An Etymological Approach to Cross-Language Orthographic Similarity. Application on Romanian

Alina Maria Ciobanu, Liviu P. Dinu

University of Bucharest
Center for Computational Linguistics
http://nlp.unibuc.ro

EMNLP 2014

Overview

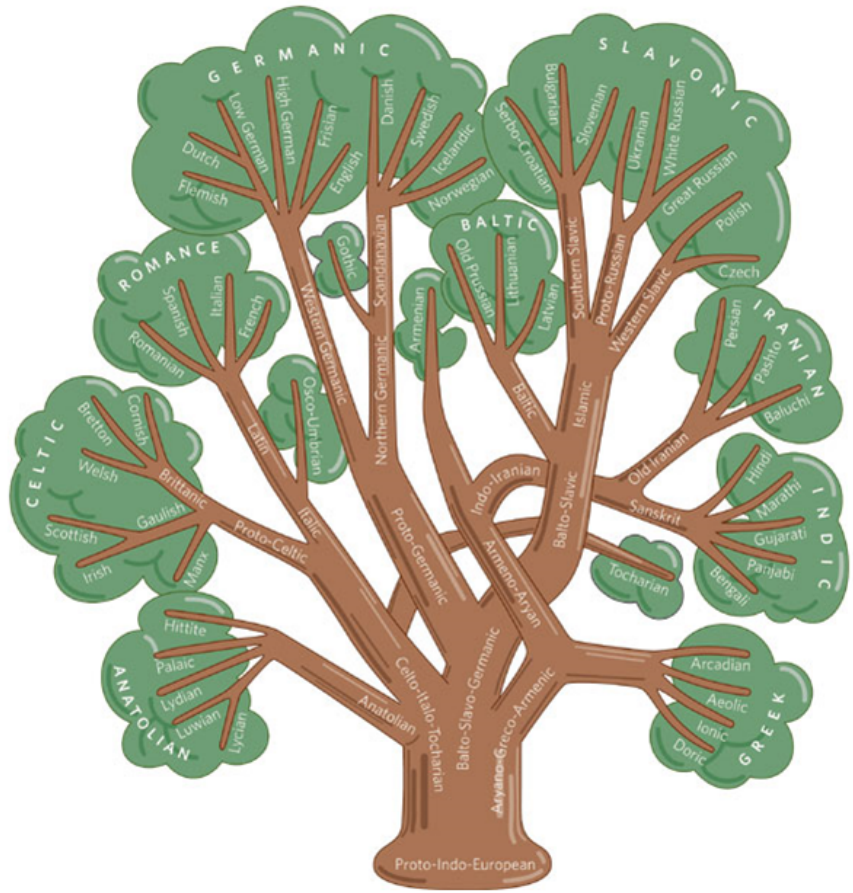
- Orthographic similarity: motivation and approach
- Identifying language relationships
- Computing degrees of similarity
  - Results on 3 Romanian corpora from different historical periods
  - Results on Europarl (Romanian subcorpus)
- Conclusions and future work

Language similarity

- The similarity of natural languages is a fairly vague notion, both linguists and non-linguists having intuitions about which languages are more similar to which others [McMahon and McMahon, 2003].
- Four types of similarity: typological, morphological, syntatic, lexical [Homola and Kubon, 2006].
- It is necessary to develop quantitative and computational methods in this field [McMahon and McMahon, 2003].

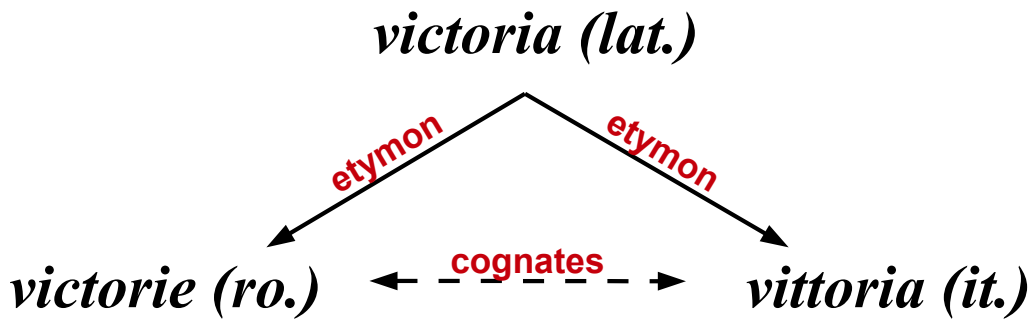
Applications

- Linguistic phylogeny reconstruction [Alekseyenko et al, 2012; Barbançon et al, 2013].
- Machine translation [Koppel and Ordan, 2011].
- Language acquisition [Benati and VanPatten, 2011].
- Language intelligibility assessment [Gooskens et al, 2008].



Our approach

- A language *L1* is closer to a language *L2* when texts written in *L2* are easier understood by speakers of *L1* without prior knowledge of *L2*.
- When people read a text in a foreign language, they first identify the words which resemble words from their native language.
- Two types of related words:
  - Word-etymon pairs
  - Cognate pairs



Orthographic similarity

- Some pairs of related words are closer than others.
- Word-etymon pairs:

  lună (ro.), ***luna*** (lat.) vs. ***bătrân*** (ro.), ***veteranus*** (lat.)

- Cognate pairs:

  vânt (ro.), ***vent*** (fr.) vs. ***castel*** (ro.), ***château*** (fr.)

Algorithm and methodology

Input: corpus C in L_1

1. Text processing
   - 1.1. Remove stop words
   - 1.2. Lemmatize
2. Language relationships identification
   - 2.1. Detect etymologies
   - 2.2. Identify cognates
   - 2.3. Cluster by language families
3. Language similarity computation
   - 3.1. Measure word distances
   - 3.2. Compute degrees of similarity

Output: similarity hierarchy for L_1

Similarity method

Definition

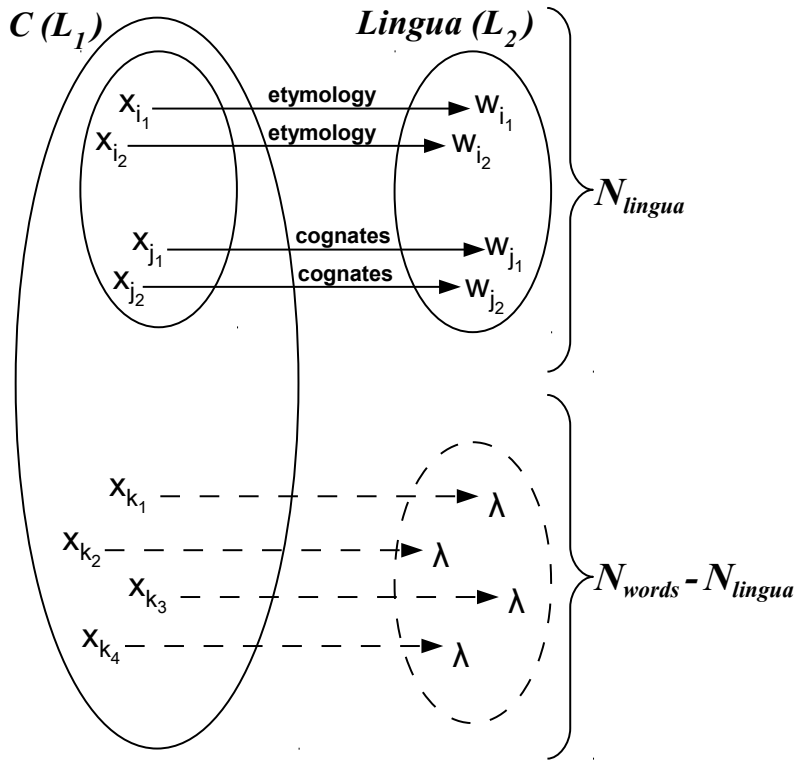
Given a string distance Δ, we define the distance between languages L_1 and L_2 (with frequency support from corpus C in L_1) as follows:

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

Definition

The similarity between L_1 and L_2 is:

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



Etymology detection

- We extract etymologies from electronic dictionaries.

Pattern

⟨abbr class="abbrev" title="limba language_name"⟩
    language_abbreviation
⟨/abbr⟩
⟨b⟩ etymon ⟨/b⟩

Entry

⟨b⟩ capitol ⟨/b⟩

⟨abbr class="abbrev" title="limba italiana"⟩
    it.
⟨/abbr⟩
⟨b⟩ capitolo ⟨/b⟩
⟨abbr class="abbrev" title="limba latina"⟩
    lat.
⟨/abbr⟩

⟨b⟩ capitulum ⟨/b⟩

Etymology detection

- We extract etymologies from electronic dictionaries.

Pattern

⟨abbr class="abbrev" title="limba language_name"⟩
    language_abbreviation
⟨/abbr⟩
⟨b⟩ etymon ⟨/b⟩

Entry

⟨b⟩ capitol ⟨/b⟩

⟨abbr class="abbrev" title="limba italiana"⟩
    it.
⟨/abbr⟩
⟨b⟩ capitolo ⟨/b⟩
⟨abbr class="abbrev" title="limba latina"⟩
    lat.
⟨/abbr⟩

⟨b⟩ capitulum ⟨/b⟩

Etymology detection

- We extract etymologies from electronic dictionaries.

Pattern

⟨abbr class="abbrev" title="limba language_name"⟩
    language_abbreviation
⟨/abbr⟩
⟨b⟩ etymon ⟨/b⟩

Entry

⟨b⟩ capitol ⟨/b⟩

⟨abbr class="abbrev" title="limba italiana"⟩
    it.
⟨/abbr⟩
⟨b⟩ capitolo ⟨/b⟩
⟨abbr class="abbrev" title="limba latina"⟩
    lat.
⟨/abbr⟩

⟨b⟩ capitulum ⟨/b⟩

Etymology detection

- We extract etymologies from electronic dictionaries.

Pattern

⟨abbr class="abbrev" title="limba language_name"⟩
    language_abbreviation
⟨/abbr⟩
⟨b⟩ etymon ⟨/b⟩

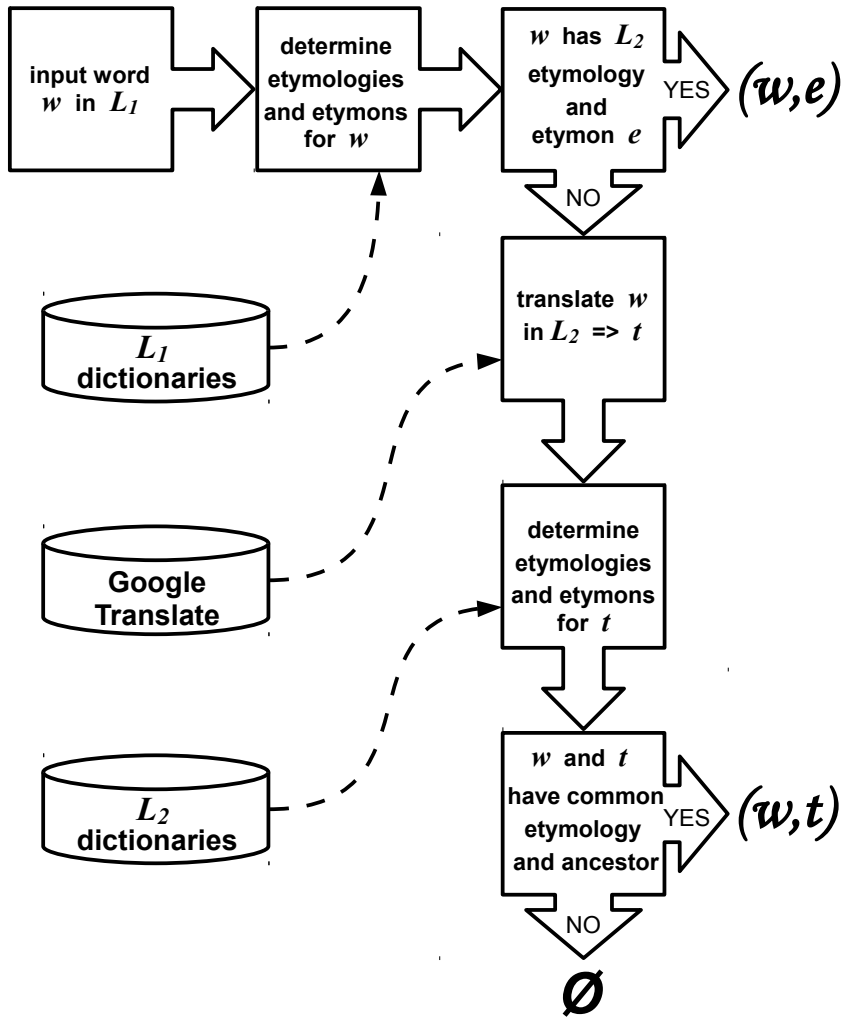
Entry

⟨b⟩ capitol ⟨/b⟩

⟨abbr class="abbrev" title="limba italiana"⟩
    it.
⟨/abbr⟩
⟨b⟩ capitolo ⟨/b⟩
⟨abbr class="abbrev" title="limba latina"⟩
    lat.
⟨/abbr⟩

⟨b⟩ capitulum ⟨/b⟩

Cognate identification



Orthographic metrics

- We use string similarity metrics to compute the orthographic similarity between related words.
- Many methods have been used so far, but we cannot say which is the most appropriate for a given task.
- We use three orthographic metrics and compare their results.

Orthographic metrics

The edit distance

$$\Delta(w_i, w_j) = \frac{LD(w_i, w_j)}{max(|w_i|, |w_j|)} \quad (3)$$

where $LD(w_i, w_j)$ is the number of operations required to transform w_i in w_j.

The longest common subsequence ratio

$$\Delta(w_i, w_j) = \frac{LCS(w_i, w_j)}{max(|w_i|, |w_j|)} \quad (4)$$

where $LCS(w_i, w_j)$ is the longest common subsequence of w_i and w_j.

The rank distance

Given two rankings $L_1 = (x_1, x_2, \ldots, x_n)$ and $L_2 = (y_1, y_2, \ldots, y_n)$, and $V(L_1)$, $V(L_2)$ their alphabets, the rank distance is defined as follows:

$$\Delta(L_1, L_2) = \sum_{x \in V(L_1) \cap V(L_2)} |ord(x|L_1) - ord(x|L_2)| + \sum_{x \in V(L_1) \setminus V(L_2)} ord(x|L_1) + \sum_{x \in V(L_2) \setminus V(L_1)} ord(x|L_2) \quad (5)$$

where $ord(x|L)$ is the rank of x in ranking L, in a Borda sense. To extend the distance to words, we index each character with a number equal to the number of its previous occurrences in the given word. For normalization, we divide the rank distance by the maximum possible value between w_i and w_j: $|w_i|(|w_i| + 1)/2 + |w_j|(|w_j| + 1)/2$.

Application: Romanian

- Romanian is a Romance language, surrounded by Slavic languages.
- Its communication with the Romance kernel was difficult.
- Its position in the Romance family is controversial, either isolated or more integrated within the group [McMahon and McMahon, 2003].

Datasets

- 17^{th} and 18^{th} century: Romanian chronicles. **(Chronicles)**
- 19^{th} century: the publishing works of the Romanian poet Mihai Eminescu. **(Eminescu)**
- 21^{st} century: the parliamentary debates held in the Romanian Parliament. **(Parliament)**

- The basic Romanian lexicon. **(RVR)**

Dataset	#words		#stop words		#lemmas
	token	type	token	type	type
Parliament	22,469,290	162,399	14,451,178	214	40,065
Eminescu	870,828	65,742	565,396	212	21,456
Chronicles	253,786	28,936	170,582	193	8,189
RVR	2,464	2,464	124	124	2,252

Etymology detection evaluation

- We compare the manually determined etymologies with the automatically obtained etymologies on samples of 500 words.
- We evaluate the languages for which we determine both etymologies and cognate pairs:
  - Romanian 95.8%
  - French 96.8%
  - Italian 97.8%
  - Spanish 96.6%
  - Portuguese 97.0%
  - Turkish 96.0%
  - English 97.2%

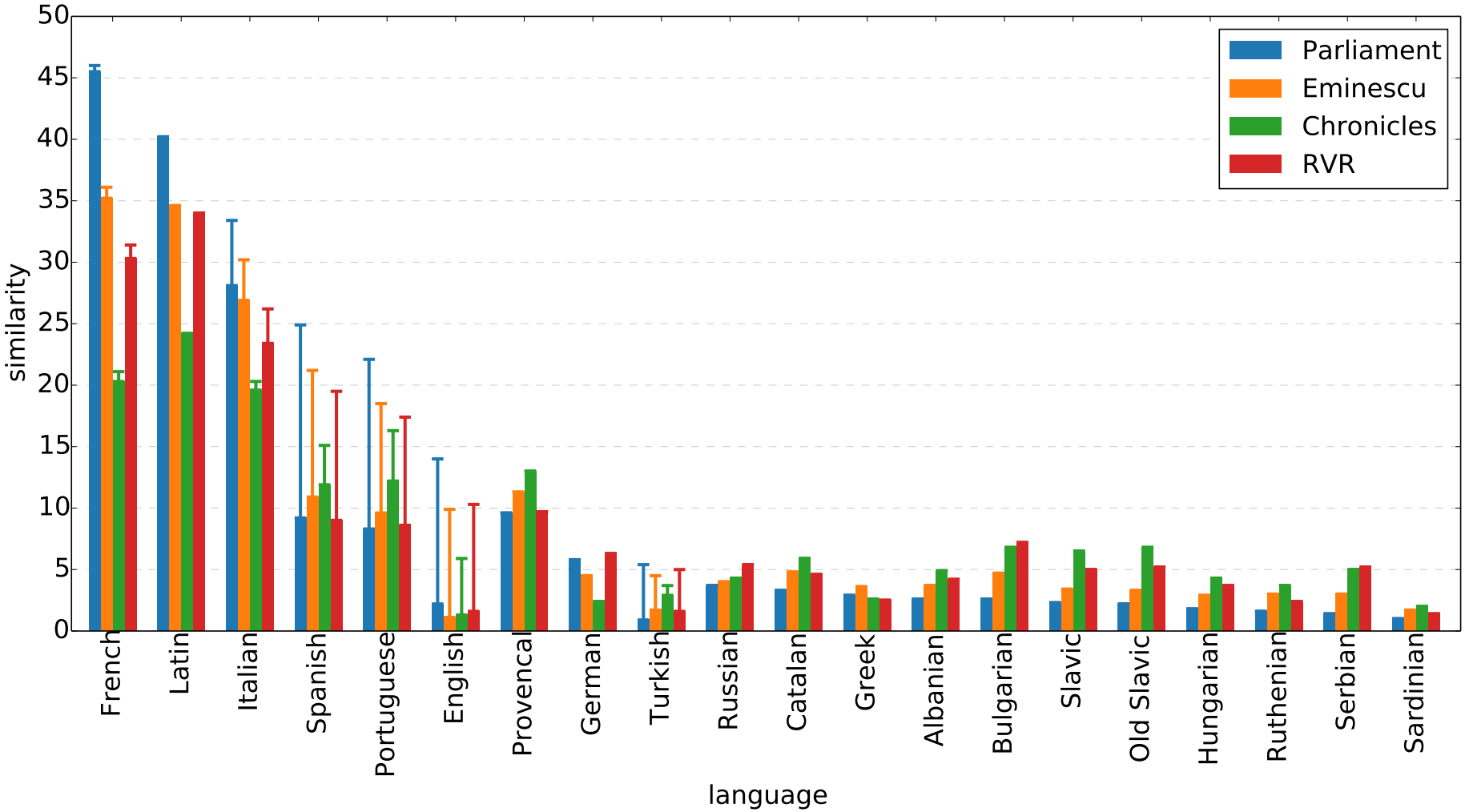
Diacritics

- Many words have undergone transformations by the augmentation of language-specific diacritics when entering a new language.
- From an orthographic perspective, the resemblance of words is higher between words without diacritics.

  amiciție (ro.), ***amitié*** (fr.) vs. ***amicitie*** (ro.), ***amitie*** (fr.)

- In Romanian, five diacritics are used today: *ă*, *â*, *î*, *ș*, *ț*.
- We create two versions of each dataset: with and without diacritics.

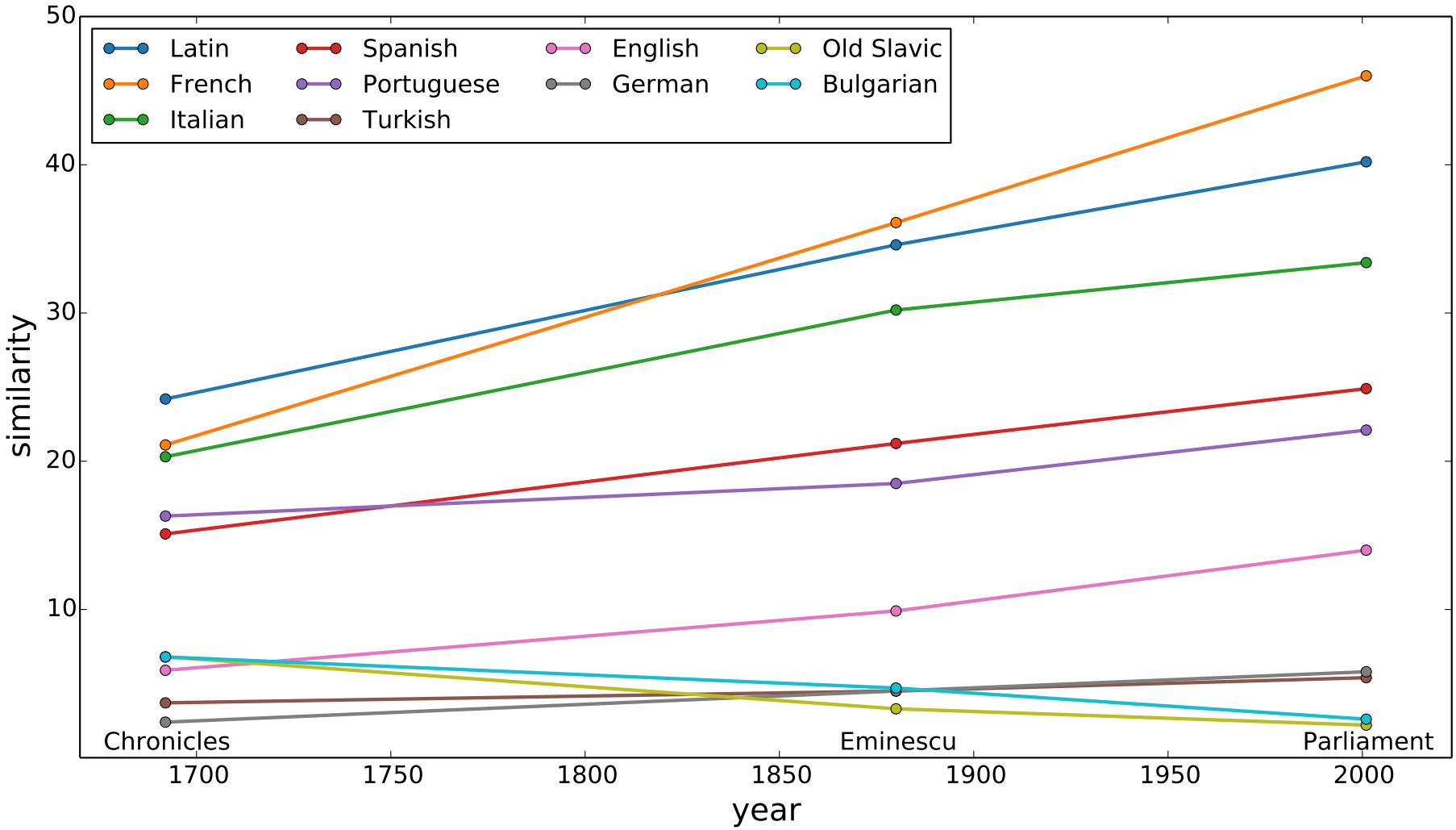
Results for the Romanian datasets



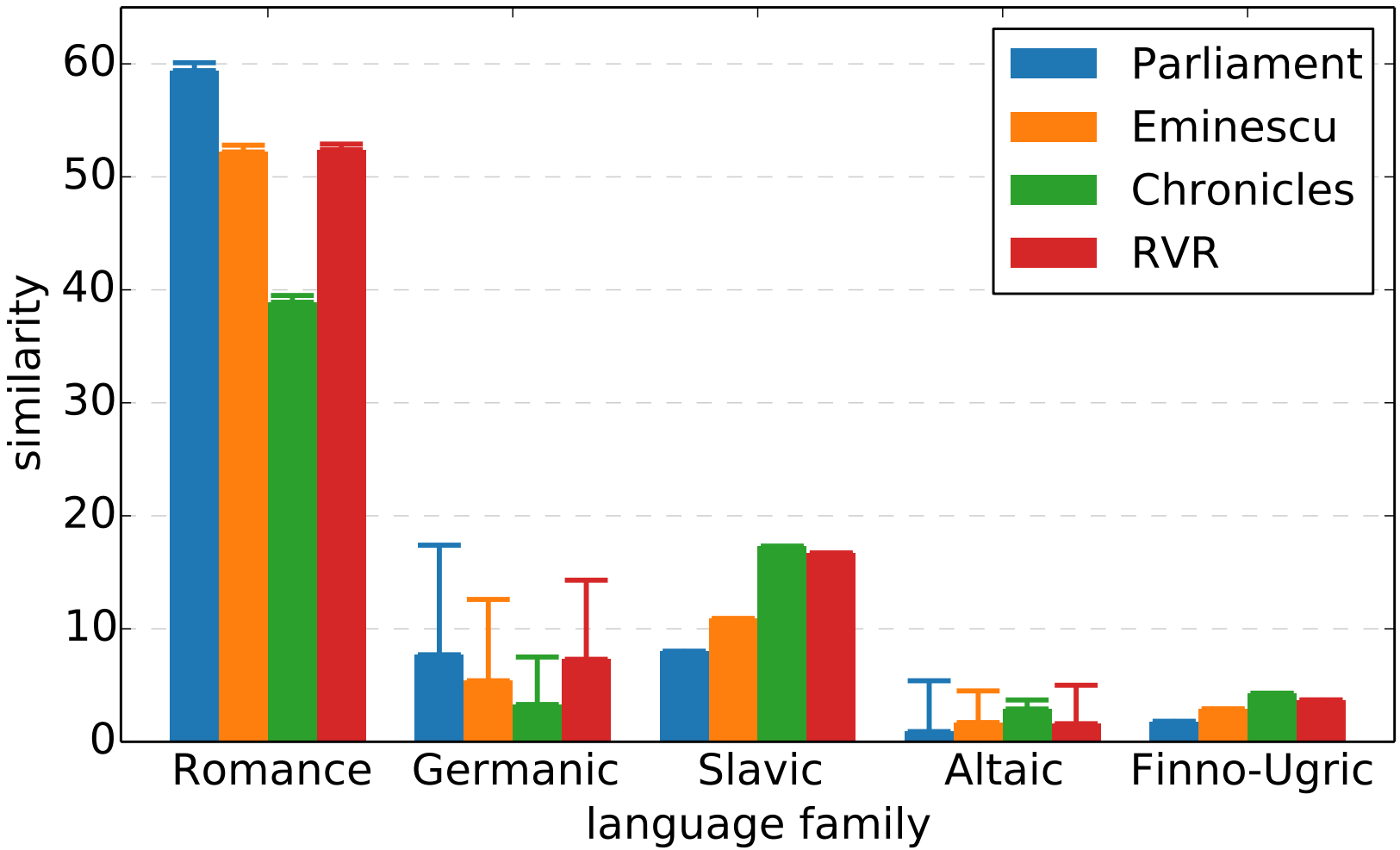
Ranking of similarity

Language	Parliament			Eminescu			Chronicles			RVR		
	%w	e	e+c	%w	e	e+c	%w	e	e+c	%w	e	e+c
French	70.6	45.5	46.0	57.2	35.2	36.1	36.7	20.3	21.1	50.6	30.3	31.4
Latin	63.7	40.2	—	59.9	34.6	—	44.9	24.2	—	56.5	34.0	—
Italian	48.5	28.1	33.4	44.7	26.9	30.2	31.7	19.6	20.3	41.4	23.4	26.2
Spanish	40.2	9.2	24.9	38.1	10.9	21.2	29.7	11.9	15.1	32.5	9.0	19.5
Portuguese	35.0	8.3	22.1	31.3	9.6	18.5	28.3	12.2	16.3	29.3	8.6	17.4
English	22.1	2.2	14.0	18.8	1.1	9.9	11.3	1.3	5.9	14.3	1.6	10.3
Provencal	17.7	9.6	—	20.7	11.3	—	21.8	13.0	—	16.8	9.7	—
German	9.2	5.8	—	6.9	4.5	—	4.9	2.4	—	10.2	6.3	—
Turkish	7.7	0.9	5.4	6.6	1.7	4.5	5.6	2.9	3.7	7.4	1.6	5.0
Russian	5.9	3.7	—	6.5	4.0	—	7.5	4.3	—	9.0	5.4	—

Romanian evolution



Language families



Surrounding languages

Language	Parliament			Eminescu			Chronicles			RVR		
	%w	d	nd	%w	d	nd	%w	d	nd	%w	d	nd
Turkish	7.7	5.4	5.6	6.6	4.5	4.7	5.6	3.7	3.9	7.4	5.0	5.3
Russian	5.9	3.7	4.0	6.5	4.0	4.4	7.5	4.3	4.9	9.0	5.4	6.2
Albanian	4.8	2.6	3.0	6.7	3.7	4.0	9.1	4.9	5.3	8.4	4.2	4.8
Bulgarian	4	2.6	3.0	7.4	4.7	5.5	10.6	6.8	7.8	11.8	7.2	8.4
Slavic	4.9	2.3	2.5	6.6	3.4	3.8	12.1	6.5	7.7	9.8	5.0	5.7
Old Slavic	3.8	2.2	2.7	6.1	3.3	4.3	11.9	6.8	8.7	9.5	5.2	6.0
Hungarian	2.9	1.8	2.0	5.1	2.9	3.3	7.5	4.3	4.7	7.4	3.7	4.6
Serbian	2.6	1.4	1.6	5.8	3.0	3.4	8.9	5.0	5.5	8.6	5.2	6.0
Polish	1.3	0.7	0.8	2.2	1.2	1.5	4.3	2.2	2.6	4.3	2.5	2.8
Serbo-Croatian	0.3	0.1	0.1	0.6	0.3	0.3	1.1	0.5	0.5	1.6	0.8	0.9
Ukrainian	0.0	0.0	0.0	0.1	0.0	0.0	0.6	0.3	0.3	0.4	0.3	0.3

Orthographic metrics

- Are the differences between the results obtained with each metric statistically significant?
- ANOVA hypothesis tests on samples of 5,000 words.
  - The mean computed values for the three metrics are not all equal.
- Pairwise t-tests with Bonferonni correction for the p-value.
  - The differences between the metrics are statistically significant, but they are small.
- There is a high correlation between the similarity rankings ($\rho > 0.98$ for each pair of metrics).

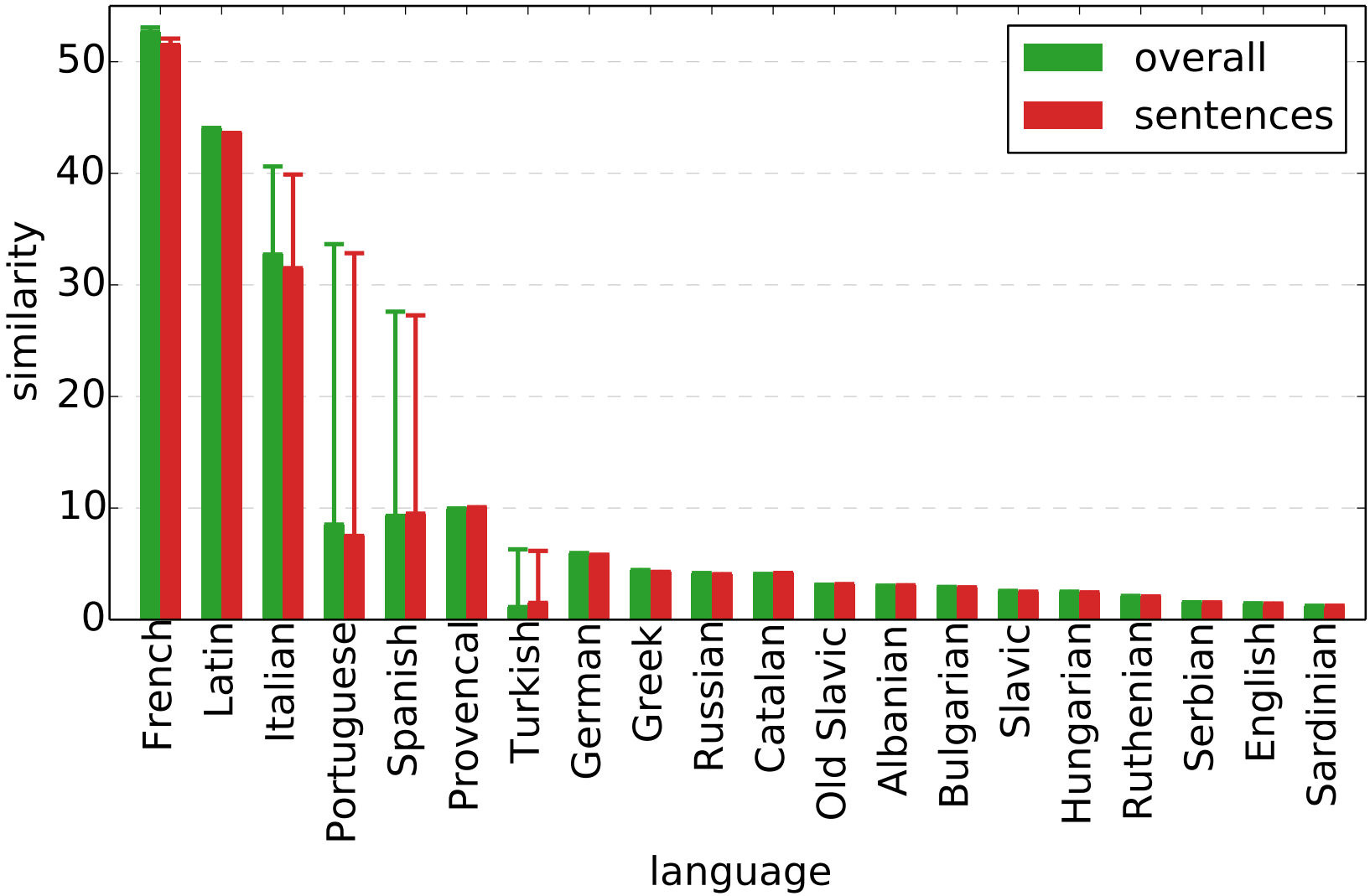
Further experiments

- We use Europarl [Koehn, 2005] - the Romanian subcorpus.
- We investigate two questions:
  - Are degrees of similarity between Romanian and other languages consistent across different corpora from the same period?
  - Are there differences between the overall degrees of similarity (the bag-of-words model) and those obtained at sentence level?

Further experiments

- We conduct four experiments:
  - **Exp. #1:** we use the bag-of-words model on Europarl.
  - **Exp. #2:** we aggregate sentence-level rankings of similarity.
  - **Exp. #3:** we remove outliers (regarding the sentence length).
  - **Exp. #4:** we remove outliers (regarding the degrees of similarity).

Results for Europarl



Results for Europarl

Language	Parl.	Exp. #1	Exp. #2	Exp. #3	Exp. #4
French	45.5	53.1	52.1	52.1	52.8
Latin	40.2	44.1	43.6	43.6	44.0
Italian	33.4	40.6	39.9	39.9	40.2
Portuguese	22.1	33.6	32.9	32.8	33.2
Spanish	24.9	27.6	27.3	27.3	26.8
English	14.0	16.0	15.7	15.7	15.1
Provencal	9.6	10.0	10.1	10.1	9.3
Turkish	5.4	6.3	6.2	6.1	5.7
German	5.8	5.9	5.8	5.8	5.3
Greek	2.9	4.4	4.3	4.3	3.8

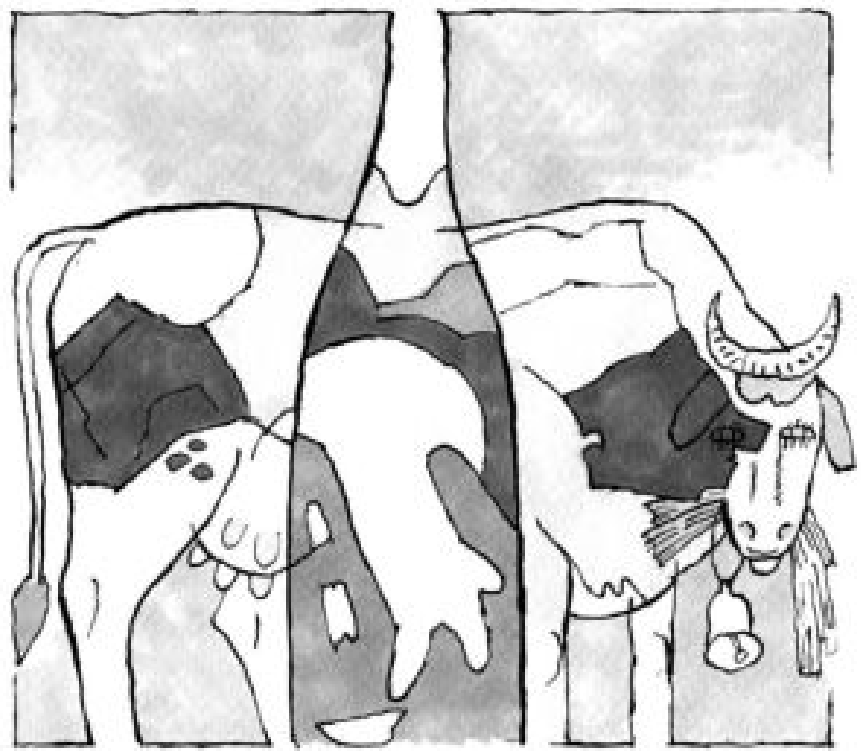
Language similarity

Cu un kil de carne de vacă nu mori de foame, cu un litru de vin nu mori de sete[1]. **(ro)**

Con un chilo di carne di vaca non morire di fame, con un litro di vino non morire di sete. **(it)**

Com um quilo de carne de vaca não morrer de fome, com um litro de vinho não morrer de sede. **(pt)**

Con un kilo de carne de vacuno no morirse de hambre, con un litro de vino no morir de sed. **(es)**



[1] With a kilo of beef one does not starve, with a liter of wine one does not die of thirst. **(en)**

Conclusions

- We proposed a computational method for determining cross-language orthographic similarity.
- We applied the method on Romanian corpora from different historical periods.
- We plan to extend our analysis to other languages as well, as we gain access to resources.
- We plan to combine the orthographic approach with syntactic and semantic evidence for a wider perspective on language similarity.

Thank you!