

# presupposition & 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.<sup>1</sup>

## background

Inspired by Zimmerer et al. (2017) we are interested in observations concerning coherence and presupposing 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 antecedents of conditions b-f whether be tagged “DET” or not. This distance should give us information structural evidence of how strong these noun occurrences 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).

---

<sup>1</sup>snc.1:h2.pb.1000char/pg.queries.cites

## questions

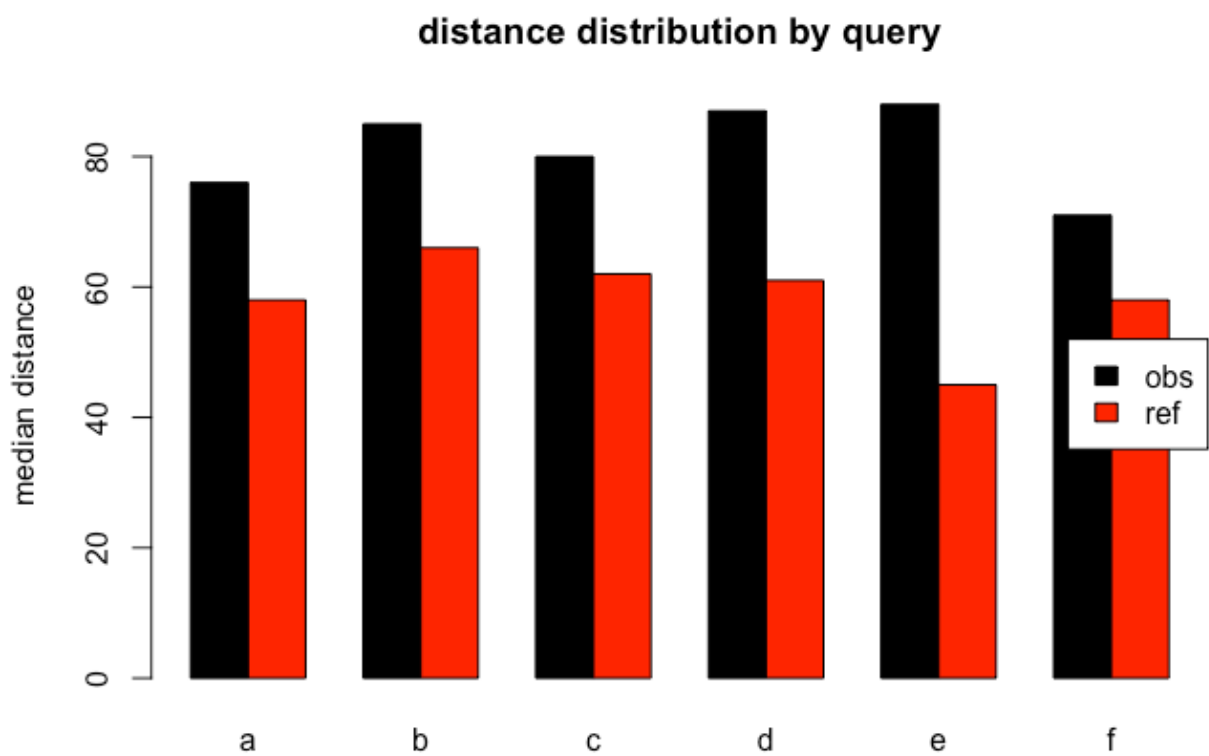
Measuring the referent-reference distance which we here assume as indicator of coherence we hope to find empirical evidence for disturbed or not world building capabilities within schizophrenia language. Premising that a large noun distance indicates a low reference-referent association we hypothesise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

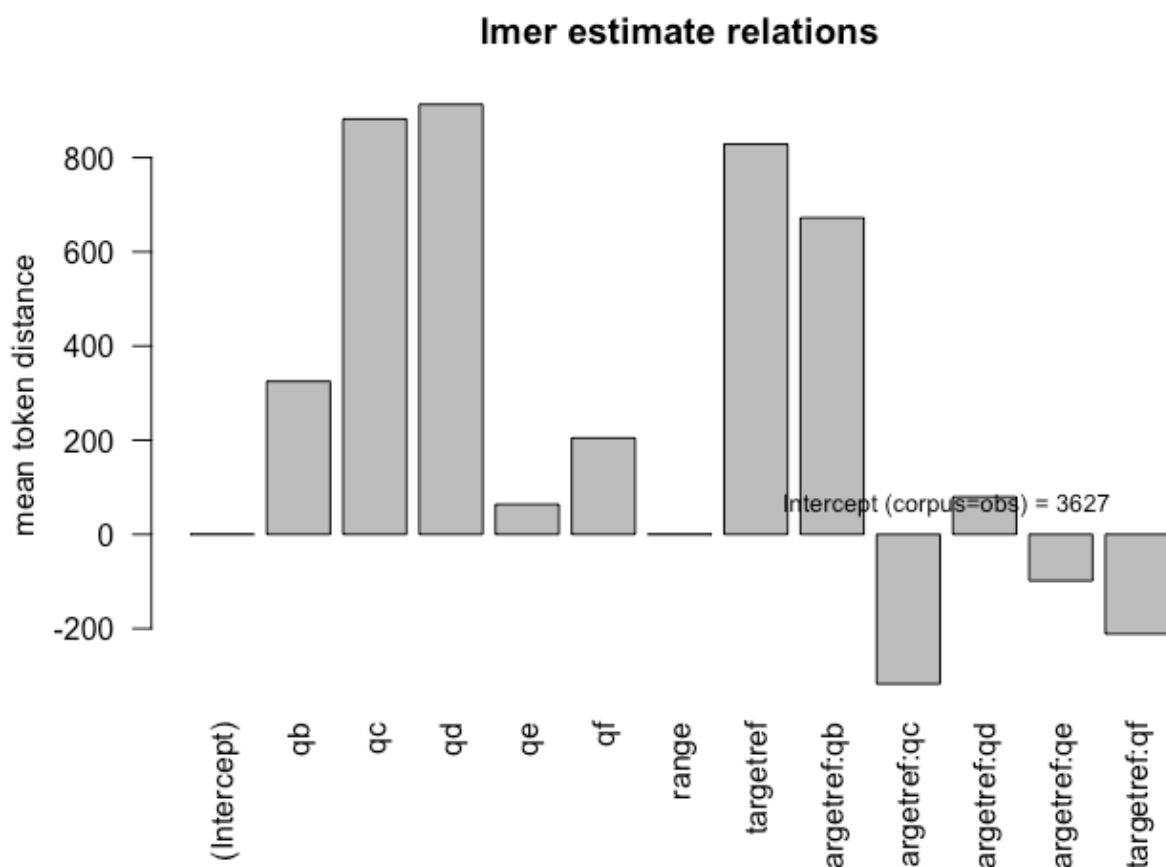
q	target	url	lemma	m	range	dist	det	pos
a	ref	36	pyramid	75	8785	66	FALSE	96640
a	obs	413	disability	19	2963	46	FALSE	227780
a	ref	8	shift	35	3119	202	TRUE	18327
b	ref	98	talent	27	2479	111	FALSE	263578
a	ref	10	people	39	4885	71	FALSE	20807
a	obs	898	fuck	141	2974	37	FALSE	663038
c	obs	697	memory	27	1345	111	FALSE	480047
a	ref	53	console	38	8169	78	FALSE	152972
a	obs	607	people	24	6036	390	FALSE	383650
a	ref	85	other	6	2468	79	FALSE	230059

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 and determiner (`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.