Universitatea din Bucureşti
Facultatea de Matematică şi Informatică
Şcoala Doctorală de Informatică

Doctoral Thesis

Problems of Similarity in Computational Linguistics

Probleme de Similaritate în Lingvistica Computaţională

Summary

Author: Alina-Maria Ciobanu
Scientific Advisor: Prof. Dr. Liviu P. Dinu

-2019-

Contents

1 Introduction **2**

2 Contributions **4**

3 Thesis Outline **6**

4 Published Papers **7**

References **9**

1 Introduction

Language relatedness and language change across space and time are fundamental concerns in historical linguistics. While both problems have been traditionally investigated with comparative linguistics instruments (Campbell, 1998), modern approaches impose the use and development of quantitative and computational methods in this field (McMahon et al., 2005a; Heggarty, 2012; Atkinson, 2013).

Although grouping of natural languages in linguistic families is generally accepted, the degrees of similarity between languages belonging to the same family are far from being certain, not only for exotic languages, but even for extensively studied languages, many of which are closely related. The similarity of languages plays a crucial role in many of the natural language processing (NLP) tasks that deal with multilingual documents, in historical linguistics and in the study of second language acquisition. It is also a valuable means of creating multilingual dictionaries and corpora for endangered languages, for whose preservation significant efforts have been made recently.

Natural languages are living eco-systems. They are subject to continuous change due, in part, to the natural phenomena of language contact and borrowing (Campbell, 1998). According to Hall (1960), there is no such thing as a "pure language" – a language "without any borrowing from a foreign language". The unprecedented contact between languages in today's context of high mobility and the explosion of communication tools lead to an inherent enrichment of languages by borrowings. "Why" and "how" the borrowing process takes place are fundamental questions which, by their nature, invite to experimental perspective (Chitoran, 2011). To answer the first question, Campbell (1998) notes that "Languages borrow words from other languages primarily because of need and prestige". Further, the author states that the result of the borrowing process depends on numerous factors, such as the length and intensity of the contact and the extent to which the populations in question are bilingual. To answer the second question, it is necessary to investigate borrowings and to analyze how words evolve from one language into another. Hence, the outcome of the contact between two populations is a challenging and interesting research problem.

Addressing the problem of language relatedness implies developing methods for identifying cognates and borrowings. Cognates are words in different languages having the same etymology and a common ancestor. Investigating pairs of cognates is very useful not only in historical and comparative linguistics (in the study of language relatedness (Ng et al., 2010), phylogenetic inference (Atkinson et al., 2005) and in identifying how and to what extent languages changed over time or influenced each other), but also in other research areas, such as bilingual word recognition (Dijkstra et al., 2012), corpus linguistics (Simard et al., 1992), cross-lingual information retrieval (Buckley et al., 1997) and machine translation (Kondrak et al., 2003). A borrowed word, also called "loanword", is defined as a "lexical item (a word) which has been 'borrowed' from another language, a word which originally was not part of the vocabulary of the recipient language but was adopted from some other language and made part of the borrowing language's vocabulary" (Campbell, 1998). Although admittedly regarded as relevant factors in the history of a language (McMahon et al., 2005b), borrowings bias the genetic classification of the languages, characterizing them as being closer than they actually are (Minett and Wang, 2003). According to Gray and Atkinson (2003), correctly determining cognates and borrowings is essential in the process of phylogenetic inference, as false cognates and unrecognized borrowings could incorrectly increase the degree of similarity between languages. False cognates are more harmful than missing valid cognates in language comparison, because they can lead to incorrect conclusions regarding the genetic relationships

between languages (List et al., 2017). Thus, the need for discriminating between cognates and borrowings emerges. Heggarty (2012) acknowledges the necessity and difficulty of the task, emphasizing the role of the "computerized approaches".

Proto-word reconstruction, which is central to the study of language evolution, consists in recreating the words in an ancient language from the words in its modern daughter languages. Bouchard-Côté et al. (2013) emphasize the important role this task plays in historical linguistics, because it enables evaluating proposals regarding the phenomenon of language change. While the main hypothesis in this research problem is that there are regularities and patterns in how words evolved from the ancestor language to its modern daughter languages, there are also words which diverged significantly from their ancestor. Take, for example, the Latin word "umbilicu(lu)s" (meaning "umbilicus"). It evolved into "buric" (Romanian), "nombril" (French), and "umbigo" (Portuguese).

Automating the process of reconstructing proto-words is a challenging task (Oakes, 2000; Bouchard-Côté et al., 2013; Atkinson, 2013), and it is closely related to two other research problems: cognate production (determining the form of a given word's cognate pair) and modern word form production (determining the form in which a proto-word evolves in a modern daughter language). We emphasize two research directions that rely on these tasks of word production: diachronic linguistics, which is concerned with language evolution over time, and the study of foreign language learning, which focuses on the learning process and on the influence of the learner's mother tongue in this process.

The rapid development of the online repositories 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 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 developing cross-language and multilingual tools and of adapting NLP tools from resource-rich to resource-poor closely related languages has been acknowledged by the research community and has been materialized through multiple events, such as the Workshop on Natural Language Processing for Similar Languages, Varieties and Dialects – VarDial (Zampieri et al., 2014) and the associated shared tasks. Despite having attracted considerable attention in the past few years, automatic language identification is not a solved problem. This research problem is particularly challenging in some situations, for example when distinguishing between language varieties, dialects or very similar languages, when the input pieces of text are very short, or when code-switching (mixing multiple languages in a single communication) occurs.

In this thesis, we focus on problems of similarity in computational linguistics. In the first part, we propose a method for determining degrees of similarity between languages. Then we address problems of computational historical linguistics, such as identifying cognates and borrowings, reconstructing proto-words, producing modern word forms and cognates. Finally, we investigate the discrimination between similar languages and language varieties. Our goal is twofold: to propose computational solutions for problems of similarity in computational linguistics, and to develop a series of tools to assist linguists and domain experts in studying the evolution of the languages.

2 Contributions

This thesis is mainly based on our research papers, which we have extended and enhanced. The complete list of 29 published papers is presented in Section 4. The main contributions of the thesis can be summarized as follows.

We propose and analyze aggregate computational measures for quantifying the lexical and syntactic similarity of the languages. Firstly, we propose a method for determining the orthographic similarity. We account for etymons and cognates and we investigate not only the number of related words, but also their forms, quantifying orthographic similarities and extending our analysis from words to languages. Our study provides evidence to be used in the investigation of the written intelligibility (the ability of people writing in different languages to understand one another without prior knowledge of foreign languages). Secondly, we propose a method for determining the syntactic similarity between languages. We investigate multiple approaches and metrics that lead to consistent results. We report results on multiple languages belonging to various language families. We show how these metrics can be used for clustering the languages based on their lexical or syntactic similarity. The results have been published and presented at RANLP 2013 (Ciobanu and Dinu, 2013), LREC 2014 (Ciobanu and Dinu, 2014e), EMNLP 2014 (Ciobanu and Dinu, 2014b) and CICLing 2017 (Ciobanu et al., 2017a).

We develop, implement and evaluate a dictionary-based approach to identifying cognates based on the etymology of the words. The proposed method can be used to create large databases of cognate pairs, and it can be easily generalized to languages for which electronic dictionaries with etymological information are available. As a case study, we apply this method on Romanian, identify the etymologies of the Romanian words and determine cognate pairs between Romanian and related languages (such as Italian, French, Spanish and Portuguese). Using these pieces of information, we build a dataset of multilingual cognates for Romanian, and develop a parallel list of over 3,000 cognate sets across five Romance languages with common Latin etymology. The results have been published and presented at LREC 2014 (Ciobanu and Dinu, 2014d).

We investigate and propose metrics for quantifying the level of readability (or text comprehensibility) for translated texts, using multiple types of features and analyzing the impact of translation. The results have been published and presented at PITR@EACL 2014 (Ciobanu and Dinu, 2014a) and RANLP 2015 (Ciobanu et al., 2015).

We propose machine learning methods for identifying cognates and for discriminating between cognates and borrowings. Firstly, we introduce a method to automatically determine if two given words are cognates. We employ an orthographic alignment method which proved relevant for sequence alignment in computational biology. We use aligned subsequences as features for machine learning algorithms in order to infer rules for linguistic changes undergone by words when entering new languages and to discriminate between cognates and non-cognates. We apply our method on a subset of the automatically extracted dataset of cognates previously built for Romanian and four related Romance languages: Italian, French, Spanish and Portuguese. Secondly, we investigate the task of discriminating between cognates and borrowings. The challenge and importance of this task is emphasized by Heggarty (2012) as follows: "What solution is there, then, if we can neither ignore the problem of distinguishing cognates from loanwords, nor overcome it in the many cases where we do not have the necessarily linguistic knowledge to do so? There is in fact a possibility: to sidestep the question entirely at the data analysis stage, and simply to identify which forms are judged to be somehow 'correlate' with each other

— whether by specialists in those languages, or more objectively by computerized approaches". Furthermore, Jäger (2018) considers the handling of language contact (and borrowings, more specifically) an "unsolved problem for computational historical linguistics". We address this research problem and propose an automatic method for identifying the type of relationship between the words (cognates or borrowings/loanwords). We show that orthographic features have discriminative power and we analyze the relevance of several underlying linguistic factors in the classification task. We run experiments on multiple pairs of languages. The results have been published and presented at ACL 2014 (Ciobanu and Dinu, 2014c) and ACL 2015 (Ciobanu and Dinu, 2015).

We develop a method for automatically producing related words, with the following sub-problems: reconstructing proto-words, producing modern word forms and producing cognates.

We begin with the reconstruction of proto-words. Given words in modern languages, the task is to automatically reconstruct the proto-words from which the modern words evolved. We address this problem in two steps. For the first step, given cognate pairs in multiple modern languages and their common ancestors, we propose a word production method based on sequence labeling for reconstructing proto-words, that we apply on each modern language individually. For the second step, we introduce an ensemble system to use information from sister languages for reconstructing their common proto-words. We run experiments for reconstructing proto-word on three datasets of cognates in Romance languages. The results have been published and presented at COLING 2018 (Ciobanu and Dinu, 2018a; Ciobanu and Dinu, 2018b).

Then, we address the production of modern word forms. Natural languages are permanently in contact, borrowing from one another. We investigate word derivation from a donor language into a recipient language, we experiment with multiple pairs of languages and we further evaluate how well our approach models the form of foreign words which have been borrowed by Romanian, and which donor language is better modeled by our method. The results have been published and presented at CICLing 2017 (Dinu and Ciobanu, 2017).

Finally, for the production of cognates, we investigate if, given a word, we can automatically determine the form of its cognate pair in a related language. We run experiments on multiple pairs of languages, some of wich are more closely related (such as English - German), while others are more remotely related (such as English - Spanish). The results have been published and presented at KES 2016 (Ciobanu, 2016) and CICLing 2017 (Dinu and Ciobanu, 2017).

We implement and evaluate linear and ensemble systems for discriminating between closely related languages. We also use the ensemble systems in two related research problems, to identify the native language of a speaker and to analyze the mother tongue in the context of author profiling. The results have been published at BEA@EMNLP 2017 (Zampieri et al., 2017), CLEF 2017 (Ciobanu et al., 2017b), VarDial@COLING 2015 (Nisioi et al., 2016) and VarDial@COLING 2018 (Ciobanu et al., 2018a; Ciobanu et al., 2018b).

We conduct an initial computational study on the dialects of Romanian. We analyze the differences between them using lists of common words in orthographic and phonetic form. We extract orthographic and phonetic features from the pairwise alignment of the words from the dialects with their translation in Romanian an we analyze the predictive power of these features, building a classification problem for word-level dialect identification. The results have been published and presented at LREC 2016 (Ciobanu and Dinu, 2016).

3 Thesis Outline

The thesis is organized as follows.

In Chapter 1 we introduce the research topic that we address, the motivation behind this work, as well as the main contributions.

In Chapter 2 we present an overview of the main concepts and theoretical notions on the similarity of languages in computational linguistics. We review the state-of-the-art methods, we briefly describe some of their applications in various areas, and their practical implications in day-to-day NLP tasks. We present several popular linguistic distances that have proven useful in NLP.

In Chapter 3 we address the problem of language similarity from a quantitative perspective. We propose a method for computing degrees of similarity between languages. We begin with the lexical level, and continue with the syntactic level. We apply our method on multiple corpora, and compare our results, in terms of similarity, with the generally-accepted language phylogeny. We also build a dataset of related words for the Romanian lexicon. Finally, we address the problem of readability in translated texts, assessing the impact of the translation process on the readability level, and investigating the discriminative power of several shallow readability features in the problem of translationese identification. The results presented in this chapter are based on the following papers: Ciobanu and Dinu (2013), Ciobanu and Dinu (2014d), Ciobanu and Dinu (2014b), Ciobanu and Dinu (2014e) Ciobanu and Dinu (2014a), Ciobanu et al. (2015) and Ciobanu et al. (2017a).

In Chapter 4 we present a method for identifying related words. We develop a machine learning approach to identifying related words automatically and we extend the study further to a finer-grained level, to identify the type of relationship between the words (cognates versus borrowings). We perform experiments on Romance languages, showing that linguistic distances and orthographic features have high discriminative power. The results presented in this chapter are based on the following papers: Ciobanu and Dinu (2014c) and Ciobanu and Dinu (2015).

In Chapter 5 we advance towards the problem of word production. We propose a method for predicting the form of the words in a target language. We evaluate our method on multiple datasets including various languages, in three subtasks: production of modern words, production of cognates and reconstruction of proto-words (for which we also use an ensemble, which proves useful when parallel data from multiple sister languages are available). The results presented in this chapter are based on the following papers: Ciobanu (2016), Dinu and Ciobanu (2017), Ciobanu and Dinu (2018a) and Ciobanu and Dinu (2018b).

In Chapter 6 we propose algorithms for discriminating between closely related languages, language varieties and dialects. We first experiment with linear classifiers and we further propose a method based on ensembles for the same task. We evaluate our results through participation in the VarDial shared task – for discriminating between similar languages – and also in the CLEF and NLI shared tasks – for two related research problems – author profiling based on the author's native language and native language identification. We conduct an initial computational study on the Romanian dialects and investigate the discriminative power of orthographic and phonetic features in word level dialect identification. The results presented in this chapter are based on the following papers: Ciobanu and Dinu (2016), Nisioi et al. (2016), Zampieri et al. (2017), Ciobanu et al. (2017b), Ciobanu et al. (2018a) and Ciobanu et al. (2018b) .

Finally, in Chapter 7 we summarize the content and contributions of this thesis, we draw conclusions and propose several ideas and research directions for future work.

4 Published Papers

1. Alina Maria Ciobanu and Liviu P. Dinu. 2018. Ab Initio: Automatic Latin Proto-word Reconstruction. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, pages 1604–1614.

2. Alina Maria Ciobanu and Liviu P. Dinu. 2018. Simulating Language Evolution: a Tool for Historical Linguistics. In *The 27th International Conference on Computational Linguistics: System Demonstrations, COLING 2018*, pages 68–72.

3. Alina Maria Ciobanu, Shervin Malmasi, and Liviu P. Dinu. 2018. German Dialect Identification Using Classifier Ensembles. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial@COLING 2018*, pages 288–294.

4. Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, Santanu Pal, and Liviu P. Dinu. 2018. Discriminating between Indo-Aryan Languages Using SVM Ensembles. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial@COLING 2018*, pages 178–184.

5. Bogdan Dumitru, Alina Maria Ciobanu, and Liviu P. Dinu. 2018. ALB at SemEval-2018 Task 10: A System for Capturing Discriminative Attributes. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018*, pages 963–967.

6. Marcos Zampieri, Alina Maria Ciobanu, and Liviu P. Dinu. 2017. Native Language Identification on Text and Speech. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@EMNLP 2017*, pages 398–404.

7. Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, and Liviu P. Dinu. 2017. Including Dialects and Language Varieties in Author Profiling. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*.

8. Alina Maria Ciobanu, Liviu P. Dinu, and Andrea Sgarro. 2017. Towards a Map of the Syntactic Similarity of Languages. In *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, Revised Selected Papers, Part I*, pages 576–590.

9. Liviu P. Dinu and Alina Maria Ciobanu. 2017. Romanian Word Production: An Orthographic Approach Based on Sequence Labeling. In *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, Revised Selected Papers, Part I*, pages 591–603.

10. Alina Maria Ciobanu. 2016. Sequence Labeling for Cognate Production. In *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 20th International Conference KES 2016*, pages 1391–1399.

11. Alina Maria Ciobanu and Liviu P. Dinu. 2016. A Computational Perspective on the Romanian Dialects. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*.

12. Sergiu Nisioi, Alina Maria Ciobanu, and Liviu P. Dinu. 2016. Vanilla Classifiers for Distinguishing between Similar Languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial@COLING 2016*, pages 235–242.

13. 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 of the Asian Federation of Natural Language Processing, ACL 2015, Volume 2: Short Papers*, pages 431–437.

14. Alina Maria Ciobanu, Liviu P. Dinu, and Flaviu Pepelea. 2015. Readability Assessment of Translated Texts. In *Recent Advances in Natural Language Processing, RANLP 2015*, pages 97–103.

15. Marcos Zampieri, Alina Maria Ciobanu, Vlad Niculae, and Liviu P. Dinu. 2015. AMBRA: A Ranking Approach to Temporal Text Classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015*, pages 851–855.

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

17. Alina Maria Ciobanu and Liviu P. Dinu. 2014. A Quantitative Insight into the Impact of Translation on Readability. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations, PITR@EACL 2014*, pages 104–113.

18. Vlad Niculae, Marcos Zampieri, Liviu P. Dinu, and Alina Maria Ciobanu. 2014. Temporal Text Ranking and Automatic Dating of Texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*, pages 17–21.

19. Alina Maria Ciobanu, Anca Dinu, and Liviu P. Dinu. 2014. Predicting Romanian Stress Assignment. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*, pages 64–68.

20. Cornelia Caragea, Jian Wu, Alina Maria Ciobanu, Kyle Williams, Juan Pablo Fernández Ramírez, Hung-Hsuan Chen, Zhaohui Wu, and C. Lee Giles. 2014. CiteSeer x : A Scholarly Big Dataset. In *Proceedings of Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014*, pages 311–322.

21. Alina Maria Ciobanu and Liviu P. Dinu. 2014. 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.

22. Liviu P. Dinu, Alina Maria Ciobanu, Ioana Chitoran, and Vlad Niculae. 2014. Using a Machine Learning Model to Assess the Complexity of Stress Systems. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 331–336.

23. Alina Maria Ciobanu and Liviu P. Dinu. 2014. Building a Dataset of Multilingual Cognates for the Romanian Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1038–1043.

24. Alina Maria Ciobanu and Liviu P. Dinu. 2014. On the Romance Languages Mutual Intelligibility. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 3313–3318.

25. Anca Dinu and Alina Maria Ciobanu. 2013. Alternative Measures of Word Relatedness in Distributional Semantics. In *Proceedings of the Joint Symposium on Semantic Processing. Textual*

Inference and Structures in Corpora, JSSP 2013, pages 80–84.

26. Alina Maria Ciobanu, Anca Dinu, Liviu P. Dinu, Vlad Niculae, and Octavia-Maria Sulea. 2013. Temporal Classification for Historical Romanian Texts. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH@ACL 2013*, pages 102–106.

27. Alina Maria Ciobanu, Liviu P. Dinu, Octavia-Maria Sulea, Anca Dinu, and Vlad Niculae. 2013. Temporal Text Classification for Romanian Novels Set in the Past. In *Recent Advances in Natural Language Processing, RANLP 2013*, pages 136–140.

28. Alina Maria Ciobanu and Liviu P. Dinu. 2013. A Dictionary-based Approach for Evaluating Orthographic Methods in Cognates Identification. In *Recent Advances in Natural Language Processing, RANLP 2013*, pages 141–147.

29. Alina Maria Ciobanu and Liviu P. Dinu. 2012. On the Romanian Rhyme Detection. In *The 24th International Conference on Computational Linguistics, Proceedings of the Conference: Demonstration Papers, COLING 2012*, pages 87–94.

References

Quentin D. Atkinson, Russell D. Gray, Geoff K. Nicholls, and David J. Welch. 2005. From Words to Dates: Water into Wine, Mathemagic or Phylogenetic Inference? *Transactions of the Philological Society*, 103(2):193–219.

Quentin D Atkinson. 2013. The Descent of Words. *Proceedings of the National Academy of Sciences*, 110(11):4159–4160.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated Reconstruction of Ancient Languages Using Probabilistic Models of Sound Change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.

Chris Buckley, Mandar Mitra, Janet A. Walz, and Claire Cardie. 1997. Using Clustering and SuperConcepts Within SMART: TREC 6. In *Proceedings of the 6th Text Retrieval Conference, TREC 1997*, pages 107–124.

Lyle Campbell. 1998. *Historical Linguistics. An Introduction*. MIT Press.

Ioana Chitoran. 2011. The Nature of Historical Change. In *The Oxford Handbook of Laboratory Phonology*. Oxford University Press.

Alina Maria Ciobanu and Liviu P. Dinu. 2013. A Dictionary-based Approach for Evaluating Orthographic Methods in Cognates Identification. In *Recent Advances in Natural Language Processing, RANLP 2013*, pages 141–147.

Alina Maria Ciobanu and Liviu P. Dinu. 2014a. A Quantitative Insight into the Impact of Translation on Readability. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations, PITR@EACL 2014*, pages 104–113.

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

Alina Maria Ciobanu and Liviu P. Dinu. 2014d. Building a Dataset of Multilingual Cognates for the Romanian Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1038–1043.

Alina Maria Ciobanu and Liviu P. Dinu. 2014e. On the Romance Languages Mutual Intelligibility. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 3313–3318.

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 of the Asian Federation of Natural Language Processing, ACL 2015, Volume 2: Short Papers*, pages 431–437.

Alina Maria Ciobanu and Liviu P. Dinu. 2016. A Computational Perspective on the Romanian Dialects. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*.

Alina Maria Ciobanu and Liviu P. Dinu. 2018a. Ab Initio: Automatic Latin Proto-word Reconstruction. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, pages 1604–1614.

Alina Maria Ciobanu and Liviu P. Dinu. 2018b. Simulating Language Evolution: a Tool for Historical Linguistics. In *The 27th International Conference on Computational Linguistics: System Demonstrations, COLING 2018*, pages 68–72.

Alina Maria Ciobanu, Liviu P. Dinu, and Flaviu Pepelea. 2015. Readability Assessment of Translated Texts. In *Recent Advances in Natural Language Processing, RANLP 2015*, pages 97–103.

Alina Maria Ciobanu, Liviu P. Dinu, and Andrea Sgarro. 2017a. Towards a Map of the Syntactic Similarity of Languages. In *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, Revised Selected Papers, Part I*, pages 576–590.

Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, and Liviu P. Dinu. 2017b. Including Dialects and Language Varieties in Author Profiling. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*.

Alina Maria Ciobanu, Shervin Malmasi, and Liviu P. Dinu. 2018a. German Dialect Identification Using Classifier Ensembles. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial@COLING 2018*, pages 288–294.

Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, Santanu Pal, and Liviu P. Dinu. 2018b. Discriminating between Indo-Aryan Languages Using SVM Ensembles. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial@COLING 2018*, pages 178–184.

Alina Maria Ciobanu. 2016. Sequence Labeling for Cognate Production. In *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 20th International Conference KES 2016*, pages 1391–1399.

Ton Dijkstra, Franc Grootjen, and Job Schepens. 2012. Distributions of Cognates in Europe as Based on Levenshtein Distance. *Bilingualism: Language and Cognition*, 15:157–166.

Liviu P. Dinu and Alina Maria Ciobanu. 2017. Romanian Word Production: An Orthographic Approach Based on Sequence Labeling. In *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, Revised Selected Papers, Part I*, pages 591–603.

Russel Gray and Quentin Atkinson. 2003. Language Tree Divergences Support the Anatolian Theory of Indo-European Origin. *Nature*, 426:435–439.

Robert Anderson Hall. 1960. *Linguistics and Your Language*. Doubleday New York.

Paul Heggarty. 2012. Beyond Lexicostatistics: How to Get More out of "Word List" Comparisons. In *Quantitative Approaches to Linguistic Diversity: Commemorating the Centenary of the Birth of Morris Swadesh*, pages 113–137. Benjamins.

Gerhard Jäger. 2018. Computational Historical Linguistics. *CoRR*, abs/1805.08099.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates Can Improve Statistical Translation Models. In *Proceedings of HLT-NAACL*, pages 46–48.

Johann-Mattis List, Simon J. Greenhill, and Russell D. Gray. 2017. The Potential of Automatic Word Comparison for Historical Linguistics. *PLOS ONE*, 12(1):1–18.

April McMahon, Paul Heggarty, Robert McMahon, and Natalia Slaska. 2005a. Swadesh Sublists and the Benefits of Borrowing: an Andean Case Study. *Transactions of the Philological Society*, 103(2):147–170.

April McMahon, Paul Heggarty, Robert McMahon, and Natalia Slaska. 2005b. Swadesh Sublists and the Benefits of Borrowing: an Andean Case Study. *Transactions of the Philological Society*, 103(2):147–170.

James W. Minett and William S.-Y. Wang. 2003. On Detecting Borrowing: Distance-based and Character-based Approaches. *Diachronica*, 20(2):289–331.

Ee-Lee Ng, Beatrice Chin, Alvin W. Yeo, and Bali Ranaivo-Malançon. 2010. Identification of Closely-Related Indigenous Languages: An Orthographic Approach. *Int. J. of Asian Lang. Proc.*, 20(2):43–62.

Sergiu Nisioi, Alina Maria Ciobanu, and Liviu P. Dinu. 2016. Vanilla Classifiers for Distinguishing between Similar Languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial@COLING 2016*, pages 235–242.

Michael P. Oakes. 2000. Computer Estimation of Vocabulary in a Protolanguage from Word Lists in Four Daughter Languages. *Journal of Quantitative Linguistics*, 7:233–243.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*.

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.

Marcos Zampieri, Alina Maria Ciobanu, and Liviu P. Dinu. 2017. Native Language Identification on Text and Speech. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@EMNLP 2017*, pages 398–404.