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 al. (2017) 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 (M7)

To compute distances we queried a corpus for matching conditions where certain (assumed) determiners appear before similar nouns. In M7 we observed all matching antecendents of conditions b-f wether be tagged “DET” or not. 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).

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

questions

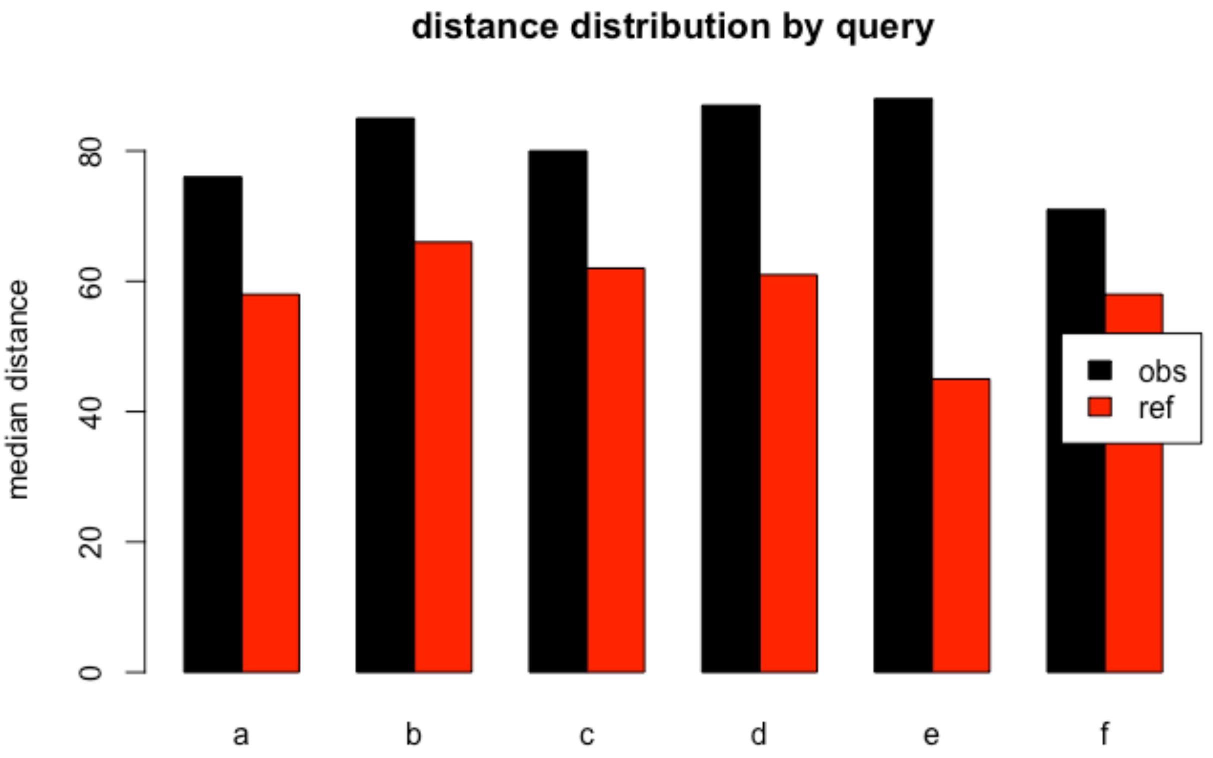
Measuring the referent-reference distance which we here assume as indicator of coherence we hope to find empirical evidence for disturbed or not world building capabilities within schizophrenia language. Premising that a large noun distance indicates a low reference-referent association we hypothesise 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`). Both were pos-tagged using the R udpipe:: package (Wijffels 2023) which tags according to the universal dependencies tagset maintained by De Marneffe et al. (2021). Still the 755074 tokens can only, within 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 M7 consists of `939879` distance datapoints (sample below) derived from the postagged corpus.

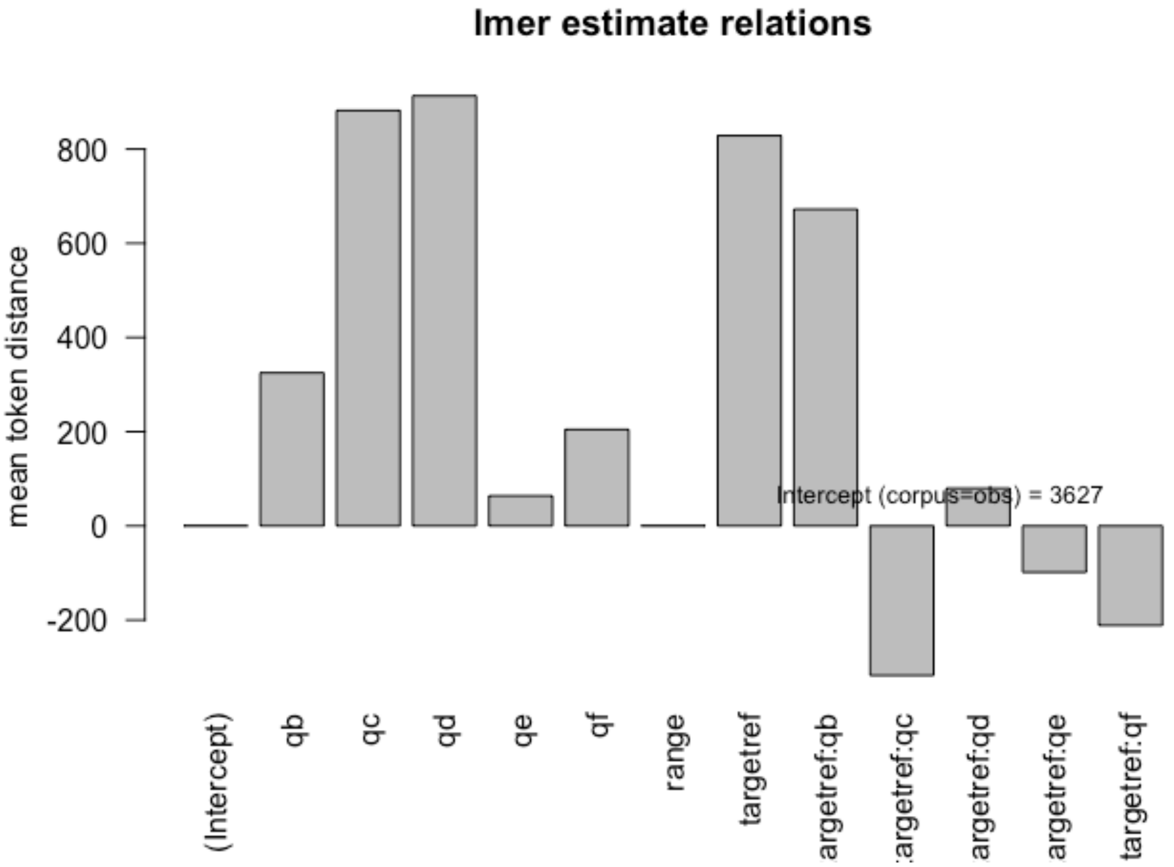
q	target	url	lemma	m	range	dist	det	pos
a	obs	760	friend	14	3808	100	FALSE	540136
a	ref	40	dude	3	5266	105	FALSE	112505
a	ref	47	food	23	3352	35	FALSE	136527
f	obs	964	brain	24	1959	45	FALSE	724599
a	ref	50	limit	59	4210	75	FALSE	140956
a	obs	887	time	13	8308	327	FALSE	644352
a	obs	631	one	20	3641	92	FALSE	409147
a	ref	73	lecture	65	6064	296	FALSE	204569
d	ref	47	dishwasher	26	3352	40	FALSE	134670
a	ref	26	bbq	47	3909	10	FALSE	70284

results



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

query conditions for preceding token



conclusion

Over conditions [`c, e, f`] we find significantly higher distance scores in the target corpus which proves our hypothesis. An ANOVA analysis of the linear regression model (cf. Bates et al. 2015) which posited a main effect of corpus*q+range and random effects of lemma (`lme4::lmer(dist~target*q+range+(1|lemma)+(1|det),df)`) gets a p-value of `p=0.0035625` for the mean difference of `829` tokens (targetref) compared to the target.

So the medium distance of nouns, preceded by one of our queries, is with `77` tokens width for the target corpus vs. `59` in the reference corpus also with respect to the covariables significantly (`p<0.01`) higher but still to be tested with growing the corpus.