Abstract

Contemporary standard language corpora are ideal for NLP. There are few morphologically and syntactically annotated corpora for Romanian, and those existing or in progress only deal with the Contemporary Romanian standard. However, the necessity to study the dynamics of natural languages gave rise to balanced corpora, containing non-standard texts. In this paper, we describe the creation of tools for processing non-standard Romanian to build a big balanced corpus. We want to preserve in annotated form as many early stages of language as possible. We have already built a corpus in Old Romanian. We also intend to include the South-Danube dialects, remote to the standard language, along with regional forms closer to the standard. We try to preserve data about endangered idioms such as Aromanian, Meglenoromanian and Istroromanian dialects, and calculate the distance between different regional variants, including the language spoken in the Republic of Moldova. This distance, as well as the mutual understanding between the speakers, is the correct criterion for the classification of idioms as different languages, or as dialects, or as regional variants close to the standard.

1 Introduction

The UAIC-RoDia-DepTb (ISLRN 156-635-615-024-0) is a balanced treebank that becomes the core of a big corpus for the Old and Regional Romanian and for its South-Danube dialects. The treebank has now 16,187 sentences, with 322,404 tokens, illustrating all the styles of communication, between which there are 5,723 sentences in old Romanian and 1,230 sentences in regional variants of the Romanian.

If we know the non-standard, regional or earlier forms, we can understand the laws of natural language evolution; we can know how it functions in the communication and process it. The use of Old or Regional Romanian should not be judged as mistaken reporting them to the standard rules, but it is in accordance with other rules which we must discover.

Linguists are increasingly interested in the study of old languages with modern tools and their demand for old language processing tools is growing worldwide. Consequently, the computational linguists are building Diachronic and Old Corpora for all the natural languages, some described in many related works, as: Borin and Forsberg (2008), Davies (2010), Prevost and Stein (2013), etc.

We have built a sub-corpus for the Old Romanian, another for the regional variants, but it is more difficult to build corpora for the South Danube Romanian dialects (because they are very different from the standard language, cannot be understood by his speakers and two of them have no written aspect).

The emergence of South-Danube dialects is historically and politically determined. Due to their isolation of the linguistic center, they are more conservative than this one and retain many archaic linguistic phenomena.

Matteo Giulio Bartoli 1925 formulates the theory of isolated area or side areas, demonstrating with some examples that these areas are more conservative than the center. The dialects conserve more forms of the old languages or of the language where they were inherited (in our case, the Latin language).

For example, Istroromanian is the dialect of
the Romanian spoken by the smallest number of speakers, in eight villages in Croatia by 1,000 people, called "Vlach", only recently recognized as a national minority. This is an isolated area, very conservative for the Old Romanian. Through its disappearance and the lack of collected and digitized testimonies we could lose important data on the evolution of the Romanian language.

The dialectal variations study is important for the history of languages and for the etymology. The big etymological dictionary of Roman languages DeRom\(^1\) has between its bibliographical sources books of specialists in dialectology as: Iosif Popovici 1909 and Richard Sarbu 1998 for the Istroromanian dialect.

Besides their historical importance, these dialects are languages of disadvantaged minorities, threatened with extinction, with limited access to culture. Their folk creation and other contributions must be conserved and the people speaking it have to be received into the family of European languages.

2 Related Work

The need for a research infrastructure for the study of historical lexical resources by digitization and implication of language technology is increasingly recognized by the historical research community. Historical documents are being digitized on a vast scale in cultural heritage and digital library projects in many countries. Modern linguistics studies pays increasing attention to diachronic and dialectal variations of languages. Similar corpora of other languages have started from contemporary language processing tools by adapting them to their old or regional variants.

Digitized historical corpora are already created for many languages: English Yanez-Bouza (2011), Davies (2012), Spanish and Portuguese Davies (2010), French Stein (2008) and so on. In another paper, France Martineau 2007 analyze the use of probabilistic parsing methods for old French texts. Unfortunately, the described probabilistic parser was trained and can be used only within the project and cannot be adapted for other language and annotation conventions.

Thus, Nuria Yanez-Bouza 2011 describes the building of a rule-based automate pre-annotator which has around 30 rules to identify complex verb forms (VCOMP), adjective phrases (AP), noun phrases (NP) and prepositional phrases (PP). The next phase is the manual annotation via a special interface. We also have a hybrid POS-tagger which permits the introduction of rules (see below) and use the manual annotation for a part of the semantic relations.

In another paper Borin et al. (2010) present an ongoing work on the building of digitalized diachronic Swedish lexical resource. The paper presents a basic research infrastructure for language technology called BLARK (Basic Language Resource Kit) which includes basic lexical resources, annotated corpora and basic NLP tools for processing these corpora. The same authors Borin and Forsberg (2008) describe the creation of the tool for the morphological analyze of Old Swedish words, which should be followed by syntactic and semantic analysis. East European languages have several preprocessing issues. For example, for some of them Cyrillic and Latin scripts were used in various periods of time; hence some documents need to be transliterated before further processing (Gruszczynski and Ogrodniczuk, 2015).

3 Tools for Romanian Standard Processing

3.1 The UAIC-Ro Hybrid Part Of Speech (POS)-Tagger

The tools for the processing of the Contemporary Romanian are the basis for creating the ones for the old or regional Romanian processing. The lexicon of the UAIC-POS-tagger for Contemporary Romanian also contains archaic words and forms, extracted from dictionaries, while Old Romanian also contains words and forms used today.

The UAIC-Ro POS-tagger is hybrid, i.e. successfully combines a statistical model with a rule based system (Simionescu, 2011). The specificity of the hybrid model is that it applies a set of rules to reduce the large set of valid pairs lemma and POS-tag (abbreviated morphological analysis) which can be applied at a word-form. In fact, there are morphological homonyms, interpretable taking into account the words in the vicinity of their occurrence. After the reduction of the set of possible analysis, the statistical system is put into operation.

The dictionary of the POS-tagger is formed of triplets: word-form, lemma (the basis form of word, found in dictionaries), POS-tag (an abbrevi-

\(^1\) http://www.degruyter.com/view/product/205712
ated morphological analysis). The amount of the POS-tagger lexicon is related with the accuracy of the tool. The tool for the Contemporary Romanian contains 1,15 million distinct words extracted from dictionaries and 100,000 proper nouns extracted from Wikipedia. The set of 406 tags are a reduced version of the tagset used by the Multext East project Erjavec (2004).

The rule of big dimensions shows that the higher the number of tags is, the greater the gold corpus for the training must be. The training corpus for Contemporary Romanian consists in the NAACL 2003 corpus (39,000 sentences), and another 28,000 sentences extracted from the JRC-ACQUIS. The corpus for evaluation was Orwell’s novel 1984, manually annotated in the Multext East project.

But these corpora were not have identical set of conventions and not use our tagset. The training corpus is not entirely manually corrected. It is possible that there are inconsistencies between the annotation of these corpora. The accuracy of the tool, evaluated on standard Romanian, is 95.12 % without rules and 96.66 % with rules.

The POS-tagger has been trained on standardized, but also on non-standardized language, before processing 2,570 sentences in Social Media communication (Romanian chat). A method to increase the accuracy on chat sentences was to double the training corpus, with and without letters with specific Romanian diacritics: ş, à, ţ, î, ă, that are not always used in chat communication.

For the diacritics only RO-POS tagger evaluated on Orwell’s 1984 novel, we obtain an accuracy of 97.03 %. For the diacritics only RO-POS tagger, evaluated on the chat corpus the accuracy was only 68.67 %, for the Mixed diacritics RO-POS tagger, evaluated on mixed 1984, the accuracy was 94.38 %, and for the Mixed diacritics RO-POS tagger, evaluated on the chat corpus, the accuracy was 84.78 %, as shown in Perez 2016.

For difficult texts (old, non-standardized, new styles of communication not yet trained) we largely use manual validation of the output of the tool, and by the bootstrapping method, the corrected sentences are added to the gold training corpus.

3.2 The Malt Parser trained on Romanian

A variant of the Malt parser trained on UAIC-RoDepTb began to operate satisfactorily. The conventions of annotation used are in FDG (Functional Dependency Grammar), with labels of classical syntax, with numerous semantic sub-classifications of modifiers. Creating the treebank in 2007, Augusto Perez had the intention of targeting the treebank for didactic purposes, for medium learning, even building a computer game to prepare students for exams, but, of course, the learning system cannot be so easily convinced to adopt the Dependency Grammar.

This system can be transposed both into the modern syntactic system of Universal Dependencies (UD) with loss of semantic information and into a semantic annotation system by adding information. This is why we will continue to use this classic format, in which the processing tools were trained, and then it will be automatically (supervised) converted into UD (Universal Dependencies) and into a semantic annotation.

The parser is named multilingual or universal in Hall et al. (2006), because its functioning is based only on dependency relations, on the training and on the morphological previously annotation in no matter what language. The accuracy of the parser is determined by the size of the training corpus, more than 10,000 sentences, by the exactitude of the morphological annotation, and by the consistency of the syntactic conventions. Thus, the syntactic parser will not cause problems if used for annotation of ancient or regional texts, respecting consistently the conventions of UAIC-RoDia Dep-treebank, if the difficult problem of correct morphological annotation will be solved.

The parser was successfully used for syntactic annotation in the 2,570 chat phrases. On this occasion we found that the parser has a better accuracy after being trained on a large corpus which contains standard and non-standard sentences. After the training only with chat sentences, the accuracy on chat was 71.74 % for head attachment, 66.08 % for label attachment and 62.31 % for both attachment.

Using the same method, after the training with 15,000 sentences from all types of texts, including 4,000 sentences from the seventeenth century, and after the creation of Old Ro POS-tagger, the accuracy of our parser evaluated on Old Romanian was, for both attachments, 77.06 %, for head attachment, 83.79 %, and for label attachment, 82.5 %. The results are better than on chat corpus, because the training corpus was bigger, and we
used the new Old Ro POS tagger (described below) with a satisfactory accuracy, with the output entirely manually corrected.

3.3 POS-tagger for the Old and Dialectal Romanian

For building a series of Old Romanian processing tools, we began by building a POS-tagger for Old Romanian, which would give us basic annotation for the syntactic and semantic parsers or for any other type of annotation. To build a new POS-tagger, a list of tags, a lexicon, and a training corpus is needed. After the elaboration of these data, we can make clones of the UAIC hybrid POS-tagger, described above, and of the POS-tagger of the Institute of Mathematics and Computer Science of Chisinau, Republic of Moldova, which can analyze also Romanian words written in Cyrillic letters.

3.3.1 The List of Tags for the Old Ro POS-tagger

To establish the list of tags for the new POS-tagger, we began with the list of the UAIC hybrid POS-tagger for Contemporary Romanian, re-introducing some tags which had been eliminated, i.e. the detailed analysis of personal and reflexive pronouns (dative and accusative case, strong and weak forms) and the complex tags for the relational words (prepositions, conjunctions, relative adverbs). The first category (personal and reflexive pronouns) is useful to differentiate the direct and indirect objects, to establish the co-references and the expletives.

A new set of tags which we introduced, original ones, were aimed at annotating language specific phenomenon of (Old) Romanian, namely the negation of non-personal synthetic modes (participle, gerund and supine) by the prefix ne-. Words as “neterminat”, “nestrictă”, “năştiind” (En: unfinished, unbroken, not knowing) have the lemma “a termina” (to finish), “a strică” (to break), “a şti” (to know), because the verbs “a netermina”, “a nestrictă”, “a neşti” do not exist. The postag of these forms will be annotated as: “Vmp–sm-z”, “Vmp–pf–z”, “Vmg–z”. (participle negative singular masculine, participle negative plural feminine, gerund negative). The opposite tag to “Vmp–pf–z” has, on the eighth position, the p that means “positive”.

The annotation of verbal participles as adjectives is not acceptable, and has been corrected everywhere in the training corpus. We intend to build a dictionary of predicate arguments and adjunct structures (Cenel-Augusto Perez, 2015), and the participle has the same possible (syntactic and semantic) dependencies as the other verbal forms. We do not accept the letters y / n for annotating categories other than +/- definiteness, because they are not transparent and can be confusing, as shown in Mărânduc et al. (2016).

Finally, if the above forms are possible, but uncommon in Contemporary Romanian, there are tags needed to annotate specific forms only of the Old language. The list of tags for the forms of Dh (emphatic determiners) was also doubled as Ph (emphatic pronouns), because these forms exist in Old Romanian independently, not only as determiners of a noun. The emphatic adjective has a lot of specific forms in Old Romanian: “eluşi”, “eişt”, “iuişi” (En: himself, themselves). Another phenomenon is the imperative formed from the long infinitive. Example: “Nu vă teameţi!” (with short infinitive) “Nu vă teameţi!” (with long infinitive) (En: Do not be afraid!)

In Contemporary Romanian, the relative pronoun “care” has a very reduced inflection, and in Old Romanian the inflection is complex. Examples: “carele” Pw3msry, “carea” Pw3fsry, “cari” Pw3mpr, “carii” Pw3mpr (En: which).

The new tagset for the Old Romanian has 540 tags. All these tags was annotated in the entire corpus and the tags which do not exist in the tagset was eliminated from all the sub-corpora and from the lexicon of the POS-tagger. In this way we assured the consistency of the training corpus with the lexicon.

3.3.2 The Lexicon for the Old Ro POS-tagger

The collection of texts in Old Romanian is quite advanced; it contains 23 documents in TXT format from the sixteenth century, 60 from the seventeenth, 76 from the eighteenth and 325 from the nineteenth century. These texts were cleaned of the meta-text, specialists comments and notes, then processed by a concordancer program that builds indexes and makes statistics of the number of occurrences for each form in all the 500 books in text format opened. The concordancer used is the Lucon 03.16, the Cătălin Mititelu’s program, available on Sourceforge site2.

However, the old Romanian texts were written

2https://sourceforge.net/directory/os:windows/?q=Lucon
in Old Cyrillic letters. The collection that we hold contains both scans with Cyrillic and transcripts in Latin letters made by specialists. There was no Optical Character Recognizer (OCR) program for such letters. The letters differ from one book to another and there are many transitional alphabets with mixed Latin and Cyrillic letters. Our colleagues from the Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova are now building an OCR which began operating satisfactorily for some texts (Cojocaru et al., 2017).

In the case of the Romanian language written with Cyrillic, we used the same lexicon for the Old-Romanian POS-tagger and for the OCR program, having two variants, with the word forms written in Latin letters and written in Cyrillic letters, trying to help our colleagues to increase the performances of the OCR for Romanian Cyrillic letters. The lexicon is obtained from the annotation of the Gospel, the first part of the New Testament (1648) having 5,028 sentences, text partially obtained by the cited above OCR program.

We have about 500 scanned old books without transcription in Latin alphabet. We hope that the new OCR will solve this problem. The solution of transcriptions made by specialists is questionable, because the specialists made sometimes interpretative transcriptions, closer to the old text, approaching it by the contemporary language (so that the POS-tagger will not recognize them in the text obtained by OCR) and introduced numerous notes and comments in contemporary language, while their elimination is time consuming.

Using the Lucon program, we created a lexicon to be introduced in the POS-tagger, with 120,000 word-forms. However, we do not know which number of actualized forms of words it contains; after the OCR processing of more printed old books, we will construct a more authentic lexicon.

Using the program DEPAR (Dictionary Parser) (Mărănduc et al., 2017) we extracted a list of 5,000 stable unanalyzable Multi Word Expressions from a dictionary (Mărănduc, 2010), and also 98,000 lexical or spelling variants extracted from the Thesaurus Dictionary3. The variants are generally old or regional, consequently their introduction in the POS-tagger lexicon is useful for the processing both of the Old and of the regional variants of the Daco-Romanian dialect (the name used by the dialectologists for the language spoken in Romania), but they are not useful for the South Danube dialects, that have special dictionaries.

The project Monumenta Linguae Dacoromanorum4, started in 1988 in cooperation with the Freiburg University, has the purpose to digitize the old religious books in Romanian. A new edition of the first Bible printed in Romanian (1688) has been completed by using manually checked automatic morphologic annotation; we have received the indices of their edition, using another system of abbreviations compatible with ours, to be introduced in our Old Ro POS-tagger lexicon.

Finally, the result will be processed with a tool that generates complete paradigms of words (adding possible forms that do not appear in the indexed texts). We have such a tool for the Contemporary Romanian, called Anamorph (Timofciuc et al., 2013), and now a variant for Old Romanian has been built and its training must begin (Gîfu and Simionescu, 2016). All variants of the word are associated with the contemporary lemma, and the inflexion is generated starting from the root of the word form chosen in the text.

3.3.3 The training corpus for the Old Ro POS-tagger

The third step is the construction of a training corpus for the Old Romanian POS-tagger. We have no other solution than to train the POS-tagger and the Malt parser on the contemporary gold corpus, then to process the old texts, and then to correct the output of the tools, at the beginning having a modest accuracy.

The manually corrected sentences will form in time the training corpus for the Old Romanian. The module attached of the POS-tagger will extract from these manually corrected annotations the forms or the analysis that does not exist in its lexicon and will add them. An increase of the accuracy of the POS-tagger is expected after the increasing of the training corpus, i. e. after the manually correction of more books automatically annotated with this tool. The training corpus (having now 18,187 sentences, including 6,882 of the 17th to 19th centuries) is smaller than the one for the Standard Romanian POS-tagger, but is consistent with the tagset. The POS-tagger for the Old Romanian has now an accuracy of 91.66 %.

3http://edtlr.info.uaic.ro/ 4https://consilr.info.uaic.ro/ mld/monumenta/
3.3.4 Building POS-taggers for processing the South Danube Dialects

Then, we will apply the same solution: to build a few clones of the UAIC-Ro-POS-tagger for each South-Danube dialect. The same steps must be completed for each South-Danube dialect, beginning with the difficult step of the acquisition of digitized sources.

The construction of the Regional lexicon begins with the acquisition of a big collection of sources: texts in each variant of the language, in editable form. Then, two problems must be solved: first, it needs a big lexicon containing all the possible word-forms in each dialect of the language, with the lemmas and with the correct morphological analysis for any form found in texts, and secondly, it needs a big gold corpus (manually checked) for the training.

The greater difficulty lies in the fact that there is a very big distance between these variants of the language, we will need to start from zero the construction of the training corpus for each dialect, without using the large corpus of the other variants of the language. It can only be used as training corpus on regional variants spoken in Romania and the Republic of Moldova.

We have a collection of texts published by specialists from each dialect and some dictionaries. In order to create a collection of sources, we should do dialectal surveys in the villages where the dialects are spoken, because two of them do not have a written aspect, to register and annotate the texts. We have not yet a collection of annotated specimens of Spontaneous spoken language in our corpus.

We can extract lemmas from the dictionaries using the program DEPAR (Dictionary Parser) (Mărănduc et al., 2017), but the inflexion must be manually introduced in the POS-tagger lexicon. For this purpose, we will have to associate specialists in the South-Danube dialects in our project. The existence of a lexicon with flexionary forms could also lead to the creation of OCR for the Aromanian dialect, that has a written aspect. The difficulty is that various phonetic transcription systems are used in different published collections.

4 Short Presentation of Romanian Dialects

4.1 The Aromanian Dialect

Aromanian is the dialect with the most speakers, approximately 250,000. They live in Albania, in Greece, (the Pindos Mountains, the province of Macedonia) or in the Republic of Macedonia, hence the name of the Macedo dialect, that can be confusing, because this is the name of a geographical region, and in that region there are also people of other origins and languages.

There are numerous texts in this dialect, because more people speaking it migrated in Romania and ones of them become specialists in dialectal linguistics. For example, Marițeanu (2006), Saramandu (2003) and Nevaci (2011), for mentioning only the contemporary ones. Other older collections are Capidan (1925), Obedenaru and Bianu (1891), Bujduveanu (2005), and so on. They have published collections of texts in more styles of the language, written or oral popular literature, dictionaries, textbooks. This dialect has been studied from the nineteen century and it has historical variants (texts in Old Aromanian). It has also different regional variants. (see Figure 1)

However, there are several difficulties. The pronunciation is differently transcribed in different books, depending on the period in which they were collected and published. This dialect is also studied by the Greek linguists that have another set of graphical conventions for the transcription. The spelling interpretive conception, nearby the standard language, exists also on these specialists, also the intention to approach the Aromanian dialect to the standard language exists, especially since it is the only one who enjoys the existence of schools.
Aromanian is no option in any OCR software, neither in the automated programs for the translation, and various letters with diacritics are not recognized. For the moment, the editable text obtained by OCR is of poor quality and should be carefully corrected by specialists in South-Danube Romanian dialects. Probably we have to introduce the three South-Danube dialects as independent options in the OCR program, each with its lexicon, including the letters from all the transcription systems.

4.2 The Meglenoromanian Dialect

The Meglenoromanians are an ethnic group living in the Meglen region of Central Macedonia, Greece. This ethnic group is less numerous than the Aromanians. The researchers estimated their number at 20,000 persons in the nineteenth century. However, Thede Kahl 2006 estimated them at 5,000 persons; the negative demographic dynamic is evident. The minorities are not recognized in Greece, and in Turkey they are submitted to an assimilation and islamization process. There are no schools in the Meglenoromanian dialect. This idiom, having no written aspect, has no cult literature. However, literary folklore texts were published by many linguists. More collections were published by Pericles Papahagi 1902, Ion Aurel Candrea 1925a, 1925b, and Theodor Capidan 1925. There is only one cult publication, a brochure about silkworm rearing with the script adapted, and terms borrowed from Romanian. The Megleno idiom is endangered, it was entered in the UNESCO Red Book on endangered languages, the "Languages in grave danger" and UNESCO Atlas of Languages in danger in the world. (Atanasov, 2014). Unlike the Aromanians, who are mostly herdsmen, the Meglenoromanians are traditionally occupied by agriculture. They are not nomads, but sedentary and therefore this dialect suffered fewer external influences and was kept as a native language spoken in the family. In the Figure 2 there is a map of their localities, drew by T. Capidan

4.3 Short Presentation of the Istroromanian Dialect

Istroromanian is an Eastern Romance idiom spoken in a few villages in the peninsula of Istria, in Croatia. The number of Istroromanian speakers is more than 500, the "smallest ethnic group in Europe". In the eighteenth century the number of speakers was 10,000, and many toponyms with origin in Istroromanian dialect demonstrate this fact. Part of speakers are migrants in Europe, USA, Canada or Australia.

It is listed among languages that are "seriously endangered" in the UNESCO Red Book of Endangered Languages. Since 2010, the Croatian Constitution recognizes Istroromians as one of 22 national minorities. However, there have not been significant changes in preserving their language, culture and ethnic identity. (see Figure 3).

Given the fact that Istroromians have long been in a gradual process of assimilation, and their language was not used in writing, it is strongly influenced by the Croatian language, and there are no documents to be processed, except a small number of texts collected by linguists: Feresini (1996), Pușcariu (1906), Cantemir (1959). Currently there are some rescue actions for the preservation of Istroromanian language, carried out by cultural associations.
A digitization of all the published texts in Istroromanian and their carefully annotation is necessary. There are some lexicons to be computerized and linked to the dictionaries of the other South-Danube dialects and to the Romanian computerized dictionaries. In the table 1, some differences between the Romanian dialects are exemplified. We ignored here the different diacritics of letters.

5 Discussion

5.1 The Linguistic Correct Definition of the Variants of the Romanian

The linguistic definition for the concept of dialect is that this variant of the language is quite different from the language standard, and mutual understanding between the speakers of the dialect and the speakers of the standard language is quite difficult or even impossible. In cases where differences are small and the understanding between speakers is easy, these are not different dialects, but simple regional variants. Such is the case of the language spoken in the Republic of Moldova, sometimes referred to as an independent language, or as a different dialect of the Romanian, for no linguistic reasons. However, to demonstrate this truth we need big corpora in the language spoken in the Republic of Moldova and in the South-Danube dialects, all morphologically and syntactically annotated, and then we can statistically calculate the indices of the approach or departure from the standard Romanian of each of these idioms. The creation of a big Romanian corpus with morphological and syntactic annotations, illustrating all the geographical and historical variations of this language is a long time project, which will continue probably since 2020 or more years. The corpus had counted in 2014 only 4,600 sentences in Romanian Standard (Perez, 2014).

5.2 Aromanian: A Dialect or an Independent Idiom?

We can establish with scientific arguments if this dialect has the tendency to become an independent natural language or to remain a dialect of Romanian, without imposing any solution, but only to ascertain if, as some experts have said, Aromanian is a language independent appertaining of the group of Romance languages.

The national consciousness of speakers, resulting on their texts, is also important for establish whether it’s an independent idiom. Aromanians are divided into three main groups: the first, living in northern and central Greek Macedonia, called Gramushtenians, the second living in the Pindos mountains, called Pindenians, the third living in the south of Epirus, and in Thessaly, called Farsherots. However, the speakers of this dialect have different conceptions about their nationality. According with the testimonies of early researchers, the first two groups consider that they are Aromanians, but the Farsherots consider that they are Romanians. By studying texts collected more soon we will see if these conceptions of Aromanians about themselves were kept or changed.

6 Future Work

The corpus which we intend to create can be useful to create computerized dictionaries for the Romanian dialects aligned with those of standard Romanian language (and of Romanian Word Net (RoWN), aligned at Princeton Word Net (PWN), and to build a machine translation system for these isolated idioms, to introduce them in the international circulation. The cultural and linguistic isolation of these speakers will cease, if a platform, such as Babel, will be able to translate into these dialects the information that the speakers find on the Internet in a well-known language.

6.1 Study of the Historical Variations and of the Evolution Tendencies

Another important utility of the corpus that we begin building now is the statistic demonstration of the evolution tendencies. The Aromanian is the most important dialect of the Romanian after the standard language, called the Dacoromanian dialect. The comparative study of Aromanian texts from different historical periods with statistical methods will allow us to know what is their evolution tendency, whether the trend of this idiom
Table 1: Differences between few words in the Romanian Dialects

| Istro-Romanian | Aromanian | Megleno-Romanian | Romanian | English |
|---------------|-----------|-----------------|----------|---------|
| klieptu       | cheptu    | klieptu         | piept    | chest   |
| bire          | ghine     | bini            | bine     | well, good |
| bliera        | azghirari | zber            | zbiera   | to roar |
| filiu         | hilj      | iliu            | fiu      | son     |
| filia         | hilje     | ieru            | fier     | iron    |
| flier         | heru      | vițal           | vițel    | calf    |
| vițelu        | yilal     | ghiarmi         | vierme   | worm    |
| (g)lierm       | iermu     |                 |          |         |

is approaching the Romanian standard language or the Greek language, whether it becomes an independent idiom or not. We plan to compare parts of our corpora using various comparison methods in order to better understand their similarities and dissimilarities.

7 Conclusions

In this paper we described an ongoing work on the creation of a big balanced corpus of Old and Regional Romanian texts, impossible without creating tools for its processing. Consequently, we described some such as tools used for the development of our corpus in several directions.

The balanced corpus does not have only a scientific interest, but also practical consequences. If the language spoken in the Republic of Moldova is easily understandable by Romanian speakers without consulting a bilingual dictionary, it demonstrates that the Moldavian is not a dialect of Romanian. However, in the case of South-Danube dialects, there are dictionaries and also final lexicons in the published books, that being necessary in order to allow Romanian readers to understand the published texts. That is proof that they are dialects and there cannot be any mutual understanding between their speakers.

References

1902. Meglenoromâni. Folk Ethnographic Study. Socec & Co publisher, Bucharest.

Peter Atanasov. 2014. The current state of meglenoromâni. megleno-romanian, an endangered idiom. Memoria Ethnologica XIV(52-53):3037.

Matteo Bartoli. 2025. Introduzione alla neolinguistica: (principi-scopi-metodi). 2. Biblioteca dell’ Archivium Romanicum - Serie II: Linguistica. https://books.google.ro/books?id=2g0MAAAAIAAJ.

Lars Borin, Marcus Forsberg, and Dimitrios Kokkinakis. 2010. Diabase: Towards a diachronic blark in support of historical studies. In Proceedings of LREC., pages 35–42.

Lars Borin and Markus Forsberg. 2008. Something old, something new: A computational morphological description of old swedish. In LREC 2008 - Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008) Conference Proceedings. pages 9–16.

Tânase Bujduveanu. 2005. The Sărăcăcians. Aromanian Book Publisher, Reading, Massachusetts.

Ion Aurel Candrea. 1925a. Meglenoromânian texts. Speech and Soul I/I:261285.

Ion Aurel Candrea. 1925b. Meglenoromânian texts. Speech and Soul I/II:100128.

Traian Cantemir. 1959. Istroromâni Texts. Romanian Academy Publisher, Institute of Linguistics Cluj.

Theodor Capidan. 1925. Meglenoromâni. Their History and Their Language. National Culture / Romanian Academy, Bucharest.

Theodor Capidan. 1928. The Meglenoromâni. Their Folk Literature. Socec & Co publisher, Bucharest.

Cătălina Mărănduc Radu Simionescu Cenel-Augusto Perez. 2015. Ro-paas a resource linked to our uaic-ro-dep-treebank. In Advances in Artificial Intelligence and Soft Computing 14th Mexican International Conference on Artificial Intelligence, MICAi 2015. pages 29–46.

Cătălina Mărănduc Radu Simionescu Cenel-Augusto Perez. 2016. Social media processing romanian chats and discourse analysis. Computacin y Sistemas 20(3):404–414.

Svetlana Cojocaru, Alexander Colesnicov, and Ludmila Malahova. 2017. Digitization of old romanian texts printed in the cyrillic script. In Proceedings of International Conference on Digital Access to Textual Cultural Heritage. pages 143–148.
Mark Davies. 2010. Creating useful historical corpora: A comparison of corde, the corpus del espanol, and the corpus do português pages 137–166.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400 million word corpus of historical american english. Corpora (7):121–157.

Tomaz Erjavec. 2004. Multext-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In Proceedings of the Fourth Conference on Language Resources and Evaluation, LREC2004. http://nl.ijs.si/ME/.

Nerina Feresini. 1996. Il Comune Iстро-romeno di Val-darsa. Edizioni Italo Svevo, Trieste.

Daniela Gifu and Radu Simionescu. 2016. Tracing language variation for romanian. In Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics, CI-CLing.

Wlodzimierz Gruszczynski and Maciej Ogrodniczuk. 2015. The electronic corpus of the 17th and 18th century polish texts (up to 1772) aims, methods, current state, problems and prospects for development. In Slavic Corpus Linguistics: The Historical Dimension. pages 21–25.

Johan Hall, Joakim Nivre, and Jens Nilsson. 2006. Discriminative classifiers for deterministic dependency parsing. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL). pages 316–323.

Thede Kahl. 2006. The islamisation of the meglen vlachs (megleno-romanians): The village of nănti (nőtia) and the năntinet in present-day turkey. Nationalities Papers 34(1):71–90. https://doi.org/10.1080/00905990500504871.

Cătălina Mărânduc. 2010. The Dictionary of Romanian Expressions, Syntagms and Phrases (DELS). Corint Publishing, Bucharest.

Cătălina Mărânduc, Ludmila Malahov, Cenel-Augusto Perez, and Alexandru Colesnicov. 2016. Rodia project of a regional and historical corpus for romanian. In Proceedings of MFOI. pages 268–284.

Cătălina Mărânduc, Cătălin Mătielu, and Radu Simionescu. 2017. Parsing romanian specialized dictionaries structured in nests. In Proceedings of Digital Access to Textual Cultural Heritage (DATECH). pages 35–41.

Matilda Caragiul Marioțeanu. 2006. Aromanians and Aromanian Dialect in Contemporary Consciousness. Romanian Academy Publisher, Bucharest.

France Martineau, Constanța Rodica Diaconescu, , and Paul Hirschbühler. 2007. Le corpus voies du français : de lelaboration à lannotation pages 121–142.

Manuela Nevaci. 2011. The Gray of the Frontier Aromanians in Dobrogea. Cartea Universitară Publishing House, Bucharest.

Gheorghiad Mihail Obedenaru and Ioan Bianu. 1891. Macedoromanans Texts. Fairy Tales and Folk Poems from Crushova. Carol GőlPublisher, Bucharest.

Cenel-Augusto Perez. 2014. Linguistic Resources for Natural Language Processing. (PhD thesis). Al. I. Cuza University, Iași.

Iosif Popovici. 1909. The Romanian Dialects in Iстria. Halle.

Sophie Prevost and Achim Stein. 2013. Syntactic Reference Corpus of Medieval French (SRCMF). ENS de Lyon; Universitett Stuttgart; Lattice, Lyon/Stuttgart/Paris.

Sextil Puşcariu. 1906. Istroromanian studies. in collaboration with m. bartoli, a. belulovici and a. byhan, vol. i. texts. Romanian Academy Annals series II,(tom. XXVIII):117–182.

Nicolae Saramandu. 2003. Aromanian and Meglenoromanian Studies. Ex Ponto Publisher, Constanța.

Richard Sârbi and Vasile Frățilă. 1998. Istroromanan Dialect. Texts and Glossary. Amarcord publisher, Timișoara.

Radu Simionescu. 2011. Hybrid pos tagger. Proceedings of Language Resources and Tools with Industrial Applications Workshop Eurolan 2011 Summer school.

Achim Stein. 2008. Syntactic annotation of old french text corpora. Corpus (7):157–172.

Ana Maria Timofciuc, Daniela Gifu, and Corina Forăscu. 2013. A simple reliable web application for public discourse analysis. In Proceedings of the International Conference on Intelligent Information Systems (IIS). pages 158–162.

Nuria Yanez-Bouza. 2011. Archer past and present (1990-2010). ICAME Journal (35):205–236.