Interferência da língua nativa na escolha das palavras durante a escrita em outra língua.

Integrantes
Iasmin
Luiz
Rafael
Thiago

CITAÇÕES

A Computational Approach to the Study of Multilingualism
Ella Rabinovich
2018

≡ VIEW 1 EXCERPT | CITES BACKGROUND

Native Language Cognate Effects on Second Language Lexical Choice
Ella Rabinovich, Yulia Tsvetkov, Shuly Wintner
Transactions of the Association for Computational Linguistics • 2018

≡ VIEW 1 EXCERPT | CITES BACKGROUND

Punctuation as Native Language Interference
Ilia Markov, Vivi Nastase, Carlo Strapparava
COLING • 2018

Sentences and Documents in Native Language Identification
Andrew Cimino, Felice Dell'Orletta, Dominique Brunato, Giulia Venturi
CLiC-it • 2018

Anglicized Words and Misspelled Cognates in Native Language Identification

Conference Paper | Full-text available | Aug 2019

Ilia Markov · Vivi Nastase · Carlo Strapparava

Automatic Native Language Identification

Thesis | Full-text available | Oct 2018

Ilia Markov

Fontes:
ResearchGate e
Semantics Scholar

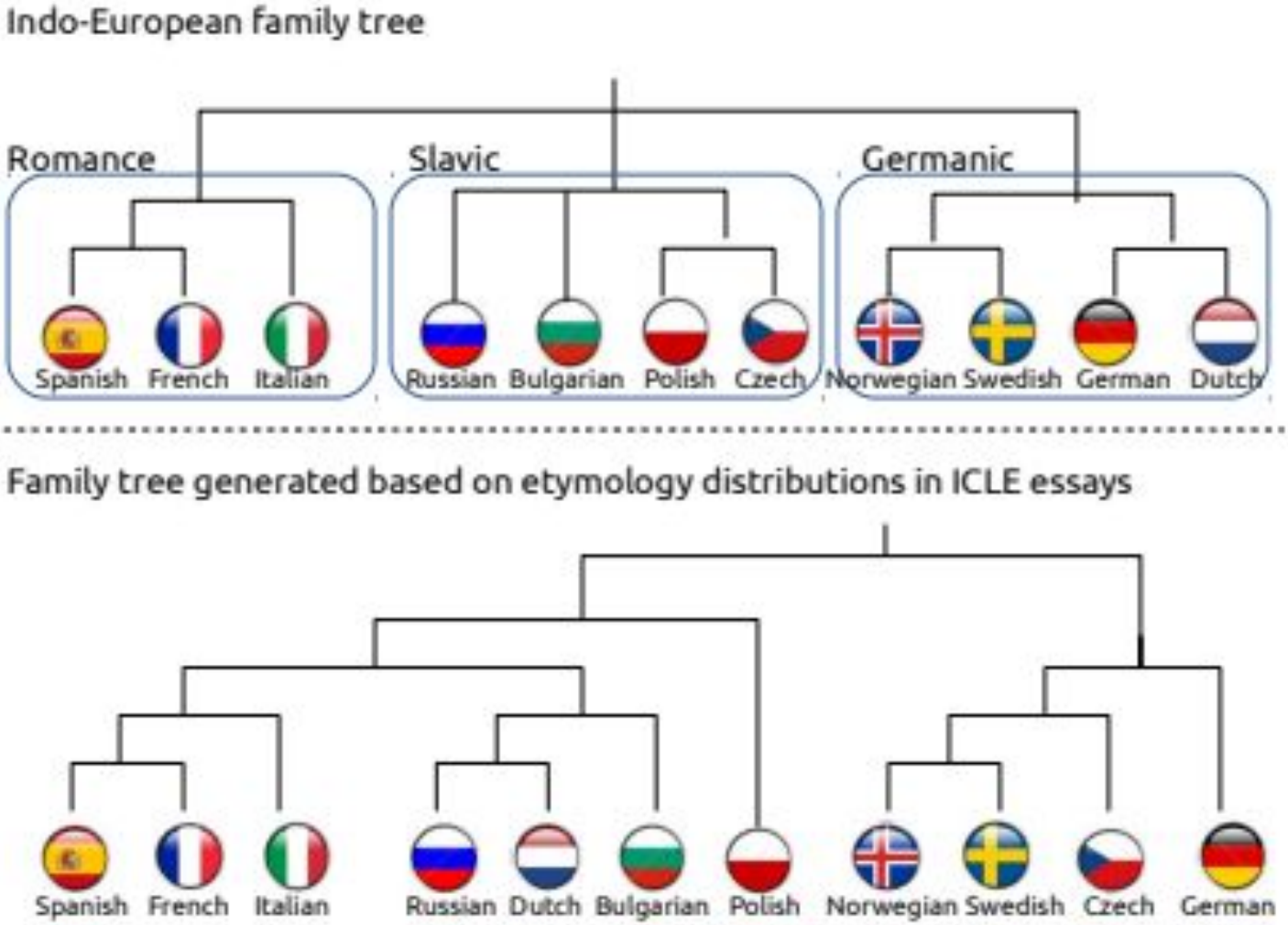
ARTIGO REFERÊNCIA: WORD ETYMOLOGY AS NATIVE LANGUAGE

- **Hipótese:** Existe interferência da língua nativa na escolha das palavras quando estudantes estão escrevendo em outra língua (neste caso Inglês).

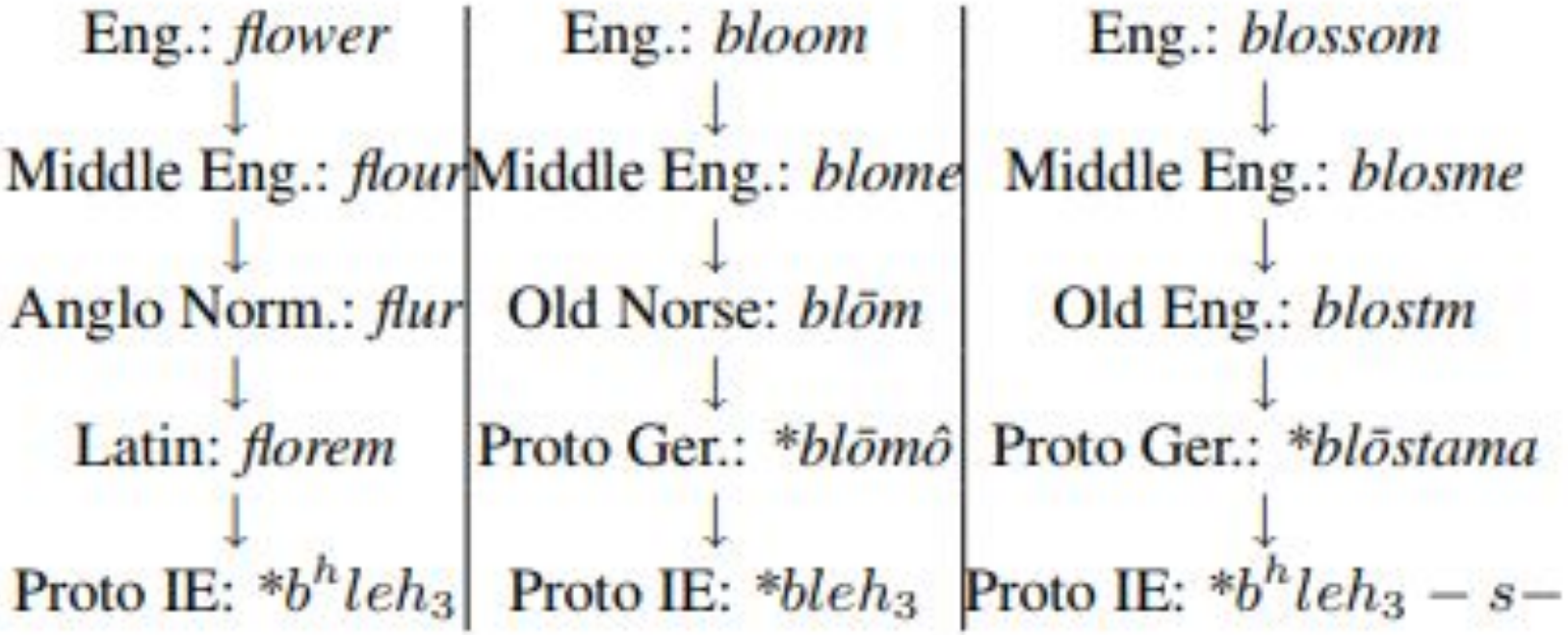
- **Método:** tratar os dados com técnicas de PLN, utilizá-los em um classificador, e observar a acurácia com e sem informação etimológica. .

- **Resultados:** não observaram diferença significante na acurácia quando adicionaram informação etimológica nos dados tratados, mas a evidência cumulativa da informação etimológica permitiu criar uma árvore genealógica de linguagens que se aproxima da real.

ARVORE DE LINGUAGEM



ARTIGO REFERÊNCIA: WORD ETYMOLOGY AS NATIVE LANGUAGE



ADAPTAÇÃO DO NOSSO TRABALHO

- Treinar um classificador SVM para que o mesmo verifique se um texto em inglês foi escrito por um estudante nativo ou não.

- Hipótese: Ao adicionar a informação etimológica das palavras a acurácia do classificador deveria sofrer alteração.

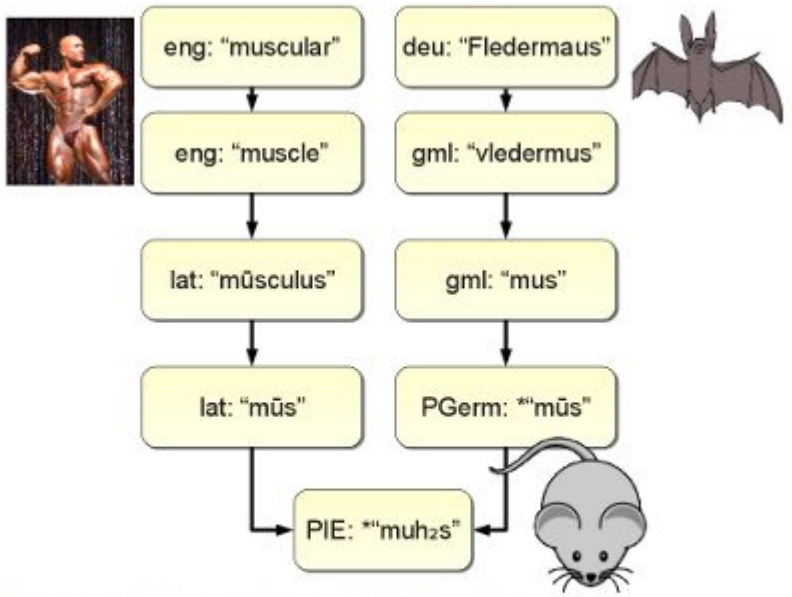
FINGERPRINT ETIMOLÓGICO

- Vetor de frequências que representa a composição etimológica de cada texto analisado.
- Representa a proporção de palavras que cada idioma ancestral possui no texto.
- Usando uma estrutura de *hash* para o dicionário etimológico utilizado, a construção tem complexidade *O(W)*, com *W* sendo W o número de palavras.

ita	0.0099
non	0.0100
frm	0.0150
grc	0.0290
fra	0.0315
xno	0.0477
fro	0.1478
ang	0.1945
enm	0.2246
lat	0.2802

ETYMOLOGICAL WORD NET

- Informações extraídas do Wiktionary.
- Mais de 3 milhões de entradas (cerca de 300 mil do inglês).
- Relações etimológicas (‘eng’,’test’ -> ‘lat’,’testum’).
- Usando Pandas com as palavras como índice, é possível fazer a busca em O(1).



Who would have thought that the English word "muscular" and the German word for the animal "bat" share the same origins?

Projeto Etywn: http://www1.icsi.berkeley.edu/~demelo/etymwn/

CONSTRUÇÃO DA BASE DE FINGERPRINTS

- Para cada documento no corpus, um vetor fingerprint é construído.
- Para cada documento pré-processado, iteramos por cada token e obtemos o idioma de origem da palavra consultado a Árvore Etimológica, fazendo uma contagem.
- Normalizamos a frequência pelo total de tokens no documento.
- O resultado é uma matriz de *n documentos x m idiomas de origem*.
- O processo de construção tem complexidade de O(W), onde W é o número total de tokens do corpus.

DADOS

- **Bases utilizadas: University of Oxford Text Archive**
  - **The Uppsala Student English Corpus**
    - Estudantes de Universidades Suecas.
    - 1489 textos com inglês como segunda língua.
  - **British Academic Written English Corpus**
    - Estudantes de Universidades Britânicas
    - 1953 textos com inglês nativo.
    - 757 textos com inglês como segunda língua.

PRÉ-PROCESSAMENTO

- Documentos separados de acordo com a língua nativa do autor
  - 1953 textos escritos por falantes nativos de inglês
  - 2246 textos escritos por falantes de outras línguas

- Remoção de tags presentes nos textos (<title>, <doc>, etc)
- Todas as letras convertidas para minúsculas
- Remoção de Stop Words
- Tokenização e Lematização
- Complexidade: O(|W|)
  - |W| = número de palavras dos documentos

DADOS PROCESSADOS

- Matriz de frequência das palavras para documentos escritos por nativos do inglês e não nativos
  - Consideradas apenas as 1000 palavras mais frequentes, em cada caso

- Matriz contento fingerprint etimológico para cada documento, cerca de 70 idiomas ancestrais identificados

TREINAMENTO

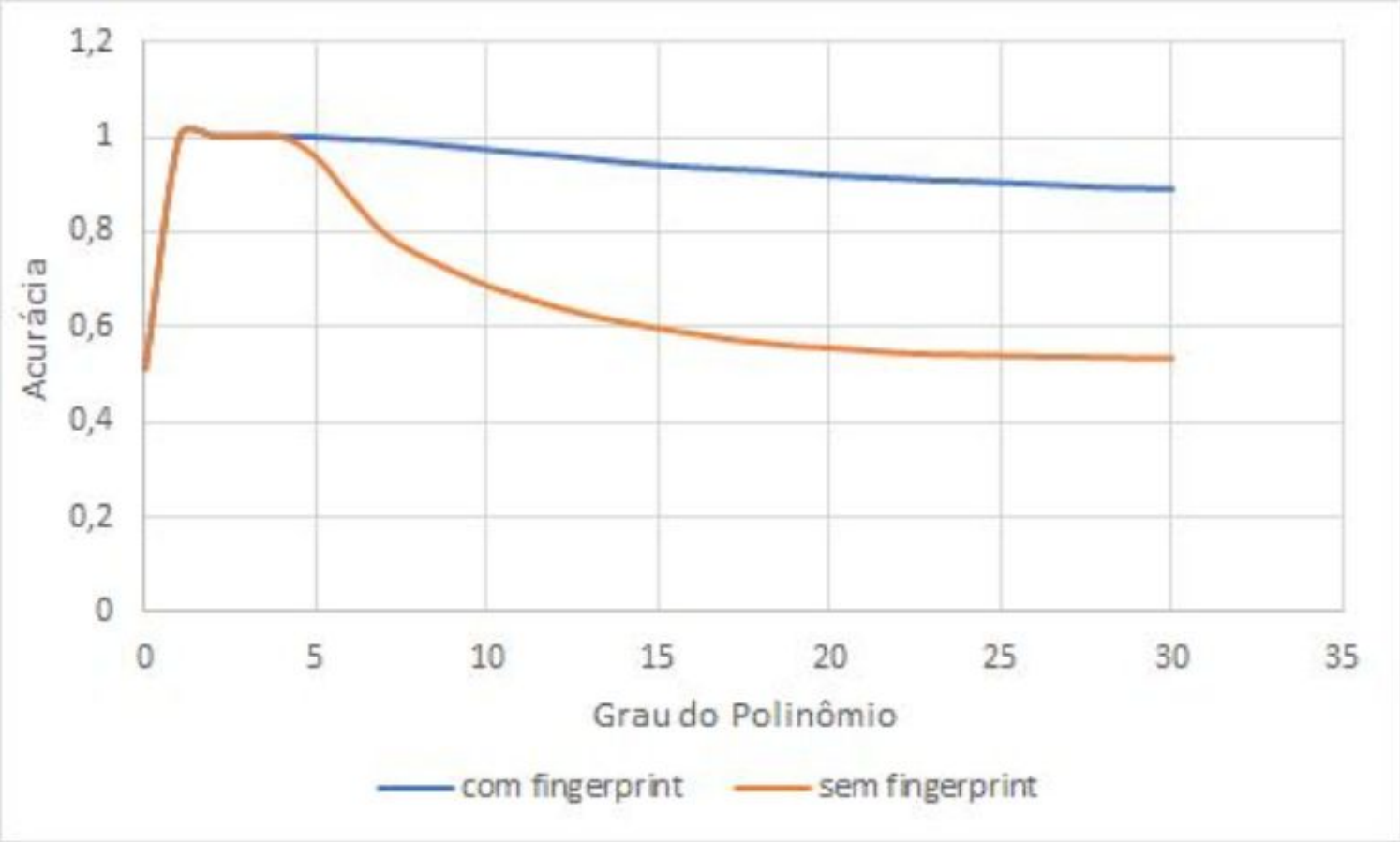
- Classificador SVM (Support Vector Machine) com Kernel polinômial
  - Treinamento apenas com palavras
  - Treinamento com palavras e fingerprint etimológico
  - Grau do polinômio variando de 0 até 30
- Validação cruzada 5-fold
- Biblioteca utilizada: scikit-learn
- Complexidade: $O(d^3)$
  - d = número de documentos

RESULTADOS

- Diferença pouco significativa usando polinômios de grau abaixo de 5

- Para polinômios de graus maiores a acurácia cai em ambos os classificadores
    - Entretanto, classificador com fingerprint etimológico mantém acurácia acima de 0.8
    - Classificador sem fingerprint etimológico cai rapidamente para cerca de 0.5 de acurácia (overfitting)

RESULTADOS


1,2
1
0,8
0,6
0,4
0,2
0
Acurácia
0
5
10
15
20
25
30
35
Grau do Polinômio
com fingerprint
sem fingerprint

CONCLUSÕES

- Aparente interferência da língua nativa na escolha lexical
    - Diferença significativa na acurácia para polinômios de graus elevados

- Necessário outros testes para verificar possíveis fatores que possam estar influenciando os resultados
    - Temas dos textos
    - Uso de base diferentes
    - Padronização nos formatos