L1 - L2 Semantic Overlaps of Learner Errors: Proof of Concept

Kanishka Misra, Hemant Devarapalli

1/17/2019

Abstract

Dataset

We use the Cambridge Learner Corpus - First Certification in English (CLC - FCE) corpus by Yannakoudakis et al (2011). The CLC-FCE corpus has 1244 short essays responses written by learners of English with 16 different first languages (L1). The corpus is annotated for errors and each error is represented based on the guidelines as stated in Nicholls (2003). There are a total of 55895 errors in the corpus.

For the purposes of measuring semantic errors, we take into account the errors involving content word corrections (n = 6837). Out of the content word based corrections, we only consider errors involving single word error corrections, reducing the total number to 5521.

Experiments

The goal of this paper is to compare two hypotheses:

1. Do Vector Semantic models reflect L1 influence in learner english (L2) errors.
2. Whether L1's influence on Learner Errors for typologically related languages is expressed by vector semantic models.

Translations

For each of the *(incorrect, correct)* pairs, we translate it into the person's L1 language using the Microsoft Azure Text translation API [1]. We use this API because unlike Google's Cloud API for translations, it provides us with alternate pronunciations which prove beneficial while querying for words that can be expressed with different parts of speech. We then remove all multi-word translations of the errors pairs, ending up with 4937 error pairs. Table 1 shows the number of error-cases for each language

language	n
Spanish	796
French	794
Greek	353
Russian	340
Italian	335
Catalan	325
Chinese (Simplified)	310
Polish	295
German	285
Portuguese	284

[1] https://docs.microsoft.com/en-us/azure/cognitive-services/Translator/reference/v3-0-reference

language	n
Turkish	272
Japanese	192
Korean	185
Thai	122
Swedish	44
Dutch	5

Hypothesis 1 (H_1): Do Vector Semantic models reflect L1 influence in learner english (L2) errors

We test H_1 by introducing a metric known as *Semantic Error Overlap* (SEO). Formally, we define SEO for language L_j as:

$$SEO_{L_j}(i,c) = \frac{1}{2k}[\sum_{c' \in N_k^j(c)} cos(i,c') + \sum_{i' \in N_k^j(i)} cos(c,i')]$$

Where i', c' are nearest neighbors in the set $N_k(.)$ of the k nearest neighbors of the words i (incorrect) and c (correct) respectively, in a given vector space. Since we have the (i, c) pairs in both languages, we compute the SEO for L1 and L2(english) for each case.

Intuitively, the SEO of a language will give us a symmetric measure of how close i and c are in their semantic space (for that language) based on their nearest neighbors, or the words they are most associated to. This is not the same as semantic similarity since vector space based on co-occurence relative to a context measure relatedness or association rather than similarity in meanings (cite).

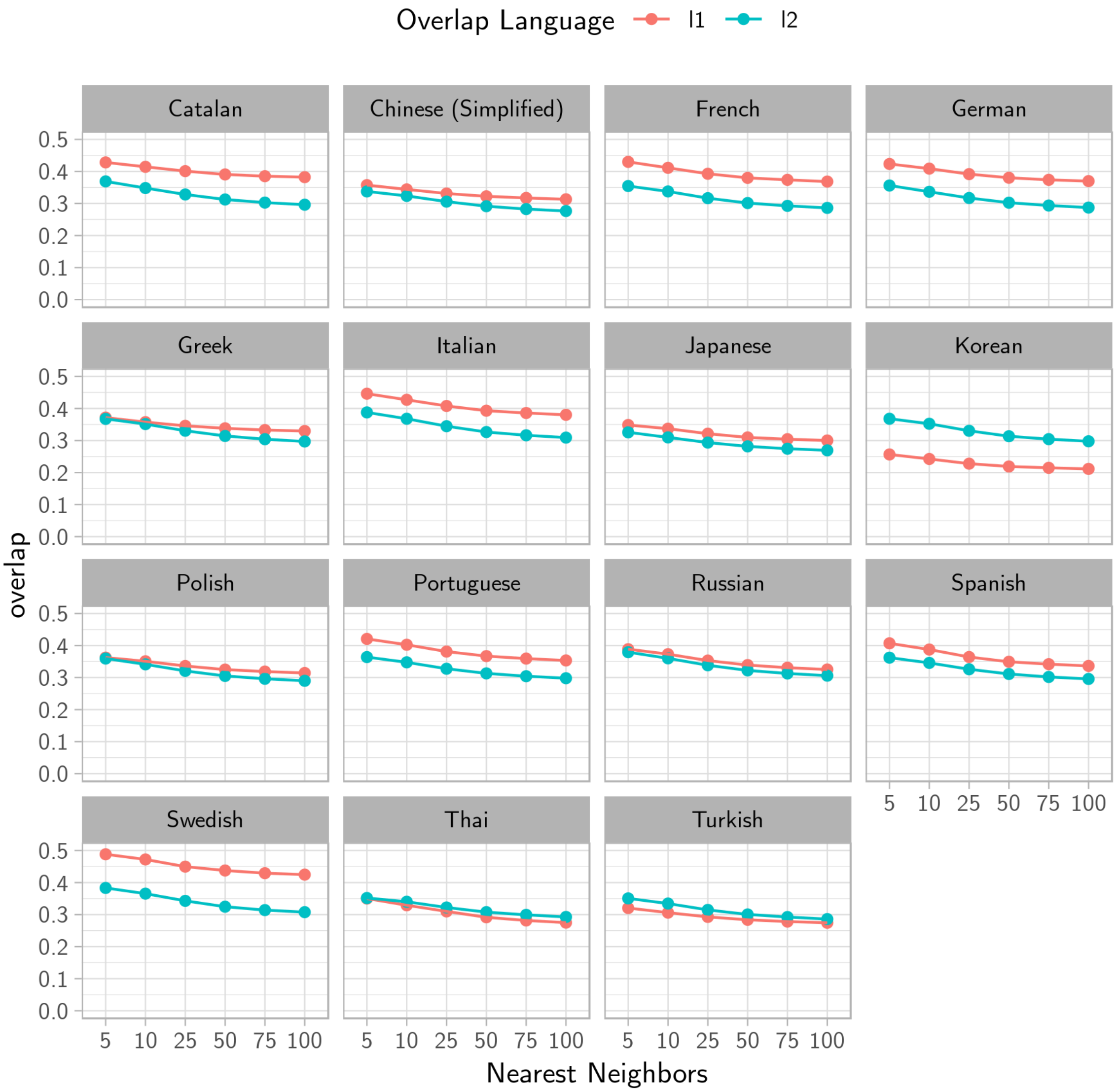
Figure 1 shows the average overlap for the L2 (English) and the respective L1 vector space for all the languages in the corpus. We test for k = 5, 10, 25, 50, 75, and 100.

We proceed with choosing k as 10 for the time being.

With k = 10, we test H_1 by computing the Spearman's ρ between the SEOs of L1 and L2 (i, c) pairs. This will capture the strength of the relationship between how related the two words are in the person's L1 vis-à-vis in L2, English. Table 1 reports the Spearman's ρ for all L1s in the corpus along with the *p-values* as well as the empirical bootstrap confidence interval estimates for 1000 resamples of the overlap values. Figure 2 shows the confidence intervals plotted for the 15 L1s.

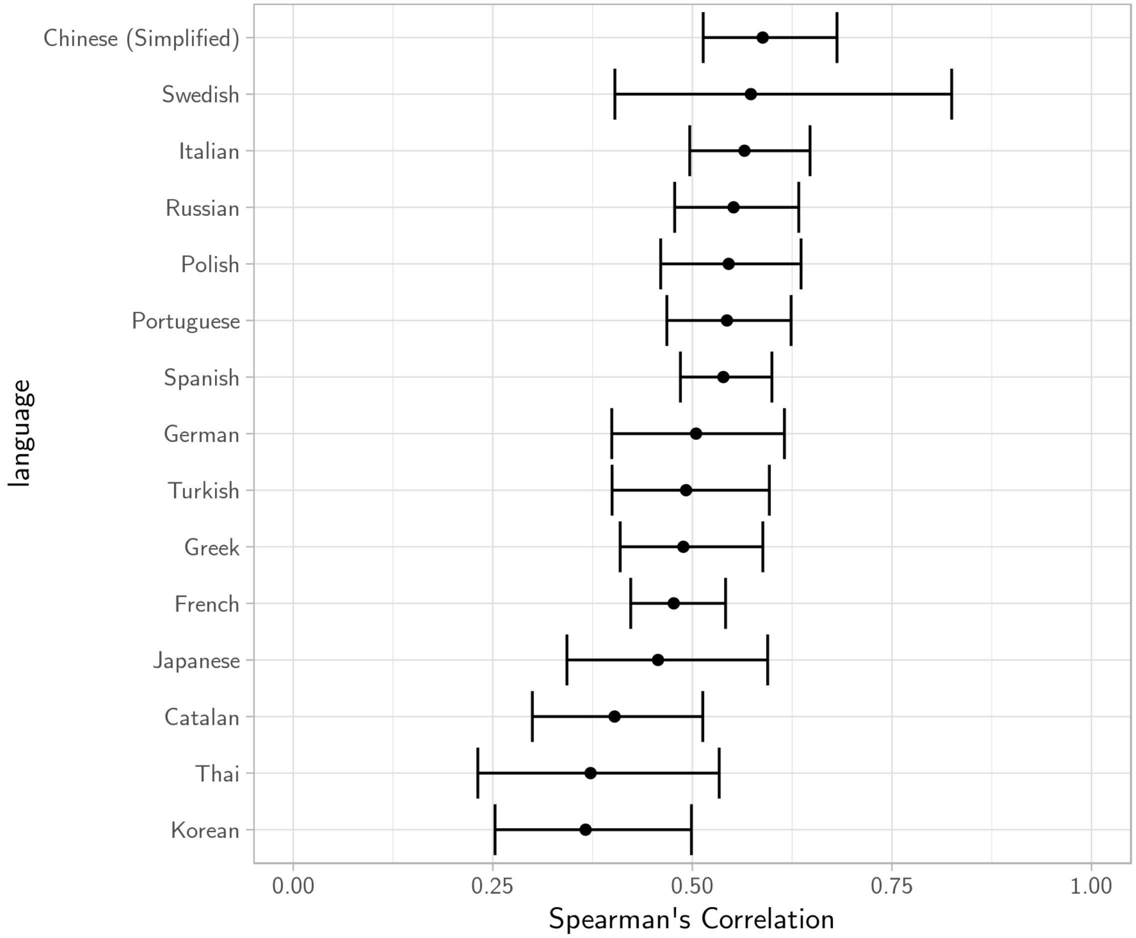
language	n	estimate	p.value	low	high
Chinese (Simplified)	310	0.5881536	0.0000000	0.5135842	0.6812029
Swedish	44	0.5733014	0.0000475	0.4029706	0.8248728
Italian	335	0.5653293	0.0000000	0.4967513	0.6474355
Russian	340	0.5516789	0.0000000	0.4778571	0.6332793
Polish	295	0.5455154	0.0000000	0.4603523	0.6361722
Portuguese	284	0.5433643	0.0000000	0.4680745	0.6237312
Spanish	796	0.5388600	0.0000000	0.4850572	0.5996513
German	285	0.5046863	0.0000000	0.3990605	0.6153570
Turkish	272	0.4921320	0.0000000	0.3994467	0.5964506
Greek	353	0.4887581	0.0000000	0.4095749	0.5883224
French	794	0.4767160	0.0000000	0.4228150	0.5416141
Japanese	192	0.4569835	0.0000000	0.3429047	0.5943805
Catalan	325	0.4026183	0.0000000	0.2995908	0.5130628
Thai	122	0.3725116	0.0000239	0.2312775	0.5336393
Korean	185	0.3662620	0.0000003	0.2528163	0.4988582

Figure 1: Overlaps for L1 (Red) and L2 (Green) with different Nearest Neighbors, for each language



We see that the Spearman's ρ between the various L1s, is in the moderate range (0.4 - 0.59), and is significant, as inferred from the bootstrapped confidence intervals as well as the p-values. Some L1s like *Thai* and *Korean* lie in the weak range, indicating that the L1 semantic influence in lexical choice isn't as pronounced in the errors produced in english.

Figure 2: Spearman's Correlation bootstrap estimates



Hypothesis 2 (H_2): Whether L1's influence on Learner Errors for typologically related languages is expressed by vector semantic models

In order to test H_2, we consider the different language families the L1s in the FCE data get segregated into - The three separate branches (among many others) of the Indo-European family of languages, *Germanic* L1s (German, Swedish), *Romance* L1s (Spanish, Catalan, French, Italian) and the *Slavic* family (Russian, Polish), the *Asian* L1s (Chinese, Korean, Japanese, Thai). We combine Turkish and Greek into the *Others* category. We then measure the Jensen-Shannon divergence between the distributions of the overlaps between each of these groups and their respective english (i, c) pairs.

(Describe Jensen Shannon Divergence)

Table 2 describes the JSD between each of the language groups in the CLC-FCE data.

Family	divergence
Germanic	0.0312044
Romance	0.0329918
Slavic	0.0364255
Other	0.0437544
Asian	0.0454337

Germanic - least (meaning closer to English).