INFERRING DIACHRONIC MORPHOLOGY USING THE ROMANIAN THESAURUS DICTIONARY

RADU SIMIONESCU[1], DAN CRISTEA[1], GABRIELA HAJA[2,3]

[1]*Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi*

[2]*"Alexandru Philippide" Institute of Philology*

[3]*Romanian Academy, the Iaşi branch*

{ radu.simionescu ,dcristea }@info.uaic.ro; gabihaja@gmail.com

Abstract

This paper presents a first step towards constructing the diachronic Romanian morphology. First, the "deformation" of a word is introduced and a classification of such deformations is proposed. The conducted research aims at detecting deformations in the roots of inflectional words (nouns, adjectives and verbs). The algorithm we present uses two important resources: a morphological dictionary of the current Romanian language, which also models the inflectional paradigms of the language, and eDTLR – the digital version of the Romanian Thesaurus Dictionary. In eDTLR each title word has associated a set of citations extracted from the Romanian literature, each having attached the year of publication. The algorithm detects root deformations in words by comparing word forms of the current language with forms extracted from the eDTLR citations. For every root change, the deformed root is deducted and all the diachronic forms are inferred. Also, using the chronology of citations, for each diachronic root a period or a year is established. The research was conducted on 4 volumes of eDTLR. The algorithm successfully detected 2,700 root deformations and inferred a total of 30,000 diachronic inflexions.

1. Morphological sources for Romanian language

eDTLR[1] (Cristea et al., 2007) is the digital version of *Romanian Language Dictionary* (DLR), edited by the Romanian Academy, between 1906 and 2010, and including its sources in digital form and the software to access them. The Dictionary describes, following lexicographical norms, all words registered in written Romanian texts, starting with *Scrisoarea lui Neacşu* (*The Neacşu's letter*), 1521, the first known text in Romanian, until today. It includes the word's etymology and quotations extracted from a large collection of texts, attributed to all social and cultural domains (2,500 titles and approx. 3,000 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 entries could display tenths of pages in the original paper dictionary (100 for a verb like

[1] Built between 2007-2010, in a project financed by the 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)

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. An external database, called *the 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 stem (root) and an ending. In its morphological variations, a word can have more stems, 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 stems, the complete set of endings and the association of different stems with endings in flexing, assembles a paradigm. Usually, a paradigm is shared by a class of words having the same part of speech. In AnaMorph, part of the paradigms have been defined manually, following a grammar of modern Romanian. The rest of them were generated automatically using a morphologic dictionary provided by DEX (in its online version[2]). The total number of paradigms is now 366, which include 150 sets of endings, completely covering the morphology of contemporary Romanian.

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 language and the one spoken 75 years ago than between today Romanian and the language spoken 160 years ago. Even 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 issued by rules that Academia imposed and which were gradually adopted by the society[3].

The approach presented in this paper consists in analyzing the set of examples contained in eDTLR to infer old forms and associate them with certain periods of time. Then, to use these periods for evidencing phenomena related to the evolution of the Romanian language. Even more, by inferring old forms, a diachronic morphologic dictionary can be built, which could, for instance, be used for POS-tagging old Romanian texts.

[2] www.dexonline.ro

[3] Romania being rather a conservative and stubborn society, sometimes the rules imposed by the Romanian Academy, the only forum that has the right to impose changes in the official orthography, are not completely observed. 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).

Since citations are paired with years, this task seems straightforward. Still, two things complicate the problem: the difficulty to recognise the morphological features of the occurrence of the title word in a citation, and the fact that more forms could have been in use in the same period.

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

- the word suffered changes in one or more of its roots;

- the word migrated to another paradigm;

- the word is a noun and changed its grammatical gender;

- the word suffered a combination of the above deformations.

In this paper we will deal strictly with detecting and inferring the forms which suffered 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 (Levenshtein, 1966).

Given a known title word (a lemma) *l*, framed under the modern paradigm *p*, and an unknown inflexion form *f* (which is not found in the list of *l*'s inflexion forms), determine if *f* can be framed under *p* and, if so, infer a root and its inflexion forms. Solving such a problem is required when, given a title word and an old inflexion 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 something differs 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 be constructed 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).*

Next, the validation follows. For each candidate root *R(f,p)[i]* generate a list of fictive inflexion 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 fictive inflexion forms in *F(R(f,p)[i], p)* which are present in any of the eDTLR citations or in the section 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

for the root having the best score, *R(f,p)[j]*, as being 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.

The examples below will put in evidence other details of the algorithm.

3.1. The verb "a dansa" (to dance)

The paradigm the title word *dansa* is part of accepts the following suffixes: {*a, am, ai, ați, au, asem, aseși, ase, aserăm, aserăți, aseră, ează, ez, ezi, ează, ăm, ați, eze, și, ă, arăm, arăți, ară, ând, ându, at, ată, ați, ate*}. For example *dansa* represents the concatenation of the root *dans* and the suffix *a*, which is associated with infinitive (as well as the homonymous form in past simple third person singular, for this particular paradigm).

The word *dănțată* (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: *dănțat* and *dănț*.

The validation of the candidates gets on like this:

- the *dănțat* root generates the forms: *dănțata*, *dănțatam*, *dănțat*, *dănțatai*, *dănțataserăm*, *dănțatezi* etc. None of these, all different from the initial unknown form *dănțată*, can be found anywhere in eDTLR – so the score is 0;
- the *dănț* root generates the forms: *dănța*, *dănțam*, *dănțau*, *dănțând*, *dănțaserăm* etc. Leaving out the found unknown form, two of the generated forms are found in eDTLR (*dănțat* and *dănțând*) – so the score is 2.

Since the root *danț* is at the origin of a score higher than 0, it is considered a deformed root and inserted in the diachronic morphologic dictionary under the same paradigm as *dansa*, so all its inflexion forms can be generated. The publication years of the citations from which *dănțată*, *dănțat* and *dănțând* were found provide enough information so that their root can be associated with a period of time.

3.2. The adjective "dator" (indebted)

The previous example has a particularity, in that the verb *dansa* displays only one root for its inflexion. This example illustrates an adjectival paradigm which accepts two roots, each associated with its own suffixes.

For the title word *dator* (adjective), the form *deatori* (masculine plural indefinite) has been found in one of the citations. The paradigm of *dator* accepts the following suffixes, grouped by the two different roots: {*VOID, ul, ului, i, ii, ilor*} {*e, ea, ei, ele, elor*}. By the *VOID* suffix we indicate the empty string, which means that if the root is *dator* then it is also a inflexion form.

There are two suffixes which match the form *deatori*: the *VOID* suffix (always matches) and the *i* suffix. So, the two candidate roots are: *deatori* (by trimming the *VOID* suffix) and *deator* (by trimming the *i* suffix).

This time, the validation of the candidates is done somehow differently, compared to the previous example. For each root, the forms which are looked up in the dictionary are formed by adding only suffixes belonging to the same group (list of suffixes) as the one to which the matching suffix belonged.

For instance, *deatori-elor* won't be searched for when validating (won't be considered a fictive form). The candidate root *deatori* was found by subtracting the *VOID* suffix. The fictive forms of *deatori* are generated by attaching only the suffixes belonging to the same group as VOID, which doesn't contain *-elor*.

- *deatori* generates the forms: *deatori*, *deatoriul*, *deatoriului*, *deatorii*, *deatoriii*, *deatoriilor*. Out of these, one form is found in the dictionary: *deatorii*.
- on the other hand, *deator* generates: *deator*, *deatorul*, *deatorului*, *deatori*, *deatorii*, *deatorilor*. This time two different forms are found: *deatorii* and *deator*.

In this case, both candidates produced scores greater than 0, still the one with the best score is considered as the actual root of the old version of this word, and that is *deator* – which is true actually.

But what would have happened if the form *deator* wouldn't have been found in the dictionary? This would end in a tie, and in such a case the shorter root chosen. This heuristic, generally, seems to guess the correct forms.

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, of a total of approximately 175,000, as the whole dictionary contains), for which the dictionary includes 205,654 citations. We have found a total of 14,782 unknown inflexion forms which have a known lemma. Out of them we inferred a total of 22,697 new inflexion forms, by using 7,295 forms that were found in the entries as pilot forms (in citations or in the morphological specifications paragraphs). 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.

Since the second example mentioned also a number of heuristics, it means that in rare cases the algorithm can fail. When a root is inferred incorrectly, it triggers a set of incorrect old, deformed inflexion forms. In order to report statistical values about its accuracy, a manual evaluation was performed on all the inferred roots (for nouns and

adjectives only). The correctors were 20 master students in Computational Linguistics[4]. Each student received a packet which contained random entries, where an entry is a word with one of its roots being automatically inferred. The other roots are 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 packets have been revealed to the students. In the next phase they discussed and negotiated their choices in order to decrease the number of contradictions. In the end, still some contradictions remained. By counting the entries which were, in the end, considered correct by both students, we got a total of 2,064 correctly inferred roots, out of the 2,120 total entries. This represents a percentage of 97.358% for the case of nouns.

We chose to leave out the verbs from this correction project because we considered that manually filling in the unknown roots of old forms of verbs is going to be too difficult and time consuming for the time we had at our disposal.

The total number of new roots manually typed in by the students, which didn't conflict among correction packets was 550. The number of roots which did conflict was 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 conflicts (32%) is explained by the fact that we were very much constrained by time in the second part, when the negotiations between correctors had to happen. Even more, almost half of the students didn't manage to contribute to this second part at all. The experiment proved that guessing a root of an old word is a tricky process and requires an extensive knowledge about the history of the language, as in the corpus of citations given for correction/completion there are words which have been in use some 400 years ago.

5. Conclusions

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

The next step is to investigate also 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 has 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 of language evolution.

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

[4] at the Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi

Acknowledgements

The research reported in this paper was partially supported by the PSP-ICT project METANET4U.

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.

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.

Levenshtein V.I. (1966). "Binary codes capable of correcting deletions, insertions, and reversals". *Soviet Physics Doklady* 10: 707–10.

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