ROMBAC: The Romanian Balanced Annotated Corpus

Radu Ion, Elena Irimia, Dan Ştefănescu, Dan Tufiş
Research Institute for Artificial Intelligence - Romanian Academy,
13, Calea 13 Septembrie, 050711, Bucharest, Romania
{radu, elena, danstef, tufis}@racai.ro

Abstract

This article describes the collecting, processing and validation of a large balanced corpus for Romanian. The annotation types and structure of the corpus are briefly reviewed. It was constructed at the Research Institute for Artificial Intelligence of the Romanian Academy in the context of an international project (METANET4U). The processing covers tokenization, POS-tagging, lemmatization and chunking. The corpus is in XML format generated by our in-house annotation tools; the corpus encoding schema is XCES compliant and the metadata specification is conformant to the METANET recommendations. To the best of our knowledge, this is the first large and richly annotated corpus for Romanian. ROMBAC is intended to be the foundation of a linguistic environment containing a reference corpus for contemporary Romanian and a comprehensive collection of interoperable processing tools.

Keywords: balanced corpus, corpus processing, annotation type, metadata, XCES, TTL, ROMBAC, Romanian

1. Introduction

METANET4U (http://metanet4u.eu/) belongs to a cluster of projects aiming at fostering the technological foundations of a multilingual European information society. These projects synergistically follow specifications and recommendations issued by the META-NET Network of Excellence (http://www.meta-net.eu) and are commonly using META-SHARE (developed by META-NET), a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources (http://www.meta-net.eu/meta-share).

As a partner in METANET4U, RACAI delivered through META-SHARE several mono- and multi-lingual resources and tools for NLP tasks on Romanian textual data. One of the most valuable resources we uploaded on the Meta-Share platform is ROMBAC, the Balanced Annotated Corpus of Romanian, which will be described in the rest of this article.

Developing a balanced corpus presupposes: defining its structure, its linguistic coverage, collecting texts according to the established structure, solving problems of copyright, processing text with linguistic technologies (segmentation, lemmatization, tagging, etc.), text indexing according to various criteria useful in exploitation, extracting statistical data, developing an exploitation platform, as friendly and flexible as possible, establishing secured access methods in order to prevent vandalism or misuse represent features and conditions indispensable for the usage of the resource in the best possible manner. In the context of public access, the hardware architecture has to be adequate to simultaneous access by more users.

Some of these problems were internally solved, as can be seen in the following sections. Others, like the exploitation platform, are assured by the METANET4U support for the development of this resource.

Section 2 of this paper is dedicated to describing the structure of the corpus, with focus on the content of each sub-domain, the modality of acquiring it and the initial pre-processing steps that were made. Section 3.1 presents the processing tool and heuristics for the tokenization, lemmatization and POS-tagging of the corpus. Section 3.2 describes the different forms and levels of annotations, both as resulted from our processing tools and as conforming to international standards in annotation (Subsections 3.2.1, 3.2.3 and 3.2.3). Section 4 presents the semi-automatic method we envisaged for the validation of the corpus annotation and section 5 presents the corpus statistics. In Section 6 we draw some conclusion and present the future work.

2. Corpus Structure

The corpus contains, discounting the punctuation, about 36,000,000 words evenly distributed into five genres: **journalistic** (news and editorials), **pharmaceutical and medical** short texts, **legalese**, **biographies** of the major Romanian writers **and critical reviews** of their works, and **fiction** (both original and translated novels and poetry). The texts are tokenized, morpho-syntactically tagged, lemmatized, shallow-parsed (chunked) and XCES-compliant encoded.

2.1 The journalistic sub-domain (News)

The journalistic sub-corpus of ROMBAC consists of the issues published daily between 2003 and 2006 of the AGENDA newspaper (http://www.agenda.ro/)[1]. The AGENDA sub-corpus is a middle-sized journalistic corpus, having about 8,500,000 words. It evolved from a very large collection of journalistic articles, initially available in various formats (doc, rtf and pdf). They were

[1] We take the opportunity to thank once again the publisher for making the data available for research purposes.

converted into ASCII format, with diacritical characters encoded initially as SGML entities and recently in UTF8.

2.2. The medical sub-domain (Medical)

The second sub-corpus of ROMBAC has been extracted from the EMEA corpus. EMEA is a parallel corpus made out of PDF documents from the European Medicines Agency, compiled by Jörg Tiedemann. All files are automatically converted from PDF to plain text. For more details about the corpus and the conversion strategy, see (Tiedemann, 2009). The Romanian-English part of the corpus was downloaded from the following web address: http://opus.lingfil.uu.se/EMEA.php. From the Romanian part, a number of 800 documents (most of the texts are drug leaflets) containing around 9,100,000 words were randomly selected to be part of the Romanian Balanced Corpus.

2.3. The legalese sub-domain (Legal)

The juridical sub-corpus has been extracted from the JRC-Acquis corpus, a collection of legislative texts representing the total body of European Union (EU) law applicable the EU Member States. It is a parallel corpus available in 22 languages: all the official languages in European Union minus Irish, the translations of which are not currently available (Steinberger et al., 2006). This is a big collection of documents, containing laws published starting from 1958 until 2006.

The Romanian files available in the corpus were initially in Microsoft Word format and they had to be converted in text format. The conversion requested some intermediary processing steps for removing the translators' comments, deleting the footnotes and headers, normalizing the diacritics usage (each of the characters "ş" and "ţ" were represented by two different codes). For our purposes, we retained only the documents published between 2003 and 2006, summing around 7,500,000 words.

2.4. The academic sub-domain (Biography)

The fourth sub-corpus of ROMBAC is based on the content of the Romanian Literature General Dictionary (DGLR, 2009), a 7 volume critical anthology which contain biographies of Romanian writers, poets, essayists as well as commentaries about their work, information about publications, literary concepts, literary trends, anonymous writings, literary institutions, translators from/in Romanian etc. This impressive dictionary, created by the Institute for Literary History and Theory "George Călinescu" [2] of the Romanian Academy, has been provided in UTF8 text format by the authors, as part of their commitments to the METANET4U project. The text contains around 4,300,000 words.

2.5. The literary subdomain (Fiction)

The fifth part of the ROMBAC corpus is a collection of novels and poems authored by 28 classical Romanian writers from the end of the 19th and beginning of the 20th centuries. This corpus was in part written with the old Romanian orthography. The orthography was updated to the current norms and the codes for the diacritical characters were unified. This sub-corpus contains about 6,800,000 words.

3. Corpus Annotation

3.1. Processing tools

The texts in the corpus were normalized at the orthographic level, cleaned of footnotes, headers and page numbers and the punctuation was separated from the words. After this preliminary phase, the corpus was subjected to an annotation process using the TTL text processing platform developed at RACAI (Ion, 2007; Tufiş et al., 2008). TTL is entirely written in Perl and performs named entity recognition, sentence splitting, tokenization, POS tagging and chunking. We have exposed it as a SOAP compliant web service with an available WSDL file[3] and also as a REST web-service for the WebLicht platform (Henrich et al., 2010).

TTL tokenizer (Ion, 2007) is language aware and recognizes Romanian multiword functional expressions, clitics and contractions. Then, the tokens were annotated at the morpho-lexical level (MSD annotation), using TTL's HMM tiered tagger. The tagset used in the ROMBAC is a large tagset: 614 MSD tags fully compatible with the MULTEXT-East morpho-lexical specifications[4] plus 20 named entity tags (Tufiş & Ion, 2007). The reduced (hidden) tagset used for tiered tagging (Tufiş, 1999; Tufiş & Dragomirescu, 2004) contains 93 tags for words and 10 tags for punctuation.

The corpus was further lemmatized through a look-up procedure in a large (more than 1,200,000 entries), human-validated Romanian word-form lexicon the entries of which have the form:

```
<word-form><TAB><lemma><TAB><tag><EOL>
```

In Romanian, as in many other languages, most of the time a word-form and its tag uniquely identify the lemma. When this is not the case, the lemmatizer selects the most frequent lemma out of the competing ones. For the tokens not in the word-form lexicon (and which are not tagged as proper names), the lemma is provided by a five-gram letter Markov Model-based guesser, trained on the correct lemmas from the word-form lexicon with the same POS tag as the token being lemmatized. The guesser scans all the known endings of the unknown word right to left and generates a list of probable lemma candidates by stripping off the recognized endings from the unknown word. Each candidate lemma is then scored according to the learned Markov Model on correct lemmas with the same POS tag

[2] http://www.institutulcalinescu.ro/

[3] http://ws.racai.ro/ttlws.wsdl

[4] http://nl.ijs.si/ME/V3/msd/html/msd.html

and the most probable one is selected as the result of the statistical lemmatization process.

The next processing step is the text chunking. This process is guided by a set of regular expression rules, defined over the MSDs and it deals with recognizing adjectival, adverbial, nominal, verbal and prepositional phrases. With respect to the verbal phrases, the chunker recognizes only the analytical forms of the verbs (compound tenses and passive constructions).

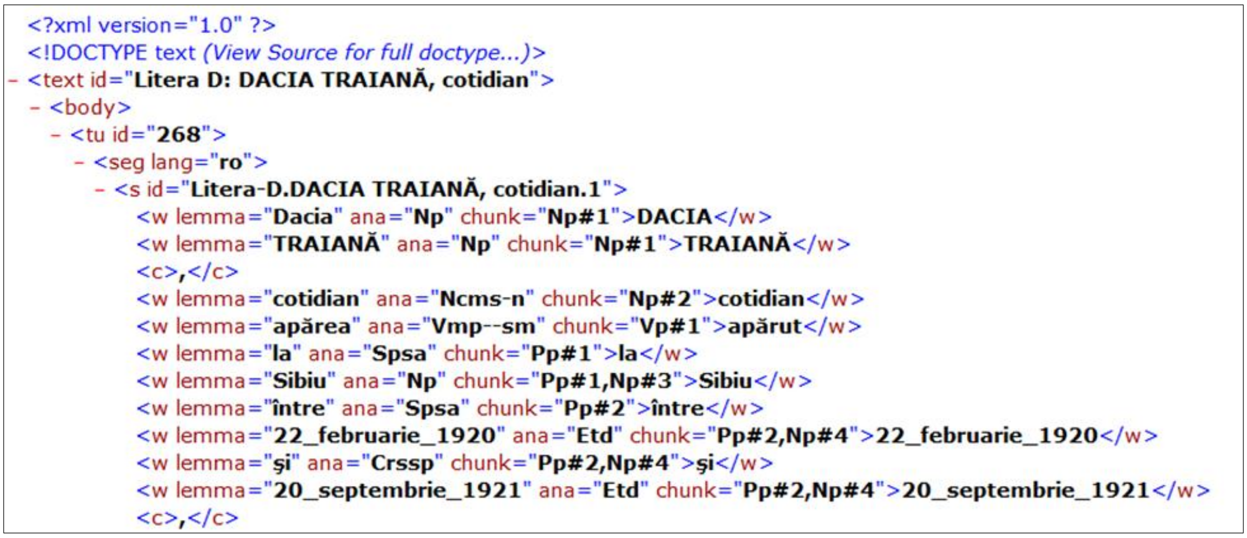
3.2. Annotation formats

3.2.1 TTL format

The output of TTL is an XML file encoding sentences (with paragraph information codified in the attribute ‘id’ of the sentence <s> element) and tokens, each token being classified either as a word (marked with the <w> element) or as a punctuation (marked with the <c> element). Each word has several attributes that will specify its lemma, its POS label (the ‘ana’ attribute), its membership to a chunk and its orthographic form given as the content of the <w> element. The Figure 1 is an example of the standard XML encoding that TTL produces.

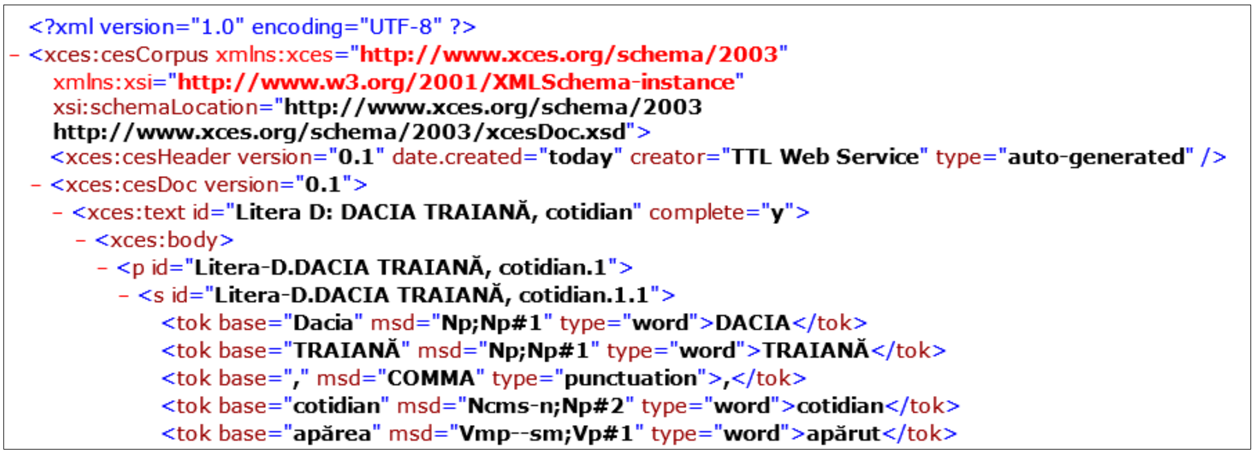
This XML format is useful for a large number of NLP applications since it conveniently delimits the units of text along with their annotations but, when clarity and standards compliance are in question, a better, more explicit and metadata aware representation is expected.

Since the Romanian Balanced Corpus had to be released as a META-NET deliverable, we chose to automatically convert our XML notation to the standard XCES Schema notation, revision 1.0.4 which is available at http://www.xces.org/schema/2003/.



```
  <?xml version="1.0" ?>
  <!DOCTYPE text (View Source for full doctype...)>
- <text id="Litera D: DACIA TRAIANĂ, cotidian">
  - <body>
    - <tu id="268">
      - <seg lang="ro">
        - <s id="Litera-D.DACIA TRAIANĂ, cotidian.1">
            <w lemma="Dacia" ana="Np" chunk="Np#1">DACIA</w>
            <w lemma="TRAIANĂ" ana="Np" chunk="Np#1">TRAIANĂ</w>
            <c>,</c>
            <w lemma="cotidian" ana="Ncms-n" chunk="Np#2">cotidian</w>
            <w lemma="apărea" ana="Vmp--sm" chunk="Vp#1">apărut</w>
            <w lemma="la" ana="Spsa" chunk="Pp#1">la</w>
            <w lemma="Sibiu" ana="Np" chunk="Pp#1,Np#3">Sibiu</w>
            <w lemma="între" ana="Spsa" chunk="Pp#2">între</w>
            <w lemma="22_februarie_1920" ana="Etd" chunk="Pp#2,Np#4">22_februarie_1920</w>
            <w lemma="şi" ana="Crssp" chunk="Pp#2,Np#4">şi</w>
            <w lemma="20_septembrie_1921" ana="Etd" chunk="Pp#2,Np#4">20_septembrie_1921</w>
            <c>,</c>
```

Figure 1: A sample of the XML output of the TTL web service run on the ‘D’ letter file of the Romanian Literature General Dictionary sub-corpus.



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <xces:cesCorpus xmlns:xces="http://www.xces.org/schema/2003"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.xces.org/schema/2003
    http://www.xces.org/schema/2003/xcesDoc.xsd">
    <xces:cesHeader version="0.1" date.created="today" creator="TTL Web Service" type="auto-generated" />
  - <xces:cesDoc version="0.1">
    - <xces:text id="Litera D: DACIA TRAIANĂ, cotidian" complete="y">
      - <xces:body>
        - <p id="Litera-D.DACIA TRAIANĂ, cotidian.1">
          - <s id="Litera-D.DACIA TRAIANĂ, cotidian.1.1">
              <tok base="Dacia" msd="Np;Np#1" type="word">DACIA</tok>
              <tok base="TRAIANĂ" msd="Np;Np#1" type="word">TRAIANĂ</tok>
              <tok base="," msd="COMMA" type="punctuation">,</tok>
              <tok base="cotidian" msd="Ncms-n;Np#2" type="word">cotidian</tok>
              <tok base="apărea" msd="Vmp--sm;Vp#1" type="word">apărut</tok>
```

Figure 2: Sample of an automatically generated document compliant with XCES Schema 1.0.4

3.2.2. Conversion to XCES format

XCES Schema has support for a wide range of annotations (including different types of alignments and the possibility to reference annotations from external files) and also for inclusion of metadata in the header of each document. This schema supports annotations on multiple layers in different files but, for our purposes we will use the types defined in the 'xcesDoc.xsd' schema. In Figure 2, we give a "translation" of the XML from Figure 1 into standard XCES.

Other than the inclusion of standard elements such as 'xces:cesHeader' and 'xces:cesDoc', the structure is pretty much similar with the XML structure from Figure 1. Now all tokens are marked up by the <tok> element and the type of the token (word or punctuation) is established with the 'type' attribute of the element. The 'base' attribute specifies the base form (lemma) of the token and the 'msd' attribute is meant (according to the authors of the XCES Schema) to hold "all relevant morpho-syntactic information". As such, with no other alternatives, we were forced to collapse the POS tagging and chunking information under the 'msd' attribute (separated by ';') by following an example of annotation of the American National Corpus[5].

In order to ensure compliance with our present web-services and workflows, we decided for in-line annotations, yet a stand-off version of encoding is simple to produce.

3.2.3 Metadata annotation

ROMBAC is distributed through the META-NET distribution network and it is available to download from the RACAI's instantiation of the MetaShare V1.1 Web Platform : http://ws.racai.ro:9191/browse/. Following the META-NET recommendations (Desipri et al., 2012), the sections that are included in the metadata XML files are:

- *IdentificationInfo*, containing the subsections: *resourceName, resouceShortName, pid*.
- *contactPerson*, with the subsections *surname, givenName, position, CommunicationInfo, affiliation*; in turn, CommunicationInfo has the following subsections: *address, zipCode, city, country, telephoneNumber, faxNumber, email, url*; affiliation, through a subsections named *OrganizationInfo*, has the following components: *organizationName, organizationShortName and CommunicationInfo* (with the subsections mentioned before);
- *DistributionInfo* contains the *availability* and *Licence Info* attributes, with LicenceInfo having the *licence, restrictionsOfUse* and *distributionAccessMedium* subsections;
- *MetadataInfo* contains the subsections *source* (which in this case is METANET) and *metaDataCreationDate*;
- *ContentInfo* has the subsections *description*, *resourceType* (value: corpus) and *mediaType* (value: text)
- *TextInfo* contains:
  - *LingualityInfo* (with lingualityType:monolingual and modalityType: writtenLanguage);
  - *LanguageInfo* (with languageCoding: ISO 639-3, languageID:ron, languageName:Romanian),
  - *SizeInfo* (with size:36000000), sizeUnitMultiplier:unit, sizeUnit:token),
  - *CharacterEncodingInfo* (with characterEncoding: UTF-8 and characterSet: other)
  - four different *AnnotationInfo* subsections; an AnnotationInfo section contains usually the subsections:
    - *annotationType* (which for our corpus has, consecutively, the values: segmentation, morphosyntactic Annotation - posTagging, lemmatization, syntactic Annotation - shallowParsing),
    - *annotationStandoff*:false,
    - *segmentationLevel*:word,
    - *annotationFormat*:text/xml,
    - *conformanceToStandardsBestPractice*:XCES,
    - *annotationTool*: TTL Web Service: http://ws.racai.ro/ttlws.wsdl,
    - *annotationMode:* automatic;
    - the annotationInfo section dedicated to POS Tagging contains two supplementary attributes
    - *tagset*: Morpho-Syntactic Descriptors or MSDs: http://nl.ijs.si/ME/V4/msd/html/index.html
    - *theoreticModel*: Hidden Markov Models.

4. Annotation Validation

Because of limited human resources, time constraints and the dimension of the corpus, hand validation of each individual token was out of question. Therefore, the validation stage was implemented as a coherent methodology for automatically identifying as many POS annotation and lemmatization errors as possible. The TTL processing workflow generates for each token occurring in the corpus a triple <word-form, lemma, MSD> and it marks each such triple in case it resulted from an out of the dictionary word (ODW). The ODW marking does not consider proper nouns, abbreviations and named entities. As a first step, we extracted, sorted and counted all triples marked as ODW and found 304,460 of them, so that they represent less than 1% (0.84%) of the total number of words in the ROMBAC corpus. We further divided the found list of ODW into frequency classes: rare word-forms (frequency 1 or 2) – 0.20%, and the rest (occurring more than twice) – 0.64% from the entire ROMBAC word content. The initial analysis concentrated on the list containing wordforms occurring more than twice (as we anticipated that the other list would contain mainly erroneous words) and found that the vast majority of them could be classified into one of the following classes: (a) words written with the old orthography (the one in force until 1992), (b) proper unknown words. (c)

[5] http://americannationalcorpus.org/FirstRelease/encoding.html

words without diacritics or with non-standard diacritics encoding.

The ODW in the category a) occurred mostly in the texts from the literary sub-domain and represented about 35% of all ODW. Practically all these texts were published before the 1992 orthography reform. Although these words are technically ODW, their tagging and lemmatization were almost perfect due to a slight modification of the lexical lookup. However, the a) category ODWs in the corpus was not yet updated. The ODW in class b) represented almost half (49.39%) of all ODW. They came almost exclusively from the Medical sub-corpus representing terms and modifiers (nouns and adjectives) specific to the domain. Interestingly enough, they occurred, to a large extent, in direct case and indefinite form which is less ambiguous in Romanian and that is why few tagging and lemmatisation errors have been noticed. The ODW in class c) were less numerous and were found in all the sub-corpora with the most part occurring in the medical sub-corpus. As the diacritics play an important role, here several tagging errors were especially on the definiteness attribute value. The corpus was not yet corrected.

The rare words were classified in two additional categories: d) typographical errors (missing space between adjacent words, inverted letters, extra letters and missing letters) and e) foreign words. The last category of ODWs appeared in the Medical sub-corpus and also in the Legal sub-corpus (although much less frequent). The explanation is related to some flaws in the sentence alignments and extraction of the Romanian data from the multilingual corpora EMEA and JRC-Acquis. The tagging and lemmatization of the ODW in categories d) and e) is unreliable but altogether they do not exceed 0.1% of the entire ROMBAC corpus. The final cleaning is not finalized by the time of this writing, but the final delivery to METANET4U will be cleared.

The complete validation methodology, described in details by (Tufiș & Irimia, 2006), showed that after 3-4 iterations of biased tagging, comparison and correction of the tagging differences, the estimated error rate in the AGENDA sub-corpus was less than 2%. For ROMBAC corpus (except AGENDA sub-corpus) this methodology was not entirely applied, due to the problems discovered and discussed above. However, as mentioned before, the final delivery of ROMBAC is expected to be as accurately annotated as AGENDA sub-corpus is today.

5. Corpus Statistics

We counted all the sentences, words (content words plus functional words), content words (nouns, main verbs, adjectives and general adverbs) and tokens (words and punctuation) in ROMBAC for each type of sub-corpus it contains. The punctuation count is computed by subtracting the words count from the tokens count. The functional word (anything that is not a content word: determiners, prepositions, conjunctions, pronouns, articles, particles, etc.) count is computed by subtracting the content words count from the words count.

	Sentences	Tokens	Words	Content words
News	651,872	10,294,016	8,558,619	4,662,528
Medical	603,161	10,950,271	9,163,029	5,226,837
Legal	659,646	9,067,516	7,482,484	4,247,737
Biogr.	314,368	5,802,961	4,298,493	2,567,427
Fiction	517,803	8,002,596	6,773,648	3,531,156
Total	**2,746,850**	**44,117,360**	**36,276,273**	**20,235,685**

Table 1: Statistics on genres of ROMBAC

Table 1 and Table 2 show that, from a purely statistical point of view, the texts included into ROMBAC corpus are balanced, in terms of sentence length, words per sentence or punctuation per sentence. The biographical sub-corpus is the most distant from the other sub-corpora: sentences are longer and it contains more punctuation. The difference is made by a more frequent use of comma and semi-colon (in strict compliance with the academic rules). An interesting fact that is easily inferable from the Table 1 is that the Medical and Legal corpora have the same ratio between content words and functional words: 1.3. The Fiction sub-corpus shows a higher use of functional words, so the ratio between content words and functional words is 1.1. At the other end, the texts in the Biographies sub-corpus use less functional words and the ratio between content words and functional words is the highest: 1.5.

	Sentences	Tokens per sentence	Words per sentence	Punct. per sentence
News	651,872	15.79	13.1	2.66
Medical	603,161	18.15	15.2	2.96
Legal	659,646	13.74	11.3	2.40
Biogr.	314,368	18.45	13.7	4.78
Fiction	517,803	15.45	13.1	2.37
Total	**2,746,850**	**16.06**	**13.2**	**2.85**

Table 2: The proportions of tokens, words and punctuation per sentence by sub-corpus type

Table 3 and 4 presents the distribution of content words among the participating parts of speech (POS) for each type of sub-corpus from ROMBAC.

	Noun	Verb	Adj.	Adv.
News	3,164,278	712,085	651,700	134,465
Medical	3,136,988	919,544	939,048	231,257
Legal	2,808,814	654,515	648,928	135,480
Biography	1,739,559	336,584	401,677	89,607
Fiction	1,691,531	1,061,464	452,210	325,951
Total	**12,541,170**	**3,684,192**	**3,093,563**	**916,760**

Table 3: POS statistics for content words in each sub-corpus of ROMBAC

	Noun	Verb	Adj.	Adv.
News	67.8%	15.3%	14%	2.9%
Medical	60%	17.6%	18%	4.4%
Legal	66.1%	15.4%	15.3%	3.2%
Biography	67.8%	13.1%	15.6%	3.5%
Fiction	48%	30%	13%	9%
Total	**62%**	**18.2%**	**15.3%**	**4.5%**

Table 4: POS distribution for content words in each sub-corpus of ROMBAC

While the fiction texts in ROMBAC make more intensive use of verbs and adverbs, the biographies texts rely more on nouns and adjectives. Quite surprising is the similar distribution of content words in the News and the Legal sub-corpora and the highest percentage of adjectives in the Medical sub-corpus.

6. Conclusion

The first version of ROMBAC, as described here has been publicly released within the METANET4U project via a local copy of the MetaShare V1.1 distribution platform (http://ws.racai.ro:9191/). A newer and more stable version MetaShare V2.0, capable of handling the different licence types will be installed soon. The ROMBAC corpus as described in this paper may be downloaded according to its associated licence.

For the final version of ROMBAC (to be delivered by September 2012) we plan the following:

- removing all the problems mentioned in section 4;
- applying the full tagging and lemmatization validation methodology as described by (Tufiş & Irimia, 2006) ensuring a minimal error rate (less than 2%);
- building a web interface allowing for remote use of ROMBAC (regular expressions based search, concordances, various statistics, etc.).

ROMBAC is the starting point of a new project carried on by the Romanian Academy aiming at building a large reference corpus of Contemporary Romanian covering more text types, both original and translations into Romanian with a significant part of multilingual corpora.

7. Acknowledgements

This work has been supported by the METANET4U SPS European project under the grant no. #270893.

8. References

Romanian Academy. (2009). *DGLR: Dicţionarul General al Literaturii Române*. Univers Enciclopedic Publishing House. Vol I-VII. 1993-2009.

Desipri, E, Gavrilidou, M., Labropoulou, P., Piperidis, S., Frontini, F., Monachini, M., Arranz, V., Mapelli, V., Francopoulo, G., Declerck, T. (2012). Documentation and User Manual of the META-SHARE Metadata Model. Editors Penny Labropoulou, Elina Desipri, in "MetaNet – A Network of Excellence Forging the Multilingual EuropeTechnology Alliance", available for download at www.meta-net.eu/meta-share/META-SHARE%20%20documentationUserManual.pdf

Henrich, V., Hinrichs, E., Hinrichs, M., and Zastrow, Th. (2010). Service-Oriented Architectures: From Desktop Tools to Web Services and Web Applications. In Dan Tufiş. Corina Forăscu (eds.): *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*. Editura Academiei. Bucureşti. ISBN 978-973-27-1972-5. pp. 69-92.

Ide, N., and Suderman, K. (2004). The American National Corpus First Release. *Proceedings of the Fourth Language Resources and Evaluation Conference* (LREC). Lisbon. 1681-84.

Ion, R. (2007). Word Sense Disambiguation Methods Applied to English and Romanian. PhD thesis (in Romanian). Romanian Academy. Bucharest. 138 p.

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) *Recent Advances in Natural Language Processing* (vol V). pp. 237--248. John Benjamins. Amsterdam/Philadelphia

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Genoa. Italy.

Tufiş D., and Irimia E. (2006). RoCo_News - A Hand Validated Journalistic Corpus of Romanian. In *Proceedings of the 5th LREC Conference*. Genoa. Italy. pp. 869--872.

Tufiş, D., Ion R., Ceauşu A., and Ştefănescu D. (2008). RACAI's Linguistic Web Services. In *Proceedings of the 6th Language Resources and Evaluation Conference – LREC'08*. Marrakech. Morocco. ELRA - European Language Resources Association.

Tufiş D., and Ion R. (2007). New Tagset Specifications. Research Report. RACAI. Bucharest. June (in Romanian).