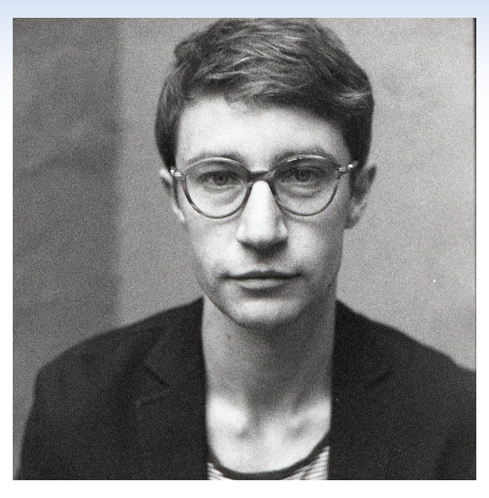

Ilia Markov
ilia.markov@uantwerpen.be

Anglicized Words and Misspelled Cognates in Native Language Identification

Ilia Markov[1], Vivi Nastase[2], Carlo Strapparava[3]

[1] University of Antwerp, Antwerp, Belgium
[2] University of Heidelberg, Heidelberg, Germany,
[3] Fondazione Bruno Kessler, Trento, Italy

Abstract

- **Native Language Identification (NLI)**: the task of identifying the native language (L1) of a person based on his/her writing in the second language (L2).
- **Language transfer effect:** L1 influences learners' second language writing.
- **Direct application:** second language teaching.

1. Phenomena

Three of the phenomena responsible for the incorrect expansion of L2's vocabulary using L1 material:

- **Cognates:** words that have the same ancestors or were derived from the same sources, e.g., SPA. *religión* and ENG. *religion*. **Misspelled cognates:** words that are misspellings from the point of view of L2, but have a very close form in L2 and L1.
- **L2-ed (anglicized) words:** words in L1 that were "adjusted" to seem and sound like legitimate L2 words: *lentaly* instead of *slowly* (SPA. *lentamente*).
- **Spelling errors** capture language-specific sound-to-spelling mappings.

2. Data and Experiment Setup

The subsets of the TOEFL11 and ICLEv2 datasets that cover languages that use the Latin script.

- **TOEFL4:** French, German, Italian, Spanish; 1,100 essays per language
- **ICLE4:** French (347 essays), German (437), Italian (392), Spanish (251)
- Tokenization
- Term frequency (*tf*) weighting scheme
- Liblinear Support Vector Machines (SVM)
- 10-fold cross-validation

3. Features

- **Part-of-speech (POS) features** capture the morpho-syntactic patterns in a text: Penn Treebank tagset (36 tags).
- **Function words (FWs)** clarify the relationships between the content-carrying elements of a sentence and introduce syntactic structures: 318 FWs from the scikit-learn package.
- **Misspelled cognates**
  1. For each misspelled English word w_m identify the intended word w_e using a spell- checking tool.
  2. For each L1:
     a) Look up the translation w_f of the intended word w_e in L1.
     b) Replace diacritics in w_f with the corresponding Latin equivalent (e.g., "é" → "e").
     c) Compute the Levenshtein distance D between w_e and w_f.
     d) If $D(w_e,w_f) < 3$ then w_f is assumed to be a cognate of w_e.
     e) If w_f is a cognate and $D(w_m,w_f) < D(w_e,w_f)$ then consider the L1 as a clue of the native language of the author.
- **L2-ed words**
  1. For each misspelled English word w_m identify its closest word in some L1:
  2. For w_f in each L1:
     a) Replace diacritics in w_f with the corresponding Latin equivalent.
     b) Compute the Levenshtein distance $D(w_m, w_f)$.
     c) Identify the L1 with the smallest $D(w_m, w_f)$ value, and if $D(w_m, w_f) < 5$ then take w_m to be an L2-ed version of w_f, and consider w_m as a clue for the native language of the author.
- **Spelling errors (SE):** misspelled words are represented through character n-grams (n = 1– 3) and added as a separate subset of the feature vector.

Table 1. Examples: POS & FW & cognates & L2-ed words features.

```
have a happy anciaṅity
```

anciaṅity (ENG. old age) → SPA. *ancianidad* → L2-ed
→ **have a JJ SPA-L2-ed**

a good *inocent* man

inocent (ENG. innocent) → SPA. *inocente* → cognate
→ **a JJ SPA-cognate NN**

N-grams (n = 1–3) from this representations are extracted.

4. Results

- **Results on the TOEFL4 and ICLE4 datasets**

Table 2. Accuracy (%) on the TOEFL4 and ICLE4 datasets.

		TOEFL4			ICLE4	
Features	**Acc. %**	**Diff**	**No.**	**Acc. %**	**Diff**	**No.**
Majority baseline	25.00			30.62		
Cognates	37.34	**12.34***	4	38.55	**7.93***	4
L2-ed	36.05	**11.05***	4	44.85	**14.23***	4
Cognates & L2-ed	39.84	**14.84***	8	46.18	**15.56***	8
Cognates & L2-ed & SE	54.55	**29.55***	7,347	56.33	**25.71***	6,391
POS & FW 1–3 grams	74.45		231,737	80.58		189,622
POS & FW 1–3 grams & cognates	75.50	**1.05***	236,716	80.72	**0.14**	192,572
POS & FW 1–3 grams & L2-ed	75.80	**1.35***	247,814	81.56	**0.98**	198,469
POS & FW 1–3 grams & cognates & L2-ed	76.20	**1.75***	253,175	81.77	**1.19**	201,623
POS & FW 1–3 grams & SE	78.23	**3.78***	238,929	82.75	**2.17***	195,869
POS & FW 1–3 grams & cognates & L2-ed & SE	78.80	**4.35***	260,367	82.61	**2.03***	207,870

- **Proficiency-level experiments**

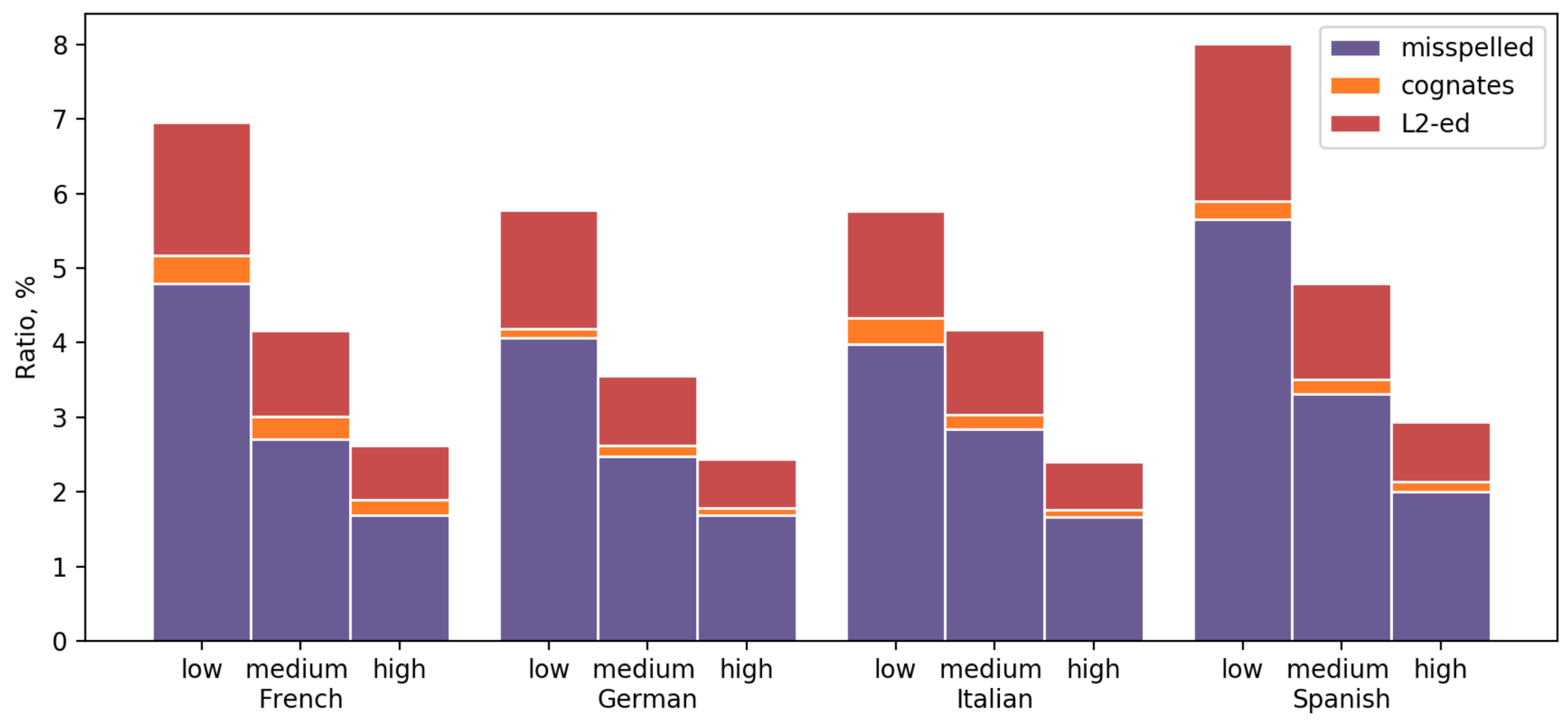

Figure 1. Ratio (%) of the misspelled words, cognates, and L2-ed words to the total number of words for each language within each proficiency level.

Table 3. Accuracy (%) for each proficiency level.

		Low			Medium			High	
Features	**Acc. %**	**Diff**	**No.**	**Acc. %**	**Diff**	**No.**	**Acc. %**	**Diff**	**No.**
Majority baseline	51.09			28.64			35.35		
Cognates	56.49	**5.40***	4	39.81	**11.17***	4	40.23	**4.88***	4
L2-ed	58.12	**7.03***	4	38.39	**9.75***	4	36.24	**0.89**	4
Cognates & L2-ed	59.24	**8.15***	8	42.57	**13.93***	8	40.18	**4.83***	8
Cognates & L2-ed & SE	60.79	**9.70***	3,241	55.26	**26.62***	6,031	45.95	**10.60***	5,366
POS & FW 1–3 grams	62.92		34,970	74.33		148,878	67.71		152,105
POS & FW 1–3 grams & cognates	62.38	**–0.54**	35,609	75.57	**1.24***	152,158	68.08	**0.37**	154,318
POS & FW 1–3 grams & L2-ed	65.16	**2.24**	37,214	76.17	**1.84***	159,508	68.03	**0.32**	160,025
POS & FW 1–3 grams & cognates & L2-ed	64.54	**1.62**	37,922	77.09	**2.76***	163,057	68.55	**0.84**	162,419
POS & FW 1–3 grams & SE	66.09	**3.17**	38,114	78.14	**3.81***	154,774	70.07	**2.36***	157,346
POS & FW 1–3 grams & cognates & L2-ed & SE	69.13	**6.21***	41,066	79.25	**4.92***	168,953	71.28	**3.57***	167,660

5. Conclusions

- ➢ All three phenomena provide useful information for identifying the L1 of the author.
- ➢ Higher results are achieved when features representing each of these are combined: they are complementary for the NLI task.
- ➢ The frequency of misspellings in general – and of L2-ed words – decreases with an increase in proficiency, but their contribution to the NLI task remains strong for all levels.

RESEARCH POSTER PRESENTATION DESIGN © 2015
www.PosterPresentations.com