Reconstructing the Diachronic Morphology of Romanian from Dictionary Citations

Dan Cristea[1,2], Radu Simionescu[1], Gabriela Haja[3]
1 Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași
2 Institute for Computer Science, Romanian Academy, the Iași branch
3 "Alexandru Philippide" Institute of Philology, Romanian Academy, the Iași
E-mail: radu.simionescu@info.uaic.ro, dcristea@info.uaic.ro, gabihaja@yahoo.com

Abstract

This work represents a first step in the direction of reconstructing a diachronic morphology for Romanian. The main resource used in this task is the digital version of the Romanian Language Thesaurus Dictionary (eDTLR). This resource offers various usage examples for its entries, citations extracted from old and modern Romanian texts. The concept of "word deformation" is introduced and classified into more categories. The research conducted aims at detecting one type of such deformations occurring in the citations – changes only in the root of the old form words, without the migration to another paradigm. An algorithm is presented which automatically infers old root forms, and which is based on a paradigmatic data model of the current Romanian morphology. Having the inferred roots and the paradigms that they are part of, old flexion forms of the words can be deduced. Even more, by exploiting the chronology of the citations, the inferred old word forms can be framed in certain periods of time, finally configuring an important linguistic resource for researchers interested in the evolution of the Romanian language.

Keywords: morphology, diachronic, language evolution

1. Morphological sources for Romanian language

eDTLR[1] (Cristea et al., 2007) is the digital version of the Romanian Language Dictionary (DLR), edited by Romanian Academy, between 1906 and 2010. Apart from XML representations of the entries, eDTLR includes also part of the sources that have been used to build the corpus of citations, in digital form, and the software to access them. The Dictionary basically describes, following the lexicographic norms of the Academy, all words registered in documents and texts (from *Scrisoarea lui Neacșu /Letter of Neacșu*, 1521, the first known text in Romanian, until today). It includes etymology and each word sense is illustrated by quotations from a large collection of texts, attributed to all social and cultural domains (2500 titles and approx. 3000 volumes).

The morphological variation in the evolution of Romanian is mirrored in the rich collection of citations that eDTLR includes (more than 1.3 million). Richly sensed words could display tenths of pages in the original paper dictionary (for instance, 100 pages for a verb like *a veni/to come*. Moreover, the citations cover all historical periods in the evolution of written and spoken Romanian language (Rosetti, et al., 1968; Gheție, 1977; Gheție and Chivu, 2000), which makes them extremely valuable as a source of data in the attempt to reconstruct a diachronic morphology. Each citation includes exactly one occurrence of the title word. Moreover, citations are paired with codes identifying uniquely the source document and the pages from where it has been extracted.

[1] Built between 2007 – 2010, in a project financed by Romanian Government and coordinated by UAIC-FII (https://consilr.info.uaic.ro/edtlr/wiki/index.php?title=Digitalizing_the_Thesaurus_Dictionary_of_the_Romanian_Language)

An external database, called `chronology`, has been compiled, as pairs code-year or code-interval, where the year/interval are publishing dates of the source. As such, a certain morphological form of the title word can be precisely located in time.

AnaMorph (Cristea, Forăscu, 2006) is a paradigmatic word flexing instrument for Romanian. It sees a word as a lexical unit made up of two morphemes, a root and an ending. In its morphological variations, a word can have more roots, as given by the irregularities in declination or conjugation. There are mainly two causes of these irregularities: inheritance of old forms and phonetic alternations. The number of roots, the complete set of endings and the association of different roots with endings in flexing, assembles a paradigm (Tufiș, 1989). Usually, a paradigm is shared by a class of words having the same part of speech. In AnaMorph, the paradigms have been defined manually, following a grammar of modern Romanian. As such, 366 paradigms, which include 150 sets of endings, completely cover the morphology of nouns, verbs and adjectives of the contemporary Romanian, as given by DEX (in its online version[2]).

2. Going back in time

If we compare the language spoken or written today with that of the first quarter of the previous century we get fewer differences than between the today Romanian and that of the middle 19th century. The more we go back in the past, the bigger the differences are. But this can be taken also in the sense that we expect to find more common word forms between today's Romanian language and the one spoken 75 years ago than between today Romanian and the one spoken 160 years ago. Even

[2] www.dexonline.ro

more, changes are not abrupt, affecting the whole vocabulary at once, but merely involve the class of words belonging to the same paradigms and sometimes only isolated words. Mainly, at one moment in time or over a certain interval, one paradigm gradually changed. Very rarely, abrupt changes may also occur, in which case they are mainly issued by rules imposed by the Romanian Academia and which were gradually adopted by the society[3].

The research presented in this paper aims at inferring old forms and associate them with certain periods of time, based on the set of examples contained in eDTLR and the use of the `chronology`. Then, the timing associated with language changes is used to put in evidence phenomena related to the evolution of the Romanian language.

Since citations are paired with years/intervals, this task seems straightforward. Still, two things do complicate it very much: the recognition of the morphological tags of the occurrences of the title word in the citations and the fact that more forms could have been in use in the same moment or over the same period.

We have detected four ways in which a word can change its paradigm over the time:

- the word underwent changes in one or more of its roots;
- the word migrated to another paradigm;
- the word is a noun and changed its grammatical gender;
- a combination of the above deformations.

This research deals only with detecting and inferring the forms which underwent a root change. For our study, we have taken into consideration only the forms which are not present in the morphologic dictionary of the current Romanian language and can be obtained in conjugation or declination from a known lemma (for which a paradigm is known).

In the present study we have considered only nouns, adjectives and verbs (the three categories with the richest morphology) in Romanian.

3. The algorithm

In the following, we refer to a word as being "known" if it is present in the morphological dictionary of the current Romanian language. The occurrence of the title word in the citations is detected imposing a one-occurrence-of-title-word-per-citation restriction, and making use of a variation of the Levenshtein distance.

Given a known title word (a lemma) *l*, framed under the modern paradigm *p*, and an unknown flexion form *f* which is extracted from a citation which corresponds to *l*, assume that *f* is an old form of a root deformation of *l*. Next, verify the assumption made. Determine if *f* can be framed under *p* and if so, infer a root and its flexion forms. Solving such a problem is required when, given a title word and an old flexion form extracted from one of its citations, we want to establish if this old form is a root change: it is part of the same paradigm as the title word but there is a change in the root.

We define *s(p)* as the list of suffixes indicated by the paradigm p.

To determine if *f* can be framed under the paradigm *p*, for every suffix *s(p)[k]* that matches *f* at the end we assume that *f* might appear from a deformed root plus the suffix *s(p)[k]*. By trimming each such matching suffix from the form, we create a set of candidate roots *R(f,p)*. This is used in the context of having a title word

Next, the validation phase follows. For each candidate root *R(f,p)[i]* generate a list of fictive flexion forms *F(R(f,p)[i], p)* by attaching the suffixes imposed by *p* to the candidate root *R(f,p)[i]*. Define a score for *R(f,p)[i]* as the number of automatically generated (above called *fictive*) flexion forms in *F(R(f,p)[i], p)* which are detected in any of the eDTLR citations or in the sections dedicated to morphological specifications. If none of the candidates have a score higher then 0, then conclude that *f* cannot be framed under paradigm *p*. Otherwise, conclude that the root having the best score, *R(f,p)[j]*, is an old deformed root. The forms F*(R(f,p)[j], p)* can now be inferred and morphologically classified due to the data model of the paradigms, which associate a part of speech for each suffix that they contain.

Since the chronology of the citation can be mapped to all words belonging to it, the inferred forms after applying the root changing algorithm, once detected in some citation, become automatically attributed to the time/period of the citation.
Some details of the algorithm, hidden in the short presentation of above, can best be understood following the examples below. But first, a short description of the paragraphs which specify morphological variations for words is given.

When a fictive form is searched in eDTLR, it is looked up in the citations, and also in paragraphs which specify morphological variations for words. Such paragraphs have a somewhat standardized format. Below is a sample for the word *dator* (the formatting was kept exactly like in

[3] Romania being rather a conservative and stubborn society, sometimes the rules imposed by Romanian Academy, the only forum that has the right to impose changes in the official orthography, are not obeyed by everyone. For instance, the 1993 new orthography regulations have divided the society in two currents: those accepting to use *â* in the inner position and those insisting to keep the old written form *î* (among other details). At least, the language taught in schools is always conformant with the academic regulations.

the source):
-óare, dătór, -oáre, (învechit) **datóriu, -oáre, datúr, -úrie, deatóriu, -ie, dătóriu, -ie,** (regional) **deatór, -oáre** (ALR SN I h 1 006/95), **ditór, -oáre** adj., subst. - Lat. **debitorius, -a, -um** (după **da**[4]).
To look up a fictive form in such a source, the words that it contains must be extracted first. For many word forms, this format specifies only their ending. A parser was made to extract complete words from this. It was not a trivial task. This format was not developed for computer parsing. It was developed for the average human reader. For a native Romanian speaker it comes natural to understand the flexion form of a word, given a suffix. But from a computational point of view, to obtain complete word forms from this format is quite a delicate problem. The suffix must be attached at the end of the first complete form from its left. *–urie* for instance must be attached to *datur*. This attachment has its own rules. Because the ending *ur* matches in the beginning of *urie* only *ie* must be attached to *datur* to form the complete word *daturie*. And of course, there are exceptions to the rules. For example, the correct way to attach *–oare* to *deator* is *deatoare*. Such exceptions were implemented using a map of common endings which require special handling.

Example 1: title word *dansa* (verb *to dance*)

The paradigm this word is part of accepts the following suffixes: *dans*-{*a am ai aţi au asem aseşi ase aserăm aserăţi aseră ează ez ezi ează ăm aţi eze aşi ă arăm arăţi ară ând ându at ată aţi ate*}. For example *dansa* has the root *dans* and the suffix *a*, which is associated with the infinitive form (homonymous with past simple third person singular for this particular paradigm).

The word *dănţată* (past participle) has been found in a citation under the *dansa* title word. In this case, two suffixes match: *ă* and *ată* so there are two candidate roots: *dănţat* and *dănţ*.

The validation of the candidates goes as follows:

- The *dănţat* root generates the forms: *dănţata dănţatam dănţat dănţatai dănţataserăm dănţatezi* etc. In counting the occurrences of the generated forms in the dictionary we, of course, do not include the original form (in our case *dănţată*). There has been no match found, so the score was 0;
- Out of the *dănţ* root, the generated forms are: *dănţa dănţam dănţau dănţând dănţaserăm* etc. Leaving out the occurrences of the original form, two of the generated forms are found in eDTLR, which gives a score of 2 (these forms are *dănţat* and *dănţând*).

Since one root yielded a score higher than 0, the root *dănţ* is considered a deformed root and inserted in the diachronic morphologic dictionary under the same paradigm as *dansa*, so all its flexion forms will also be generated. Out of the three occurrences of the forms *dănţată*, *dănţat* and *dănţând*, only one belongs to a citation (the original form *dănţată*), the other two being examples of form variation, and the `chronology` indicates the year 1854 for this form.

Example 2: title word *dator* (adjective *indebted*)

The word from our previous example, *dansa*, has only one root for all its modern and old flexion forms. In this example we illustrate an adjectival paradigm which accepts two roots, each with its own suffixes.

For the title word *dator*, the occurrence *deatori* (masculine plural with no definiteness) has been found in one of its citations. The paradigm for *dator* accepts two groups of suffixes, each one in combination with a different root:

- *dator*-{*Ø ul ului i ii ilor*}
- *datoar*-{*e ea ei ele elor*}.

We noted with Ø the empty ending.

The matching of the *deatori* form against the endings in the two groups, succeeds only in the first group in the positions 1 and 4. The matches: *deatori-Ø* and *deator-i*, trigger, respectively, two candidate roots: *deatori* and *deator*. However, the paradigm of the modern word *dator* imposes all endings of the first group be combined with the same root. As said, we will oblige all virtual form to stick to the restrictions of the modern paradigm. Therefore, for the *deatori* candidate root, the virtual generated forms out of the endings in the first group are: {*deatori, deatoriul, deatoriului, deatorii, deatoriii, deatoriilor*}. Out of these, *deatorii* is found once in the dictionary. Secondly, for the *deator* candidate root, the corresponding virtual forms would be: {*deator, deatorul, deatorului, deatori, deatorii, deatorilor*}. This time two different forms are found: *deatorii* and *deator*. For the reasons explained, forms like, for instance, *deatori-elor* won't be searched for when validating.

In this case both candidates yielded scores greater than 0, still the one with the best score is considered as the actual root of the deformed version of this word, and that is *deator* – which is correct actually. Moreover, let's notice that the only form among those generated from the root *deatori* which had one occurrence in the dictionary is also among the forms derived from the second root, *deator*, and this ensures that the solution is safe.

But what would have happened if the form *deator* wouldn't have been found in the dictionary? This would be a case of equal scores, which is resolved in the benefit of the shorter root. This heuristic seems to guess in most of the cases the correct root. Of course, when a root is inferred incorrectly, a set of incorrect old, deformed flexion forms, are expected to be inferred incorrectly.

Example 3: title word *deschide* (verb *open*)

This example is a more complex one, for which the

shorter root heuristic happens to be applied. The flexion form *dăşchise* was found in a citation.
The suffixes which are accepted by the paradigm of *deschide*, grouped by different roots, are :

- *deschid-*{ *e eam eai ea eaţi eau eţi Ø em ă*};
- *deschi-*{*sesem seseşi sese seserăm seserăţi seseră sei seşi se serăm serăţi seră s să şi*};
- *deschiz-*{*iând ându*}.

There are three endings which match at the end of *dăşchise*: *Ø*, *e* and *se*. They belong to two different groups of suffixes. They generate the following candidate roots, with the corresponding virtual forms (which will be looked up in the entire eDTLR):

1. *dăşchise* (trimmed *Ø*) with the virtual forms: *dăşchisee dăşchiseeam ... dăşchise dăşchiseem dăşchiseă*
2. *dăşchis* (trimmed -*e*) with the virtual forms: *dăşchise dăşchiseam ... dăşchis dăşchisem dăşchisă*
3. *dăşchi* (trimmed -*se*) with the virtual forms: *dăşchisesem dăşchiseseşi... dăşchis dăşchisă dăşchişi*

The first root yielded a score of 0, because none of the virtual forms were found anywhere in eDTLR. For the second root, *dăşchis-Ø* and *dăşchis-ă* were found (the representation *–ă* is used to illustrate the manner in which these virtual forms were generated) resulting in a score of 2. The third root also yielded a score of 2 because the virtual forms *dăşchi-s* and *dăşchi-să* were found. The forms which determined the score for the second and third candidate roots are the same. It is an unfortunate coincidence that such a situation occurred. Not only the paradigm of *deschide* permits this, but also the citations and morphologic variations paragraphs from eDTLR used to validate this case contained only the two forms *dăşchis* and *dăşchisă*. Maybe if there was one more such old form present in eDTLR, it would have led to a clearer result.

Still, the shorter root heuristic applies resulting in the root *dăşchi* to be considered the correct one – which is true actually.

4. Results

The morphologic dictionary of the current Romanian language, which is used for determining if a word is "known" or not, contains a total of 1.15 million forms, corresponding to approx. 145,000 distinct lemmas.

The algorithm described above was applied for 41,911 entries (the letters D, P, S, V) out of the total of approximately 175,000 title words, as the whole dictionary contains. For these entries the dictionary includes 205,654 citations. We have found a total of 14,782 unknown flexion forms which have a known lemma. Out of them we inferred a total of 22,697 new flexion forms, by using 7,295 forms that were found in the entries as pilot forms (citations of the morphological specification paragraph). In total, we have classified morphologically 29,870 old, unknown words. The total number of new roots inferred was 2,705 for 1,938 known lemmas.

The algorithm relies on the shortest root heuristic in only 35 cases, which means 1.29% of the total cases.

We manually evaluated all the inferred roots for the nouns and adjectives only. The correctors were 20 master students in Computational Linguistics. Each student received a packet which contained random entries, where an entry is a word with one of its roots automatically inferred by the presented algorithm. The other roots were unknown. The correctors' job was to identify the roots which were inferred erroneously and also to type in the unknown roots.

Each entry was randomly distributed to 2 correctors. After the first phase of the correction, the contradictions between the pachets have been reveald to the students. In the next phase they discussed and negociated upon the correctness of their choises, and the number of contradictions decreased. At the end there were still some contradictions left. By counting the entries which were in the end considered correct by both students, we got a total of 2,064 correct inferred roots, out of the total of 2,120 entries. This represents a percentage of 97.36% for the case of nouns.

We have chosen to leave out the verbs from this evaluation/correction student project, because the task was considered too difficult and prone to many errors for subjects without proper linguistic training.

The total number of new roots manually typed in by the students and which proved to be consistent in different correction packets was 550. The number of roots which did conflict was, however, 181. The small number of roots inserted manually is explained by the fact that only a third (36%) of the nouns and adjectives contained more than one root.

The big ratio of conflict (24.76%) is explained by the fact that we were very much constrained by time in the second part (the confrontation part), when the negotiations between correctors had to happen. About half of the students didn't manage to contribute to this second part at all.

Guessing a root of an old word, is a tricky process and requires an extensive knowledge about the history of the language, provided that there are words which were written 400 years ago in the data given for correction/completion.

5. Conclusions

Determining the forms the words had over time, anchored in deformation of roots and the paradigmatic morphology, is the first step in inferring general rules of evolution of the Romanian language. Out of this study we aim to reconstruct the general trends that governed the evolution of Romanian language.

The following steps would be dedicated to the investigation of the other cases of variation of word paradigms, mentioned in the first section. After precisely defining the paradigms associated with each title word and the interval of time each paradigm had been in use, we intend to build chronological records of each title word, by arranging their paradigms on the time axis. Then we will correlate these chronological records in search for patterns of variation, with the intent to infer the rules that govern the language evolution.

Various resources will be built in the process, which could be used for creating fascinating tools, such as a diachronic part of a speech tagger, or a tool which would automatically predict the interval in which a text has been written.

6. Acknowledgements

The research conducted in this article was partially supported by the ICT-PSP projects MetaNet4U and Atlas projects.

7. Bibliography

Cristea, D., Forăscu, C. (2006). Linguistic Resources and Technologies for Romanian Language. In Journal of Computer Science of Moldova, Academy of Science of Moldova, Institute of Mathematics and Computer Science, vol. 14, nr. 1(40), pp. 34-73, ISSN 1561-4042.

Cristea, D., Răschip, M., Forăscu, C., Haja, G., Florescu, C., Aldea, B., Dănilă, E. (2007): The Digital Form of the Thesaurus Dictionary of the Romanian Language. In Proceedings of SpeD 2007 *Speech Technology and Human - Computer Dialogue*, Iasi, May 10-12, 2007.

Rosetti, A. et al., (1968 – 1973). Istoria literaturii române, București: Editura Academiei Republicii Socialiste România.

Gheție, I. (1977) (coord.) Istoria limbii române literare. Epoca veche (1532-1780), București, Editura Academiei Române.

Gheție, I., Chivu, G. (2000) (coord.) Contribuții la istoria limbii române literare. Secolul al XVIII-lea (1688-1780), București, Editura Academiei Române.

D.Tufiş. "It Would Be Much Easier If WENT Were GOED", in Proceedings of the 4th European Conference of the Association for Computational Linguistics, Manchester, 1989.