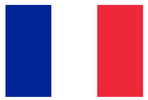
L1 Influence on Content Word Errors in Learner English Corpora: Insights from Distributed Representation of Words.

Kanishka Misra, Hemanth Devarapalli, Julia Taylor Rayz
Purdue University, West Lafayette, IN
{kmisra, hdevarap, jtaylor1}@purdue.edu

Introduction

- Errors made by non-native speakers of a language are often a result of a transfer of properties from the speaker's first language (L1) during Second Language (L2) acquisition.



Incorrect Usage		Correct Usage
scene (scène)	->	stage (scène)
possibility (possibilitat)	->	opportunity (oportunitat)

- This work builds on the findings in two recent studies (Kochmar and Shutova, 2016, 2017; K&S hereafter) that explore differences in lexico-semantic models of a person's L1 and L2 and test their hypotheses within the framework of two multilingual word embedding models.

Distributed Representations of Words

- Distributional Hypothesis: ***similar words tend to occur in similar contexts*** (Harris, 1954; Firth 1958)
- Words are represented as dense vectors that capture certain semantic information. We use **fasttext** (Bojanowski et al. 2016) and **polyglot** (Al-Rfou et al. 2013) vectors in our experiments.

Research Questions

1. Do distributed representation of words reflect L1 influence on learner English error words?
2. Do distributed representation of learner English error words exhibit similar relationships between genealogically similar languages?

Corpus

First Certificate in English (FCE) corpus (Yannakoudakis et al. 2011))containing 2488 error annotated essays written in english by learners representing 16 different L1s (Dutch left out due to very low count).

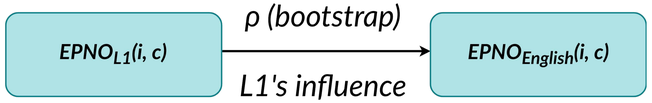
Table 1. Number of Errors made by people representing various L1s

L1	*Errors*	*L1*	*Errors*	*L1*	*Errors*
Spanish	796	Catalan	325	Turkish	272
French	794	Chinese (Simplified)	310	Japanese	192
Greek	353	Polish	295	Korean	185
Russian	340	German	285	Thai	122
Italian	335	Portuguese	284	Swedish	44

Methodology

- Translate the incorrect-correct word pair *(i, c)* into learner's L1 using Microsoft Azure API
- Compute **Error Pair Neighbor Overlap (EPNO)** - measures how close the incorrect and correct word are in vector space in terms of the words related to them (nearest neighbors), for each (i, c) pair in English and the learner's L1.

Experiment 1



Distributed Representation of words tend to capture the **influence of L1 on Learner English Error** words*.

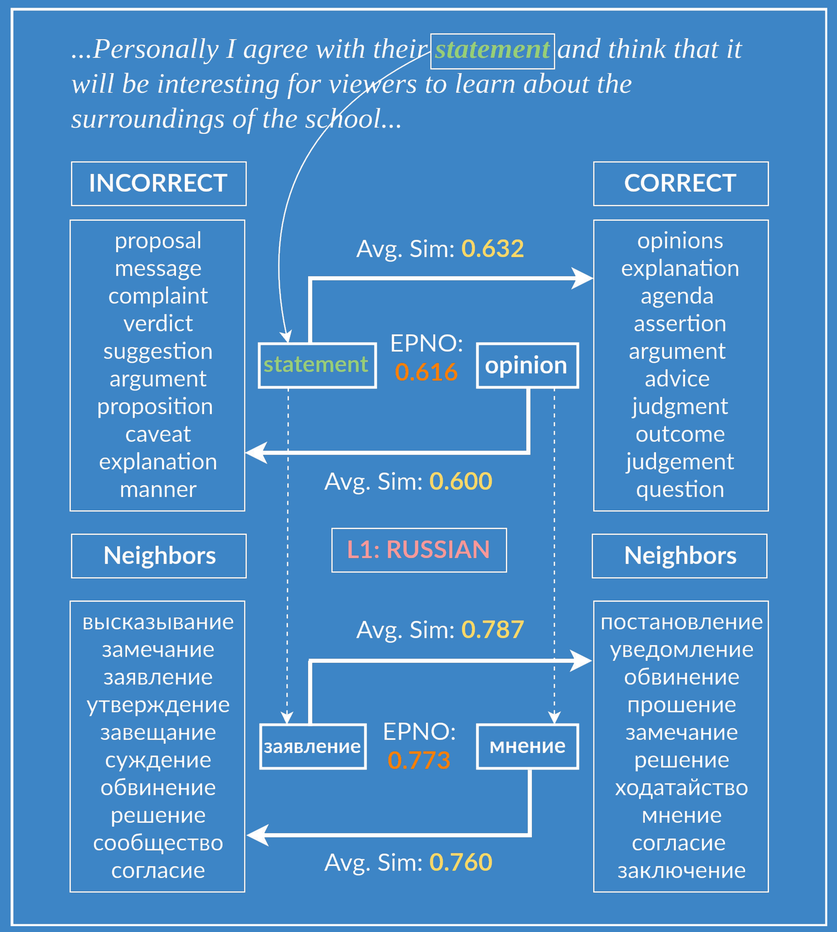


Figure 1. An Example of the EPNO value calculation with polyglot vectors

$$EPNO_L(i,c) = \frac{1}{2k}[\sum_{c' \in NN_k^L(c)} cos(i,c') + \sum_{i' \in NN_k^L(i)} cos(c,i')]$$

Where $NN_k^L(x)$ is the nearest neighbor function with k-nearest neighbors in language *L*, here: $k = 10$

* Within multi-lingual fasttext and Polyglot vector spaces, Languages on which experiments were performed on described in Table 1

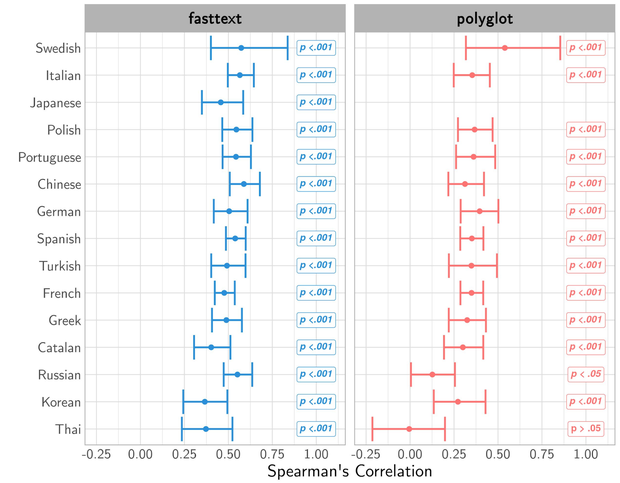


Figure 2. ρ estimates between EPNO between English and L1s with Bootstrapped CIs and p-values.

Experiment 2

Languages were grouped within genealogical groups and the differences between $EPNO_{English}$ and $EPNO_{L1}$ were compared for 10000 resamples within the group

Table 2. Resampled Differences between EPNOs between Language groups in the two Vector Spaces

Group	*Languages*	$\Delta_{fasttext}$	$\Delta_{polyglot}$
Germanic	German Swedish	0.135	**0.184**
Romance	Spanish Catalan Italian French Portuguese	0.129	0.188
Slavic	Russian Polish	0.127	0.226
Asian	Chinese Japanese* Korean Thai	**0.123**	0.217
Other	Turkish Greek	0.128	0.195

Results and Discussion

- Significant **positive correlation** between $EPNO_{English}$ and $EPNO_{L1}$, for both **fasttext** and **polyglot** vectors (for all languages except Thai within **polyglot**).
- Contrasting results between fasttext and polyglot:
  - $\Delta_{fasttext}$ - agreement with the results of K&S - Asian EPNO values are more similar to English.
  - $\Delta_{polyglot}$ - aligning with the initial assumptions of our work as well as K&S Germanic EPNOs are more similar to English.
- The Contrast between fasttext and polyglot results can be attributed to:
  - **Dimension size:** 300 in **fasttext** vs 64 in **polyglot**
  - **Vocab Size:** Order of million in **fasttext** vs 10k - 100k in **polyglot**
  - **Objective:** **fasttext** -> subword + word, **polyglot** -> word only.

References

Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.
Firth, J. R. (1958). Papers in Linguistics, 1934-1951. Oxford University Press.
Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011, June). A new dataset and method for automatically grading ESOL texts. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 180-189). Association for Computational Linguistics.
Kochmar, E., & Shutova, E. (2016, August). Cross-lingual lexico-semantic transfer in language learning. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 974-983).
Kochmar, E., & Shutova, E. (2017, September). Modelling semantic acquisition in second language learning. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 293-302).

PURDUE UNIVERSITY