A Computational Perspective on the Romanian Dialects

Alina Maria Ciobanu, Liviu P. Dinu

Faculty of Mathematics and Computer Science, University of Bucharest
Center for Computational Linguistics, University of Bucharest
alina.ciobanu@my.fmi.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

In this paper we conduct an initial study on the dialects of Romanian. We analyze the differences between Romanian and its dialects using the Swadesh list. We analyze the predictive power of the orthographic and phonetic features of the words, building a classification problem for dialect identification.

Keywords: Romanian, dialects, language similarity.

1. Introduction and Related Work

The rapid development of the online repositories has lead to a significant increase in the number of multilingual documents, allowing users from all over the world to access information that has never been available before. This accelerated growth created the stringent need to overcome the language barrier by developing methods and tools for processing multilingual information. Nowadays, NLP tools for the official languages spoken in the European Union and for the most popular languages are constantly created and improved. However, there are many other language varieties and dialects that could benefit from such NLP tools. The effort for building NLP tools for resource-poor language varieties and dialects can be reduced by adapting the tools from related languages for which more resources are available. The importance of adapting NLP tools from resource-rich to resource-poor closely related languages has been acknowledged by the research communities and has been materialized through multiple events, such as the workshop on Language Technology for Closely Related Languages and Language Variants (Nakov et al., 2014) or the workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (Zampieri et al., 2014).

A related problem occurs when researchers are interested in the cultural heritage of small communities, who developed their own techniques of communication and prefer using dialects instead of the official language of the region they live in. In some situations, these dialects are close enough to the standard language, but in other situations the difference is consistent, so much so that some dialects have become languages of their own (for example Friulian, spoken in the North-East of Italy). These matters raise interesting research problems, since many such dialects are used only in speaking. Moreover, they often tend to be used only in very specific situations (such as speaking in the family), very rarely being taught in schools. Thus, many dialects are in danger of extinction, according to the UNESCO list of endangered languages (Moseley, 2010).

In this paper, we conduct an initial study on the dialects of Romanian. This investigation has the purpose of providing a deeper understanding of the differences between dialects, which would aid the adaptation of existing NLP tools for related varieties. The aim of our investigation is to assess the orthographic and phonetic differences between the dialects of Romanian. In this paper, we quantify only the orthographic and phonetic differences, but the morphology and the syntax are other important aspects which contribute to the individualization of each variety, that we leave for further study.

Previously, Tonelli et al. (2010) proposed such an adaptation for a morphological analyzer for Venetan. Similarly, Kanayama et al. (2014) built a dependency parser for Korean, leveraging resources (transfer learning) from Japanese. They showed that Korean sentences could be successfully parsed using features learnt from a Japanese training corpus. In the field of machine translation, Aminian et al. (2014) developed a method for translating from dialectal Arabic into English in which they reduce the OOV ratio by exploiting resources from standard Arabic. Although dialects and varieties have been investigated for other languages, such as Spanish and Portuguese (Zampieri and Gebre, 2012; Zampieri et al., 2013), Romanian dialects did not receive much attention in NLP. To our knowledge, while the syllabic structure of Aromanian has been previously investigated (Nisioi, 2014), this is one of the very first computational comparative studies on the Romanian dialects.

2. The Romanian Dialects

Romanian is a Romance language, belonging to the Italic branch of the Indo-European language family, and is of particular interest regarding its geographic setting. It is surrounded by Slavic languages and its relationship with the big Romance kernel was difficult. According to Tagliavini (1972), Romanian has been isolated for a long period from the Latin culture in an environment of different languages. Joseph (1999) emphasizes the reasons which make Romanian of special interest to linguists with comparative interests. Besides general typological comparisons that can be made between any two or more languages, Romanian can be studied based on comparisons of genetic and geographical nature. Joseph further states that, regarding genetic relationships, Romanian can be studied in the context of those languages most closely related to it and that the well-studied Romance languages enable comparisons that might not be possible otherwise, within less well docu-

mented families of languages. Romanian is of particular interest also regarding its geographic setting, participating in numerous areally-based similarities that define the Balkan convergence area.
Romanian is spoken by over 24 million people as native language, out of which 17 million are located in Romania, and most of the others in territories that surround Romania (Lewis et al., 2015). According to most Romanian linguists (Puşcariu, 1976; Petrovici, 1970; Caragiu Marioţeanu, 1975), Romanian has four dialects:

- Daco-Romanian, or Romanian (RO) - spoken primarily in Romania and the Republic of Moldova, where it has an official status.
- Macedo-Romanian, or Aromanian (AR) – spoken in relatively wide areas in Macedonia, Albania, Greece, Bulgaria, Serbia and Romania.
- Megleno-Romanian (ME) – spoken in a more narrow area in the Meglen region, in the South of the Balkan Peninsula.
- Istro-Romanian (IS) – spoken in a few villages from the North-East of the Istrian Peninsula in Croatia. It is much closer to Italy than to Romania, from a geographical point of view, but shows obvious similarities with Romanian. It seems that the community of Istro-Romanians exists here since before the 12th century. Istro-Romanian is today on the "selected" list of endangered languages, according to the UNESCO classification.[1]

Romanian was originally a single language, descendant of the oriental Latin, spoken in the regions around the romanized Danube: Moesia Inferior and Superior, Dacia and Pannonia Inferior (Rosetti, 1966). The period of common Romanian begun in the 7th-8th century and ended in the 10th century, when a part of the population migrated to the South of the Danube, beginning the creation of the dialects. Densuşianu (1901) places the migration to the South even earlier in time, in the 6th and 7th century. Thus, starting with the 10th century, given a series of political, military, economical and social events, the 4 dialects of Romanian were born: Daco-Romanian (to the North of the Danube), Aromanian, Megleno-Romanian and Istro-Romanian (to the South of the Danube). Among these dialects, only Daco-Romanian could develop into a national standard language, in the context of several political and historical factors, leading to the Romanian language that is spoken today inside the borders of Romania. The other three dialects are spoken in communities spread in different countries. An explanation for this fact is the setting of the Slavic people at the South of the Danube, which has lead, among others, to the dispersion of the groups that spoke the three dialects to the South of the Danube. According to the Ethnologue (Lewis et al., 2015), the three dialects to the South of the Danube were developed between the 5th and the 10th century, while according to Rosetti (1966), this process took place after the 10th century. Thus, according to Rosetti (1966), Aromanian and Megleno-Romanian developed in the 11th century, while Istro-Romanian developed in the 13th century. Rosetti (1966) states that there are, actually, two main dialects of Romanian: Daco-Romanian and Aromanian, the other two being derived from them (Megleno-Romanian derived from Aromanian and Istro-Romanian derived from Daco-Romanian).

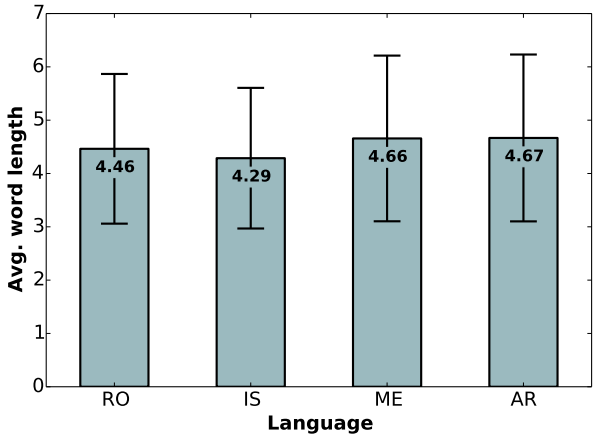

Figure 1: Average word length, using the orthographic form of the words.

3. Experiments

In this section we describe our investigations and experiments on the dialects of Romanian. We are mainly interested in assessing the differences between the dialects from the South of the Danube and Daco-Romanian. We henceforth refer to Daco-Romanian, the standard language spoken in Romania, as Romanian.

3.1. Data

We use a dataset of 108 words comprising the short Swadesh list for the Romanian dialects.[2] The Swadesh list has been widely used in lexicostatistics and comparative linguistics, to investigate the classification of the languages (Dyen et al., 1992; McMahon and McMahon, 2003). The dataset is provided in two versions: orthographic and phonetic. In Figure 1 we represent the average word length (considering their orthographic form) for the Romanian dialects. Istro-Romanian has the shortest words, followed by Romanian. Megleno-Romanian and Aromanian have slightly longer words, on average, but the differences are not significant.
The orthographic or phonetic distance has been widely used for analyzing related words and for reconstructing phylogenies (Kondrak, 2004; Delmestri and Cristianini, 2012). We use the edit distance to observe how close the Romanian dialects are to one another (Table 1).

[1] UNESCO Interactive Atlas of the World's Languages in Danger (Moseley, 2010) provides for Istro-Romanian the following information: severely endangered, with an estimation of 300 first-language and 100 second-language speakers in Istria, plus 1000 others living outside of Istria.

[2] http://starling.rinet.ru/new100/main.htm

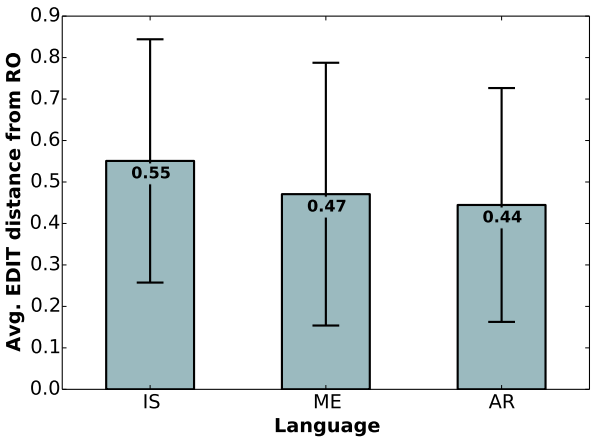

Figure 2: Average edit distance from Romanian, using the orthographic form of the words.

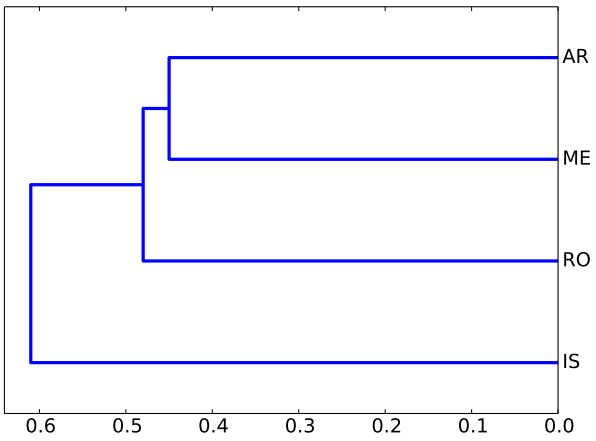

Figure 3: Dendrogram representing the hierarchical clustering using the farthest neighbor algorithm and the orthographic form of the words as input.

The edit distance (Levenshtein, 1965) 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.

	ME	**AR**	**IS**
ME	–		
AR	0.44	–	
IS	0.61	0.58	–
RO	0.47	0.44	0.55

(a) Orthographic

	ME	**AR**	**IS**
ME	–		
AR	0.40	–	
IS	0.54	0.54	–
RO	0.39	0.40	0.42

(b) Phonetic

Table 1: The average edit distance between the words.

Using the orthographic form of the words (see also Figure 2), Aromanian words are closest to the Romanian words (0.44), followed by Megleno-Romanian (0.47) and Istro-Romanian (0.55). When using the phonetic form of the words, Megleno-Romanian words are closest to the Romanian words (0.39), followed by Aromanian (0.40) and Istro-Romanian (0.42). At the phonetic level, the distance between Romanian and the other three dialects is much smaller than the same distance measured at the orthographic level. In both situations, Istro-Romanian is farthest from the other dialects. One possible reason could be the geographical regions in which Istro-Romanian is spoken, farther from the regions where the other dialects are spoken. In Figure 3 we represent the dendrogram for the Romanian dialects, based on the computed distances on the orthographic version of the dataset.

	1	2	3	4	5
ME	un	re	in	ar	ca
AR	ár	re	eá	in	oá
IS	re	că	ur	če	år
RO	ár	oá	ri	ti	ts

Table 2: The most common 2-grams for each dialect.

3.2. Dialect Identification

We are interested to see if the orthographic or phonetic differences between Romanian and the other Romanian dialects (spoken at the South of the Danube) are dialect-specific (i.e., if they have enough discriminative power to identify the dialect to which a word belongs).

To this end, we build a classification problem as follows: given the parallel list of 108 words (in all the Romanian dialects), we extract pairs having the form *(romanian-word, dialect-word)*, where *dialect* $\in$ {Istro-Romanian, Megleno-Romanian, Aromanian}. We obtain, thus, a dataset of 324 such input pairs.

The goal is to automatically decide to which dialect the *dialect-word* belongs. The dialect identification problem is not trivial and our goal, in this paper, is not to improve on the state-of-the-art methods in this research area, but to investigate the predictive power of the orthographic and phonetic differences between Romanian and its dialects. We use a methodology that has been previously used for discriminating between related and unrelated words, and for distinguishing the type of relationship between the words (Ciobanu and Dinu, 2014b; Ciobanu and Dinu, 2015).

We align the words using the Needleman-Wunsch alignment algorithm (Needleman and Wunsch, 1970) and we extract n-gram features from the alignment of the words. Additionally, we also extract n-grams of characters from the *dialect-word*. We search for the optimum n-gram size in $\{1, 2, 3, 4\}$, both for the n-grams extracted from the alignment and for the n-grams extracted from the *dialect-word*. We train a Logistic Regression classifier, using the imple-

Dialect	Word pair	Alignment				
ME	roşu - roş	r	o	ş	u	
		r	o	ş	-	
AR	roşu - aróşŭ	-	r	o	ş	u
		a	r	ó	ş	ŭ
IS	roşu - rójsu	r	o	-	ş	u
		r	ó	j	s	u

Table 3: Alignment of the Romanian word *roşu* (meaning *red*) with its translations in the other Romanian dialects.

mentation provided by Weka (Hall et al., 2009). Since our dataset is small, we evaluate the performance of the model with 5-fold cross-validation. For both experiments (orthographic and phonetic), $n = 2$ proves to be the optimal n-gram size. In Table 2 we report the most common 2-grams for each dialect (using the orthographic version of the words) and in Table 3 we show examples of word pairs aligned with the Needleman-Wunsch algorithm.

3.3. Results

In Table 4 we report the cross-validation results for dialect identification, for the orthographic version of the dataset (Table 4a) and for the phonetic version of the dataset (Table 4b). For the former, the best results, in terms of F-score values, are obtained for Istro-Romanian (0.70), followed by Aromanian (0.60) and Megleno-Romanian (0.56). This shows that the Istro-Romanian dialect can be identified easier, and the orthographic features of the Istro-Romanian words have the highest predictive power. For the later, the ranking is different: Aromanian is identified with the highest F-score (0.71), followed by Istro-Romanian (0.70), Megleno-Romanian being on the last position (0.53). At the phonetic level, we notice that the Megleno-Romanian dialect is the most difficult to identify.

In Table 5 we report the confusion matrix for both experiments (orthographic and phonetic). We report the number of instances that are correctly classified and misclassified for each dialect. In both versions of the dataset, the maximum number of correctly classified instances is reported for Istro-Romanian (with a maximum of 88 for the phonetic version of the dataset). While for the phonetic version of the dataset only 3 Istro-Romanian words are classified as Aromanian, for the orthographic version of the dataset we notice an increase, with 10 Istro-Romanian words being classified as Aromanian. For Aromanian, most of the misclassified instances are labeled as Megleno-Romanian, in both versions of the dataset. For Megleno-Romanian, most of the misclassified instances in the orthographic version of the dataset are labeled as Aromanian (25), while for the phonetic version of the dataset most of the misclassified instances are labeled as Istro-Romanian (38).

Dialect	Precision	Recall	F-score
ME	0.53	0.60	0.56
AR	0.63	0.57	0.60
IS	0.71	0.69	0.70

(a) Orthographic

Dialect	Precision	Recall	F-score
ME	0.57	0.50	0.53
AR	0.78	0.64	0.71
IS	0.62	0.81	0.70

(b) Phonetic

Table 4: Cross-validation results for dialect identification using the orthographic (a) and the phonetic (b) form of the words.

	ME	AR	IS
ME	**65**	25	18
AR	34	**62**	12
IS	23	10	**75**

(a) Orthographic

	ME	AR	IS
ME	**54**	16	38
AR	23	**70**	15
IS	17	3	**88**

(b) Phonetic

Table 5: Confusion matrix for dialect identification using the orthographic (a) and the phonetic (b) form of the words. We report the number of correctly classified and misclassified instances.

4. Conclusions

In this paper we conducted an initial study on the Romanian dialects. We analyzed the orthographic and phonetic differences between the Romanian dialects, using the Swadesh list and building a classification problem for dialect identification. The results obtained so far show that Istro-Romanian has more dialect-specific differences from Romanian, followed by Aromanian and Megleno-Romanian. The next steps in our investigation will be to conduct a similar study on corpora (Ciobanu and Dinu, 2014a) instead of word lists, as far as resources are available, and to assess the mutual intelligibility of the Romanian dialects. The necessity of such a study is increased by the fact that at least one of the Romanian dialects (namely Istro-Romanian) is today on the "selected" list of endangered languages, according to the UNESCO classification (Moseley, 2010).

Acknowledgements

We thank the anonymous reviewers for their helpful and constructive comments. The research of Liviu P. Dinu was supported by a grant of the Romanian National Authority for Scientific Research, CNCS UEFISCDI, project number PN-II-ID-PCE-2011-3-0959.

5. References

Maryam Aminian, Mahmoud Ghoneim, and Mona Diab. 2014. Handling OOV Words in Dialectal Arabic to English Machine Translation. In *Proceedings of the Workshop on Language Technology for Closely Related Languages and Language Variants, LT4CloseLang 2014*, pages 99–108.

Matilda Caragiu Marioţeanu. 1975. *Compendiu de Dialectologie Română*. Editura Ştiinţifică şi Enciclopedică, Bucureşti.

Alina Maria Ciobanu and Liviu P. Dinu. 2014a. An Etymological Approach to Cross-Language Orthographic Similarity. Application on Romanian. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1047–1058.

Alina Maria Ciobanu and Liviu P. Dinu. 2014b. Automatic Detection of Cognates Using Orthographic Alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, volume 2: Short Papers, ACL 2014*, pages 99–105.

Alina Maria Ciobanu and Liviu P. Dinu. 2015. Automatic Discrimination between Cognates and Borrowings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, volume 2: Short Papers, ACL-IJCNLP 2015*, pages 431–437.

Antonella Delmestri and Nello Cristianini. 2012. Linguistic Phylogenetic Inference by PAM-like Matrices. *Journal of Quantitative Linguistics*, 19(2):95–120.

Ovid Densuşianu. 1901. *Histoire de la Langue Roumaine*, volume 1. E. Leroux.

Isidore Dyen, Joseph B. Kruskal, and Paul Black. 1992. An Indoeuropean Classification: a Lexicostatistical Experiment. *Transactions of the Americal Philosophical Society*, 82(5):1–132.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

Brian D. Joseph. 1999. Romanian and the Balkans: Some Comparative Perspectives. In Sheila Embleton, John E. Joseph, and Hans-Joseph Niederehe, editors, *The Emergence of the Modern Language Sciences*. John Benjamins Publishing Company.

Hiroshi Kanayama, Youngja Park, Yuta Tsuboi, and Dongmook Yi. 2014. Learning from a Neighbor: Adapting a Japanese Parser for Korean Through Feature Transfer Learning. In *Proceedings of the Workshop on Language Technology for Closely Related Languages and Language Variants, LT4CloseLang 2014*, pages 2–12.

Grzegorz Kondrak. 2004. Combining Evidence in Cognate Identification. In *Proceedings of the 17th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2004*, pages 44–59.

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

Paul Lewis, Gary Simons, and Charles Fennig. 2015. *Ethnologue: Languages of the World. 18th edition*. Summer Institute of Linguistics, Dallas, Texas.

April McMahon and Robert McMahon. 2003. Finding Families: Quantitative Methods in Language Classification. *Transactions of the Philological Society*, 101(1):7–55.

Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger, 3rd edition*. UNESCO Publishing, Paris.

Preslav Nakov, Petya Osenova, and Cristina Vertan, editors. 2014. *Proceedings of the Workshop on Language Technology for Closely Related Languages and Language Variants, LT4CloseLang 2014*. Association for Computational Linguistics.

Saul B. Needleman and Christian D. Wunsch. 1970. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of two Proteins. *Journal of Molecular Biology*, 48(3):443–453.

Sergiu Nisioi. 2014. On the Syllabic Structures of Aromanian. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities LaTeCH 2014*, pages 110–118.

Emil Petrovici. 1970. *Studii de Dialectologie şi Toponimie*. Editura Academiei, Bucureşti.

Sextil Puşcariu. 1976. *Limba Română*. Editura Minerva.

Alexandru Rosetti. 1966. *Istoria Limbii Române*. Editura Ştiinţifică.

Carlo Tagliavini. 1972. *Le Origini delle Lingue Neolatine*. Casa editrice Patron.

Sara Tonelli, Emanuele Pianta, Rodolfo Delmonte, and Michele Brunelli. 2010. VenPro: A Morphological Analyzer for Venetan. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, pages 866–870.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic Identification of Language Varieties: The Case of Portuguese. In *Proceedings of the 11th Conference on Natural Language Processing, KONVENS 2012*, pages 233–237.

Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. 2013. N-Gram Language Models and POS Distribution for the Identification of Spanish Varieties. In *Proceedings of the 20th Conférence du Traitement Automatique du Langage Naturel, TALN 2013*, pages 580–587.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann, editors. 2014. *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, VarDial 2014*. Association for Computational Linguistics and Dublin City University.