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 (M5)

To compute distances we queried a corpus for matching conditions where certain (assumed) determiners appear before similar nouns. In M5 no restrictions concerning the matching antecedents to be tagged “DET” were accounted for. 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

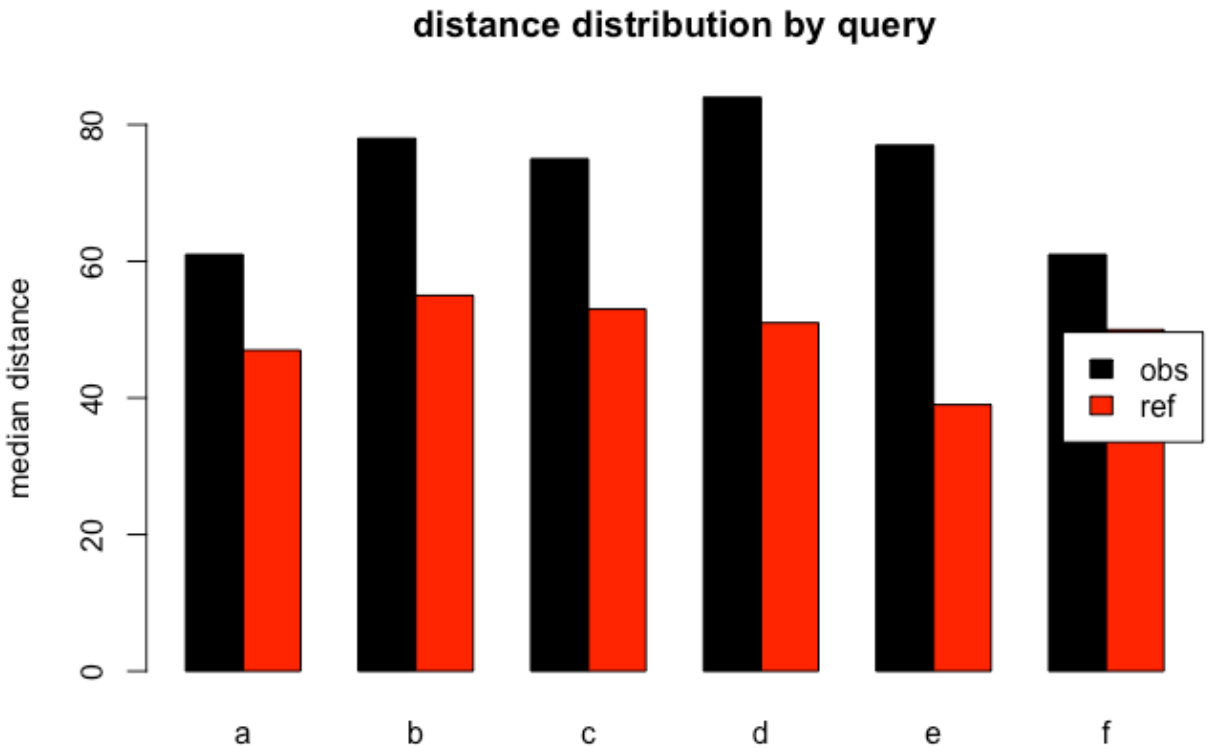
Measuring the referent-reference distance which we here assume as indicator of coherence we hope to find empirical evidence for disturbed or not world building capabilities within schizophrenia language. Premising that a large noun distance indicates a low reference-referent association we hypothesise 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 M5 consists of 259044 distance datapoints (sample below) derived from the postagged corpus.

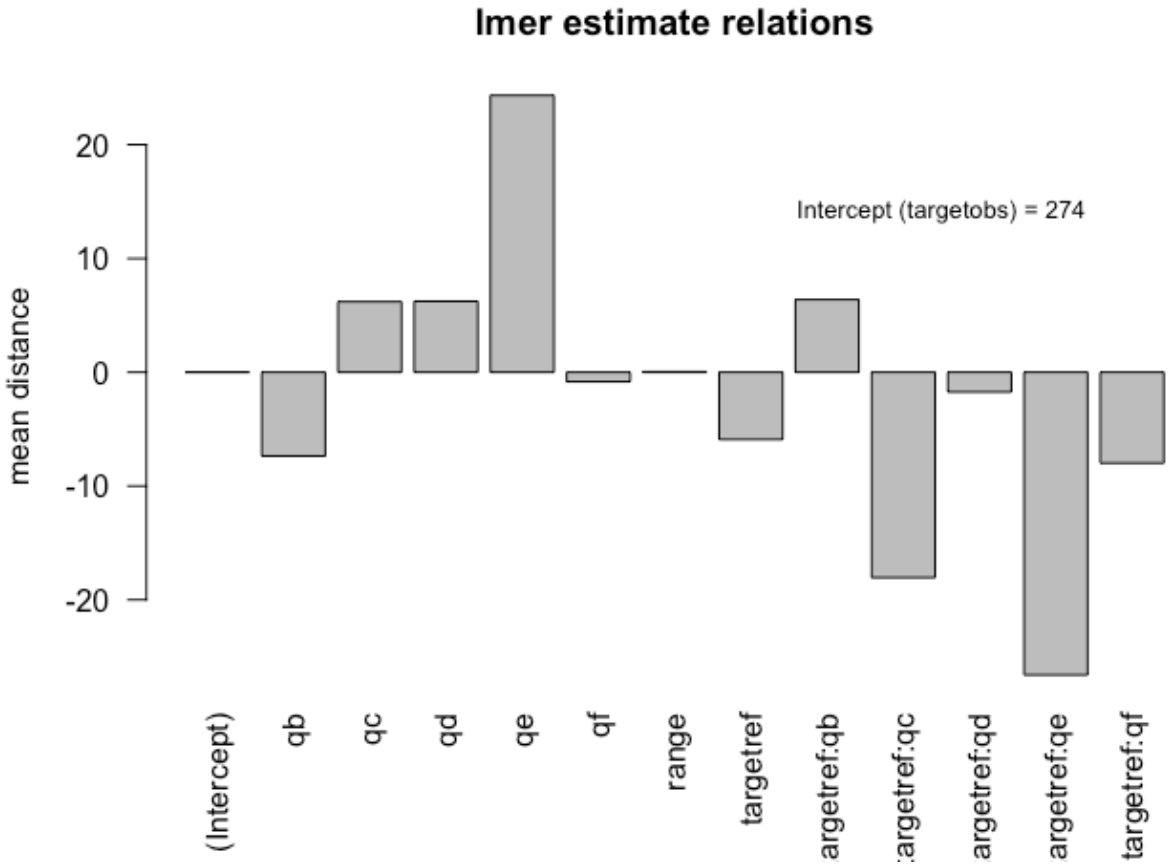
q	target	url	lemma	m	range	dist	det	pos
a	ref	44	burger	178	6207	15	FALSE	5996
b	obs	806	psychosis	8	945	17	FALSE	698
d	obs	622	horror	25	1181	134	FALSE	1084
a	ref	94	game	26	1612	145	FALSE	1549
a	ref	53	device	5	8169	5	FALSE	3850
c	ref	53	system	10	8169	50	FALSE	894
a	obs	172	network	4	289	59	FALSE	89
c	ref	53	price	64	8169	119	FALSE	1541
e	obs	340	head	5	215	25	FALSE	60
a	ref	32	song	10	1687	200	FALSE	218

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 all conditions 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),df)`) gets a p-value of `p=0.0000004` for the mean difference of `-6` tokens (targetref) compared to the target.
So the medium distance of nouns, preceded by one of our queries, is with `73` tokens width for the target corpus vs. `50` in the reference corpus also with respect to the covariables significantly (`p<0.001`) higher but still to be tested with growing the corpus.