification in the RELATE platform.

Following the execution through the system, the platform presents the associated identifiers. These are then converted into MT labels and finally into top-level domains. Depending on the document entered, the number of identifiers is usually larger than the number of MT labels, which in turn is larger than the number of top-level domains. An example is presented in the next image. The transformation is handled automatically within the platform, using the EuroVoc hierarchy.

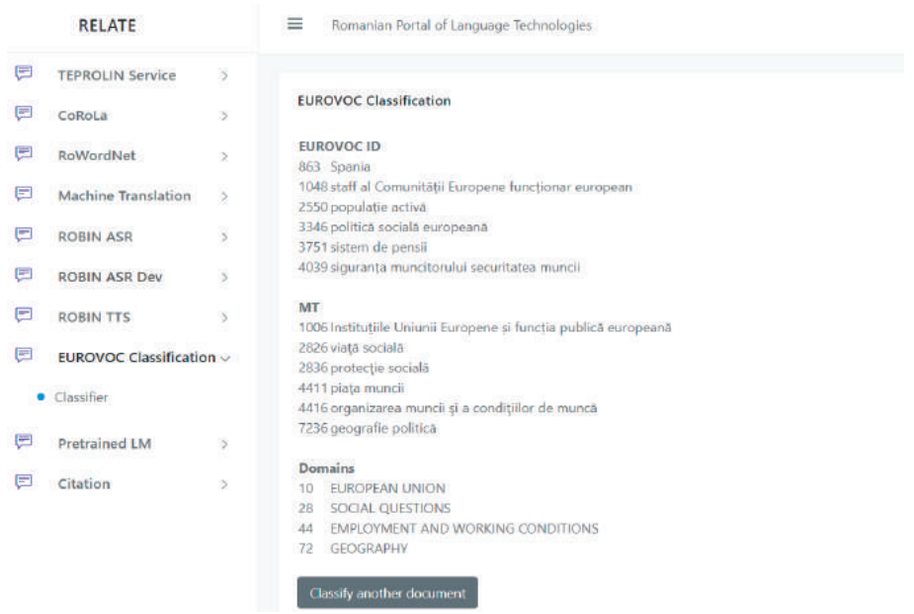

Figure 2. Results from single document EuroVoc classification in the RELATE platform.

The second integration is realized in the internal part of the platform, used for corpora annotation. This already provides mechanisms for uploading large corpora and performing sentence splitting, tokenization, part-of-speech tagging, dependency parsing. Furthermore, the integration of EuroVoc classification is available at the end of the processing pipeline as an additional task. Invocation of the new task is presented in Figure 4.

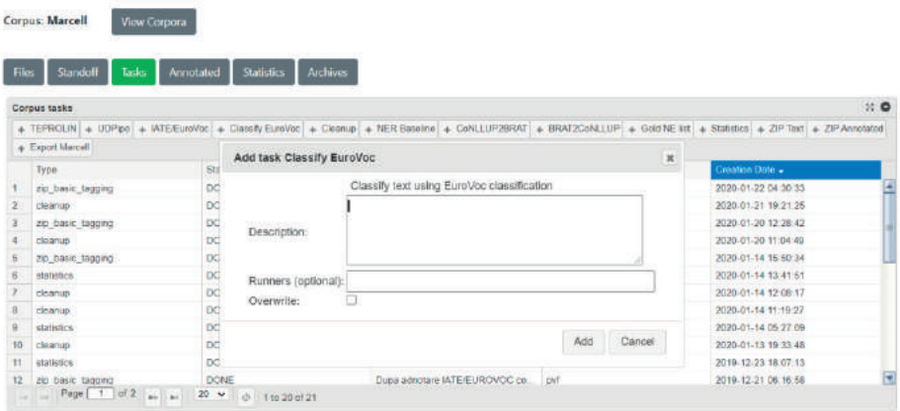

Figure 3. Corpora processing integration in the RELATE platform.

According to the Marcell specification, the results of EuroVoc classification is available in the metadata fields of each annotated file. Since we use CoNLLU-Plus format for the annotations, we first defined the columns like this "# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC MARCELL:IATE MARCELL:EUROVOC", thus considering the last two columns to correspond to IATE and EuroVoc terms. The EuroVoc classification is given by the line "# eurovoc_domains = 04 08 10 24". An example is presented in the following figure.

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC MARCELL:IATE MARCELL:EUROVOC
# eurovoc_domains = 04 08 10 24
# sent_id = 1
# text = ORDIN nr. 3.154 din 24 octombrie 2008 pentru modificarea si completarea Normelor metodologice de aplicare a prevederilor Ordonantei
1   ORDIN   ordin   ADP   Spsa   AdpType=Prep|Case=Acc   2   case   _   SpacesBefore=\n   _   _
2   nr.   nr.   NOUN   Yn   Abbr=Yes   0   root   _   _   _   _
3   3.154   3.154   NUM   Mc-p-d   Number=Plur|NumForm=Digit|NumType=Card   2   nummod   _   _   _   _
4   din   din   ADP   Spsa   AdpType=Prep|Case=Acc   6   case   _   _   _   _
5   24   24   NUM   Mc-p-d   Number=Plur|NumForm=Digit|NumType=Card   6   nummod   _   _   _   _
6   octombrie   octombrie   NOUN   Ncms-n   Definite=Ind|Gender=Masc|Number=Sing   2   nmod   _   _   _   _
7   2008   2008   NUM   Mc-p-d   Number=Plur|NumForm=Digit|NumType=Card   6   nummod   _   SpacesAfter=\n   _   _
8   pentru   pentru   ADP   Spsa   AdpType=Prep|Case=Acc   9   case   _   _   _   _
9   modificarea   modificare   NOUN   Ncfsry   Case=Acc,Nom|Definite=Def|Gender=Fem|Number=Sing   2   nmod   _   _
10   si   si   CCONJ   Crssp   Polarity=Pos   11   cc   _   _   _   _
11   completarea   completare   NOUN   Ncfsry   Case=Acc,Nom|Definite=Def|Gender=Fem|Number=Sing   9   conj   _   _
12   Normelor   normă   NOUN   Ncfpoy   Case=Dat,Gen|Definite=Def|Gender=Fem|Number=Plur   11   nmod   _   _   _
13   metodologice   metodologic   ADJ   Afpfp-n   Definite=Ind|Degree=Pos|Gender=Fem|Number=Plur   12   amod   _   _   _
14   de   de   ADP   Spsa   AdpType=Prep|Case=Acc   15   case   _   _   _   _
```

Figure 4. A Romanian legal document annotated according to Marcell specifications.

6. Marcell project sustainability

Most projects end providing new data once their allocated project duration expires. In the case of Marcell we considered a sustainability scenario in which new data could be provided even after the project ended. In order to ensure future operation, each annotation pipeline was embedded into a Docker container with all the required resources. Furthermore, a basic graphical user interface (GUI) was constructed in the form of a web site. This GUI was also dockerized with all the configuration and needed resources. The entry point of the Marcell sustainability GUI is presented in Figure 5.

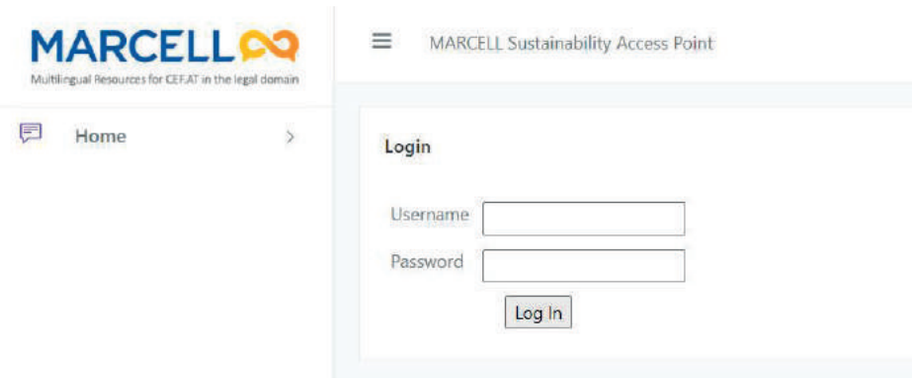

Figure 5. Marcell sustainability GUI entry point.

In order to construct the sustainability framework, each partner provided docker scripts for their pipelines. These were initially staged on the RELATE server and tested. The GUI itself was initially a stripped down version of RELATE which was further augmented with Marcell specific customizations as well as integration of the different language specific pipelines. The resulting platform allows uploading new raw text archives (as new legislation becomes available), then starting an annotation process which calls the corresponding pipeline depending on the specified language. Depending on the number of dockers available for each language, the pipeline invocation can happen in parallel, thus reducing the overall time required for annotation. Finally, the results are stored back in the GUI and can be exported in the MARCELL specific format. Figure 6 presents different corpora loaded in the GUI during the sustainability framework testing.

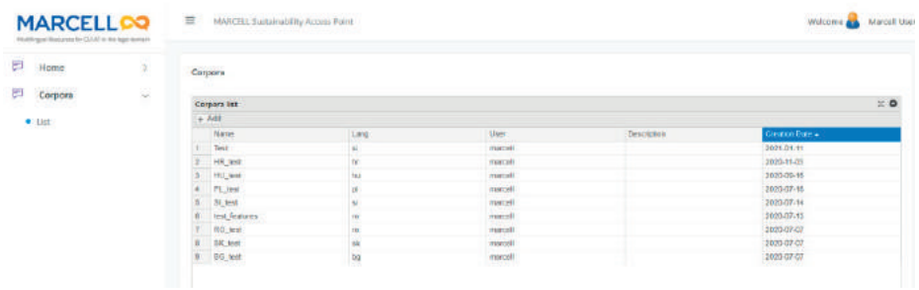

Figure 6. Corpora loaded in the Marcell sustainability framework GUI.

The GUI offers additional functionality such as computing statistics, allowing to see how the corpus grows over time, archiving of raw and annotated text, allowing to store different versions of the corpus. Resulting archives can be downloaded as ZIP files containing Marcell formatted documents.

6. Instead of Conclusions

The MARCELL consortium includes some veterans of the CEE Language Technology area, who have known and respected each other for more than 20 or 30 years. They are accompanied by younger and experienced researchers who will take over the responsibilities for the technologies of our national languages and accomplish the roadmap for all European Language Equality. Dr. Tamás Váradi has led the way for many years in a professional and elegant manner, being attentive to all details implied by his coordination of large scientific consortia. Dr. Tamás Váradi is an ideal project coordinator, diligent to observe the deadlines and milestones, harmonize efforts of his partners towards a successful accomplishment of the objectives.

The ICIA team wishes him a long and prosperous life with many more accomplishments.

A big and whole-hearted THANK YOU, Tamás!

References

Aho, A., Corasick, M.: Efficient string matching: An aid to bibliographic search. Commun. ACM 18/6, 333–340 (1975)

Barbu Mititelu, V., Tufiș, D., Irimia, E.: The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In: Calzolari, N., Choukri, Kh., Cieri, Ch., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the International Conference on Language Resources and Evaluation. pp. 1178–1185 ELRA (2018)

Boella, G., Di Caro, L., Lesmo, L., Rispoli, D.: Multi-label Classification of Legislative Text into EuroVoc. In: Schäfer, B. (ed.) Legal Knowledge and Information Systems: JURIX 2012: the Twenty-fifth Annual Conference Vol. 250. pp. 21. IOS Press, Amsterdam (2012)

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. In: Transactions of the Association for Computational Linguistics Vol. 5. pp. 135–146 (2017)

Boroș, T., Dumitrescu, S. D, Burtică, R.: NLP-Cube: End-to-End Raw Text Processing With Neural Networks. In: Zeman, D., Hajič, J. (eds.) Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 171–179. Association for Computational Linguistics, Brussels, Belgium (2018)

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P.: Natural language processing (almost) from scratch. Journal of machine learning research 12, 2493–2537 (2011)

Coman, A., Mitrofan M, Tufiș D.: Automatic identification and classification of legal terms in Romanian law texts. In: Onofrei, M., Bibiri, A-D., Dragoș Nicolae, C., Tufiș, D., Cristea, D. (eds.) Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2019). pp. 39–49. Editura Universității "Alexandru Ioan Cuza", Iași (2019)

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning Word Vectors for 157 Languages. In: Calzolari, N., Choukri, Kh., Cieri, Ch., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the International Conference on Language Resources and Evaluation (2018)

Ion, R.: TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian. In: Păiș, V., Gîfu, D., Trandabăț, D., Cristea, D., Tufiș, D. (eds.) Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018) Editura Universității "Alexandru Ioan Cuza", Iași (2018)

Ion. R.: Word Sense Disambiguation Methods Applied to English and Romanian. PhD Thesis. pp. 148. Romanian Academy, Bucharest (2007)

Joachims, T.: Text categorization with Support Vector Machines: Learning with many relevant features. In: Machine Learning: ECML-98. pp. 137–142. Springer Verlag, Berlin, Heidelberg (1998)

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification. In: Lapata, M., Blunsom, Ph., Koller, A. (eds.) Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics, Valencia, Spain (2017)

T. Mikolov, K. Chen, G. Corrado, J. Dean: Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 (2013)

Paiș, V., Tufiș, D.: Computing distributed representations of words using the CoRoLa corpus. In: Proceedings of the Romanian Academy. Series A Vol. 19, No. 2, pp. 403–409. Romanian Academy, Publishing House of the Romanian Academy, Bucharest (2018)

Păiș, V., Tufiș, D., Ion, R.: Integration of Romanian NLP tools into the RELATE platform. In: Onofrei, M., Bibiri, A-D., Dragoș Nicolae, C., Tufiș, D., Cristea, D. (eds.) Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2019). pp. 181–192. Editura Universității "Alexandru Ioan Cuza", Iași (2019)

Pouliquen, B., Steinberger, R. and Ignat, C.: Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. In: Cristea, D., Ide, N., Tufiș, D. (eds.) Proceedings of the Workshop Ontologies and Information Extraction at the Summer School The Semantic Web and Language Technology –Its Potential and Practicalities (EUROLAN 2003). Bucharest, Romania (2003)

Šarić, F., Dalbelo Bašić, B., Moens, M. F., Šnajder, J.: Multi-label classification of croatian legal documents using EuroVoc thesaurus. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of SPLeT-Semantic processing of legal texts: Legal resources and access to law workshop. ELRA, Reykjavik (2014)

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. and Varga, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+languages. In: Calzolari, N., Choukri, Kh., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., Tapias, D. (eds.) Proceedings of the 5th international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy (2006)

Steinberger, R., Ebrahim, M. and Turchi, M.: JRC EuroVoc Indexer JEX - A freely available multi-label categorisation tool. In: Calzolari, N., Choukri, Kh., Declerck, T., Ugur Dogan, M., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012). pp. 798–805. Istanbul, Turkey (2012)

Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) 3(3), 1–13 (2007)

Tufiș, D., Mitrofan, M., Păiș, V., Ion, R., Coman, A.: Collection and Annotation of the Romanian Legal Corpus. In: Calzolari, N., Béchet, F., Blache, P., Choukri, Kh., Cieri, Ch., Declerck, Th., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 12th Language Resources and Evaluation Conference. pp. 2766–2770. European Language Resources Association, Marseille, France (2020)

Tufiș, D., Mititelu, V. B., Irimia, E., Păiș, V., Ion, R., Diewald, N., Mitrofan, M., Onofrei, M.: Little strokes fell great oaks. creating CoRoLa, the reference corpus of contemporary Romanian. In: Revue roumaine de linguistique no.3. pp. 227–240. Editura Academiei Romane, Bucarest (2019)

Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pęzik, P., Verginica, B.M., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., Janez, B.: The MARCELL Legislative Corpus. In: Calzolari, N., Béchet, F., Blache, P., Choukri, Kh., Cieri, Ch., Declerck, Th., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020). pp. 3754–3761. Marseille, France (2020)