Analysis of cross-lingual semantic change in professional discourse with quantitative methods

Joanna Byszuk joanna.byszuk@ijp.pan.pl @jbyszuk Institute of Polish Language Polish Academy of Sciences

BACKGROUND

WHY STUDY THAT?

1. to discover mother-tongue dependent stylistic variation between texts of similar topic and use;

2. to reveal cross-lingual semantic shifts, e.g. differences in the use of terminology;

3. to evaluate possible limitations of used methods,

WHAT'S INSIDE THE DATASET?

Blog posts by: programmers from Polish-, English-, German- and Spanish-speaking countries written in English, marked for authors' nationality.

HOW TO APPROACH THAT?

supervised machine learning classification
+
network analysis
+
word vector models
+
Levenshtein's distance for spelling variation

IS ENGLISH THE LINGUA FRANCA FOR PROGRAMMERS?

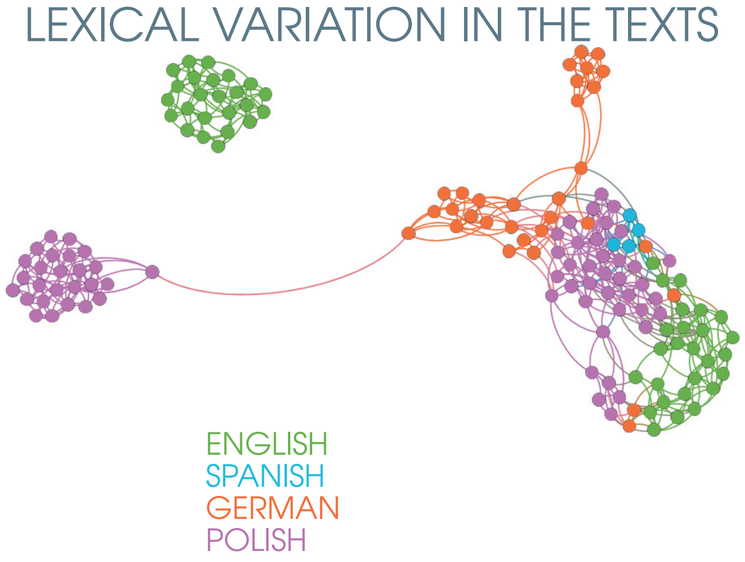
CAN WE SEE THE NATIONALITY FACTOR IN COMMUNICATION STYLE?

SPELLING TRENDS:

* general prevalance of forms in AmE spelling, especially -ize, -yze forms
* especially in PL, noticable slight tendency to mix BrE and AmE terms
* PL, ES and DE more likely to use more hyphenated forms, e.g. co-workers,

* troublesome words:
favo(u)rite, colo(u)r, behavio(u)r, travel(l)ing
* interestingly 'cancelled' is only spelled with double 'l' in EN corpus

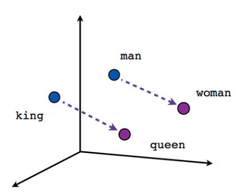
LEXICAL VARIATION IN THE TEXTS



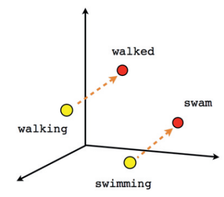
SYNTACTICAL VARIATION OF THE TEXTS

ANSWERING SOCIOLINGUISTIC QUESTIONS WITH WORD VECTORS

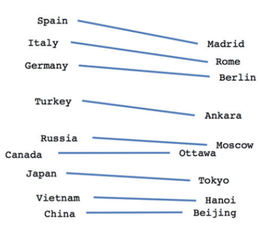
vector representation of words can help unravel semantic similarity between terms - what if we compare results for various languages?



Male-Female

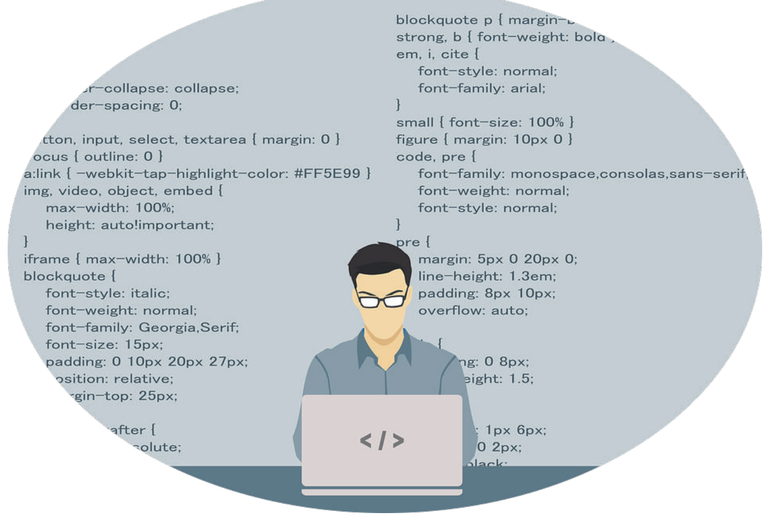


Verb tense



Country-Capital

I CODE, SO WHO AM I?



a developer
(92.5% EN, 82% DE, 72% PL, only 15% ES)
a programmer
(55% ES, 19% PL, only 7% DE and 1.5% EN)
an engineer
(28.5% ES, 11% DE, 9% PL, 6% EN)

WHAT DOES IT MEAN TO 'WORK'? NEAREST NEIGHBORS:

EN:
work it but so done not get also some with we they s that just was even all great this
DE:
work licensed this that do but so is well which just it also they on means your out something change
ES:
work they how things with it and will on this to are together me all can these do that be
PL:
work working and they but we their it when even will people them usually not make your

Bibliography:
Bakarov, A. (2018). A Survey of Word Embeddings Evaluation Methods. arXiv:1801.09536 (cs), 21 January 2018. http://arxiv.org/abs/1801.09536.
and semantic relations in domain-specific discourses. Revista Alicantina de Estudios Ingleses 24. 213-233.
Ruder, S., Vulić. I. and Søgaard A. (2017). A Survey Of Cross-lingual Word Embedding Models. arXiv:1706.04902 (cs), 15 June 2017. http://arxiv.org/abs/1706.04902.
Solly, M. (2015). The Stylistics of Professional Discourse. Edinburgh: Edinburgh University Press.
Stewart, I., Chancellor, S., De Choudhury, M. and Eisenstein, J. (2018). #anorexia, #anarexia, #anarexyia: Characterizing Online Community