# 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.[1]

## 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 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

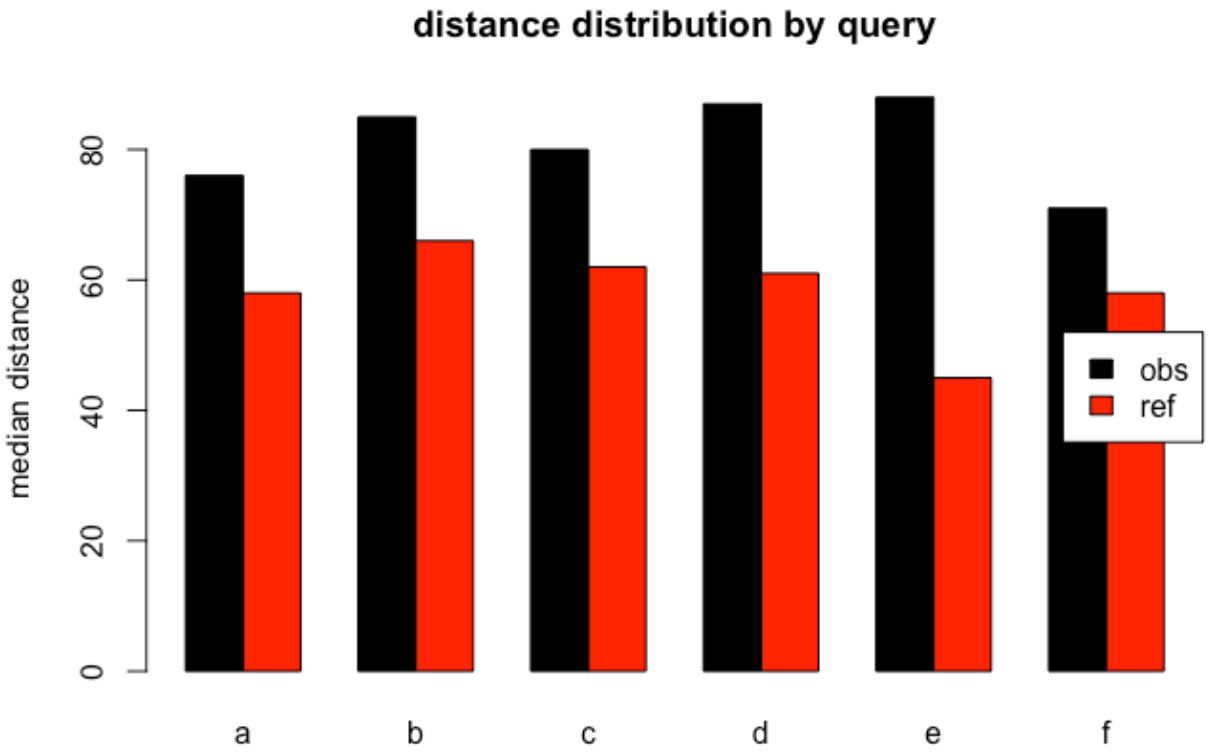
Measuring the referent-reference distance which we here assume as indicator of coherence we hope to find empirical evidence for disturbed or not world building capabilities within schizophrenia language. Premising that a large noun distance indicates a low reference-referent association we hypothesise 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.

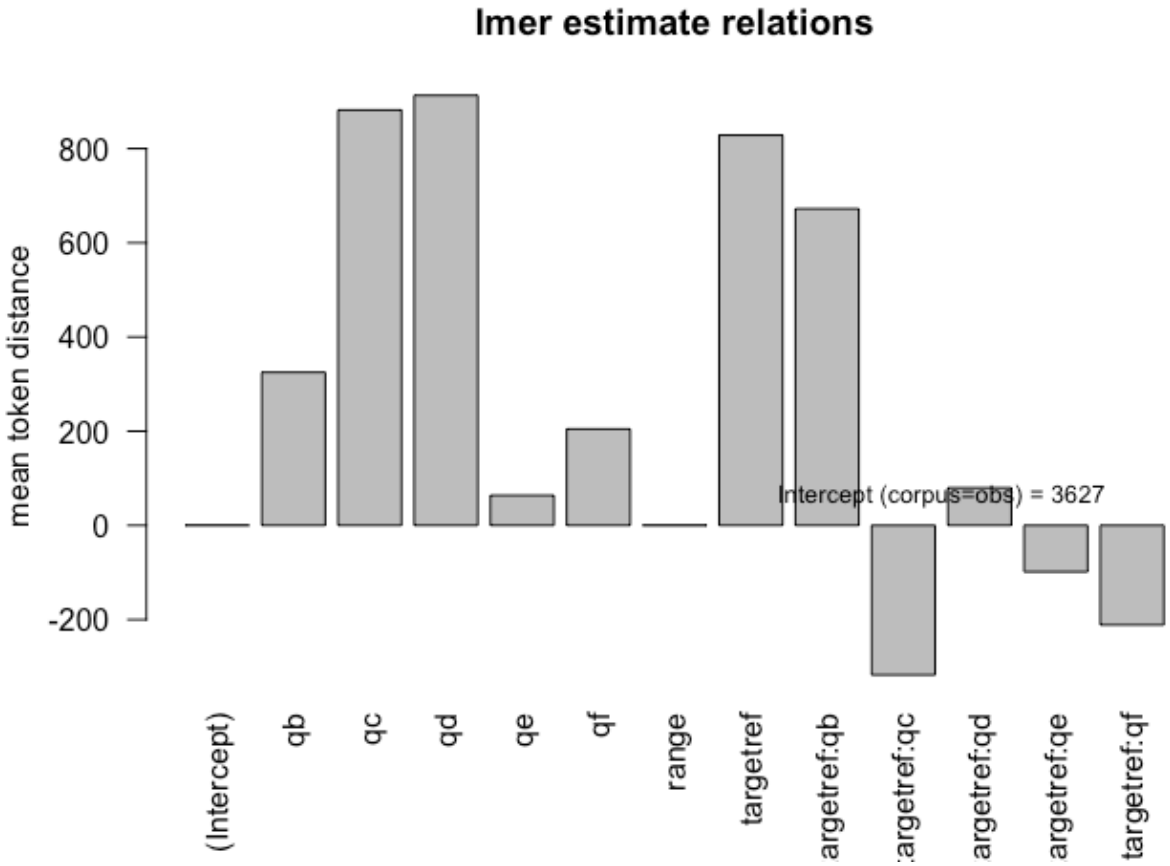
| q | target | url | lemma | m | range | dist | det | pos |
|---|---|---|---|---|---|---|---|---|
| a | obs | 676 | think | 31 | 5411 | 167 | FALSE | 459436 |
| b | ref | 92 | game | 80 | 5392 | 72 | FALSE | 247420 |
| a | ref | 34 | ad | 36 | 1519 | 195 | FALSE | 91608 |
| a | obs | 887 | symptom | 52 | 8308 | 29 | FALSE | 641962 |
| a | obs | 676 | schizophreni-form | 2 | 5411 | 320 | FALSE | 459627 |
| a | ref | 12 | series | 22 | 5221 | 509 | TRUE | 32797 |
| a | obs | 618 | people | 14 | 451 | 13 | FALSE | 398025 |
| a | obs | 932 | post | 18 | 3756 | 170 | FALSE | 697157 |
| a | obs | 897 | people | 78 | 5685 | 199 | FALSE | 659168 |
| a | obs | 337 | be | 91 | 1570 | 9 | FALSE | 174466 |

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.