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).

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

questions

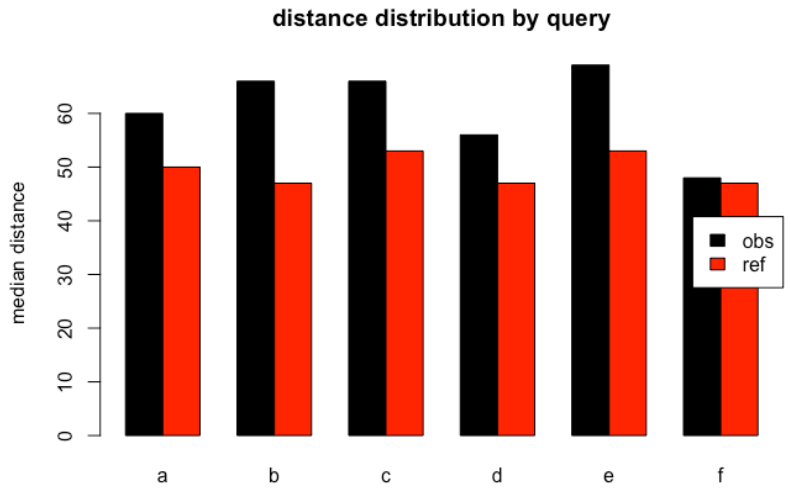
Measuring the referent-reference distance which we here assume as indicator of coherence we hope to find empirical evidence for disturbed or not world building capabilities within schizophrenia language. Premising that a large noun distance indicates a low reference-referent association we hypothesise 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=755074` tokens) and a reference corpus of r/unpopularopinion (`n=271563`). 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 755074 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 `87145` distance datapoints derived from the postagged corpus.

	dist	q	target	url	lemma	range	corpsize	det
24812	854	d	obs	723	drug	5636	755074	TRUE
22571	65	d	obs	412	psychia-trist	2917	755074	TRUE
25287	83	d	obs	805	test	5703	755074	TRUE
17671	75	c	obs	455	weight	2637	755074	TRUE
44012	33	a	ref	44	burger	6207	271563	TRUE
67066	5	c	ref	44	burger	6207	271563	TRUE
75815	22	d	ref	24	movie	2885	271563	TRUE
78384	15	d	ref	44	burger	6207	271563	TRUE
85166	99	f	ref	35	city	2876	271563	FALSE
10176	487	a	obs	805	intelli-gence	5703	755074	TRUE

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

Over all conditions 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+range and random effects of lemma (`lme4::lmer(dist ~ corp*q+range + (1|lemma)`) gets a p-value of `p=0.0000066` for the mean difference of `-25` tokens (targetref) compared to the target.
So the median distance of nouns, preceded by one of our queries, with `60` tokens width for the target corpus and `50` in the reference corpus, is also with respect to the covariates significantly (`p<0.001`) higher but still to be tested on a larger corpus.

B. REF: