ROMANIAN ETYMOLOGICAL CHAINS – A PRELIMINARY ANALYSIS

RALUCA MOISEANU¹, DAN CRISTEA²

¹ Alexandru Ioan Cuza University of Iasi, Computer Science Faculty, Computational Linguistic Department
² Romanian Academy, Institute for Theoretical Computer Science;
{raluca.moiseanu, dcristea}@info.uaic.ro

Abstract

In this paper the origin of describe the preliminary steps towards a recursive reconstruction of Romanian words together with the positioning of their loans within a time frame, as reflected in the European Linguistic Thesauri. A pilot application accepts as input a Romanian word and accesses online linguistic resources, such as eDTLR – The Thesaurus Dictionary of the Romanian Language in electronic form, displaying etymological information. The etymology of a word is subsequently searched in foreign sources (for the time being only French and Italian online dictionaries), in order to compute its etymological trajectory. Import years, where available, are used to place on the time axes the approximate time of imports. The research intends to highlight a methodological framework on which a future real scale investigation could be anchored.

Keywords: etymon, online dictionaries, database, parser

1. Introduction

This project has been triggered by the need of having a dynamic and complex structured database able to provide the etymological information of any Romanian word (except the ones with unknown etymology). In our attempt to recreate the etymological chain of a word we shall, first of all, provide an insight of what etymology as a science is, as well as the main features of the Romanian etymology. Once the theoretical background is established we shall move on to the linguistic resources and technologies used to support the generation of etymological chains.

An etymological chain is a string of one or more etymons along with their origin language and entry year. As data structure, etymological chains are graphs (Alt, 2006) that have a root word in the studied languages (Romanian, in our case) and one or more descendants from source languages (Central and Eastern languages, in our case, with whom Romanian languages has had contact throughout the years).

The paper describes the beta version of the application used to automatically extract the information from online Italian and French dictionaries, version that has been tested on a number of 2000 XML files from the eDTLR – the Romanian Thesaurus Dictionary in electronic form (Cristea et al., 2007), corresponding to the same number of dictionary entries.

2. Etymology as a science

Derived from the Greek etymon meaning “true sense” and the suffix, logia, denoting “the study of”, etymology as a science studies the origin of words. Etymology considers words as having either an internal origin (therefore, in the target language, by applying transformation
rules specific to the lexicon or the grammar of that language, through affixation, compounding and conversion) or an external origin (through borrowings/loans from one or more languages). Regardless the acceptance channels, the etymology has to decipher the phonetic and morphological transformations from the original word to the actual word.

The Linguistic calque is to be situated at the border between the internal and external generation of words as the new words are formed within the source language by imitating an external structure. An etymon can come from two or more languages either during the same period of time or throughout different periods of time. This is called multiple etymology. Most of the Romanian words have multiple etymology, Latin being referred to as an indirect source. Romanian is a Romance language, belonging to the Italic branch of the Indo-European language family, having much in common with languages such as French, Italian, Spanish and Portuguese. However the closest to Romanian are the other Eastern Romance dialects, spoken south of Danube: Aromanian/Macedo-Romanian, Megleno-Romanian and Istro-Romanian dialects. An alternative name for Romanian used by linguistics to disambiguate with the other Eastern Romance languages is Daco-Romanian, referring to the area where it is spoken (which corresponds roughly to the onetime Roman province of Dacia).

Marius Sala et al (1988) considered 2581 words as being representative for the Romanian vocabulary. The etymological structure of this vocabulary is shown below:

- Romance elements 71.66%, out of which:
  - 30.33 % Latin
  - 22.12 % French
  - 15.26 % Classical Latin
  - 3.95 % Italian
- Internally formed 3.91 % (most from Latin etymons)
- Slavic 14.17 %, out of which:
  - 9.18 % Old Slavic
  - 2.6 % Bulgarian
  - 1.12 % Russian
  - 0.85 % Serbian-Croatian
  - 0.23 % Ukrainian
  - 0.19 % Polish
- German 2.47 %
- Neo-Greek 1.7 %
- Thracian – Dacian, a sub-layer, 0.96 %
- Hungarian 1.43 %
- Turkish 0.73 %
- English 0.07 % (and growing)
- Onomatopoeias 0.19 %
- Unknown origin 2.71 %

The data listed above has been used to establish the first two Latin languages of focus for this preliminary study.
3. Data collection

The collection of resources (online dictionaries) and simulation, trials of manually generated etymological chains represented the starting point of the project. The manually gathered etymological chains were also used as validators for the application (Burhui, 2013) that has been put together for the automatic generation of etymological chains.

The quest for online resources has proved itself rather sinuous as many of the online etymology dictionaries or online dictionaries did not display etymological data. For the purpose of this paper we have narrowed down the area of research to only Italian and French, which sum up (Sala, 1988) 26.07% of the representative vocabulary of the Romanian language.

Once we have identified the two online sources, for Italian – http://www.sapere.it, and for French – http://www.cnrtl.fr/etymologie, that seemed to best fit our purposes, we have extracted from these dictionaries a list of notations used to mark the etymon (such as: fr., fr.ant., it., ital., lat.class., lat.vulg., lat.mediev. etc.). This list has been included as an external resource onto the program.

What follows below is a list of examples of etymological chains, manually extracted from the two online dictionaries (Italian and French). In these examples, the details that we would want the application to return upon interrogation are also indicated: POS, gender, entry year, and source language.

- **bastard** s.m. din it. bastardo;

*IT. bastardo sec XV dal fr. ant. batard;*

*FR. batard 1150 l'orig. de bastard est obsc.;*

\[
\text{ro.bastard} \leftrightarrow \text{it.bastardo} \leftrightarrow \text{fr.batard} \leftrightarrow \text{unknown etymon;}
\]

- **ciment** s.n. sec XIX din it. cimento, fr. Ciment;

*IT. Cimento s.m. dal lat. Caementum;*

*FR. Ciment s.m. 1165-70 du lat. class. caementum;*

\[
\text{ro.ciment} \leftrightarrow \text{it.cimento} \leftrightarrow \text{lat.caementum}
\]

- **cortina** s.f. din it. Cortina;

*IT. Cortina n.f. dal lat. tardo Cortina;*

\[
\text{ro.cortina} \leftrightarrow \text{it.cortina} \leftrightarrow \text{lat.tardo.cortina;}
\]
- **paladin** s.m. din fr. paladin, it. paladino;
  
  FR. paladin s.m. 1552 empr. a l’ital. paladino;
  
  IT. paladino n.m. dal lat. mediev. palatinum;

- **sopran** s.m. din fr., it. soprano;

  FR. soprano 1768 du lat. vulg. superanus;

  IT. soprano dal lat. vulg. superanus;

- **vaccin** s.n. 1827 din fr. vaccin, lat. vaccinus, cf. it. vaccine;

  FR. vaccin 1801 du lat. vaccinu(s);

  IT. vaccino dal lat. vaccinu(m);

- **vagabond** s.m. 1795 din fr. vagabond, lat. vagabundus, cf. it. vagabond;

  FR. vagabond 1382 du lat. vagabundu(s);

  IT. vagabondo dal lat. vagabundu(m).
As shown in the above examples we have manually extracted the entry year, where available and also listed the etymons with unknown origins. Most of the above examples have double etymology, the etymon being both Italian and French, both pointing to Latin as being an indirect origin for the Romanian words.

From this early stage four types of etymological chains can already be seen:

- **type1**: root \(\leftrightarrow\) orig1 \(\leftrightarrow\) orig2 \(\leftrightarrow\) orig*

- **type2**: root \(\leftrightarrow\) orig1 \(\leftrightarrow\) orig2 \(\leftrightarrow\) orig3

- **type3**: root \(\leftrightarrow\) orig1 \(\downarrow\) orig2 \(\rightarrow\) orig3

- **type4**: root \(\leftrightarrow\) orig1 \(\leftrightarrow\) orig2 \(\leftrightarrow\) orig3

### 4. The application and a comparison with other approaches

A beta version of the application (Burhui, 2013) allows a user to input an entry Romanian word, out of which it generates one or more linear etymological chains. At this stage the application searches the entry in the Romanian lexicographic thesaurus (eDTLR) and, once found, it extracts the etymological information. If the etymological sources indicate a French or an Italian origin, it directs the search onto the corresponding French ([http://www.cnrtl.fr/etymologie](http://www.cnrtl.fr/etymologie)) or Italian ([http://www.sapere.it](http://www.sapere.it)) online dictionaries, parses the etymological information and displays it. The year of the import is filled in as the year of the first citation.

![Figure 1: The general architecture of the system](image)
A high level overview of the application design is shown in Figure 1. A graphical interface able to display the four schemes put in evidence in the previous section remains to be implemented.

Susan Alt (2006) describes the etymons as being words, located in time and space, which stand in a particular diachronic relation to other words, and etymological links as being the etymological relations between linguistic units. In her attempt to define a model of etymological structures she uses the TLFI (http://www.tlfi.fr) as the primary linguistic material to recover data.

The nodes of her graph are lexical entries in diverse lexicographic sources. In linear chains, the first entry is the anchoring word, the second one represents its direct etymon, the third one – the etymon of the first etymon, a.s.o. In case of compound words, her graphs diverge towards two entries, each one continued with their corresponding sources.

Alt pays a particular attention to the type of links between the entries (such as loan word relations or compound word relations). This type of information will be inserted also in our graphs once parsers would become refined enough to be able to distinguish this type of information in the source dictionaries.

5. Conclusions

Our preliminary manual investigations, as well as the first experiments done with the tool have brought to light a high number of entries with unknown or uncertain etymons, which can easily turn into the subject of some statistics drawn based on this project. Moreover, the attachment of the import dates makes it possible to detect some incorrectly dated etymons (which, as mentioned are extracted in our primary source from the date of the first mention of the imported word). Among the peculiar etymological chains that we have obtained during our manual trials we have stumbled across entries for which the first or second etymon entry year is subsequent to the one of the target language. What we believe to be incorrectly dated etymons would have to be validated against a collection of online dictionaries, rather than just one dictionary.

The main difficulty that we have faced so far is the lack of online resources that would contain the etymon and entry year as well, or the lack of online resources altogether (Bulgarian, Slavic, etc.). Among the resources that we have found so far (English, German, Spanish), differences in notation of the etymon in each dictionary makes the parsing challenging. However, in the future we aim to increase the number of online dictionaries accessed, the most wanted for studying the origins of Romanian being the German, Bulgarian, Russian, Turkish, Greek, English, Polish, Ukrainian, Hungarian and Latin dictionaries.

The language barrier is not to be neglected as well. The Greek, Turkish, Russian and Slavic dictionaries pose a real issue upon retrieving the required information.

Although the first steps have been done, the project is far from being completed and also raising more questions than solutions.

We believe that the research, only an inception of which is described in this paper, would rather convincingly motivate the birth of an international consortium that would look into the development of this project at European scale. Let’s note that similar initiatives have been suggested already for other languages: (Alt, 1996) for French or the Etymology explorer (http://roots.robestone.com) for English. Agreeing on some common conventions of notation
of etymological chains, sharing lexicographic resources, parsing technologies for dictionaries and the software that builds the etymological graphs itself, could result in a reconstruction of interchangeable etymological graphs that would configure more and more dense parts of a map of linguistic influences. Their correlation with historical events could bring into light new insights over cultural interferences, could correct errors and reveal unknown linguistic and historical facts.

Acknowledgments: We are grateful to Gabriela Haja, from the “A.Philippide” Institute of the Romanian Academy for coining the idea of etymological chains, and to Andrei Scutelnicu and Alin Placinta – Salaru, from the Computational Linguistics at the Faculty of Computer Science of the “Alexandru Ioan Cuza” University of Iași, and Anca Bibiri, from the Department of Interdisciplinary Studies of the same University, for contributing to the elaboration of software and for acquiring information about dictionaries.

References

Dictionaries:
eDTLR – The Thesaurus Dictionary of Romanian Language in electronic form

http://www.cnrtl.fr/etymologie;

http://www.sapere.it;

http://dexonline.ro