20 May 2021

Present: DF, SD, VDS

Topic: steps for analyzing texts

* Remove grammatical morphemes
* Remove vowels find word to match it with 🡪 potentially count word movement
* Levenshtein distances (LD) 🡪 versus bag of consonants method
* Bag of consonants method
  + Don’t remove grammatical morphemes, just vowels
  + Figure out proportion (%) of consonants are from grammatical morphemes
  + Need to identify morphemes by hand at first

TO DO:

DF – send VDS VBA code and draft of my norming study paper

VDS – keep working on code for comparing texts

Maybe add LW

16 June 2021

Present: SD, DF, LW, VD’S

Oxford book chapter: ways in which languages can differ systematically (orthography, phonology, syntax, morphology, etc)

Research project:

* VD’S work so far:
  + should we use the LD form from Schrepens?
  + Currently using the stemmer to get a word’s root, but this may not be great
    - Is it correctly stemming?
      * Any vowels that it chops off are okay bc we don’t care about them anyway
    - Removes inflections but not derivations
      * We want it to keep derivations
      * Not clear whether speakers store derivational morphemes in their mental lexicon or not
      * DF needs to share her stemmer (i.e., lemmatizer, same thing, diff software)
  + Aligning translation equivalents by hand for method 1 (traditional LD of word by word)
  + We want to remove clitics, determiners, pronouns
    - Currently being done bc frequent words are removed
    - May not want to do this in method 2
  + Score between EN + SP = .6, score between EN and RS = .2
    - Cyrillic is being transliterated into latin letters (good enough for now, may be a topic we want to revisit)
* Method 2 (bag of consonants method) needs some revision
  + Currently looks more like method 1 (using straightforward LD)
  + Instead, want to count number of consonants
    - E.g. total number of Ps, Ts, Ks, etc
      * Determine number of difference between languages
      * For example, English has 4 Ps and 1 T, while Spanish has 1P and 2 Ts = 4 total different consonants
    - Then apply consider string length from that total # of difference
  + We will also want to figure out from this if a certain % of consonants belong to grammatical morphemes consistently (e.g., plural ‘s’, past tense ‘d’, etc)
    - So for this **don’t use stemmer**
* Eventual goal
  + Will potentially want to include manner of articulation, other phonological differences
    - Complicated by the fact that speakers of languages with the same alphabet use both orthographic and phonological information, and orthography may be a better predictor than phoneme difference (see DF’s paper with Mila Tasseva)
  + Will want to also include larger number of sources from a variety of domains
    - Just using UN documents for now, will want documents with casual language
    - Lexical frequency and borrowing rates affected by domain (topic) of the language
  + Will want to include direction translations, **not** different editions
    - Eg, novels change content to market to different audiences
* Next steps
  + VD’S + LW will meet sometime before next week to discuss the code
  + SD will create a teams channel
  + Next meeting on Wednesday 6/23 depending on VD’S’s availability

23 June 2021

Present: DF, SD, VDS, LW

Topic: progress on methods

* Bag of consonants method:
  + Gotten through calculating how many of each consonant
    - But removed all grammatical morphemes
    - So need to go back and return those and recalculate
    - DON”T use stemmer/lemmatizer for this method (but DO use with LD method)
  + How to calculate the proportion of grammatical morphemes
    - Got to start out by hand for this
    - Start with a 500 word text and mark out the grammatical morphemes
    - Count all consonants and figure out how many of each are grammatical
* Sources of text
  + Stick with EU/UN documents for now, since they’re direct translations without much editing
  + Go on to news sources next
  + Then phone transcript translations
    - SD will ask Amy Irbach at Green Dragon if she has any we can use
  + Translation/transliteration method matters
    - Russian, French, Spanish, german and english for now, with maybe Arabic as a control>
    - Chinese pinyin has been normalized with spelling conventions, which may be of issue because Chinese L1 learners are taught pinyin in school
  + Eventually we will also want to consider the effect of genre/register, but not right now
* Text length
  + Also see if text length makes a difference by calculating language matches by text length
  + Start with 20 word (ish), 50 word, and 300 word
    - ISH- because number of words may vary cross-linguistically depending on how languages lexicalize concepts

TO DO:

-DF: background sources

-SD: contact Amy about transcripts

-LW & VDS: work on perfecting the methods to calculate in a manner we all agree on

Wiley Blackwell handbook – proposal going in week after July 4