* Shared LCP poster
* Open subtitles
  + Pulled mostly from movies/tv shows
  + Rip subtitles from files
  + Word lists – team who does it using opensubtitles (not directly opensubtitles)
  + [**https://github.com/hermitdave/FrequencyWords/**](https://github.com/hermitdave/FrequencyWords/)
    - [**http://www.lrec-conf.org/proceedings/lrec2016/pdf/947\_Paper.pdf**](http://www.lrec-conf.org/proceedings/lrec2016/pdf/947_Paper.pdf)
    - [**https://opus.nlpl.eu/OpenSubtitles2018.php**](https://opus.nlpl.eu/OpenSubtitles2018.php)
    - Includes number of files, tokens, and sentences per language
    - BLEU scores – Evaluation of quality of Translation (TL)
    - Haven't done – TL words from lang A -> Lang B to get most freq TL of word
      * Rapido -> rapid or fast
      * How does the direction of learning affect understanding
      * Eng rapid -> Rapido is easier cause they know rapid
      * ES rapido -> Fast isn’t as good bc similarity is not clear
        + Asymmtry of similarity
* Frequency list taken from opensubtitles (movie/tv show genre)
  + Github repo frequency list from?– credit if used
  + Our Freq list – devowelized, not stemmed, removed stop words
  + Txt file - see teams drive
  + Words lists TL b4 devowelization
  + Cleaning out conjugations of words (make, makes, making etc)
    - How to account for other languages where grammatical affixes are no as clear cut?
    - Should focus on yet?
  + Removing stop words should mostly be left with nouns, adjectives etc.
  + **Clean stop words and use top 2000 (accounts for conjugations of words etc)**
* LCP poster – read row to column direction of TL
  + Eng –> ES 0.31
  + ES -> ENG 0.40
  + A Spanish speaker to learn English is harder?
  + How does asymmetry of similarity affect learning
  + Divergence of distance based on frequency list is a product divergent frequency of elements
    - List from 1 to 1000 is not the same words from another language
    - Words that appear in both may be valued differently
* Distance b/t cognate connections
  + How to add this in?
  + TO DO FOR LATER
* Working with genres – can look at newspaper texts and scrape for freq lists
  + FOR LATER
* Morphological overlap
  + Inflections, declension, case words
  + Grammar and syntax
* **Pluck word lists from open subtitles**
  + **Try 1000 and try 2000 (see if it changes anything – worth it?)**
    - **Do we need more words?**
* **Now doing ENG, ES, FR, DE (12 combos)** 
  + **2 sets of comparisons 1000 vs 2000**
  + **Graph/Table data**
* **KNOWN FACTS**
  + **Eng close to FR bc of borrowing**
  + **ENG is Germanic lang**
  + Latin core of ENG and FR
* **LIVERPOOL Paper to Present – 30 mins?**
  + **Intro – 5 min**
  + **10 min Language freedom index - Gareth**
  + **10 min Linguistic distance measure - Vivian**
  + **5 min wrap up – what we've found, applications, where we’re going/goals**
  + **Just giving raw nums – general observations**
    - **ENG to ES is diff from ES to ENG**
  + **Give specific examples of linguistic asymmetry**
  + **Historical context of development of language – numbers match what we expect – follow the etymology – notes from Anyssa**
* The numbers
  + Serbia -> Croation vice versa 1
  + Pirahã –> Albanian -> 0
  + What do decimal values between 0 and 1 actually mean?
  + What value of statistical divergence is significant
* TO DO for next meeting
* Get 1000 and 2000 wordlists with raw data (numbers) examine
* Then examine – anomalies, find examples of asymmetry etc.