Discovering argumentatively linked lexicon

# Corpora used

1. Open subtitles corpus (“moses” single language file format):
   1. De-En , 22 M sent; 278 M words <http://opus.nlpl.eu/download.php?f=OpenSubtitles/v2018/moses/de-en.txt.zip>
   2. De-Fr , 16 M sent; 203 M words <http://opus.nlpl.eu/download.php?f=OpenSubtitles/v2018/moses/de-fr.txt.zip>
   3. En-Fr, 42 M sent; 535 M words <http://opus.nlpl.eu/download.php?f=OpenSubtitles/v2018/moses/en-fr.txt.zip>
2. European Medicines Association (EMEA corpus)
3. Europarl corpus

# Example workflow (English)

## Extracting patterns: “however”

grep -hr "however" < OpenSubtitles.de-en.en | awk 'BEGIN {FS="[ :]";}{for(i=1;i<=NF;i++){if($i ~ /^however/){printf $0; printf "\tI:"; printf i; printf "\tNF:"; printf NF; printf "\n"}}}' > OpenSubtitles.de-en.en--however3.txt

grep -hr "however" < OpenSubtitles.de-en.en | awk 'BEGIN {FS="[ :]";}{for(i=1;i<=NF;i++){if($i ~ /^however/ && i / NF >0.25 && i / NF<0.75){printf $0; printf "\n"}}}' > OpenSubtitles.de-en.en--however4.txt

grep -hr "however" < OpenSubtitles.de-en.en | awk 'BEGIN {FS="[ :]";}{for(i=1;i<=NF;i++){if($i ~ /^however/ && i / NF >0.1 && i / NF<0.9 && NF > 7){print $0}}}' > OpenSubtitles.de-en.en--however4.txt

sort < OpenSubtitles.de-en.en--however4.txt | uniq -c | sort -nr >OpenSubtitles.de-en.en--however4uq-std.txt

uniq < OpenSubtitles.de-en.en--however4.txt >OpenSubtitles.de-en.en--however4uq.txt

## Parsing

lexparser-lang.sh English 150 edu/stanford/nlp/models/lexparser/englishPCFG.ser.gz out /Users/bogdan/elisp/hd2011arg/\_data/corpArgumentation/c04openSubtitles/opensub-deen/OpenSubtitles.de-en.en--however4uq.txt

# Example workflow (German)

## Extracting patterns: “jedoch”

grep -hr "jedoch" < OpenSubtitles.de-en.de | awk 'BEGIN {FS="[ :]";}{for(i=1;i<=NF;i++){if($i ~ /^jedoch/ && i / NF >0.1 && i / NF<0.9 && NF > 7){print $0}}}' > OpenSubtitles.de-en.de--jedoch4.txt

uniq < OpenSubtitles.de-en.de--jedoch4.txt > OpenSubtitles.de-en.de--jedoch4uq.txt

iconv -f UTF-8 -t ISO-8859-1//IGNORE < OpenSubtitles.de-en.de--jedoch4uq.txt > OpenSubtitles.de-en.de--jedoch4uq-iso.txt

lexparser-lang.sh German 150 edu/stanford/nlp/models/lexparser/germanPCFG.ser.gz out /Users/bogdan/elisp/hd2011arg/\_data/corpArgumentation/c04openSubtitles/opensub-deen/OpenSubtitles.de-en.de--jedoch4uq-iso.txt

# Results

Parsed files on

<https://heibox.uni-heidelberg.de/d/5c6a33a93c854911a6f7/>

(S

(NP

(NP (DT Another))

(SBAR

(WHNP (WP who))

(S

(VP (VBZ deserves)

(NP (NN praise))))))

(, ,)

(ADVP (RB however))

(, ,)

(VP (VBZ is)

(VP (VBG spat)

(PP (IN upon))))

(. .))

**goal: to extract the link**

deserves prise // is spat upon

**Suggestion**:

same highest-level constituents on both sides, e.g., VP