Towards an Etymological Map of Romanian

Alina Maria Cristea,[1] **Anca Dinu,**[1] **Liviu P. Dinu,**[1]
Simona Georgescu,[1] **Ana Sabina Uban,**[1,2] **Laurentiu Zoicas**[1]

[1]University of Bucharest
[2]Universitat Politècnica de València
alina.cristea@fmi.unibuc.ro, ancaddinu@gmail.com
liviu.p.dinu@gmail.com, simona.georgescu@lls.unibuc.ro
ana.uban@gmail.com, laurentiu.zoicas@lls.unibuc.ro

Abstract

In this paper we investigate the etymology of Romanian words. We start from the Romanian lexicon and automatically extract information from multiple etymological dictionaries. We evaluate the results and perform extensive quantitative and qualitative analyses with the goal of building an etymological map of the language.

1 Introduction

Located at crossroads between East and West, the Romanian language presents a kaleidoscopic etymological picture. Originated from Latin, it suffered the influence of many cultures with which the other Romance languages did not have (much or any) contact, hence its physiognomy became, from a certain point on, different from that of its cognate languages (cf. Niculescu (1965, 1978, 1999, 2003)). The Romanian lexicographers, having to deal with this miscellaneous etymological structure of the language, must perform a fairly complicated task which not rarely ends up by giving in to the difficulty of identifying a word's origin.

Our analysis, based on a computational systematization of the origins of words, aims to evaluate quantitatively and qualitatively Romanian's etymological composition. We propose a socio-cultural interpretation of the semantic domains most permeable to borrowings from the various languages with which Romanian had a stronger contact, considering that a systematic perspective on the lexicon's etymological structure, doubled by statistics on the permeability and needs of the various onomasiological fields, may provide clues for future research concerning still unknown etymologies.

1.1 Preliminaries. Peculiarity of Romanian vs. the Other Romance Languages

Apart from its genetically belonging to the Romance linguistic family, the Romanian language shares certain phonological, morpho-syntactic and lexical features with the Balkan languages, as a consequence of its geographical position. For this reason, it was also included in the so-called "Balkansprachbund" or "Balkan language area" (cf. Rosetti (1968)) , together with Eastern South Slavic languages (Bulgarian, Macedonian and Serbian), Albanian and Greek.

There are two significant differences between Romanian and the other Romance languages:

1. According to Sala et al. (1988), the fundamental lexical core of Romanian comprises less words inherited from Latin than the other Romance languages (Ro. 30% vs. It. 44%, Fr. 36%, Sp. 40%, Pt. 45%).

2. At the same time, while the Italo-Occidental Romance languages make use, in their basic lexicon, of at least 25% loanwords from Latin (It. 28%, Fr. 27%, Sp. 27%, Pt. 25%), the Romanian language only counts little more than 1% words borrowed directly form Latin; thus, even if we add the 8% Latin words borrowed via French and Italian, the most Eastern Romance language still does not reach the Occidental proportion of the "cultural superstrate" (cf. Reinheimer Ripeanu (2004)).

By combining these two components (inherited and borrowed words from Latin), considering their proportion in the representative lexicon of the Romance languages, we obtain It. 72%, Fr. 63%, Sp. 67%, Pt. 70%, while in Romanian the Latin element hardly reaches 32% (or 39% if we also consider the Latin words penetrated via French and

Proceedings of Recent Advances in Natural Language Processing, pages 315–323
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_037

Italian). A reason for this considerable etymological divergence could be, on the one hand, its late integration and early separation from the Roman Empire: conquered at the beginning of the 2nd century, Dacia was left unconnected with the Empire in the second half of the 3rd century). This could explain the lower proportion of inherited words. On the other hand, the different geographical context had a significant influence on the further development of the Romanian language, because, while the Italo-Occidental Romance languages were passing through a period of re-latinisation, massively borrowing words from Latin, the Oriental Latin descendant had strong contact with the Slavic, Greek and Turkish languages, all of which have left deep marks on the Romanian lexicon.

We must also briefly describe here another particularity of the Romanian lexicon, namely the external multiple etymology, defined as "the provenance of a single word from two or more lending languages, at the same time and on the same territory, or in different times and in different territories" (Celac, 2020). This situation resides in multiple internal and external factors that influenced the Romanian language simultaneously, especially during its modernization period (the 19th century). The "cultural loanwords" (i.e., words related to technology, science, cultural life, mostly corresponding to the international vocabulary items, cf. Moroianu (2015); Celac (2020)) could penetrate more or less at the same time from different source-languages, depending on the foreign language that was used as a source in the borrowing process. As the languages that were used as source for the neological enrichment of Romanian are multiple – besides French and Italian we also count Latin, Modern Greek, Russian and German –, it is not infrequent the case where a word has three, four or even five etymologies.

Moreover, one should take into account the dialectal fragmentation of Romanian before its cultural unification and standardization (starting not before the second half on the 19th century), which led to the same situation of multiple-source borrowing, depending on the contact language of each Romanian province: for example, the Romanian speakers in Moldavia would borrow from the Ukrainian language, while Southern Romania would use Bulgarian or Serbian as source-languages. Thus, one and the same Slavic word could have penetrated through different channels, which results as well in multiple etymology.

While the concept of "multiple etymology" is rather unusual for the other Romance languages, this peculiar situation being almost absent in the rest of the Latin descendants, the Romanian language has a significant number of lexical units borrowed more or less simultaneously from various sources, that reach a proportion of almost 18% of the fundamental lexical core (cf. Sala et al. (1988)).

This situation represents one of the main difficulties that Romanian lexicographers have to face. In our approach, we will provide a statistic of words having from one up to six etymologies. It goes without saying that the possibility of errors cannot be overlooked, as many lexicographers have also dealt with this particular Romanian lexical characteristic by placing at the same level several etymologies, whenever they were simply unsure about the immediate origin of a word.

1.2 Romanian Lexicography – A Brief Survey

In this section we offer a brief overview of the main resources one can use for etymological information concerning the Romanian lexicon. We also present the dictionaries we used for this research, explaining the reasons for our choices.

By comparing the lexicographical resources for Romanian with those created for the other main Romance languages (Italian, French, Catalan, Spanish and Portuguese), one can notice the absence of a substantial etymological dictionary of Romanian, equivalent to the lexicographic instruments we can use, for instance, for French (FEW (Wartburg, 1922–2002)), Catalan (DECat (Coromines, 1980–2001)) or Spanish (DCECH (Coromines and Pascual, 1980–1991)).

Despite various attempts to provide reliable etymological dictionaries, the results have been either incomplete (e.g., *Etymologicum Magnum Romaniae* (Hasdeu, 1886–1898), ceased at the letter B, or Candrea and Densusianu (1907–1914) – comprising only the words of Latin origin, besides not going further than the letter P), or not fully trustworthy (DER, cf. Hristea (2009)). The thesaurus dictionary of Romanian, DA (Puşcariu, 1913–1949) / DLR (Iordan, 1965–2010), is not consistent in the etymological descriptions it offers: while the first volumes, A-De and F-Lojniță (Puşcariu, 1913–1949), offer solid etymological descriptions, the remaining volumes (Iordan, 1965–2010) – reduce

the etymological explanations to a minimum. The ongoing project of a new complete etymological dictionary, DELR ((Academia Română, 2011–), covering so far the letters A-C), has been criticized not only for punctual shortcomings (cf. Celac (2012)), but for its whole design, being destined only to review the tradition of the etymological research on the lexemes (cf. Ernst (2013); Schweickard (2013)).

Somewhat more reliable sources for Romanian etymology, despite not having been designed to meet this purpose, but as explanatory dictionaries of the language, are Șăineanu (1929), Scriban (1939) and DEX ((Academia Română, 1996 [1975]), second edition *bis* 2009, second edition *ter* 2012).

Since one of the requirements for this research was the use of complete and consistent sources that are, at the same time, available online, we resorted to the following dictionaries, listed below in order of their priority: DEX '16, DEX '09, DER (Cioranescu, 1966), Scriban (1939), Șăineanu (1929), DEX '12, DEX '98, DEX '96, DEX '84, DEX '75, DEX-S (Academia Română, Institutul de Lingvistică din București, 1988), DN (Marcu and Maneca, 1986), DLRLV (Costinescu et al., 1987). The order of the sources in our analysis was shaped according to their relative reliability, which was established following the empirical observations regarding the accuracy of the data provided.

2 Extracting and Processing the Data

In this section we describe our procedure for automatically extracting and processing etymological information for the Romanian lexicon.

Dictionary	Match
DEX '09	– Din #fr.# @abat-jour.@
DN	[[...] / < #fr.# $abat-jour$]
Scriban	(fr. $abat-jour,$ [...])
DER	< #Fr.# $abat-jour,$

Table 1: Examples of different formats for representing etymological information in dictionaries covered by *dexonline* for the Romanian word *abajur* (meaning *lampshade*), which is borrowed from the French word *abat-jour*.

2.1 Data

We identify the etymologies and etymons of Romanian words using *dexonline*,[1] a machine-readable dictionary which aggregates information from over 30 Romanian dictionaries. Some of these are restricted by license and copyright, but others are publicly available. *Dexonline* provides the public data as an SQL dump, which we import in a local database server for querying.[2] By parsing the definitions from the etymological dictionaries listed in the previous section, we automatically extract information regarding words' etymologies. The definitions are partly formatted, with different formats for different dictionaries. We extract the relevant information using regular expressions. In Table 1 we provide examples of different formats for representing etymological information in *dexonline*.

When more options are possible for explaining a word's etymology, *dexonline* provides several hypothesis. We account for all the given alternatives, enabling our method to issue more accurate results, both when a dictionary considers a word to have multiple etymology (e.g., DEX '09 provides both French *vérisme* and Italian *verismo* as etymologies of *verism*, meaning "a literary and musical movement developed at the end of the 19th century") and when different dictionaries provide different languages of origin (e.g., DEX '09 provides Russian *koleaska* as etymology for *caleașcă* (meaning *carriage*), while Scriban provides French *calèche* as etymology for the same word).

We introduce the order of priority mentioned in Section 1.2 in case different dictionaries provide different etymons (or different orthographic forms of the same etymon) for a certain word and language of origin (e.g., DEX '09 provides French *abattis* as etymology for *abatiză* (meaning *abatis*), while DN spells the word *abbatis*).

In cases of homonymy, we take into account all the separate dictionary entries. By *homonyms* we mean words that have the same form, but different origins (e.g., *lac*1 meaning *lake*, and *lac*2 meaning *lacquer*; according to DEX '09, the first lexeme is inherited from the Latin word *lacus*, while the second one is borrowed from the German word *Lack*; the form coincidence derives from the historical phonetics of Romanian). All the values reported henceforth refer to words as conjunctions between a phonetic form and a conceptual content, taking into account their origin and history, and not only as raw word forms.

[1] https://dexonline.ro

[2] We use the database backup available on January 17, 2021.

2.2 Processing

We employ several post-processing steps for the etymological information, mainly for cleaning and normalization. For etymons, we keep both the processed forms and the original ones, for future reference. We provide below some processing rules along with motivations.

For extracted source languages:

- Grouping together different abbreviations for source languages used by different dictionaries (e.g., *tc*, *tur*, *turc*, *turk* all refer to Turkish).
- Conflating different periods of some languages (e.g., we group *vlat*, *mlat*, *nlat* – Old, Medieval, Neo-Latin under Latin), while keeping separated languages such as Old Slavic *vs* Slavic or Ancient Greek *vs* Neo-Greek.

For extracted etymons:

- Removing some diacritical symbols that mark the stressed syllable or vowel length and are not regularly rendered in the source language (e.g., Italian *abáte* becomes *abate* after removing the diacritical mark of the stress; Latin *abbattĕre* becomes *abbattere* after removing the diacritical mark of a short vowel, which shows that the stressed syllable is the antepenultimate).
- Replacing the rough breathing mark ‘ with the letter *h* for Greek etymons. This diacritical mark is rendered, in the transcription from Ancient Greek into Latin, by the letter *h*, and we apply the same transformation (e.g., the Greek etymon *‘omóphonos* of the word *omofon* (meaning *homophonous*) becomes *homophonos* after removing the stress mark and replacing the rough breathing mark).
- Removing endings for the oblique cases of Latin or Greek etymons (e.g., *marmor*, ***-oris*** for *marmură* (meaning *marble*), Neo-Greek *ároma, arómа**tos*** for *aromat* (meaning *aromatic*)) or secondary forms provided (e.g., Latin *adnotare,* ***annotare*** for *a adnota* (meaning *to annotate*), French *ballerin,* ***ballerine*** for balerin (meaning *ballet dancer*)).
- Removing the asterisk symbol that marks unattested etymons (e.g., the Latin etymon **conquerire* of *a cuceri*, meaning *to conquer*).
- Removing letters provided between round or square brackets. The former represent the spelling from the cultured language for Latin (e.g., *invol(u)tus*, etymon of *învolt*, meaning *abundant*), but Romanian inherited words do not originate from the cultured language. The latter have different meanings, such as reconstructing an intermediary form of the word (e.g., Latin *eccum-[i]lloc*), but in any case the information is not relevant for this study.
- Filtering out proper names, since they are not relevant for this study.

# etymologies	# words
6	4
5	25
4	209
3	1,675
2	9,923
1	37,051
0	45,357

Table 2: Number of words per number of automatically identified etymologies.

Source language	#words	#verbs	#adjectives	#nouns
French	35,511	2,533	8,219	23,610
Latin	9,313	1,203	2,215	5,302
Italian	3,358	384	471	1,960
German	2,767	73	300	2,331
English	2,064	41	253	1,700
Greek	1,754	1	380	1,141
Turkish	1,293	3	73	1,092
Slavic	1,155	236	86	803
Neo-Greek	1,026	54	51	836
Russian	896	9	62	777
Old Slavic	836	1	95	652
Bulgarian	650	60	33	533
Hungarian	622	50	35	472
Serbian	532	48	20	428
Ukrainian	270	19	10	235
Spanish	220	1	10	193
Polish	181	1	7	161
Ruthenian	151	3	6	124

Table 3: The number of Romanian words that originate from each source language. We report only the languages from which at least 100 words originated.

3 Quantitative Analysis

In Table 2 we report the number of Romanian words having zero, one or multiple etymologies identified automatically. 48,887 words out of a total of 94,244 words have at least one automatically

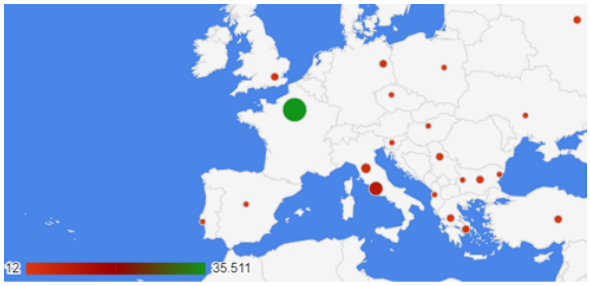

Figure 1: Map of the Romanian words' etymologies.

identified etymology and this set will constitute our data from now on. In Table 3 we report the number of Romanian words that originate from each source language and how many of these words are verbs, adjectives and nouns. In Figure 1 we illustrate the distribution of Romanian words of different etymology, by proportion mapping. The bigger the bullet, the more Romanian words originated from the language in that geographical region. Note that we do not dispose of dated etymologies and so we lack the diachronic dimension; thus, different languages or different evolution stages of the same language are represented on the same territory (e.g., the two circles from Italy represent Latin and Italian and the three circles located in present day Bulgaria correspond to Old Slavic, Slavic and Bulgarian).

We also compared the data we obtained with the information we knew concerning the fundamental lexical core from Sala et al. (1988), namely 7.5% French borrowings and 30% Latin inherited words. We easily notice not only an inverted ratio between the quantity of words of Latin origin and those originated from French, but also a hugely expanded proportion of French borrowings. About 38% of the whole lexicon and almost 73% of the words that have at least one etymology attested in the dictionaries have French origin (versus only 7.5% of the representative lexicon), while the quantity of words of Latin origin (most of them inherited) hardly reaches 9% of the whole lexicon and 19% of the words that have at least one etymology attested in the dictionaries (versus the proportion of 30% for the fundamental lexical core). The gap is explainable by the distinction between the basic, common lexical core (covering 80% of everyday speech) and the cultured lexicon and specialised terminology, developed in the last century by massively borrowing lexical items from French.

Sorting the borrowings of each language by parts of speech highlights the significant quantitative breach between the nominal parts (by far the majority) and the verbal ones. But, while the inherited lexicon shows a ratio of 1 verb to 6 nominal parts (noun+adjective), the borrowing process has considerably enriched the quantity of nominal parts of speech, in detriment of the verbal ones: e.g., the French borrowings encapsulate a ratio of 1 verb to 12 nominal parts, the English ones display a ratio of 1 verb to 48 nominal parts, and the Turkish loanwords enclose a correlation of 1 verb to 388 nominal parts. This situation allows a deeper insight into the language structure, showing that expressing an action, state or occurrence requires a higher degree of internalized lexicon or of acquaintance with the language: we can deduce, on the one hand, that the speakers are able to express their experiences by using a fairly small number of verbs, but need to constantly increase the amount of nouns to designate the new objects they observe or concepts they acquire; on the other hand, morpho-syntactic restriction may also play a part: while a nominal part of speech is easily adaptable to the morphological system of Romanian, the complex verbal conjugation may impede its immediate adoption. Also, it seems that the more related the source language is to Romanian, the easier is the morpho-syntactic adaptation of verbs, which might explain the above ratio order, French, English, Turkish.

By classifying the lexicon in parts of speech, we also notice a shortcoming in the Romanian lexicography, namely the inconsistency of lexicographers when establishing the period when a word of Slavic origin entered the Romanian language: from this categorization it results that only one verb was borrowed from Old Slavic, while more than 200 come from Slavic. It is, however, evident that many fundamental verbs of Slavic origin have penetrated during the period of early contact between the two communities, thus, originate from Old Slavic (e.g., *a iubi* (meaning *to love*), *a citi* (meaning *to read*), *a greși* (meaning *to make a mistake*)). In this case, we only highlight a terminological misunderstanding.

In order to quantify the resemblance between Romanian words and their etymons, for different source languages, we compute the edit distance (Levenshtein, 1965) for {word, etymon} pairs, using the post-processed etymon form (see Section 2.2). The edit distance counts the minimum number of operations (insertion, deletion and substitution) required to transform one string into another. We use a normalized version of this metric, dividing the edit distance by the length of the

longest string. The obtained values are between 0 and 1; the lower the values, the closer the Romanian words are to their etymons. In Figure 2 we report the average edit distance between the Romanian words and their etymons, per language. Overall, Romanian words borrowed from English are closest to their etymons. For 990 out of 2,064 words with English etymology, the edit distance is 0, meaning that those Romanian words have not undergone any transformations when entering the language (e.g., *marketing*, *management*, *avocado*). For Latin, 633 out of 9,313 words are identical to their etymon (e.g., *vultur*, meaning *vulture*).

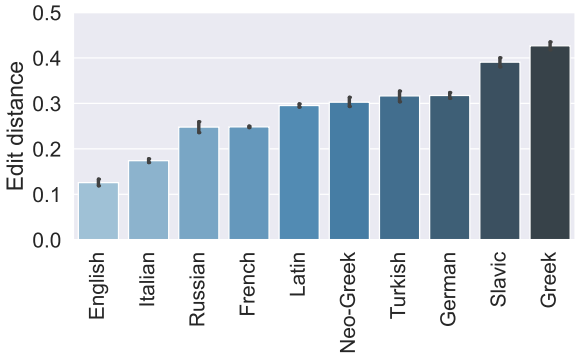

Figure 2: Average normalized edit distance between Romanian words and their etymons for the top 10 languages from Table 3.

.

4 Qualitative Analysis

4.1 Analysis of Lexicographical Errors

In order to evaluate our automatic method for extracting etymologies, we excerpt a sample of 1,000 words. We manually determine the etymologies of the words in the sample using the web interface of *dexonline*, we compare these results with the automatically obtained etymologies, and we report an accuracy of 99.2%.

The main error source is the recording of erroneous etymologies in the dictionaries. One of the most common errors is to consider the ultimate origin (either Latin or Ancient Greek) as the immediate etymology of a Romanian word, without taking into account its form and sometimes meaning, which point to a different source language. To take just an example, *apotropéu* meaning "magic remedy to ward off evil" is considered to be directly originated from Ancient Greek *apotrópaion*, but neither the form nor the meaning allows such supposition: on the one hand, it is not usual for a word borrowed as a proparoxytone (a word stressed on the ante-penultimate syllable) to become an oxytone (stressed on the ultimate syllable) in Romanian, on the other hand, the Greek word functioned as an adjective, *apotrópaios*, whose meaning *tutelary / expiatory / abominable* does not precisely match the Romanian significance. Nonetheless, if we take a look at the European modern languages, we can easily find the German lexeme *Apotropäum*, meaning exactly "magic remedy to ward off evil", as a term circumscribed to archaeology, both formally and semantically able to account for the Romanian word. Thus, it would be correct to indicate the German noun as the immediate origin of the Romanian word, and not the Ancient Greek adjective, with which it only has a distant connection.

A quite frequent error consists of almost automatically labelling a "cultural loanword" as French. For instance, the origin of Ro. *helipot* (meaning *helipot*) is attributed to a nonexistent French word "hélipot". Similarly, certain dictionaries invent a French word *acquisiteur* (for *acquéreur*) in order to explain Ro. *achizitor* (meaning *acquirer*). A similar example is that of Ro. *național* (meaning *national*), explained as a borrowing from Latin *nationalis* (to which the French word *national* is added, by virtue of the concept of multiple etymology). Nonetheless, the supposed Latin word *nationalis* is not documented in Latin, the concept being a modern one.

4.2 A Semantic Insight into the Romanian Lexicon's Structure

In this section we provide an analysis of the etymological composition of the Romanian lexicon based on semantic fields.

We start by building a list of conceptual domains, based on the Romanian linguistic atlases (Puscariu, 1938–1942; Petrovici, 1956–1972), which provide a list of semantic fields that covers the vocabulary, containing as well the most usual terms belonging to each of these onomasiological fields. We select a subgroup of these, and merge a few together, resulting in a final list of 10 semantic fields. We then manually extract a selection of prototypical terms for each of the resulted groups, on average 36 terms per group. We employ these terms as seeds for automatically populating the semantic clusters, using semantic similarity metrics based on word embeddings, a standard method for measuring lexical semantic similarity in the field of computational

Semantic field	Top languages
Agriculture	French, Latin, Italian, Slavic, Old Slavic, German, Turkish, Greek, English, Neo-Greek
Animals	French, Latin, Slavic, Italian, Bulgarian, German, Old Slavic, Greek, English, Turkish
Occupations/administration	French, Latin, Italian, German, English, Greek, Russian, Slavic, Neo-Greek, Turkish
Transportation	French, Latin, Italian, English, German, Greek, Russian, Slavic, Turkish, Neo-Greek
Time	French, Latin, Italian, German, Greek, English, Slavic, Old Slavic, Neo-Greek, Russian
Food & drink	French, Latin, Italian, German, English, Turkish, Slavic, Neo-Greek, Greek, Old Slavic
Domestic, clothing & hygiene	French, Latin, Italian, English, German, Greek, Slavic, Turkish, Neo-Greek, Old Slavic
Colors & patterns	French, Latin, Italian, German, Greek, Neo-Greek, Russian, English, Turkish, Slavic
Personality & emotions	French, Latin, Italian, Greek, Slavic, German, Old Slavic, English, Neo-Greek, Bulgarian
Education	French, Latin, Italian, Greek, German, English, Russian, Neo-Greek, Slavic, Turkish

Table 4: Semantic fields and etymologies.

analysis of semantic change. In our study, we make use of word embeddings computed using the FastText algorithm, pre-trained on Wikipedia for the top six languages Romanian borrowed from. The vectors have 300 dimensions and were obtained using the skip-gram model described by Bojanowski et al. (2016) with default parameters. These embeddings have previously been used in studies on semantic similarity of cognate sets in Romance languages (Uban et al., 2019, 2021). To group the terms in our dataset into the different semantic fields, we apply a KNN classifier (k=7) trained on the pre-defined list of semantic groups and prototypical terms. We then retrieve for each semantic cluster the distribution of etymologies for the words it contains. In Table 4 we show the top languages found in the etymologies of words belonging to each cluster.

One can easily observe in Table 4 that in 9 out of 10 semantic domains the first 3 source languages are invariably French, Latin, and Italian, precisely in this order. In 6 out of 10 onomasiological fields, the fourth position is occupied by a Germanic language (either German or English), while in 2 cases it is the Greek language holding this position. In 8 out of 10 domains, at least one Slavic language is represented among the first 8 source languages. It is also noteworthy that the Slavic (probably Old Slavic, see the comment above in Section 3) is the third most represented language in the onomasiological field of animals, and the fourth in the domain of agriculture and fifth in personality / emotions. The Turkish language reaches its highest position (the sixth) in the semantic field of food and drink, which reflects the predominance of trade relations between the two communities. The constant presence of French and Italian (putting aside Latin, which is mostly the source for inherited, not borrowed words) as top source languages in the borrowing process, clearly shows that the genetic relations, on the one hand, and the cultural connections, on the other hand, prevail over the geographical contiguity in the selection of source languages for the lexical enrichment.

5 Conclusions

For historical, geographical and linguistic reasons, Romanian presents a complex lexicographic picture, especially in terms of etymology. While reliable etymological dictionaries for Romanian are still missing, we proposed a computer-assisted etymological analysis doubled by a linguistic manual verification and interpretation, using the available dictionaries via *dexonline*. The comparison between the obtained data and previous knowledge about the Romanian fundamental lexicon revealed an inverted proportion between French and Latin and a surprisingly high percentage of French borrowings. We visualized the Romanian etymologies per source language on a geographic map and we also minded their part of speech proportions and interpreted them. Error analysis showed that the automatic extraction was performed with high accuracy, while the remaining errors are due to erroneous etymologies from the dictionaries. Finally, we experimented with the etymological composition of the Romanian lexicon based on semantic fields. Starting from a list of conceptual domains, adapted from Romanian linguistic atlases, we automatically obtained 10 onomasiological fields containing Romanian words in our dataset and their etymologies. For each of these categories, we ordered the source languages and interpreted the results from a sociocultural and historical point of view.

Ethics Statement

All our data are extracted from publicly available sources. There are no ethical issues in our work.

Acknowledgments

We would like to thank the reviewers for their helpful comments. All authors contributed equally to this work. This research is supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI UEFISCDI, project number 108, COTOHILI, within PNCDI III.

References

Academia Română. 1996 [1975]. *Dicționarul explicativ al limbii române (DEX)*. Univers enciclopedic, Bucharest.

Academia Română. 2011–. *Dicţionarul etimologic al limbii române (DELR)*. Bucharest, Editura Academiei Române.

Academia Română, Institutul de Lingvistică din București. 1988. *Supliment la Dicționarul explicativ al limbii române*. Editura Academiei, Bucharest.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Ion-Aurel Candrea and Ovid Densusianu. 1907–1914. *Dicţionarul etimologic al limbii române. Elementele latine (a–putea)*. Socec, Bucharest.

Victor Celac. 2012. Observaţii privind tratarea dialectelor limbii române, problema datării lexemelor şi valorificarea surselor în noul Dicţionar etimologic al limbii române. *Fonetică şi dialectologie*, 31:205–226.

Victor Celac. 2020. Methods and practice in Romanian etymological research (with particular focus on multiple etymology). In Marinela Burada and Raluca Sinu, editors, *Dictionary Research, Practice, and Use in Romania*, pages 21–67. Cambridge Scholars Publishing.

Alejandro Cioranescu. 1966. *Diccionario etimológico rumano*. Universidad de la Laguna, Tenerife.

Joan Coromines. 1980–2001. *Diccionari etimològic i complementari de la llengua catalana*. Curial, Barcelone.

Joan Coromines and José Antonio Pascual. 1980–1991. *Diccionario crítico etimológico castellano e hispánico*. Gredos, Madrid.

Mariana Costinescu, Magdalena Georgescu, and Florentina Zgraon. 1987. *Dicționarul limbii române literare vechi (1640-1780) - Termeni regionali*. Editura Științifică și Enciclopedică, Bucharest.

Gerhard Ernst. 2013. Review of DELR. *Revue de linguistique romane*, 77:554–557.

Bogdan Petriceicu Hasdeu. 1886–1898. *Etymologicum Magnum Romaniae: dicționarul limbei istorice și poporane a românilor*. Socec and Teclu, Bucharest.

Theodor Hristea. 2009. Considerații pe marginea unui dicționar etimologic: CDER. *Limba română*, LVIII:481–498.

Iorgu Iordan. 1965–2010. *Dicţionarul limbii române. Serie nouă (D-E; L-Z)*. Bucharest, Academia Română/Editura Academiei Române.

Vladimir I. Levenshtein. 1965. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10:707–710.

Florin Marcu and Constant Maneca. 1986. *Dicționar de neologisme*. Editura Academiei, Bucharest.

Cristian Moroianu. 2015. *Etimologie și lexicologie românească. Convergențe sincronice și diacronice*. Editura Universității din București.

Alexandru Niculescu. 1965. *Individualitatea limbii române între limbile romanice. Vol. I Contribuții gramaticale*. Editura Științifică, Bucharest.

Alexandru Niculescu. 1978. *Individualitatea limbii române între limbile romanice. Vol. II Contribuții socio-culturale*. Editura Științifică și Enciclopedică, Bucharest.

Alexandru Niculescu. 1999. *Individualitatea limbii române între limbile romanice. Vol. III. Noi contribuții*. Ed. Clusium, Cluj-Napoca.

Alexandru Niculescu. 2003. *Individualitatea limbii române între limbile romanice. Vol. IV. Elemente de istorie culturală*. Ed. Clusium, Cluj-Napoca.

Emil Petrovici. 1956–1972. *Atlasul lingvistic român, serie nouă, 7vol*. Editura Academiei Romane, Bucharest.

Sextil Puscariu. 1938–1942. *Atlasul lingvistic român, 2 parties, 3vol.* Muzeul Limbii Române/Harrassowitz, Sibiu/Leipzig.

Sextil Puşcariu. 1913–1949. *Dicţionarul limbii române (A-De, F-Lojniţă)*. Academia Română/Socec/Universul, Bucharest.

Sanda Reinheimer Ripeanu. 2004. *Les emprunts latins dans les langues romanes.* Editura Universității din București.

Alexandru Rosetti. 1968. *Istoria limbii române, de la origini până în secolul al XVII-lea*. Editura pentru Literatură, Bucharest.

Marius Sala, Mihaela Bîrlădeanu, Maria Iliescu, Liliana Macarie, Ioana Nechita, Mariana Ploae-Hanganu, Maria Theban, and Ioana Vintilă-Rădulescu. 1988. *Vocabularul reprezentativ al limbilor romanice*. Editura Științifică şi Enciclopedică, Bucharest.

Wolfgang Schweickard. 2013. Review of DELR. *Zeitschrift für romanische Philologie*, 129:858–866.

August Scriban. 1939. *Dicționarul limbii românești*. Institutu de Arte Grafice "Presa Bună".

Ana Sabina Uban, Alina Maria Ciobanu, and Liviu P. Dinu. 2019. Studying Laws of Semantic Divergence across Languages using Cognate Sets. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 161–166.

Ana Sabina Uban, Alina Maria Ciobanu, and Liviu P. Dinu. 2021. Cross-lingual Laws of Semantic Change. *Computational Approaches to Semantic Change*, pages 219–260.

Walther von Wartburg. 1922–2002. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes*. Klopp/Winter/Teubner/Zbinden, Bonn/Heidelberg/Leipzig-Berlin/Bâle.

Lazăr Șăineanu. 1929. *Dicționar universal al limbei române (4th edition)*. Scrisul Romanesc.