On the Romance Languages Mutual Intelligibility

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

We propose a method for computing the similarity of natural languages and for clustering them based on their lexical similarity. Our study provides evidence to be used in the investigation of the written intelligibility, i.e., the ability of people writing in different languages to understand one another without prior knowledge of foreign languages. We account for etymons and cognates, we quantify lexical similarity and we extend our analysis from words to languages. Based on the introduced methodology, we compute a matrix of Romance languages intelligibility.

Keywords: Romance languages, etymology, cognates, string similarity

1. Introduction and Related Work

Determining degrees of similarity between the world's languages is an intensely debated issue (Lebart and Rajman, 2000), many of the controversies in historical and comparative linguistics being centered on language classification (McMahon and McMahon, 2003). In spite of the fact that linguistic literature abounds in claims of classification of natural languages, McMahon and McMahon (2003) argue for the necessity of development of quantitative and computational methods in this field. Methods for comparing languages are constantly developed and periodically reassessed (Ringe et al., 2002; Alekseyenko et al., 2012; Atkinson et al., 2005; Barbancon et al., 2013) and many of them have crossed the discipline boundaries by borrowing computational tools from different fields (Bortolussi et al., 2011). Dyen et al. (1992) investigate the classification of Indo-European languages by applying a lexicostatistical method. Campbell (2003) analyzes various approaches used over time for establishing relationships between languages, emphasizing the popularity of the comparative method. Barbancon et al. (2013) show that the difficulty in the evaluation of the results regarding phylogenetic trees reconstruction resides in the variety of computational methods used and in the differences in datasets. McMahon and McMahon (2003) point out that in many situations the similarity of natural languages is a fairly vague notion, both linguists and non-linguists having rather intuitions about which languages are more similar to which others; in some cases, they are based on the very subjective opinions of the authors. If grouping of languages in linguistic families is generally accepted, the relationships between languages belonging to the same family are still controversial and are periodically investigated. Degrees of similarity between languages are far from being certain; values vary considerably from one researcher to another, not only for exotic languages, but even for extensively studied languages, many of which are closely related.

According to Gooskens (2007), some genetically related languages are so close to each other, that the speakers are able to communicate without prior instruction. Gooskens et al. (2008) analyze several phonetic and lexical predictors of intelligibility and, to determine the relevance of each linguistic level, they correlate the intelligibility scores with lexical and phonetic distances. Their analysis leads to the conclusion that the two levels are to a large extent independent and that linguistic distances can successfully predict intelligibility between closely related languages. Regarding lexical distances, they account for the number of non-cognates, arguing that these words are basically unintelligible to listeners without prior knowledge of the considered language and that intelligibility is inversely related to the number of non-cognates. The language intelligibility problem is also mentioned in the report published in 2007 at the European Commission by the High Level Group on Multilingualism (HLGM), which emphasizes "a lack of knowledge about mutual intelligibility between closely related languages in Europe and the lack of knowledge about the possibilities for communicating through receptive multilingualism, i.e., where speakers of closely related languages each speak their own language". In today's context of European multilingualism and massive population mobility, a deeper insight into this matter might not have only a theoretical, cultural, communicative, educational or scientific impact, but an economic or business impact as well. In this paper we investigate the similarity of natural languages with respect to their written intelligibility, i.e., the ability of people writing in different languages to understand one another without prior knowledge of foreign languages. The written form of a language is found not only in literature, but in other various forms as well: movie subtitles, on-line news or communication networks (chats, for example). In a broadly accepted sense, a language L_1 is closer to a language L_2 when texts written in L_2 are easier understood by speakers of L_1 without prior knowledge of L_2. The reverse is also true. In other words, the higher the intelligibility degree between two languages, the closer they are.

1.1. Our Approach

Although there are multiple aspects that are relevant in the study of language relatedness, such as orthographic, phonetic, syntactic and semantic differences, in this paper we focus only on lexical similarity. The orthographic ap-

proach 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). The motivation of our approach is that, when people encounter a language for the first time in written form, it is more likely that they can distinguish and individualize words which resemble words from their native language. These words are probably either inherited from their mother tongue (etymons), or have a common ancestor with the words in their language (cognates). We propose a dictionary-based approach to automatically extract related words and a method for computing the lexical similarity of natural languages. Our approach implies a detailed investigation which comprises, besides quantitative aspects, a qualitative insight into the relatedness of languages and accounts not only for the number of related words, as it is usually done in lexicostatistics, but also for their forms, quantifying lexical similarity.

As a case study we choose a good candidate for European languages, namely the Romance family. We investigate the written intelligibility of the following Romance languages: Romanian, Italian, French, Spanish and Portuguese. Even though they are closely related, not only do degrees of similarity between any two Romance languages vary from one author to another, but their classification is also controversial. For example, McMahon and McMahon (2003) report two different results for the classification of Romanian within the Romance family, either marginal or more integrated within the group.

The paper is organized as follows: in Section 2 we present our approach and methodology. In Section 3 we report and analyze the results obtained for the Romance languages, and Section 4 is dedicated to conclusions and future work.

2. Algorithm and Methodology

In this section we briefly describe our method for determining lexical similarity between related languages, based on linguistic relationships identification and string similarity computation (see also Figure 1). The basic steps of our methodology are detailed below.

2.1. Algorithm

2.1.1. Preprocessing

Given a corpus C, we start by preprocessing the text.

Step 1. Data Cleaning. We perform basic word segmentation, using whitespace and punctuation marks as delimiters and we lower-case all words. We remove from our corpora tokens that are irrelevant for our investigation, such as dates, numbers and non-textual annotations marked by non-alphanumeric characters.

Step 2. Stop Words Removal. We focus on analyzing word content and, in order to obtain relevant results, we remove stop words from our corpora. We use the lists of stop words for Romance languages provided by the *Apache Lucene*[1] text search engine library.

[1] http://lucene.apache.org

Step 3. Lemmatization. We use lemmas for identifying words' definitions in dictionaries and for computing adequate distances between related words. We use the *FreeLing*[2] language analysis tool suite (Padró and Stanilovsky, 2012; Padró, 2011; Padró et al., 2010; Atserias et al., 2006; Carreras et al., 2004) to lemmatize French, Italian, Spanish and Portuguese words and the *DexOnline*[3] machine-readable dictionary to lemmatize Romanian words. *DexOnline* provides information regarding the words' inflected forms and enables us to correctly identify lemmas where no part-of-speech or semantic ambiguities arise (in this case we consider the first occurred lemma).

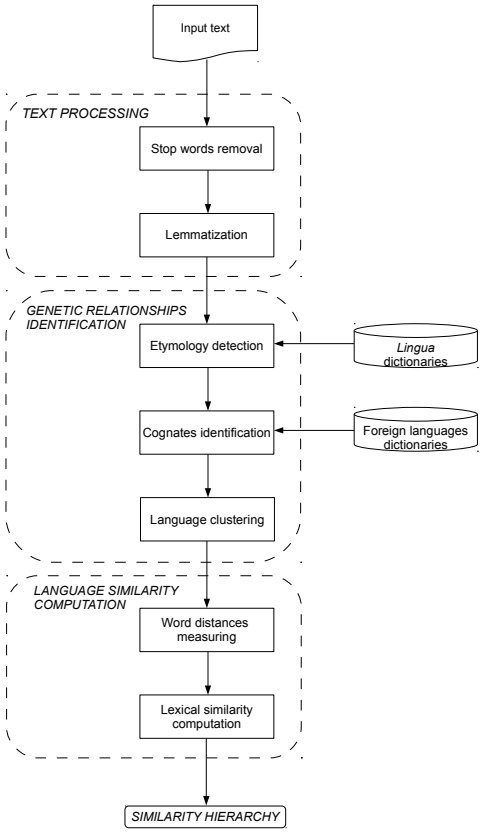


Figure 1: Algorithm for determining the lexical similarity between a corpus in language $Lingua$ and related languages

2.1.2. Relationships Identification

Step 1. Etymology Detection. For most words, etymological dictionaries offer a unique etymology, but when more options are possible for explaining a word's etymology (there are words whose etymology was and remains difficult to ascertain), dictionaries may provide multiple alternatives. We account for all the given etymological hypotheses, enabling our method to provide more accurate results. We use electronic dictionaries to extract information regarding words' etymologies and etymons for Romanian[3],

[2] http://nlp.lsi.upc.edu/freeling

[3] http://dexonline.ro

Italian[4], French[5], Spanish[6] and Portuguese[7].

Step 2. Cognates Identification. Cognates are words in different languages having the same etymology and a common ancestor. The task of cognates identification is widely used 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. The main problem is the automatic detection of cognate pairs. For this reason, most previous works regarding languages' intelligibility report results on a small number of cognate pairs, which are usually manually determined. We introduce an automatic strategy for detecting pairs of cognates between two given languages, which enables the identification of all cognate pairs for the studied corpora. Considering a set of words in a given language L_1, to identify the cognate pairs between L_1 and a related language L_2, we first determine the etymologies of the given words. Then we translate in L_1 all words without L_2 etymology[8]. We consider cognate candidate pairs formed of input words and their translations. Using electronic dictionaries, we extract etymology-related information for the translated words. To identify cognates we compare, for each pair of candidates, the etymologies and the etymons. If they match, we identify the words as being cognates. We assume that etymons match even when they are different inflected forms of the same word.

In order to evaluate our automatic method for extracting etymology-related information and for detecting related words, we randomly excerpt 500 words for each of the considered languages (Romanian, French, Italian, Spanish and Portuguese) and we manually determine their etymologies. Then, we compare these results with the automatically obtained etymologies and compute the accuracy for etymology extraction for each language. We obtain the following results: 95.45% accuracy for Romanian, 98% for Italian, 96.6% for French, 98.2% for Spanish and 99,8% for Portuguese.

2.1.3. Linguistic Distances

We are interested in determining lexical similarity, but not only regarding the number of words having etymons or cognates in foreign languages, but also regarding the resemblance degrees between the related words. For measuring orthographic distances between words we use the following three metrics: edit distance (Levenshtein, 1965), longest common subsequence ratio (Melamed, 1995) and rank distance (Dinu and Dinu, 2005).

Let $C = \{w_1, w_2, ..., w_{N_{words}}\}$ be a corpus in L_1 and let L_2 be a related language. We assume, without any loss of generality, that the elements of C are ordered such that $C_L = \{w_1, w_2, ..., w_{N_{lingua}}\}$ is the subset of C containing all the words that have an etymon or a cognate pair in L_1.

We use the following notations: N_{words} is the number of token words in C, N_{lingua} is the number of token words in C_L, λ is the empty string and x_i is the etymon or cognate pair of w_i in L_2. Given a string distance Δ, we define the distance between L_1 and L_2 (with frequency support from corpus C) as follows:

$$\Delta(L_1, L_2) = 1 - \frac{N_{lingua}}{N_{words}} + \frac{\sum_{i=1}^{N_{lingua}} \Delta(w_i, x_i)}{N_{words}} \quad (1)$$

Hence, the similarity between languages L_1 and L_2 is defined as follows:

$$Sim(L_1, L_2) = 1 - \Delta(L_1, L_2) \quad (2)$$

3. Application: Romance Languages

3.1. Corpora

A major problem in our investigation was the lack of an agreed corpus for Romance languages on which to apply our method. We decided to use three multilingual corpora:

- **George Orwell's "1984" Novel**, translated in a large number of languages and widely investigated in NLP applications;
- **Europarl**, a parallel corpus extracted from the European Parliament web site with the main intended use as aid for statistical machine translation research (Tiedemann, 2012);
- **Content of Wikipedia**, the well-known multilingual web-based encyclopedia, collaboratively edited and comprising about 30 million articles in 287 languages.

In this paper, we focus on the most representative set of words for each language, as mutual intelligibility is highly related to the words' level of usage. We assume that words which are part of the basic lexicon of a language *L* are more relevant in our study, as they are more likely to be recognized by non-speakers of *L* and contribute to the languages' perceived proximity to a higher extent than the rest of the vocabulary. Therefore, we account only for words in our corpora that are part of each language's basic lexicon. The basic lexicon represents the nucleus of a language's vocabulary and it is generally regarded as an important source of supporting evidence in this field (Campbell, 2003). Some of the criteria that have been used over time for determining the composition of a basic lexicon are stated by Dinu (1996): the polysemy, the productivity (the number of derived words), the abundance of expressions and locutions in which the word is used, the degree of usage throughout the country's territory and layers of society. For our experiments we use the representative vocabularies of Romance languages proposed by Sala (1988), which contain 2,588 distinct words for Romanian, 2,608 for Italian, 2,613 for Spanish, 2,309 for Portuguese and 2,561 for French. The words from the basic lexicons cover, in average, 55% of the total numbers of words in each corpus. In Table 1 we provide the number of type words, token words and lemmas for the three corpora and in Table 2 we report the number of type and token related words for each corpus.

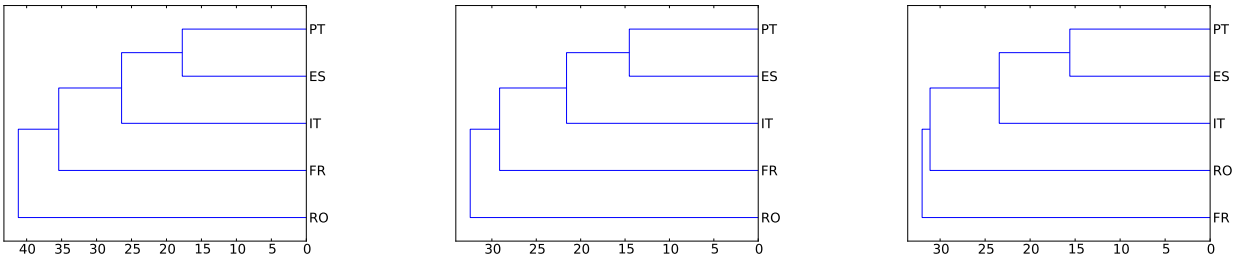
[4] http://www.sapere.it/sapere/dizionari

[5] http://www.cnrtl.fr

[6] http://lema.rae.es/drae

[7] http://www.infopedia.pt/lingua-portuguesa

[8] We translate Romanian words using *Google Translate*: http://translate.google.com



(a) "1984" (b) Europarl (c) Wikipedia

Figure 2: Dendrograms which illustrate the arrangement of the clusters, for each corpus, produced by the neighbor joining algorithm for hierarchical clustering, using the output of our method as input matrices

	Type Words	Token Words	Lemmas
RO	13,090	107,001	6,844
IT	14,234	103,864	8,816
FR	11,808	110,523	7,293
SP	11,515	96,026	5,978
PT	12,305	87,978	7,019

(a) "1984"

Type Words	Token Words	Lemmas
34,797	8,851,292	12,655
52,250	42,403,116	18,475
48,278	47,457,557	17,951
55,853	46,560,579	17,841
56,010	42,115,216	18,676

(b) Europarl

Type Words	Token Words	Lemmas
123,203	23,435,933	57,224
90,307	98,759,935	57,555
97,220	101,434,165	176,150
135,275	180,682,358	80,254
85,439	37,732,789	53,366

(c) Wikipedia

Table 1: Number of type words, token words and lemmas for multilingual corpora

	RO	IT	FR	ES	PT
RO	–	18,171	20,059	15,591	14,463
IT	14,640	–	5,956	7,361	8,888
FR	19,067	5,406	–	7,235	8,831
ES	15,549	7,414	7,752	–	14,195
PT	13,411	5,980	7,190	10,964	–

(a) "1984"

RO	IT	FR	ES	PT
–	1,687,300	2,034,434	1,325,784	1,519,814
7,085,677	–	3,508,904	4,084,724	5,059,015
8,740,955	2,480,460	–	3,653,705	4,612,181
6,510,964	3,868,224	3,781,382	–	6,830,413
4,517,895	2,700,410	3,580,361	4,823,357	–

(b) Europarl

RO	IT	FR	ES	PT
–	3,976,075	4,581,886	3,330,607	3,678,648
15,685,118	–	7,519,960	9,697,581	10,820,392
16,321,331	4,658,257	–	7,517,925	9,644,569
24,967,547	15,370,195	14,679,919	–	26,033,820
4,702,649	2,762,838	3,286,173	5,111,693	–

(c) Wikipedia

Table 2: Number of related words for multilingual corpora

3.2. Results Analysis

We apply the method described in Section 2 on words from each corpus which are part of each language's basic lexicon. We compute pairwise similarity between related words and we extend the analysis from words to languages. Due to space constraints, we report in Tables 3, 4 and 5 the average values of the three metrics described in Section 2, computed as percentages. The meaning of the tables is the following: the value situated at the intersection of line *X* and column *Y* represents the degree of comprehensibility for a speaker of language *Y* from a corpus written in language *X*.

The matrices reported in Tables 3, 4 and 5 are not symmetrical. In order to compute the similarity between L_1 and L_2, our method accounts for both cognates and etymons. For this reason, the similarity method we propose is not symmetrical, because the set of words inherited by L_1 from L_2 is different from the set of words inherited by L_2 from L_1. In other words, cognate pairs shared between L_1 and L_2 are symmetrical, but word-etymon pairs aren't.

One can notice that, generally, for speakers of Romance languages, Romanian is the least intelligible language. In other words, the degree of intelligibility of Romanian for another Romance language is lower than the intelligibility degree of any two other Romance languages. The closest languages are Spanish and Portuguese, followed by Italian and Spanish. The highest dissimilarity is found between Romanian and Spanish: Romanian is the least intelligible to Spanish (this is the lowest degree of intelligibility between any two Romance languages). However, Spanish is much more intelligible to speakers of Romance languages. In general, Romance languages are more accessible to Romanian than is Romanian accessible to those languages. This is due to the development of Romanian far from the Romance kernel. However, these differences are not very significant.

In Figure 2 we plot the dendrograms which illustrate the arrangement of the clusters, for each corpus, produced by the neighbor joining algorithm for hierarchical clustering, using the output of our method as input matrices [9].

4. Conclusion and Future Work

In this paper we propose an automatic method for determining natural languages intelligibility. We compute the intelligibility of Romance languages using 3 multilingual corpora: George Orwell's "1984" novel, the Europarl parallel corpus and Wikipedia content. The positioning of French is interesting, especially in the Wikipedia data, where the

[9]Since the clustering method works with a symmetric matrix as input and our method provides an asymmetric matrix, we build a symmetric matrix by using, for languages L_1 and L_2, the average value of $Sim(L_1, L_2)$ and $Sim(L_2, L_1)$.

	RO	IT	FR	ES	PT
RO	–	54.79	56.96	46.98	56.07
IT	61.45	–	65.58	73.15	74.05
FR	54.88	63.60	–	63.84	63.20
ES	61.74	73.61	64.68	–	81.42
PT	61.54	73.11	60.25	83.10	–

Table 3: Matrix of similarity for the "1984" novel

	RO	IT	FR	ES	PT
RO	–	65.97	65.12	59.97	64.54
IT	69.15	–	72.93	77.37	75.33
FR	62.27	68.82	–	68.84	67.11
ES	70.17	79.45	69.98	–	84.56
PT	65.57	75.65	63.15	86.40	–

Table 4: Matrix of similarity for the Europarl corpus

	RO	IT	FR	ES	PT
RO	–	67.23	65.83	59.86	65.28
IT	70.54	–	68.60	77.27	75.50
FR	67.13	66.84	–	68.08	65.68
ES	69.25	75.88	67.92	–	83.89
PT	68.17	73.42	62.35	84.92	–

Table 5: Matrix of similarity for Wikipedia content

Western Romance languages are "split" by Romanian. The "1984" and Europarl data are closer to linguistic expectations, basically confirming decreasing similarity with Romanian from east to west. The different pattern shown by Wikipedia may be due to the fact that Wikipedia data does not provide a parallel corpus.

In our future work, we intend to develop the analysis of the basic lexicons to the entire corpora and to investigate the relationships between Romance languages and other language families. We also intend to develop a semi-automatic module for the word translation step in our method, based on a thorough preliminary analysis of the existing tools, such as GIZA++ (Och and Ney, 2003) or Moses (Koehn et al., 2007). We plan to investigate the semantic and part-of-speech ambiguities for the Romanian words and to improve the preprocessing step of our method with regard to lemmatization.

Acknowledgements

We thank the anonymous reviewers for their helpful and constructive comments. We thank Ovidiu Dobroiu for the help with processing the corpora used in our experiments. The contribution of the authors to this paper is equal. Research 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

Alexander V. Alekseyenko, Quentin D. Atkinson, Remco Bouckaert, Alexei J. Drummond, Michael Dunn, Russell D. Gray, Simon J. Greenhill, Philippe Lemey, and Marc A. Suchard. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. *Science*, 337(6097):957–960.

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.

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pages 2281–2286.

Francois Barbancon, Steven N. Evans, Luay Nakhleh, Don Ringe, and Tandy Warnow. 2013. An Experimental Study Comparing Linguistic Phylogenetic Reconstruction Methods. *Diachronica*, 30(2):143–170.

Luca Bortolussi, Andrea Sgarro, Giuseppe Longobardi, and Cristina Guardiano. 2011. How Many Possible Languages Are There? In *Biology, Computation and Linguistics - New Interdisciplinary Paradigms*, pages 168–179. IOS Press.

Lyle Campbell, 2003. *How to Show Languages are Related: Methods for Distant Genetic Relationship*. Blackwell.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, pages 239–242.

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.

Anca Dinu and Liviu P. Dinu. 2005. On the Syllabic Similarities of Romance Languages. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2005*, pages 785–788.

Mihai Dinu. 1996. *Personalitatea Limbii Române*. Cartea Romaneasca, Bucureşti.

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.

Charlotte Gooskens, Wilbert Heeringa, and Karin Beijering. 2008. Phonetic and Lexical Predictors of Intelligibility. *International Journal of Humanities and Arts Computing*, 2(1-2):63–81.

Charlotte Gooskens. 2007. The Contribution of Linguistic Factors to the Intelligibility of Closely Related Languages. *Journal of Multilingual and Multicultural Development*, 28(6):445.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of*

the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions, ACL 2007, pages 177–180.

Ludovic Lebart and Martin Rajman. 2000. Computing Similarity. In *Handbook of NLP*. Dekker: Basel.

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

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

Dan Melamed. 1995. Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons. In *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 184–198.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pages 2473–2479.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. FreeLing 2.1: Five Years of Open-source Language Processing Tools. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, pages 931–936.

Lluís Padró. 2011. Analizadores Multilingües en FreeLing. *Linguamatica*, 3(2):13–20.

Don Ringe, Ann Taylor, and Tandy Warnow. 2002. Indo-European and Computational Cladistics. *Transactions of the Philological Society*, 100(1):59–129.

Marius Sala. 1988. *Vocabularul Reprezentativ al Limbilor Romanice*. Editura Academiei, Bucureşti.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pages 2214–2218.