xTitle

proposition & coherence in :schizophrenia: threads

stephan schwarz / a. stefanowitsch:16827_25S:sprache und psychose

subject

Investigate reference marking, coherence and information structure in schizophrenia language by measuring distance of similar nouns within range of comment thread preceded by certain determinants.[1]

background

Inspired by Zimmerer et alii (#REF) we are interested in observations concerning coherence and propositional conditions in schizophrenia language, as these linguistic markers appear underinvestigated in research while they seem to play a crucial role within target group language. (As such seen as asset of thinking or world building capacity which might suffer from linguistic deficits within the range of positive symptoms.)

method

To compute distances we queried a corpus for matching conditions where certain (assumed) determiners appear before similar nouns. This distance should give us information structural evidence of how strong these noun occurences are connected, i.e. if a noun appears out of the blue mostly or if it somewhere before has been introduced to the audience. In information structure definitions this would be termed with **given and new information** Prince (1981#REF).

questions

Measuring the referent-reference distance which we here assume as indicator of coherence we hope to find empirical evidence for disturbed or not world building capabilities within schizophrenia language. Premising that a large noun distance indicates a low reference-referent association we hypothesise that in a language/TOM setting where the speakers estimation of the audiences context understanding capacities is disturbed we will find higer medium scores for the distance under matching conditions.

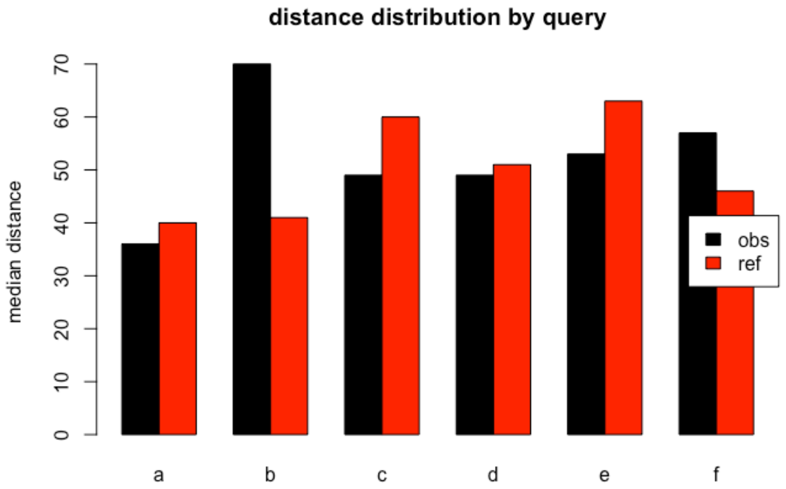
daten

We built a corpus of the reddit r/schizophrenia thread (`n=747089` tokens) and a reference corpus of r/unpopularopinion (`n=265670`). The corpus has been pos-tagged using the R udpipe:: package #REF which tags according to the universal dependencies tagset maintained by #REF. Still the 747089 tokens can only, with the workflow of growing the corpus and devising the noun distances developed be just a starting point from where with more datapoints statistical evaluation becomes relevant first.
The dataframe used for modeling consists of `17794` distance datapoints derived from the postagged corpus.

```
##       dist q target url   lemma range mf_rel     ld
## 42      48 a    obs  17  friend   287 0.0348 0.5192
## 10919   26 c    ref  75     fun   731 0.0027 0.4555
## 6303    12 b    obs 699  people  1052 0.0124 0.4059
## 6160   213 b    obs 633     day  2442 0.0061 0.3071
## 13172   52 d    ref   5   world  2415 0.0037 0.3097
## 14373   12 d    ref  97  burger  2515 0.0179 0.2831
## 579     51 a    obs 232    face   258 0.0233 0.5659
## 13180  177 d    ref   7 opinion   928 0.0075 0.4655
## 8470   213 c    obs 364    time   753 0.0027 0.4648
## 8100    20 c    obs  58     med   146 0.0342 0.6301
## 899      4 a    obs 353  moment   396 0.0051 0.5177
## 11461   26 d    obs 320 symptom   603 0.0265 0.4544
## 12695  203 d    obs 874   thing   992 0.0071 0.3861
## 11249   74 d    obs 155 emotion  1103 0.0100 0.3545
## 13330   33 d    ref  19 morning  5000 0.0056 0.2480
```

[1] snc.1:h2.pb.1000char/pg.queries

results



```
## ## conditions:
```

q	precedent	pos
a	ALL (.*)	NOUN
b	this,that,these,those	NOUN
c	the	NOUN
d	a,an,some,any	NOUN
e	my	NOUN
f	your,their,his,her	NOUN

conclusion

In condition **B** (`this, that, these, those`) which we hold for the most speaking determinants illustrating the speakers idea, that the information about a reference is already **given** we find significantly higher distance scores in the target corpus which proves our hypothesis. An ANOVA analysis of the linear regression model which posited a main effect of corpus*q and random effects of url range width, match frequency of the query and type/token-ratio within the range (`lme4::lmer(dist ~ corp*q +(1|range) + (1|mf_rel) + (1|ld))`) gets a p-value of p=0.02277 for target-corp:q.
So even if the median distance of nouns, preceded by one of our queries, is just `47` tokens wide for the target corpus and `46` in the reference corpus, it's still with respect to the covariates significantly (p<0.05) higher and yet to be tested on a larger corpus.

B. REF: