Simulating Language Evolution: A Tool for Historical Linguistics

Alina Maria Ciobanu, Liviu P. Dinu
Faculty of Mathematics and Computer Science, University of Bucharest
Human Language Technologies Research Center, University of Bucharest
alina.ciobanu@my.fmi.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

Language change across space and time is one of the main concerns in historical linguistics. In this paper, we develop a language evolution simulator: a web-based tool for word form production to assist in historical linguistics, in studying the evolution of the languages. Given a word in a source language, the system automatically predicts how the word evolves in a target language. The method that we propose is language-agnostic and does not use any external knowledge, except for the training word pairs.

1 Introduction

Natural languages are living eco-systems, they are constantly in contact and, by consequence, they change continuously. Two of the fundamental questions in historical linguistics are the following (Rama and Borin, 2014): i) *How are languages related?* and ii) *How do languages change across space and time?*. In this paper, we focus on the second question. More specifically, we investigate how words enter a target language from a source language.

Traditionally, both problems were investigated with comparative linguistics instruments (Campbell, 1998) and required a manual process. Most of the previous approaches to word form production relied on phonetic transcriptions. They built on the idea that, given the phonological context, sound changes follow certain regularities across the entire vocabulary of a language. The proposed methods (Eastlack, 1977; Hartman, 1981) required a list of known sound correspondences as input, collected from dictionaries or published studies.

Modern approaches impose the use and development of quantitative and computational methods in this field (McMahon et al., 2005; Heggarty, 2012; Atkinson, 2013), or even cross-disciplinary methods (such as those borrowed from biology). Nowadays, given the development of the machine learning techniques, computers are able to learn sound or character correspondences automatically from pairs of known related words. Beinborn et al. (2013) proposed such a method for cognate production, using the orthographic form of the words, and applying a machine translation method based on characters instead of words. The orthographic approach relies on the idea that sound changes leave traces in the orthography and alphabetic character correspondences represent, to a fairly large extent, sound correspondences (Delmestri and Cristianini, 2010). Aligning the related words to extract orthographic changes from one language to another has proven very effective, when applied to both the orthographic (Gomes and Lopes, 2011) and the phonetic (Kondrak, 2000) form of the words . For the task of cognate production based on the orthography of the words, besides the character-based machine translation approach mentioned above, another contribution belongs to Mulloni (2007), who introduced an algorithm for cognate production based on edit distance alignment and the identification of orthographic cues when words enter a new language. Another probabilistic approach to word form production is based on building generative models from the phylogenetic tree of languages, modeling the evolution of the languages and capturing various aspects of language change (Bouchard-Côté et al., 2009; Hall and Klein, 2010).

Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 68–72
Santa Fe, New Mexico, USA, August 20-26, 2018.

2 Simulating Language Evolution

We propose a method for word form production based on the orthography of the words, building on the idea that orthographic changes represent sound correspondences to a fairly large extent (Delmestri and Cristianini, 2010). Given the form of a word *u* in a source language L_1, our system predicts the form *v* of the word *u* in a target language L_2, in the hypothesis that the word *v* will be derived in L_2 from the word *u*.

From the alignment of the related words in the training set we learn orthographic cues and patterns for the changes in spelling. We use the alignment as input for a sequence labeling system (assigning a sequence of labels to a sequence of tokens), based on an approach that has been proven useful for cognate production (Ciobanu, 2016; Dinu and Ciobanu, 2017), proto-word reconstruction (Ciobanu and Dinu, 2018) and for generating transliterations (Ammar et al., 2012).

We conduct our experiments on Romanian as a target language, and experiment with 10 source languages from which words entered in Romanian.

2.1 Word Alignment

To align pairs of words we employ the Needleman-Wunsch global alignment algorithm (Needleman and Wunsch, 1970), with the orthographic form of the words as input sequences and a very simple substitution matrix, which gives equal scores to all substitutions, disregarding diacritics (e.g., we ensure that *e* and *é* are matched). For example, for the Romanian word *descifrabil* (meaning *decipherable*), borrowed from the French word *déchiffrable*, the alignment is as follows:

d	é	-	c	h	i	f	f	r	a	b	-	l	e
d	e	s	c	-	i	f̌	-	r	a	b	i	l	-

2.2 Sequence Labeling

The words in the source language are the sequences, and the characters are the tokens. Our purpose is to obtain, for each input word, a sequence of characters that compose its related word in the target language. To this end, we use first- and second-order conditional random fields (CRFs) (Lafferty et al., 2001). For each character in the source word (after the alignment), the corresponding label is the character which occurs on the same position in the target word. In case of insertions, the characters are added to the previous label. We account for affixes separately: we add two extra characters B and E, marking the beginning and the end of an input word. In order to reduce the number of labels, for input tokens that are identical to their labels we replace the label with $\star$. For the previous example, the labels are as follows:

B	d	é	c	h	i	f	f	r	a	b	l	e	E
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
$\star$	$\star$	es	$\star$	-	$\star$	$\star$	-	$\star$	$\star$	bi	$\star$	-	$\star$

As features for the sequence labeling system, we use character n-grams in a window of size w around the current token.

2.3 Experiments

We run experiments on a dataset of word-etymon pairs (Ciobanu and Dinu, 2014), from which we extract Romanian words having etymons in 10 languages. The dataset was built from an aggregation of machine-readable dictionaries[1] that contains information about the etymology of the words.The dataset is structured as a list of word pairs having the form: $w_1(L_1) \rightarrow w_2(L_2)$, where word w_2 entered L_2 from the L_1 word w_1. Example: *victoria (Latin)* $\rightarrow$ *victorie (Romanian)*. We use subsets of 800 word pairs for each language, to have an equal size that allows a comparison between source languages. The results are reported in Table 1. In Table 2 we show examples of our system's output.

We split the datasets in subsets for training, development and testing with a ratio of 3:1:1. We use the CRF implementation provided by the Mallet toolkit for machine learning (McCallum, 2002). We perform

[1] https://dexonline.ro

Source language	Baseline		Our system	
	EDIT (un-normalized)	EDIT (normalized)	EDIT (un-normalized)	EDIT (normalized)
English	2.04	0.23	1.33	0.15
French	2.16	0.24	1.42	0.15
Italian	2.60	0.32	1.62	0.23
Latin	2.75	0.34	1.76	0.22
Neo-Greek	2.39	0.29	1.82	0.24
Old Slavic	2.34	0.33	1.84	0.27
German	2.36	0.32	2.00	0.29
Turkish	1.88	0.27	2.01	0.29
Portuguese	2.95	0.52	2.50	0.43
Spanish	3.22	0.53	3.06	0.50

Table 1: Word form production for Romanian words.

a grid search for the number of iterations in $\{1, 5, 10, 25, 50, 100\}$ and for the size of the window w in $\{1, 2, 3\}$. We use a "majority class" type of baseline that does not take context into account, as described by Ciobanu (2016).

We use the edit distance (Levenshtein, 1965) between the produced words and the gold standard to evaluate the performance of our method. We use both an un-normalized and a normalized version of the edit distance. To obtain the normalization, we divide the edit distance by the length of the longer string.

We use lemmas (dictionary word forms) as input. We further experiment with some additional pre-processing steps on the input data (diacritics removal and stemming). The results are slightly improved when diacritics are not taken into account. Stemming does not improve performance, which shows that Romanian is a complex language, and foreign influences, in the case of new words entering the language, occur in the root of the words as well. Our system obtains the best results for English and French as source languages. The languages ranked higher are those with which Romanian had the most intense cultural collaboration, either more recently (English, for example), or in the past (Italian and French). The word production performance is lower even for related languages (as Portuguese and Spanish); these languages are more remote from Romania, from a geographical point of view, and this might have made the contact between languages more difficult.

Source language	Word	5-best productions
English	immunopathology	**imunopatologie**, immunopatologie, imunopafologie, imunopatologi, imunopathologie
French	opaliser	opalizare, **opaliza**, opalizară, opalizat, opalizăre
Italian	nivellazione	nivellaţie, nivellaţieu, **nivelaţie**, nivellaţiea, nivellaţiune
Latin	desideratum	desiderat, **deziderat**, desiderati, desideratu, deziderati
Neo-Greek	atherina	atherină, **aterină**, atherina, aterina, atherinire
Old Slavic	stihija	**stihie**, stihii, stihi, stihij, stihij
German	schabotte	şabottă, şabot, şabott, **şabotă**, şabotte
Turkish	peškeš	peşkeş, pešcheş, **peşcheş**, peşkşš, peşkeš
Portuguese	terneça	**tinereţe**, tinereçă, tinereţă, tinereçe, tereţe
Spanish	sainete	sainet, **sainetă**, sainete, săinet, saine

Table 2: Examples of word form production for Romanian words. We highlight the correct productions in bold.

3 A Tool for Historical Linguistics

We built a web application[2] to expose our system for word production. Its purpose is to assist linguists studying language evolution and language change, by providing n-best lists of possible word productions, when words enter a target language from a given source language. Its main impact is that it will narrow down the possibilities worth investigating when reconstructing a language, or when investigating language evolution.

The users of the application enter the source word, select the source language (from the possible 10 languages) and the system simulates the evolution of the word in Romanian. The web interface is rendered in Figure 1, along with an example produced by our system: given the source French word *documentaire*, the system produces a 10-best list of word forms in Romanian, having the correct word (*documentar*) on the first position.

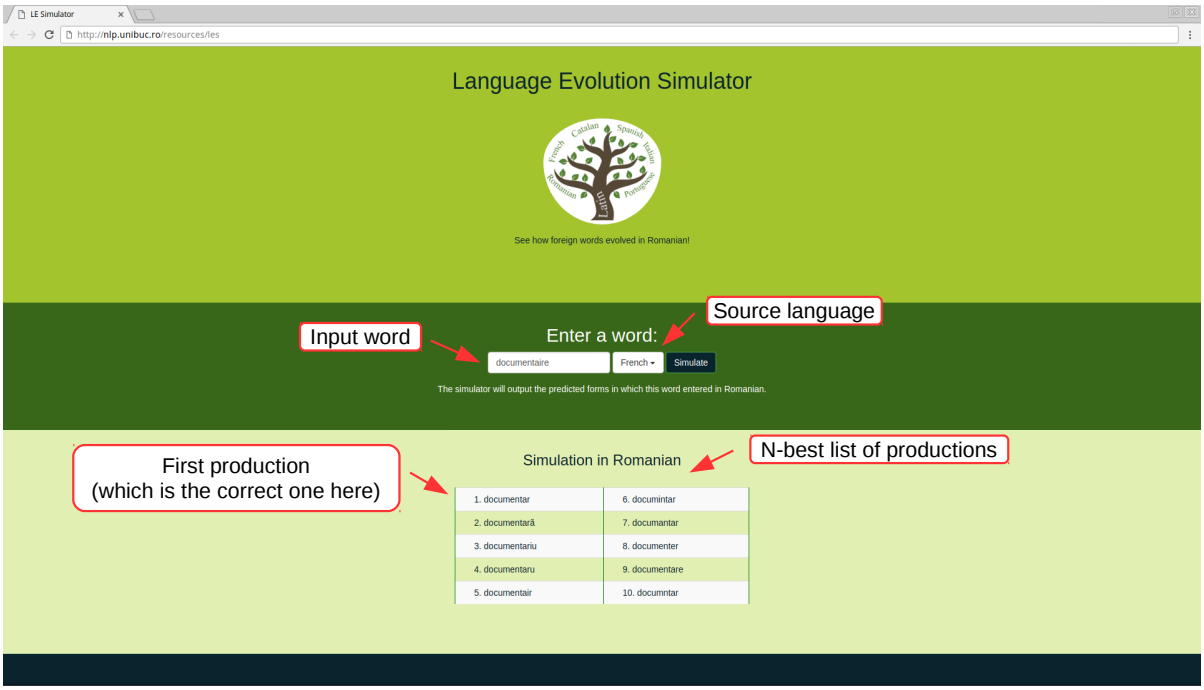

Figure 1: Language evolution simulator tool.

4 Conclusions

In this paper, we presented an automatic method for word form production, based on the orthography of the words. We experimented with Romanian as a target language and multiple source languages. We developed a language evolution simulator: a tool to be used in historical linguistics, to help in the investigation of language evolution. Given words in a source language, the system automatically predicts how they evolve in a target language.

As future work, we intend to enhance the system with more target languages, as we gain access to more data, to extend the user interface to handle blocks of text, not only single words as input, and to incorporate more types of relationships between words (cognate production and proto-word production) into the application.

Acknowledgments

We thank the anonymous reviewers for their helpful and constructive comments. The contribution of the authors to this paper is equal. Research supported by UEFISCDI, project number 53BG/2016.

[2] http://nlp.unibuc.ro/resources/les

References

Waleed Ammar, Chris Dyer, and Noah A Smith. 2012. Transliteration by sequence labeling with lattice encodings and reranking. In *Proceedings of the 4th Named Entity Workshop*, pages 66–70.

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

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate Production using Character-based Machine Translation. In *Proceedings of IJCNLP 2013*, pages 883–891.

Alexandre Bouchard-Côté, Thomas L. Griffiths, and Dan Klein. 2009. Improved Reconstruction of Protolanguage Word Forms. In *Proceedings of NAACL 2009*, pages 65–73.

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

Alina Maria Ciobanu and Liviu P. Dinu. 2014. An Etymological Approach to Cross-Language Orthographic Similarity. Application on Romanian. In *Proceedings of EMNLP 2014*, pages 1047–1058.

Alina Maria Ciobanu and Liviu P. Dinu. 2018. Ab Initio: Latin Proto-word Reconstruction. In *Proceedings of COLING 2018*.

Alina Maria Ciobanu. 2016. Sequence Labeling for Cognate Production. In *Proceedings of KES 2016*, pages 1391–1399.

Antonella Delmestri and Nello Cristianini. 2010. String Similarity Measures and PAM-like Matrices for Cognate Identification. *Bucharest Working Papers in Linguistics*, 12(2):71–82.

Liviu P. Dinu and Alina Maria Ciobanu. 2017. Romanian Word Production: an Orthographic Approach Based on Sequence Labeling. In *Proceedings of CICLing 2017*.

Charles L. Eastlack. 1977. Iberochange: A Program to Simulate Systematic Sound Change in Ibero-Romance. *Computers and the Humanities*, 11:81–88.

Luís Gomes and José Gabriel Pereira Lopes. 2011. Measuring Spelling Similarity for Cognate Identification. In *Proceedings of EPIA 2011*, pages 624–633.

David Hall and Dan Klein. 2010. Finding Cognate Groups Using Phylogenies. In *Proceedings of ACL 2010*, pages 1030–1039.

Steven Lee Hartman. 1981. A Universal Alphabet for Experiments in Comparative Phonology. *Computers and the Humanities*, 15:75–82.

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.

Grzegorz Kondrak. 2000. A New Algorithm for the Alignment of Phonetic Sequences. In *Proceedings of NAACL 2000*, pages 288–295.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML 2001*, pages 282–289.

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

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

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

Andrea Mulloni. 2007. Automatic Prediction of Cognate Orthography Using Support Vector Machines. In *Proceedings of the ACL Student Research Workshop*, pages 25–30.

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.

Taraka Rama and Lars Borin. 2014. Comparative Evaluation of String Similarity Measures for Automatic Language Classification. In George K. Mikros and Jn Macutek, editors, *Sequences in Language and Text*. De Gruyter Mouton.