

Measuring the Influence of L1 on Learner English Errors in Content Words within Word Embedding Models

Kanishka Misra, Hemanth Devarapalli, Julia Taylor Rayz

Applied Knowledge Representation and Natural Language Understanding Lab
Purdue University



Motivation

Errors made in Natural Language = Lexical Choice of the author.

Motivation

Errors made in Natural Language = Lexical Choice of the author.

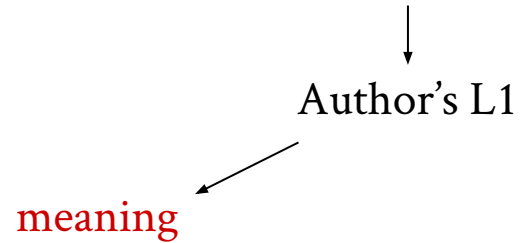


Author's L1

Groot, 1992; Koda, 1993; Groot & Keijzer, 2000; Hopman, Thompson, Austerweil, & Lupyan, 2018

Motivation

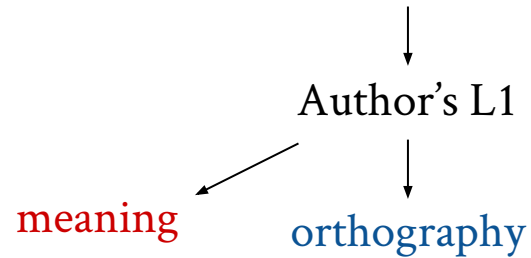
Errors made in Natural Language = Lexical Choice of the author.



Groot, 1992; Koda, 1993; Groot & Keijzer, 2000; Hopman, Thompson, Austerweil, & Lupyan, 2018

Motivation

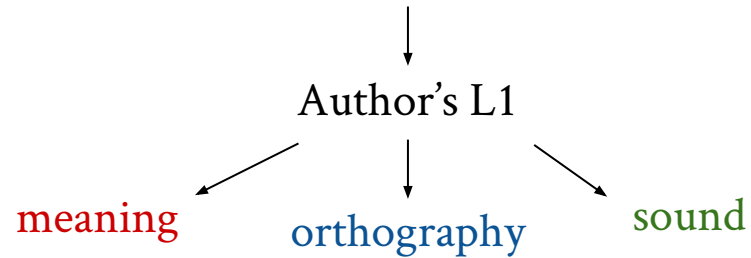
Errors made in Natural Language = Lexical Choice of the author.



Groot, 1992; Koda, 1993; Groot & Keijzer, 2000; Hopman, Thompson, Austerweil, & Lupyan, 2018

Motivation

Errors made in Natural Language = Lexical Choice of the author.

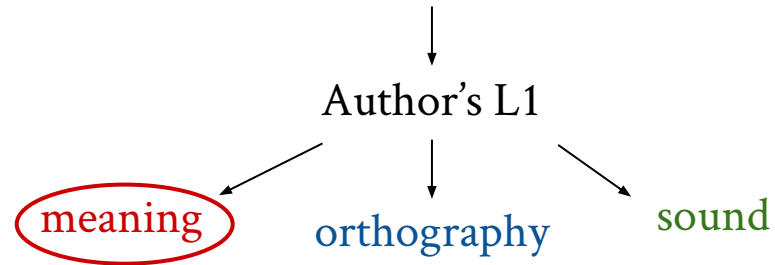


Cognate Effects

Groot, 1992; Koda, 1993; Groot & Keijzer, 2000; Hopman, Thompson, Austerweil, & Lupyan, 2018

Motivation

Errors made in Natural Language = Lexical Choice of the author.







Cognate Effects

Groot, 1992; Koda, 1993; Groot & Keijzer, 2000; Hopman, Thompson, Austerweil, & Lupyan, 2018





Motivation

Errors made in Natural Language = Lexical Choice of the author.

Incorrect usage		
		scene (scène)
		possibility (possibilitat)

Motivation

Errors made in Natural Language = Lexical Choice of the author.

Incorrect usage			Correct replacement
		scene (scène)	→ stage (scène)
		possibility (possibilitat)	→ opportunity (opportunitat)

Goals and Contributions

1. Build on research investigating errors in lexical choice of English learners.
2. Investigate how distributional semantic vector spaces can help extract the influence of a learner's native language (L1) on errors made in English.
3. Investigate whether distributional semantic vector-space based measure of L1 influence can show patterns within genealogically related languages.

Background - Influence of L1 in Lexical Choice

Influence of L1 studied as

1. Translation Ambiguity.

- Semantic overlap correlated with translation choice.
- Ambiguity causes confusion in lexical choice - errors.
- Used as predictor in estimating learning accuracy.

Prior et al., 2007; Degani & Tokowicz, 2010; Boada et al., 2013; Bracken et al., 2017; *inter alia*.

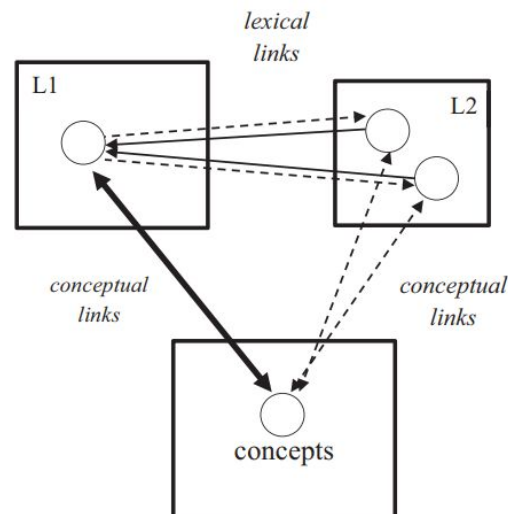
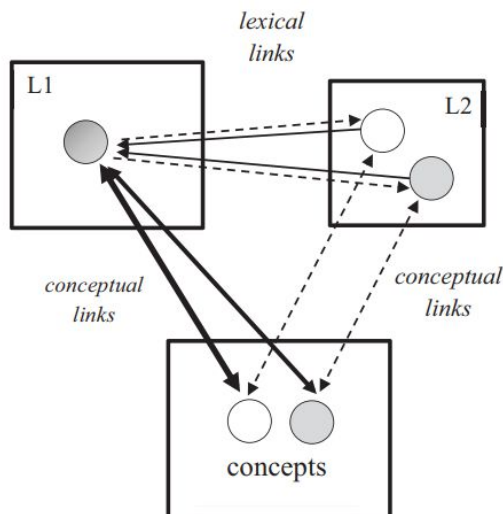


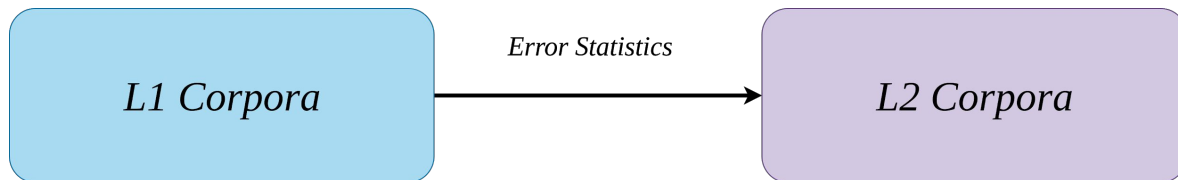
Figure Source: Bracken et al., 2017 pg. 3

Background - Influence of L1 in Lexical Choice

Influence of L1 studied as

2. Error Detection and Correction

- L1 error probabilities improved error correction of L2 preposition usage.
- Parallel corpora led to improvements in detecting and correcting mis-collocations.



Chang 2008; Rozovskaya & Roth, 2010, 2011;
Dahlmeier & Ng, 2011; Kochmar & Shutova, 2016,
2017; *inter alia*.

Background - Influence of L1 in Lexical Choice

Influence of L1 studied as

3. Large scale L2 (English) Learning analysis

- *Why are some words harder to learn for speakers of certain languages than others?*
- Cognate level features to estimate word learning accuracy on large data (Duolingo)
- Languages covered: **Spanish, Italian, Portuguese.**
- Leveraged distributional semantic vectors to estimate ambiguity between correct word and word as used by the learner (translation distance) that was found to correlate negatively with Learning accuracy.

Hopman et al. 2018

Kochmar & Shutova (2016, 2017)

Analysis of L1 effects in L2 semantic knowledge of content word combinations (Adjective-Noun, Verb-Direct Object, Subject-Verb) → *Leverage semantic features induced from L1 data to improve error detection in learner English.*

Our paper is related to three out of five Hypotheses covered in K&S:

1. L1 lexico-semantic models influence lexical choice in L2
2. L1 lexico-semantic models are portable to other typologically similar languages
3. Typological similarity between L1 and L2 facilitates semantic acquisition of knowledge in L2.

Kochmar & Shutova (2016, 2017)

Main Findings:

1. Semantic models of lexical choice from L1 helped in improving error detection.
2. The improvement was also observed when the L1 belonged to the same family (i.e., Germanic in this case).
3. **Lexical distributions of content word combinations were found to be closer to native English for typologically distant L1s rather than closer L1s.**

Kochmar & Shutova (2016, 2017)

Lexical distributions of content word combinations were found to be closer to English for typologically distant L1s rather than closer L1s.

- Learners from typologically distant languages prefer to use prefabricated phrases (eg. Asian L1s) since they like to “*play-it-safe*”, as noted in previous works.
- Those from typologically similar L1s tend to feel more confident and adventurous -> experiment with novel word combinations.

Hulstijn and Marchena (1989); Gilquin and Granger 2011

Background - Word Embeddings

Operationalize the **Distributional Hypothesis**:

Background - Word Embeddings

Operationalize the **Distributional Hypothesis**:

“The complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously.” - Firth (1935)

“Words that occur in similar contexts have similar meaning” ~ Harris (1954)

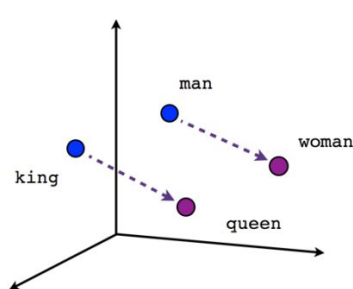
“You shall know a word by the company it keeps” - Firth (1957)

Background - Word Embeddings

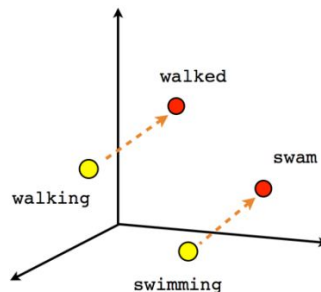
d -dimensional dense vectors (\mathbb{R}^d), commonly learned using models that leverage the context words surrounding the focus word.

1. **PMI-SVD**: Operate on Pointwise Mutual Information between words.
2. **word2vec (Mikolov et al. 2013)**: shallow neural network that is trained to predict the context words from a given input word.
3. **GloVe (Pennington et al. 2014)**: shallow neural network that operates on global co-occurrence statistics between words.

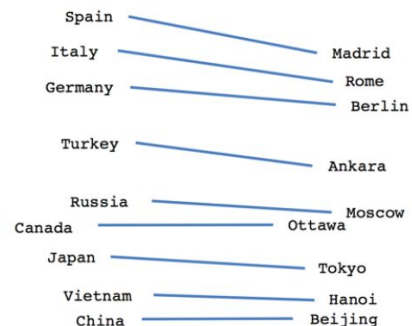
Background - Word Embeddings



Male-Female



Verb tense



Country-Capital

Source: <https://www.tensorflow.org/tutorials/representation/word2vec>

Background - Word Embeddings

Nearest Neighbors in word2vec

apple	france	January
apples	French	February
pear	Belgium	October
fruit	Paris	December
berry	Germany	November
pears	Italy	August
strawberry	Spain	September
peach	Nantes	March
potato	Marseille	April
grape	Montpellier	June
blueberry	Les_Bleus	July

Linear Analogies in word2vec (a:b::c:d)

Table 1: Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Mikolov et al. 2013

Background - Word Embeddings

fasttext: word2vec applied on subwords (3-6 character *n-grams*) → Easy to construct vectors for unknown words.

this = $\langle th + thi + his + is \rangle + \langle thi + this + his \rangle + \langle this + this \rangle$

polyglot: trained to predict higher score for original context window of a word vs. a corrupted sample (replace middle word with a random word).

imagination is greater than detail vs *imagination is wikipedia than detail*

Al-Rfou et al. 2013; Bojanowski et al. 2016

Background - Word Embeddings

fasttext: word2vec applied on subwords (3-6 character *n-grams*) → Easy to construct vectors for unknown words.

this = $\langle th + thi + his + is \rangle + \langle thi + this + his \rangle + \langle this + this \rangle$

polyglot: trained to predict higher score for original context window of a word vs. a corrupted sample (replace middle word with a random word).

*imagination is **greater** than detail* vs *imagination is **wikipedia** than detail*

Advantage: Both vector spaces available for multiple languages.

Al-Rfou et al. 2013; Bojanowski et al. 2016

Experiments

Corpus

Cambridge First Certification in English (FCE; Yannakoudakis et al. 2011)

- 2488 short-essay based responses written by English Learners.
- B2 proficiency under the Common European Framework of Reference for Languages (CEFR).
- **Error Annotated - with correct replacements for incorrect language.**
- Annotation following the scheme of Nicholls (2003).
- Learners represent 16 different L1 backgrounds.
- Only include errors involving a content word (Nouns, Adjectives, Verbs, Adverbs).
- Total Instances: **5521**

Preprocessing

- Translation of incorrect - correct pairs (i, c) into learner's L1 using Microsoft Azure API.
- Discarded multi-word translations and errors made by Dutch L1 learners (only 5 instances).
- Total Instances: 4932

Preprocessing

Table 1. Number of Errors made by learners representing various L1s in the corpus

<i>L1</i>	<i>Errors</i>	<i>L1</i>	<i>Errors</i>	<i>L1</i>	<i>Errors</i>
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

Influence of L1

Error Pair Neighborhood Overlap (EPNO): Quantifies the semantic relatedness between (i, c) word pairs based on their nearest neighbors for a given language vector space. Here, $k = 10$.

$$EPNO_L(i, c) = \frac{1}{2k} \left[\sum_{c' \in NN_k^L(c)} \overset{\text{Avg sim between } \mathbf{i} \text{ and neighbors of } \mathbf{c}}{\cos(i, c')} + \sum_{i' \in NN_k^L(i)} \overset{\text{Avg sim between } \mathbf{c} \text{ and neighbors of } \mathbf{i}}{\cos(c, i')} \right]$$

\swarrow k -nearest neighbor function \searrow

*...Personally I agree with their **statement** and think that it will be interesting for viewers to learn about the surroundings of the school...*

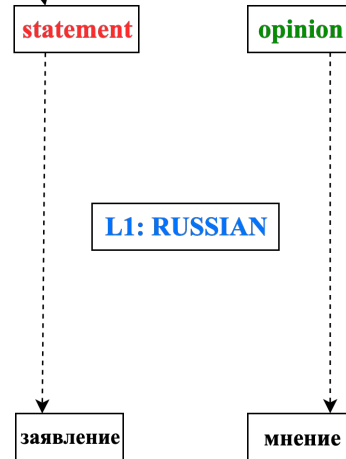
*...Personally I agree with their **statement** and think that it will be interesting for viewers to learn about the surroundings of the school...*

statement

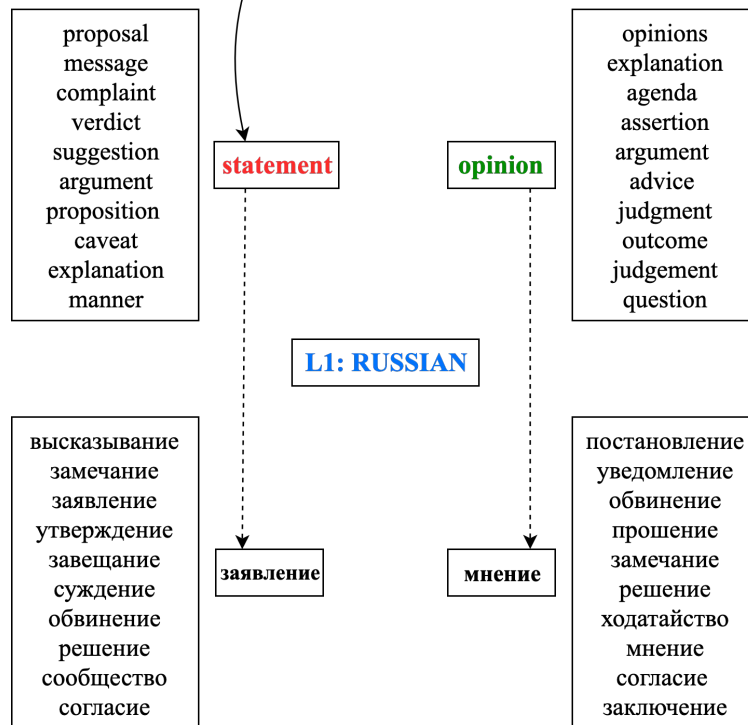
opinion

L1: RUSSIAN

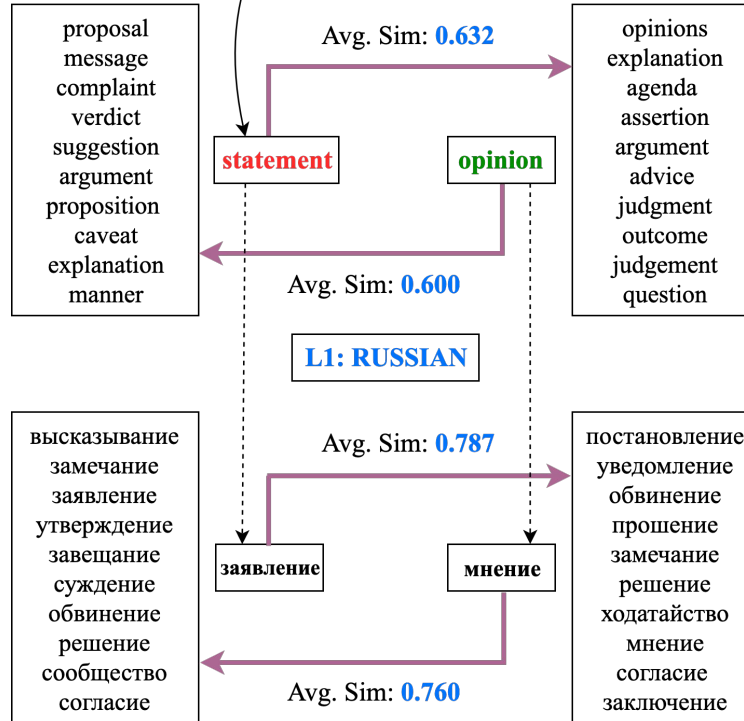
*...Personally I agree with their **statement** and think that it will be interesting for viewers to learn about the surroundings of the school...*



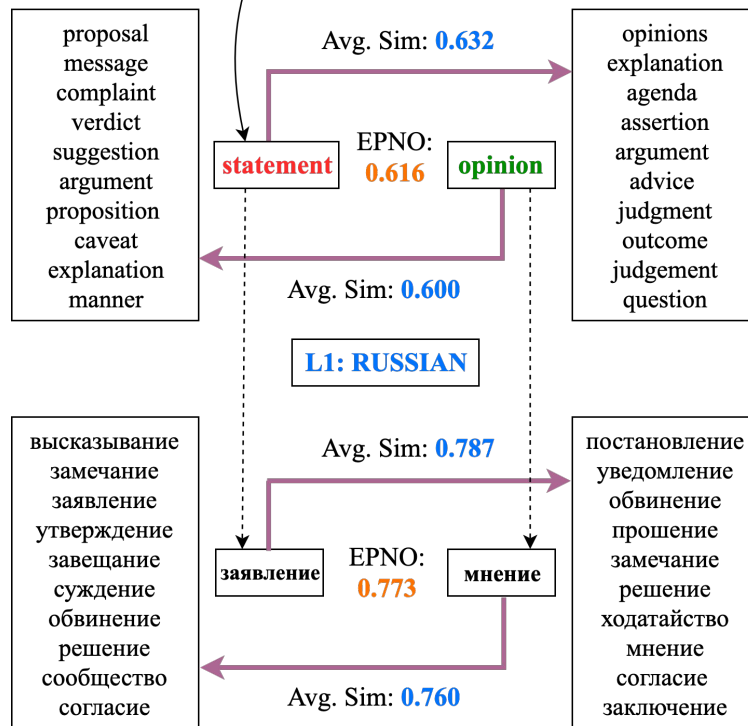
...Personally I agree with their **statement** and think that it will be interesting for viewers to learn about the surroundings of the school...



*...Personally I agree with their **statement** and think that it will be interesting for viewers to learn about the surroundings of the school...*



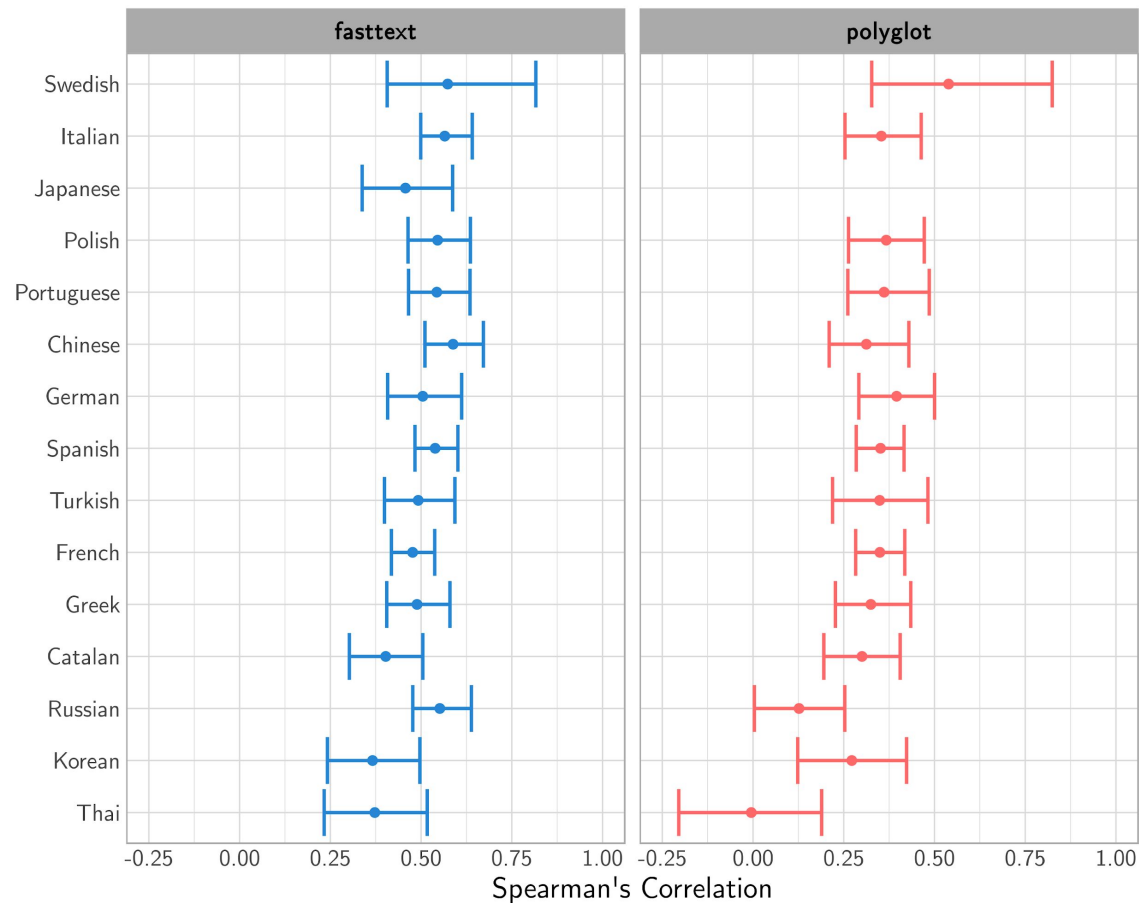
...Personally I agree with their **statement** and think that it will be interesting for viewers to learn about the surroundings of the school...



Experiment 1: Measuring L1 Influence

Whether distributional representation of words reflect L1 influence on learner English Error Words.

- Spearman's Rank Correlation Statistic (ρ) between $EPNO_{English}$ and $EPNO_{L1}$ for all L1s.
- Positive value \rightarrow Association between L1 and English content word errors based on semantic relatedness.
- Significance is tested using a non-parametric bootstrap for 1000 resamples in each language.



<i>L1</i>	$\rho_{fasttext}$	$\rho_{polyglot}$
Swedish	0.573 (<.001)	0.516 (<.001)
Italian	0.565 (<.001)	0.355 (<.001)
Japanese	0.457 (<.001)	NA
Polish	0.546 (<.001)	0.356 (<.001)
Portuguese	0.543 (<.001)	0.369 (<.001)
Chinese (Simplified)	0.588 (<.001)	0.322 (<.001)
German	0.505 (<.001)	0.384 (<.001)
Spanish	0.539 (<.001)	0.351 (<.001)
Turkish	0.492 (<.001)	0.369 (<.001)
French	0.477 (<.001)	0.373 (<.001)
Greek	0.489 (<.001)	0.351 (<.001)
Catalan	0.403 (<.001)	0.312 (<.001)
Russian	0.552 (<.001)	0.129 (<.025)
Korean	0.366 (<.001)	0.281 (<.001)
Thai	0.373 (<.001)	0.006 (.953)

Experiment 1: Results

- Significant, Positive ρ values between all L1s and English.
- **Exceptions: Thai (non-significant) and Japanese (not included) within Polyglot.**
- *Word Embedding models reflect L1 influence over learner English errors to some extent.*

Experiment 2: L1 Influence and Language Families

Whether distributional representation of words exhibit similar relationships between genealogically similar languages.

- Group L1s into Genealogical groups:
 - **Germanic**: German, Swedish
 - **Romance**: French, Spanish, Catalan, Italian, Portuguese
 - **Asian**: Chinese (simplified), Japanese, Korean, Thai
 - **Slavic**: Russian, Polish
 - **Other***: Turkish, Greek

*Other computed but not included in analysis

Experiment 2: L1 Influence and Language Families

Whether distributional representation of words exhibit similar relationships between genealogically similar languages.

- Compute differences between $EPNO_{English}$ and $EPNO_{L1} \rightarrow \Delta_{fasttext}$ and $\Delta_{polyglot}$ within groups.
- Δ computed for 1000 (i, c) resamples within each group averaged over 10,000 iterations.
- *A lower Δ would indicate similarities in error word pairs between the group and English.*
- Measure significance of difference in Δ between groups using ANOVA.

<i>Group</i>	<i>L1</i>	$\Delta_{fasttex}$ <i>t</i>	$\Delta_{polyglot}$
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

Experiment 2: Results

- Contrasting results between $\Delta_{fasttext}$ and $\Delta_{polyglot}$:
 - $\Delta_{polyglot}$ tends to agree with the initial assumptions of K&S (2016, 2017) → Languages closer to English ($EPNO_{Germanic}$) are least different from $EPNO_{English}$.
 - $\Delta_{fasttext}$ tends to agree with the findings of K&S (2016, 2017) → Languages farther from English ($EPNO_{Asian}$, $EPNO_{Slavic}$) are least different from $EPNO_{English}$.
- One-way ANOVA test revealed significant differences between language groups for both **fasttext** ($F(4, 49995) = 16539, p < 2 \times 10^{-16}$), and **polyglot** ($F(4, 49995) = 128751, p < 2 \times 10^{-16}$).

Experiment 2: Vector Differences

fasttext	polyglot
300 dimensional	64 dimensional
Vocabulary size of 1m - 10m	Vocabulary size of 10k - 100k
Trained using a subword level + contextual objective	Trained using only contextual objective

Experiment 2: Vector Differences influence NN choice

Nearest neighbors of *almost* in fasttext and polyglot embeddings

fasttext	polyglot
nearly	nearly
practically	once
virtually	roughly
almsot	just
Almost	equally
amost	virtually
almost	somewhat
alomst	less
damn-near	absolutely
pretty-much	slightly

Conclusion

- Analysis of L1 effect on content word errors based on semantic relatedness using two multilingual word embedding models: **fasttext** and **polyglot**.

Conclusion

- Analysis of L1 effect on content word errors based on semantic relatedness using two multilingual word embedding models: **fasttext** and **polyglot**.
- Association of L1 with English error word pairs.

Conclusion

- Analysis of L1 effect on content word errors based on semantic relatedness using two multilingual word embedding models: **fasttext** and **polyglot**.
- Association of L1 with English error word pairs.
- Analysis of patterns when L1s grouped into Genealogical groups.

Conclusion

- Analysis of L1 effect on content word errors based on semantic relatedness using two multilingual word embedding models: **fasttext** and **polyglot**.
- Association of L1 with English error word pairs.
- Analysis of patterns when L1s grouped into Genealogical groups.
- Conflicting results between:

Conclusion

- Analysis of L1 effect on content word errors based on semantic relatedness using two multilingual word embedding models: **fasttext** and **polyglot**.
- Association of L1 with English error word pairs.
- Analysis of patterns when L1s grouped into Genealogical groups.
- Conflicting results between:
 - **fasttext** (similar L1s most semantically different than English)
 - **polyglot** (distant L1s most semantically different than English)

Conclusion

- Analysis of L1 effect on content word errors based on semantic relatedness using two multilingual word embedding models: **fasttext** and **polyglot**.
- Association of L1 with English error word pairs.
- Analysis of patterns when L1s grouped into Genealogical groups.
- Conflicting results between:
 - **fasttext** (similar L1s most semantically different than English)
 - **polyglot** (distant L1s most semantically different than English)
 - Difference in results attributed to inherent differences between vector spaces.

Limitations

- Highly dependent on translation quality.

Limitations

- Highly dependent on translation quality.
- Small corpus → might not be representative.

Limitations

- Highly dependent on translation quality.
- Small corpus → might not be representative.
- How much positive correlation between semantic overlap is sufficient to explain variation?

Limitations

- Highly dependent on translation quality.
- Small corpus → might not be representative.
- How much positive correlation between semantic overlap is sufficient to explain variation?
- Not a “default” ICCM work...

Future Work

- Take into account bilingual lexicons for better translation. BabelNet, Multilingual Wordnet, etc.

Future Work

- Take into account bilingual lexicons for better translation. BabelNet, Multilingual Wordnet, etc.
- Contextualized word vectors: word's vector dependent on the context it occurs in (different vectors for different senses & occurrences of the word)

*I would like to **book** an appointment. vs I enjoyed reading that **book**.*

Future Work

- Take into account bilingual lexicons for better translation. BabelNet, Multilingual Wordnet, etc.
- Contextualized word vectors: word's vector dependent on the context it occurs in (different vectors for different senses & occurrences of the word)

*I would like to **book** an appointment. vs I enjoyed reading that **book**.*

- Collection of a larger, more representative error annotated corpus:
 - Can be used to fit a model to estimate error rates of content words in the corpus.
 - Model can use Semantic features such as word vector dimensions.
 - Analysis of model estimates → better explanation power.



Thank You!
Questions?



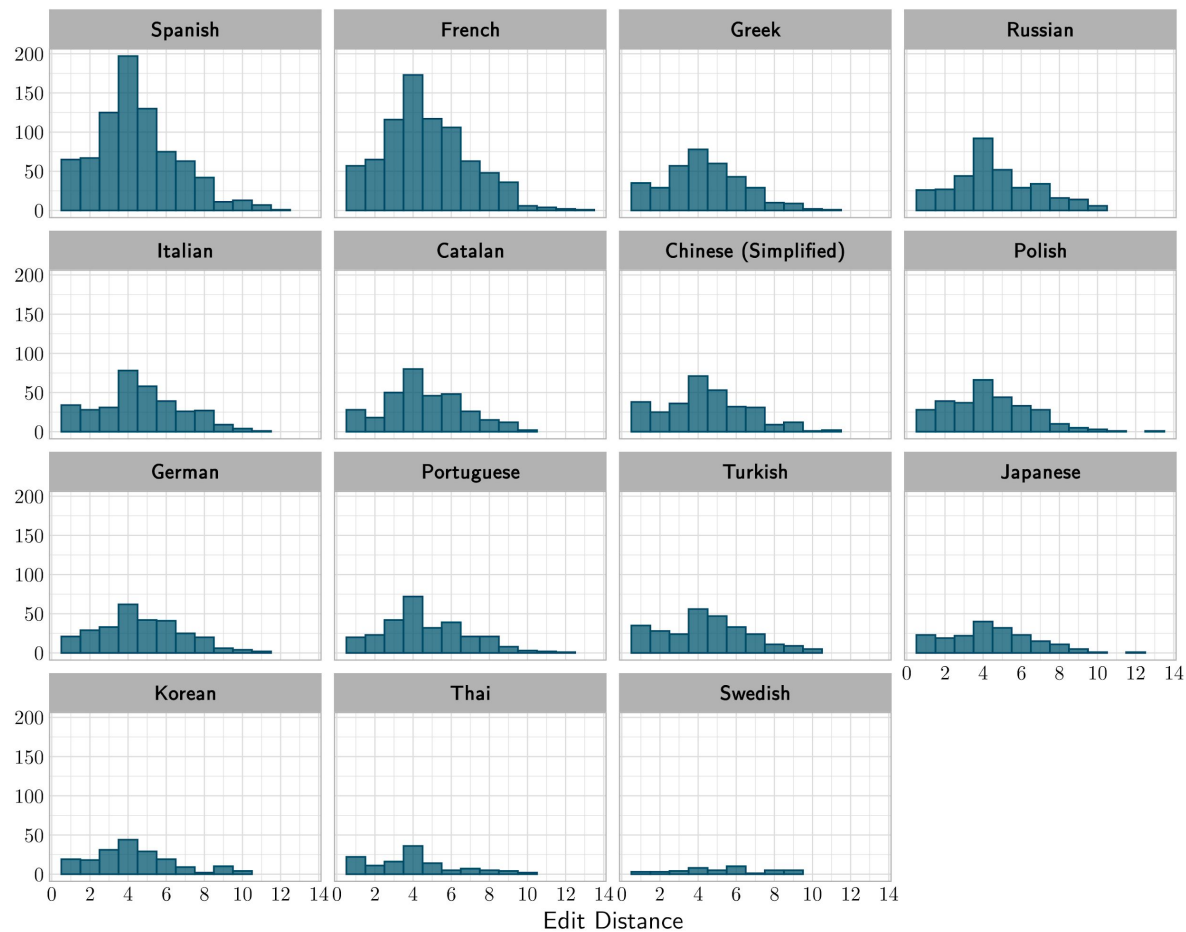
Kanishka - @iamasharkskin
Hemanth - @daemon92



kmisra@purdue.edu
hdevarap@purdue.edu
jtaylor1@purdue.edu

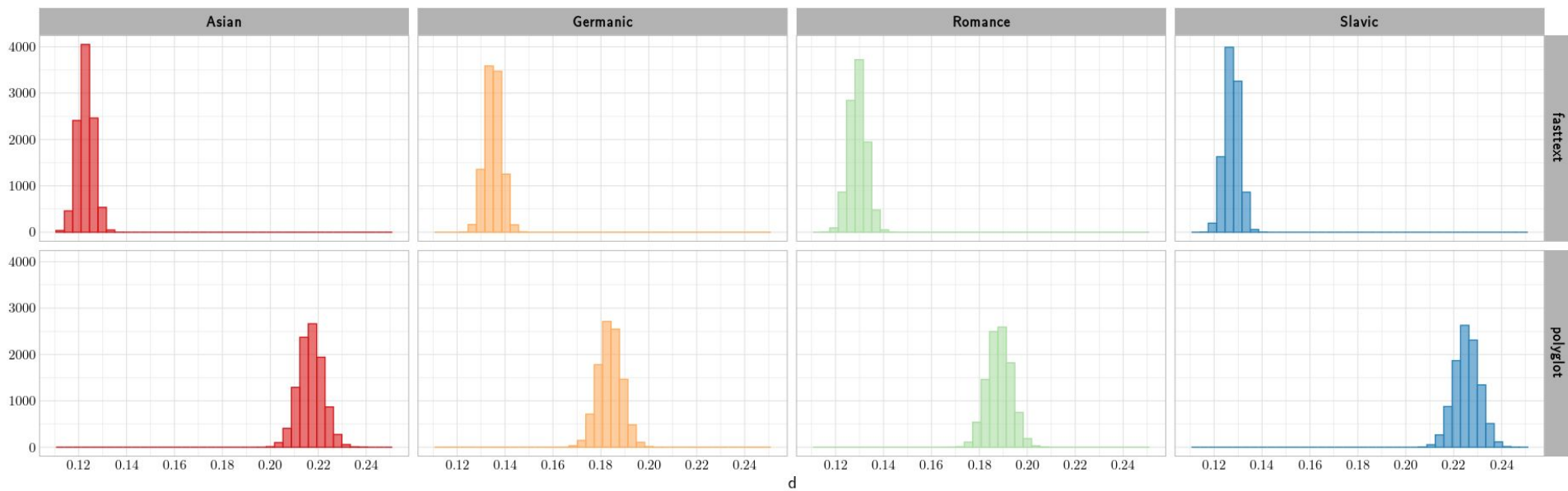


Coming soon..





Group ■ Asian ■ Germanic ■ Romance ■ Slavic



Agenda

- Motivation
- Goals and Contributions of the Research
- Literature
 - Word Embeddings
 - L1 Influence on Content Word Errors
- Measuring L1 influence Within Word Embeddings
- Investigating differences in
- Questions and Discussions