

xtitle: coherence & propositions observations in
:schizophrenia: threads

st. schwarz

2025-10-16

index

Hausarbeit im Seminar: *Sprache und Psychose (PHILGEIST_S_16827_25S)*
Dozent: Anatol Stefanowitsch
abgegeben von: placeholder
MtrNr: placeholder
abgegeben am: 2025-10-16
an: Freie Universität Berlin

Selbständigkeitserklärung

Hiermit versichere ich,

- dass ich die von mir vorgelegte Arbeit mit dem Titel

coherence & proposition observations in :schizophrenia: threads

selbständig abgefasst habe und

- dass ich keine weiteren Hilfsmittel verwendet habe als diejenigen, die im Vorfeld explizit zugelassen und von mir angegeben wurden und
- dass ich die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken (dazu zählen auch Internetquellen und KI-basierte Tools) entnommen sind, unter Angabe der Quelle kenntlich gemacht habe und
- dass ich, sollte ich explizit im Vorfeld zugelassene Hilfsmittel (z.B. KI-Tools) verwendet haben, diese in einer tabellarischen Übersicht (@ref(tab:kitable)) in meiner schriftlichen Arbeit integriert habe und
- dass ich die vorliegende Arbeit noch nicht für andere Prüfungen eingereicht habe.

Mir ist außerdem bewusst,

- dass ich diese Prüfung nicht bestanden habe, wenn ich die mir bekannte Frist für die Einreichung meiner schriftlichen Arbeit versäume und
- dass ich im Falle eines Täuschungsversuchs diese Prüfung nicht bestanden habe und
- dass ich im Falle eines schwerwiegenden Täuschungsversuchs ggf. die Gesamtprüfung endgültig nicht bestanden habe und in diesem Studiengang

bzw. Studienangebot nicht mehr weiter studieren darf und

- dass ich, sofern ich zur Erstellung dieser Arbeit KI-basierte Tools verwendet habe, die Verantwortung für durch die KI generierte eventuell fehlerhafte oder verzerrte (bias) Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate trage.

Ich bestätige mit meiner Unterschrift die Richtigkeit dieser Angaben.

2025-10-16,

```
#dataset<-7  
#prelim
```


15303.ha.draft

subject

In this paper we want to explore **reference marking, coherence and information structure in schizophrenia language** by measuring distance of similar nouns preceded by specified determinants.¹

Inspired by Zimmerer et al. (2017) we are interested in observations concerning coherence and propositional statement conditions in schizophrenia language. Nenchev et al. (2024) consider investigating in coherence as important “linguistic aspect in psychotic language addressing the relatedness between word chunks or sentences” an approach to “capture formal thought disorders (disorganisation, tangentiality, derailment and poverty of speech) in schizophrenia” (cf. Nenchev et al. (2024), 2.1: linguistic markers of schizophrenia).

This linguistic marker seems to play a crucial role within target group language features. (As such seen as asset of thinking- or world building capacity which might deviate from standard within the range of negative symptoms.) With our approach we add to the research done concerning frequency based analyses of how typical patients language might appear and how that language deviates in terms of keywords or word fields, while our interest is more dedicated to the structural layer of the language which can not be described by raw frequencies. In our opinion disturbances on that layer might be hidden and not to grasp easily such that a listener would not always be able to precisely figure out what the disturbing factor is. Missing **coherence**, which we will investigate, may be a too narrow explanation to many impressions that schizophrene language leaves the listener with. But it seems to be a good starting point to unveiling structural patterns of patients language.

¹which can be considered as a control condition as it should naturally allow wider distances between the following noun and the reference than all other conditions.

definitions, terminology, assumptions

coherence

There are several preliminary affordances to a successful communication. One is the *coherence* of a text = way of communication, which accounts for the partner being able to follow the topic and relate subjects and objects referenced. There can be more or less *common* references and such, that need to be embedded in context to be understood. The underlying network of informations to create that context is what we call *information structure* of a text. The level of complexity of that network defines how simple it would be to gather the reference from the given information. We might have to go back many sentences or even infer reference from metaphors or such to be able to understand what is said while in the other case simply recall the subject of the last sentence to get the meaning (reference) of the pronoun in **also {she} said thisandthat....** The capacity to imagine or have in mind, what concrete information is accessible to the addressee (what he actually knows or can infer) is key to a successful communication, since factors like common-ness, weltwissen and shared knowledge between addressant and addressee and informations accessible from the text itself vary depending on topic, setting, intimacy of the partners and such. So one cannot always be sure that the information provided is sufficient but the grade to which one can give a correct estimate to this sufficiency should here be the measure assuming that coherence in disturbed language is deficient which lets an utterance be more difficult to understand within the frame of given information. Now one indicator of coherence we assume is *reference distance* where according to our hypothesis a larger distance would be observed in places where the addressant overestimates² the ability of the partner to follow a reference. That would mean that we find a medium shorter distance between referent and reference in the reference corpus³ and larger distances in the target corpus. The references we are interested in are nouns that appear as anaphors i.e. here as noun analogies. The assumption is that if a noun is repeated *and* is combined with certain preceding determiners, the speaker assumes that the addressee has some knowledge of what is talked about, depending on the strength of the determination. So e.g. this, that, those, these would be rather strong determiners requiring that the noun was introduced before.

premises

deictic anchoring and propositional complexity

Zimmerer et al. (2017) consider “Deictic anchoring [...] an inherent part of the process by which we make references to aspects in the world including entities, events, locations, and time.” and define propositions as being “statements about the world which can be true or false.” They mention, according to (Kuperberg

²this is as of model 1-6 not yet integrated in the analysis

³only according to the LLM training data, which is still a blackbox

2010) “that in people with schizophrenia, cortical activity to semantic abnormalities in sentences is particularly small compared to controls if interpretation requires integration of several sentences” which can mean, that patients are not realising if their utterances are somehow disturbed on the semantics level. If “Delusions and thought disorder can be considered disruptions of propositional meaning” then the patients feeling for their stated propositions (required to the addressee) and further the estimation about what he/she can assume as familiar to the addressee can be wrong. Following Klaus Konrad (Mishara 2010) who “described the onset of a delusion as the loss of ability to transcend an experience and see it with the eyes of others” Zimmerer et al. (2017) assume that “in thought disorder, the ability to express coherent propositions can be severely impaired.” We take that as premise for our research question.

questions

Measuring the referent-reference distance which we assume as an indicator for coherence we hope to find empirical evidence for disturbed or not world building capacities 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. An environment which has potential to test our hypothesis is a reddit.com subreddit where the majority of participants describe themselves as being diagnosed with schizophrenia.⁴ As reference corpus we chose reddit r/unpopularopinion. The distance measured should give us information structural evidence of how strong the noun occurrences⁵ are connected, i.e. if a noun appears out of the blue mostly or if it somewhere before has been introduced to the audience and thus would be more or less legitimated to be determined by an antecedent. Our basic assumptions rely on the *taxonomy of given end new information* coined by Prince (1981). She develops a hierarchy of references⁶ with specific relations to each other, where each item is attributed in terms of *familiarity*⁷, that defines ranges of 1. givenness in the sense of predictability/recoverability, 2. givenness in the sense of saliency, 3. givenness in the sense of “shared knowledge”. (cf. Prince (1981), pp. 226) We base our hypothesis of *reference distance as indicator for coherence* on this model assuming that the reference/association strength⁸ determines the level of text coherence.

⁴to spare resources

⁵where “obs” comes first

⁶informations in a text

⁷cf. Prince: speaker assumptions about hearer familiarity = assumed familiarity

⁸which should be weaker with growing distance between reference-referent

data

We built a corpus of the reddit r/schizophrenia thread (**n =1500371 tokens**) and a reference corpus of r/unpopularopinion (**n =980731 tokens**). Both were pos-tagged using the R udfpipe package (Wijffels (2023)) which tags according to the universal dependencies tagset maintained by De Marneffe et al. (2021). Still the available data can only, within the pipeline of steadily growing the corpus and devising the noun distances developed be just a starting point from where with more datapoints statistical evaluation becomes relevant.

The dataframe used for our model (actual: dataset 13) consists of **142321** distance datapoints (sample cf. Tab. @ref(tab:data1) below) derived from the postagged corpus. Because the ranges of the url threads vary heavily between target and reference corpus, the distances are (in evaluation M1) normalised to the target corpus (cf. Fig. @ref(fig:gplot1) for the raw vs. normalised distances comparison.) Outliers are excluded from the analysis since they very probably do not fulfill to can be counted as anaphoric references. We silently assume that all of the noun distances which are not by value excluded as outliers occur as anaphoric references. A manual annotation and close reading of the text would be necessary to exactly determine whether the references are associated at all. This may be the task for another qualitative evaluation of our quantitative study.

Table 1:

token	upos	target	pos	prepos	url_id	range	q	det	aut_id	total_me
mg	NOUN	obs	832763	PUNCT	1299	962	a	0	224	
problem	NOUN	ref	466158	ADP	2036	3001	a	0	6171	
mph	NOUN	ref	489910	NOUN	2045	1454	a	0	3301	
opinion	NOUN	ref	207821	PRON	1948	5392	a	0	82	
difference	NOUN	ref	48044	NOUN	1876	1798	a	0	3856	
schizophrenics	NOUN	obs	164940	ADP	322	1801	a	0	23	
hole	NOUN	ref	697692	VERB	2165	2888	a	0	9151	
flavor	NOUN	ref	170549	PRON	1928	1169	a	0	5025	
day	NOUN	ref	431041	PUNCT	2025	1755	a	0	7203	
morning	NOUN	obs	978838	DET	1503	982	c	1	457	

methods

To compute distances we queried the corpus for matching conditions where certain (probable) determiners appear before analogue nouns (anaphors).

condition	value
a	any !(b,c,d,e,f)
b	this, that, those, these

condition	value
c	the
d	a, any, some
e	my
f	his, her, their, your

We decided for these 5 sets of determiners in order to see whether distances maybe influenced if the duplicated nouns are preceded by them. We would expect condition **b** to show different if not reziproke effects as condition **d**⁹ and yet the texts in the reference corpus show the expected behaviour while in the target corpus not.

For each datapoint we collect variables as:

- thread url
- author (anonymised)
- thread length (tokens)
- lexical diversity (type/token ratio)
- lemma
- distance (to the preceding occurrence, e.g. for three occurrences of dog we collect 2 distance datapoints)

The main function to determine the distances runs on a subset of the corpus with only including all nouns and their position in the corpus. It finds all duplicated nouns per url thread and computes their distances by token position.

reflections

range

Evaluating with a growing corpus we interestingly find our basic hypothesis tested again, showing an overall larger distance of analogue nouns within the range of 1 thread url for the target corpus. While earlier we devised distances from a manually assigned url identifier we saw the necessity to define our “range of interest” according to the original http url of the thread, since with a growing corpus the old url ids - derived from the `get_thread_url()` method of the `redditExtractoR` package (Rivera (2023)) used for fetching the reddit content - there are no new url ids created since one url fetch gets each time always only around 1000 urls. To ensure unique url ranges within the corpus we assigned the range (within which the noun distance is calculated) to the real thread url. The corpus itself is after each fetch sorted after url and timestamp so it represents the real flow of conversation within one thread which is important since our distance model is based on the token distances within that thread and they have

⁹which can be considered as a control condition as it should naturally allow wider distances between the following noun and the reference than all other conditions.

to follow their natural occurrence in time.

The url range is an important variable which we used for normalising the distance values since the mean distances could also depend on the overall thread length. For that we calculated for each normalisation method as are 1. per target, 2. within target and 3. cross target a range factor by which the distance values are divided. The final regression model posits fixed effects of condition, target, determiner, range and embed score (where target, condition and determiner are interacting) and random effects of thread and author.

author trace id

An important integrated feature can be the aut_id variable which represents the comment author and is unique to that. In the base .sqlite database the authors are already anonymised, so there should be no way from the published data back to the original author name of the comment. And as expected, including aut_id as random effect in the linear regression model, the significance level for the covariables of interest as are

1. q = the condition matching of the noun-preceding token
2. det = whether that match has postag “DET”
3. target = obs or reference corpus

finally increases.

lexical diversity

We thought about some serious caveats in modeling the evaluation: If (lucky for our hypothesis) the target corpus has significantly higher distance scores over nearly all conditions, does that automatically indicate a less coherent reference-referent association within what is expressed in the comments? Couldn't we also assume that if the analogue nouns appear more distanced in general that a topic which is including these nouns is simply expanding over a wider range resp. timeframe? What does that mean for our assumptions in terms of coherence? A good way here could be to integrate a general lexical diversity factor per url as fixed effect because we can assume that a higher type/token ratio logically decreases the probability of a noun appearing multiple times within a range and we could take that effect into account.¹⁰

semantics, word field, embeddings

Further we created another covariable possible to integrate in the evaluation model: The semantic embedding of one specific noun appearing on its specific position in the thread range, computed with help of an open LL word embedding model (Nussbaum et al. (2024).) This is a common AI way of devising semantic relations in a corpus which exceeds a just frequency based keyword analysis. Using an LLM here allows for a distinctive identification of word field embeddings

¹⁰this is as of model 1-6 not yet integrated in the analysis

of the noun in question. In that way we get another variable linguistic feature extracted which may give general insights into the level of standardisation that applies to the corpora. So if a noun is found to be embedded with a high score into its context (the url thread) then it can be much expected to be found there and appears less out-of-context.¹¹

statistics

In this context we thought about what it means statistically, if a high-score embedded word also ranks high in (distance) significance i.e. generally what the relations of the covariates in the context of the linear regression evaluation express. Let us picture this:

1. a word receives a high embed score if it is highly semantically related to the context within which it appears, here the comment thread.
2. therefore the necessity to introduce/elaborate on it sinks, since it may be considered a “known” or “inferable” entity within the context given.
3. now if a person is using this word, the determined use appears less incoherent by itself.
4. the reference distance thus may increase without losing in coherence.
5. **conclusion:** if we for our linear regression use a (base) formula like `distance ~ corpus`, a continuous `embed_score` predictor between -1 and 1 should correlate positive with the estimates for `dist` if applied correctly?

caveats

Since devising the word embed score does take much computing resources we had a script run on a server that solves the computing. But the first essay to integrate the new var into the evaluation model failed due to levels < 2. Why? Because in the beginning we ran the script just over a few chunks of the complete url ranges in the corpus¹² which is sorted after target,¹³ we did not compute any values for the reference corpus. So we learned this way again on linear regression models which require that a variable has more than one level (which would not be the case if the `lmer()` function excludes all NA rows: there simply would be no observations left with target=ref since all its embed.score values are NA (not yet calculated) and so all target.ref rows will be removed during regression.) The issue is solved since we found a resource saving method of computing the embed scores with a local instance of ollama that provides an API to use the model.

¹¹only according to the LLM training data, which is still a blackbox

¹²to spare resources

¹³where “obs” comes first

model evaluations

covariances

Effects of the same direction for target OBS and REF are observed in `qc`, `range` (with positive effects in `qc`) while contrary effects are observed in `qb`, `qd`, `qe`, `qf`, `det`, `embed.score`, `qb:det`, `qd:det` (with negative effects in target=obs and vcvs.)

In words:

- the antecedents **the** seem to allow a wider distance between referent and reference in both target=OBS and target=REF.
- the antecedents **this, that, these, those** - **my** - **your, their, his, her** decrease distance in target=OBS and increase distance values in target=REF; condition d (**a, an, some, any**) vcvs.
- higher `embed.score` values (better embedded noun) decrease distance in target=OBS and increase distance values in target=REF. (cf. par 3.7.5.4, better embedding allows wider distance > the expectation seems only valid for the reference corpus!)

sidenote: Positing the url range only as fixed effect instead of normalising the distances still estimates smaller distances for the reference corpus, but with no significance, the only significant difference with that regression formula shows in target=REF under condition e (antecedents: **my**).

model fazit

As you can cf. in the appendix with the separate coefficient tables for each evaluation model, we find over all normalised subsets (vs. obs/ref/all) significantly smaller distances in the reference corpus with varying effects for the conditions. In the subsets, where we didn't normalise or remove outliers, we find the opposite effect; the raw data does not prove our hypothesis. But just looking into the (raw) mean values plot of Fig. @ref(fig:barplot-mean2) we clearly see that normalising and removing outliers is necessary since mean distances there extend up to over 2000 tokens thus we wouldn't like to count all analogue noun occurrences here as anaphora.

conclusion

After evaluating over the different approaches we find our hypothesis proved, that anaphora distances in the target corpus (target=OBS) stretch over a significantly ($p < 0.001$) wider range of tokens between reference and referent in contrast to the chosen reference corpus. With our assumptions this could prove a less appropriate estimate for the coherence of the own texts produced in schizophrenic language still having in mind, that a wider distance is not stating incoherence in general but instead just that these speakers allow for a wider anaphora distance

in their text production. If these distances indeed lead to less coherent texts compared to the reference corpus must be subject to close reading and annotating samples manually and questioning them in terms of coherence by skilled readers though annotation may vary strongly depending on the disposition of readers and their general capacities of inferring references. But if we agree that shorter reference distances increase text coherence then we might say the texts produced in the target corpus are less coherent than those in the reference corpus which aligns with the common classification of patients language in psychiatry.

limitations

We had to do some silent assumptions, but the main limitation is that we will have to base our specification of the target corpus as being one that is containing schizophrene language mainly on the statements of the reddit users in our target corpus which do describe themselves as being diagnosed schizophrene to a large amount. To what extend these statements and assignments or identifications are true we cannot say and therefore limit the value of our findings only to that group of speakers.

appendix

overall wordcount of paper: 3102.

legende

Table 3: model vars

variable	explanation	values
target	corpus	obs,ref
q	condition	a,b,c,d,e,f
det	antecedent POS==DET	TRUE,FALSE
aut_id	author	author hash
lemma	lemma	noun lemma
range	url range of distance devised	1..maxlength(urlthread)
embed.score	semantic similarity score lemma vs. thread	0..1
q:a	query condition	.*
q:b	query condition	this,that,those,these
q:c	query condition	the
q:d	query condition	a,an,any,some
q:e	query condition	my
q:f	query condition	his,her,their,your

fixed effects in M1

```
## 10 x 19 Matrix of class "dgeMatrix"
##           (Intercept)  targetref          qb          qc          qd
## qb      -0.656706044  1.207097617  2.331532e+02  4.972520e-02 -1.123020e-01
## qc          0.218792760  0.035239530  4.972520e-02  3.023656e+01  2.139728e+01
## qd          0.213024201 -0.057995645 -1.123020e-01  2.139728e+01  4.951568e+04
## qe         -1.181562271  1.364128579  1.204280e+00  1.804251e-02 -3.122280e-02
## qf         -0.999266163  1.466700457  1.321795e+00  7.762191e-02 -1.314356e-03
```

```

## det      -1.792510882  1.297436814  1.285782e+00 -2.140515e+01 -2.147507e+01
## range    -0.003430091 -0.004456289 -5.684751e-05 -1.771291e-06 -3.063886e-05
## embed.score -0.304346598  0.003347917 -9.761775e-03 -5.578520e-03 -5.250995e-03
## qb:det    0.848433419 -0.992845655 -2.331005e+02  2.141809e+01  2.160561e+01
## qd:det    -0.311056815  0.114084229 -1.000347e-02 -4.340561e-03 -4.948523e+04
##           qe           qf           det           range
## qb      1.204280e+00  1.321795e+00  1.285782e+00 -5.684751e-05
## qc      1.804251e-02  7.762191e-02 -2.140515e+01 -1.771291e-06
## qd     -3.122280e-02 -1.314356e-03 -2.147507e+01 -3.063886e-05
## qe      1.471083e+01  1.256452e+00  1.298793e+00  1.601799e-05
## qf      1.256452e+00  2.393642e+01  1.298587e+00  2.964445e-05
## det      1.298793e+00  1.298587e+00  2.279447e+01  1.314184e-05
## range    1.601799e-05  2.964445e-05  1.314184e-05  3.267330e-06
## embed.score -5.662857e-03 -1.123025e-02  1.088365e-02  9.235581e-06
## qb:det    -1.229842e+00 -1.099137e+00 -2.275232e+01  7.012525e-05
## qd:det    -6.757926e-03 -7.304582e-03  1.082133e-02  3.758041e-05
##           embed.score targetref:qb targetref:qc targetref:qd
## qb     -9.761775e-03 -2.331546e+02 -5.243550e-02  1.179001e-01
## qc     -5.578520e-03 -4.868618e-02 -3.024022e+01 -2.139480e+01
## qd     -5.250995e-03  2.337161e+00 -1.144880e+02 -1.572741e+02
## qe     -5.662857e-03 -1.204734e+00 -2.043544e-02  3.386100e-02
## qf     -1.123025e-02 -1.320686e+00 -8.345716e-02  6.049089e-03
## det      1.088365e-02 -1.286510e+00  2.141138e+01  2.146985e+01
## range     9.235581e-06  5.866084e-05  7.093443e-06 -1.261267e-05
## embed.score  6.497771e-03  9.771409e-03  8.618392e-03  2.039760e-03
## qb:det    1.842562e-03  2.331055e+02 -2.142179e+01 -2.160838e+01
## qd:det    6.309983e-03 -2.213302e+00  9.309397e+01  1.268228e+02
##           targetref:qe targetref:qf targetref:det qb:det
## qb     -1.203769e+00 -1.319105e+00 -1.286549e+00 -2.331005e+02
## qc     -1.838955e-02 -7.419652e-02  2.140615e+01  2.141809e+01
## qd      4.821807e-01  5.162239e-01  1.153981e+02  2.160561e+01
## qe     -1.470978e+01 -1.257490e+00 -1.298512e+00 -1.229842e+00
## qf     -1.257873e+00 -2.392701e+01 -1.297449e+00 -1.099137e+00
## det     -1.298964e+00 -1.301955e+00 -2.279625e+01 -2.275232e+01
## range    -9.407763e-06 -4.502403e-05 -1.113598e-05  7.012525e-05
## embed.score  5.516501e-03  9.146142e-03 -1.136072e-02  1.842562e-03
## qb:det    1.225973e+00  1.110890e+00  2.275511e+01  2.814856e+02
## qd:det    -4.451085e-01 -5.050676e-01 -9.393291e+01  9.338525e-03
##           qd:det targetref:qb:det
## qb     -1.000347e-02  2.331077e+02
## qc     -4.340561e-03 -2.141755e+01
## qd     -4.948523e+04 -1.180455e+02
## qe     -6.757926e-03  1.233269e+00
## qf     -7.304582e-03  1.101642e+00
## det      1.082133e-02  2.274935e+01
## range     3.758041e-05 -7.935980e-05

```

```
## embed.score 6.309983e-03 -4.612123e-03
## qb:det      9.338525e-03 -2.814960e+02
## qd:det      4.948523e+04  9.642721e+01
```

evaluation model: 1

meta

eval output data: 13, normalised to obs, distance ceiling = outliers removed

parameter setting

```
##          value
## norm_target _rel_obs
## det.t       TRUE
## limit       TRUE
## author      TRUE
## url         TRUE
## embed1      TRUE
## embed2      f
## range1      TRUE
## range2      f
## rel         TRUE
## lme         FALSE
## lemma       FALSE
```

anova analysis

anova plain

formula: [dist_rel_obs ~ target*q*det]

```
##          Df      Sum Sq   Mean Sq    F value    Pr(>F)
## target      1  452303747 452303747 7336.4625 < 2.2e-16 ***
## q           5   12320667   2464133   39.9688 < 2.2e-16 ***
## det         1    1636109   1636109   26.5380 2.588e-07 ***
## target:q     5    2747371    549474    8.9126 1.786e-08 ***
## target:det   1     251297    251297    4.0761 0.043496 *
## q:det        2     905292    452646    7.3420 0.000648 ***
## target:q:det 1      717222    717222   11.6335 0.000648 ***
## Residuals 126209 7780971239    61651
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

anova of linear regression model

```
[anova(summary(lmer))]
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##              Sum Sq  Mean Sq NumDF  DenDF    F value    Pr(>F)
## target          1144219   1144219      1    3519    23.4567  1.333e-06 ***
## q                737483   147497      5   122421     3.0237  0.0098706 **
## det              12165     12165      1   118425     0.2494  0.6175055
## range          50399647  50399647      1    1025  1033.2042 < 2.2e-16 ***
## embed.score    25101881  25101881      1   122690   514.5942 < 2.2e-16 ***
## target:q        776335   155267      5   123486     3.1830  0.0070933 **
## target:det      541078   541078      1   123325    11.0922  0.0008672 ***
## q:det           359520   179760      2   120804     3.6851  0.0250971 *
## target:q:det    219844   219844      1   123315     4.5068  0.0337615 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

linear regression coefficients

formula: [dist_rel_obs ~ target*q*det+(1|aut_id)+range+(embed.score)+(1|url_id)]

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: eval(expr(lmeform))
## Data: dfa
##
## REML criterion at convergence: 1727648
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.8643 -0.5282 -0.1721  0.2469  6.9244
##
## Random effects:
## Groups   Name                Variance Std.Dev.
## aut_id   (Intercept)         2856     53.44
## url_id   (Intercept)         8187     90.48
## Residual                    48780    220.86
## Number of obs: 126226, groups: aut_id, 8238; url_id, 2145
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   4.625e+02  5.159e+00  8.969e+03  89.651 < 2e-16 ***
## targetref    -4.342e+01  6.299e+00  1.300e+03  -6.893 8.50e-12 ***
## qb           -2.013e+01  1.527e+01  1.218e+05  -1.318 0.187483
## qc           -2.207e+01  5.499e+00  1.226e+05  -4.014 5.98e-05 ***
## qd           -3.178e+01  2.225e+02  1.184e+05  -0.143 0.886426
## qe            2.492e+01  3.835e+00  1.247e+05   6.498 8.14e-11 ***
## qf           -1.891e+01  4.892e+00  1.244e+05  -3.866 0.000111 ***
## det            1.273e+01  4.774e+00  1.229e+05   2.667 0.007662 **
```

```
## range          -5.810e-02  1.808e-03  1.025e+03 -32.143 < 2e-16 ***
## embed.score    -1.829e+00  8.061e-02  1.227e+05 -22.685 < 2e-16 ***
## targetref:qb    1.862e+01  1.719e+01  1.225e+05   1.083 0.278599
## targetref:qc    2.281e+01  1.279e+01  1.237e+05   1.784 0.074435 .
## targetref:qd    4.413e-01  1.254e+01  1.238e+05   0.035 0.971935
## targetref:qe   -2.321e+01  9.511e+00  1.239e+05  -2.441 0.014662 *
## targetref:qf    1.801e+01  1.210e+01  1.238e+05   1.488 0.136766
## targetref:det   -1.478e+01  1.084e+01  1.239e+05  -1.363 0.172784
## qb:det          5.915e+01  1.678e+01  1.219e+05   3.526 0.000423 ***
## qd:det          3.648e+01  2.225e+02  1.184e+05   0.164 0.869736
## targetref:qb:det -5.198e+01  2.448e+01  1.233e+05  -2.123 0.033761 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 7 columns / coefficients
## Some predictor variables are on very different scales: consider rescaling
```

plots

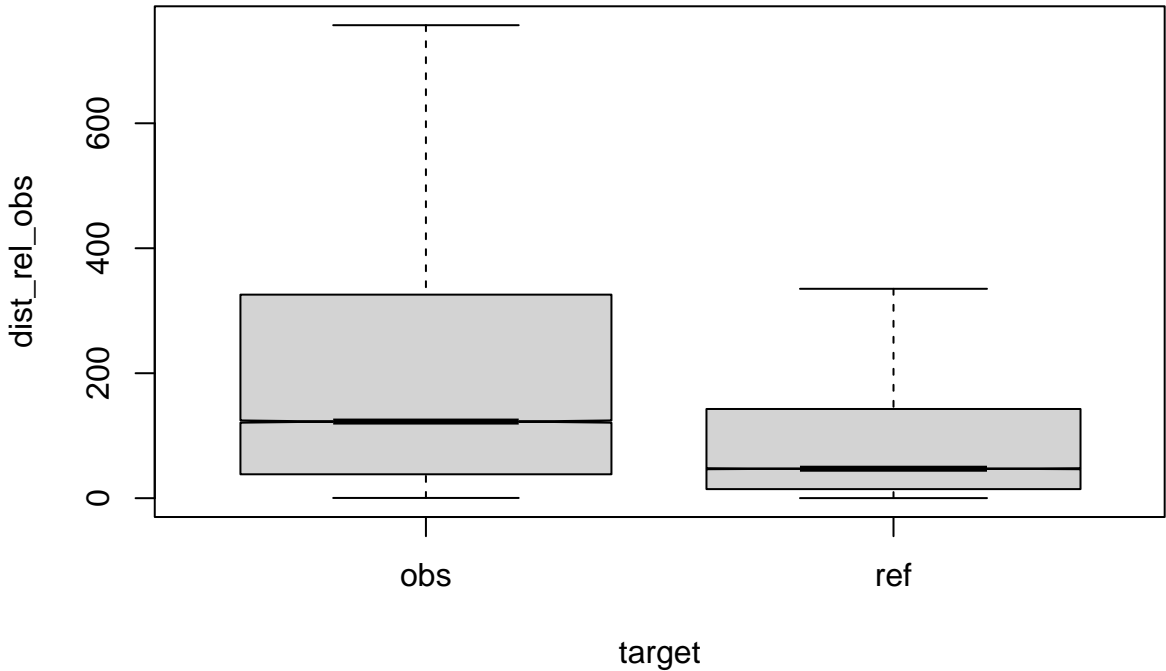


Figure 1: compare distances by corpus, normalised to obs, distance ceiling = outliers removed

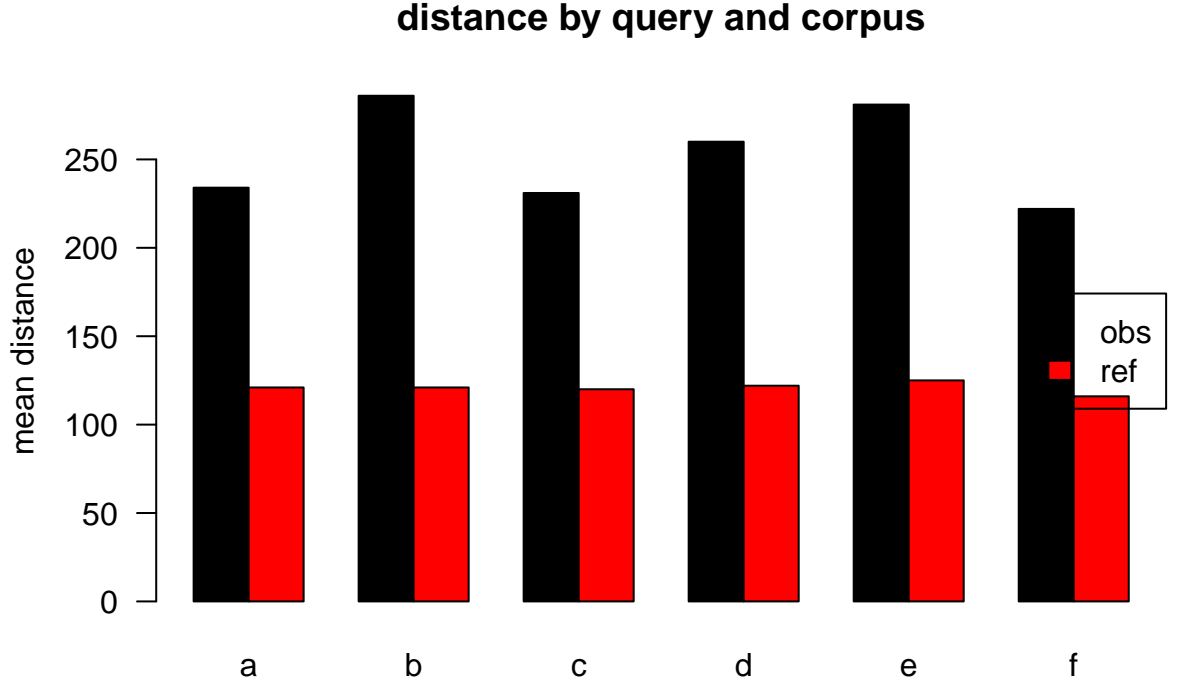


Figure 2: mean distances over query/corpus, normalised to obs, distance ceiling = outliers removed

Table 4: mean/median table for model: 1

target	q	n	mean	median
obs	a	42836	234	117
ref	a	58615	121	47
obs	b	2116	286	165
ref	b	1130	121	44
obs	c	5770	231	114
ref	c	1274	120	48
obs	d	5654	260	144
ref	d	1525	122	49
obs	e	3911	281	147
ref	e	671	125	45
obs	f	2311	222	133
ref	f	413	116	47

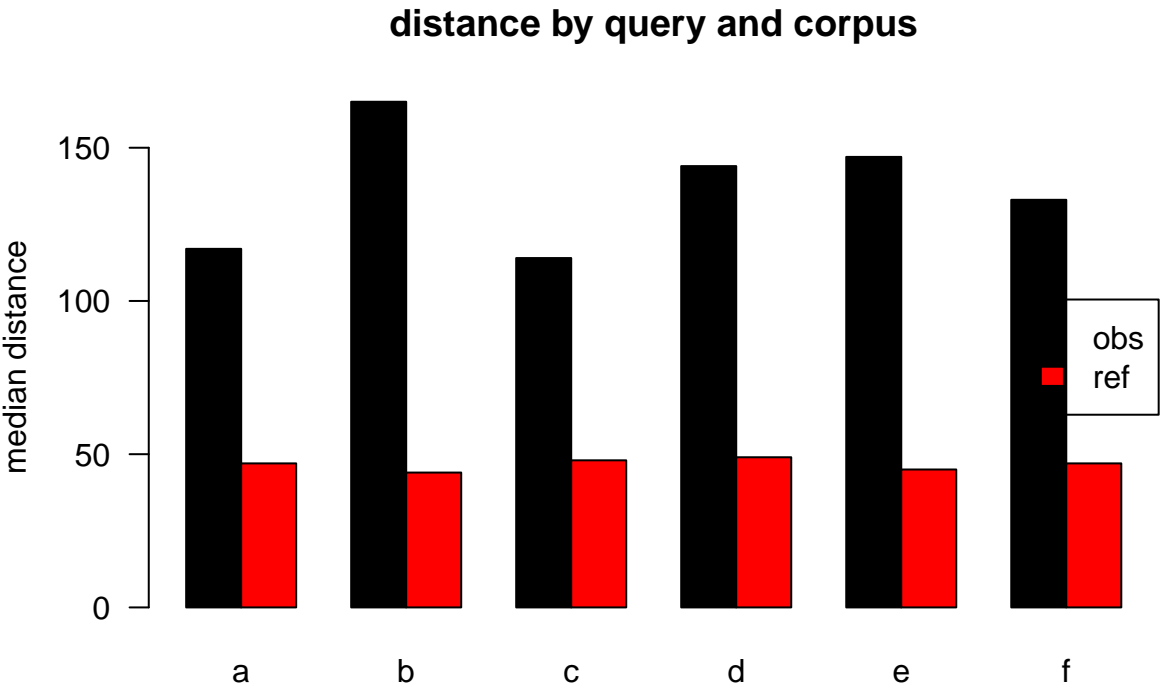


Figure 3: median distances over query/corpus, normalised to obs, distance ceiling = outliers removed

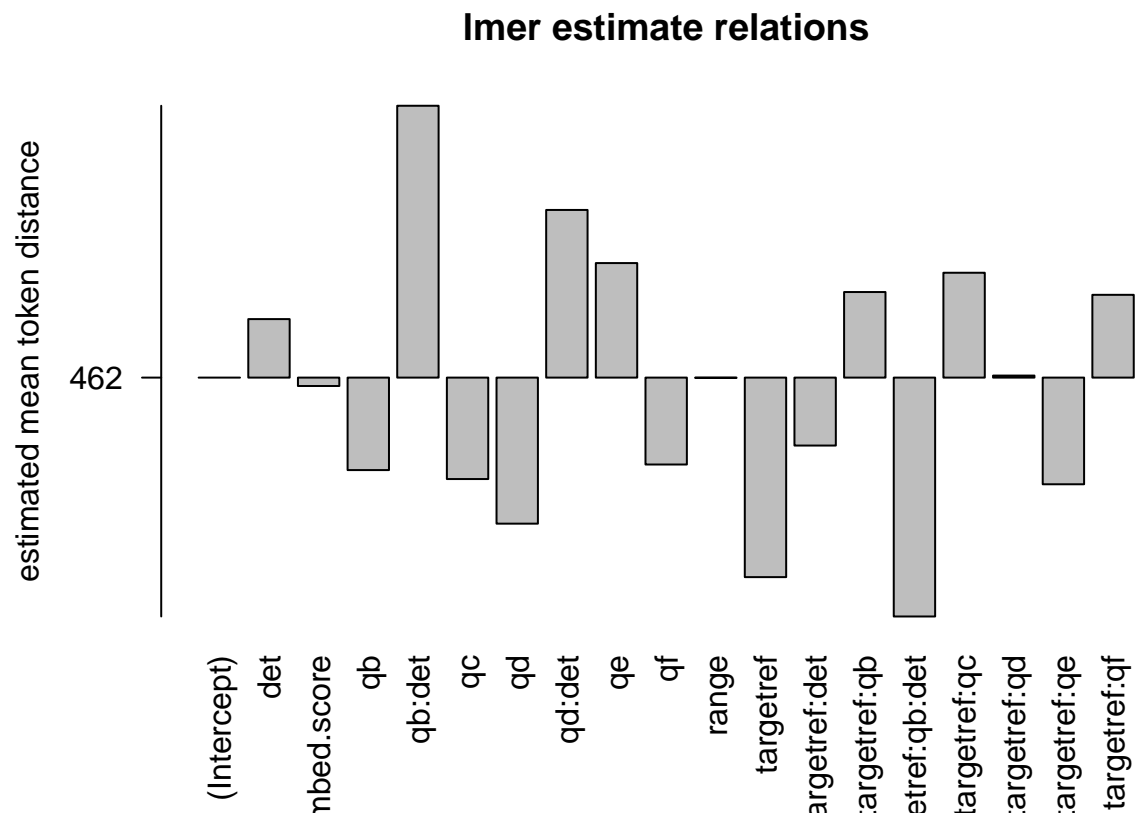


Figure 4: distances relation, normalised to obs, distance ceiling = outliers removed

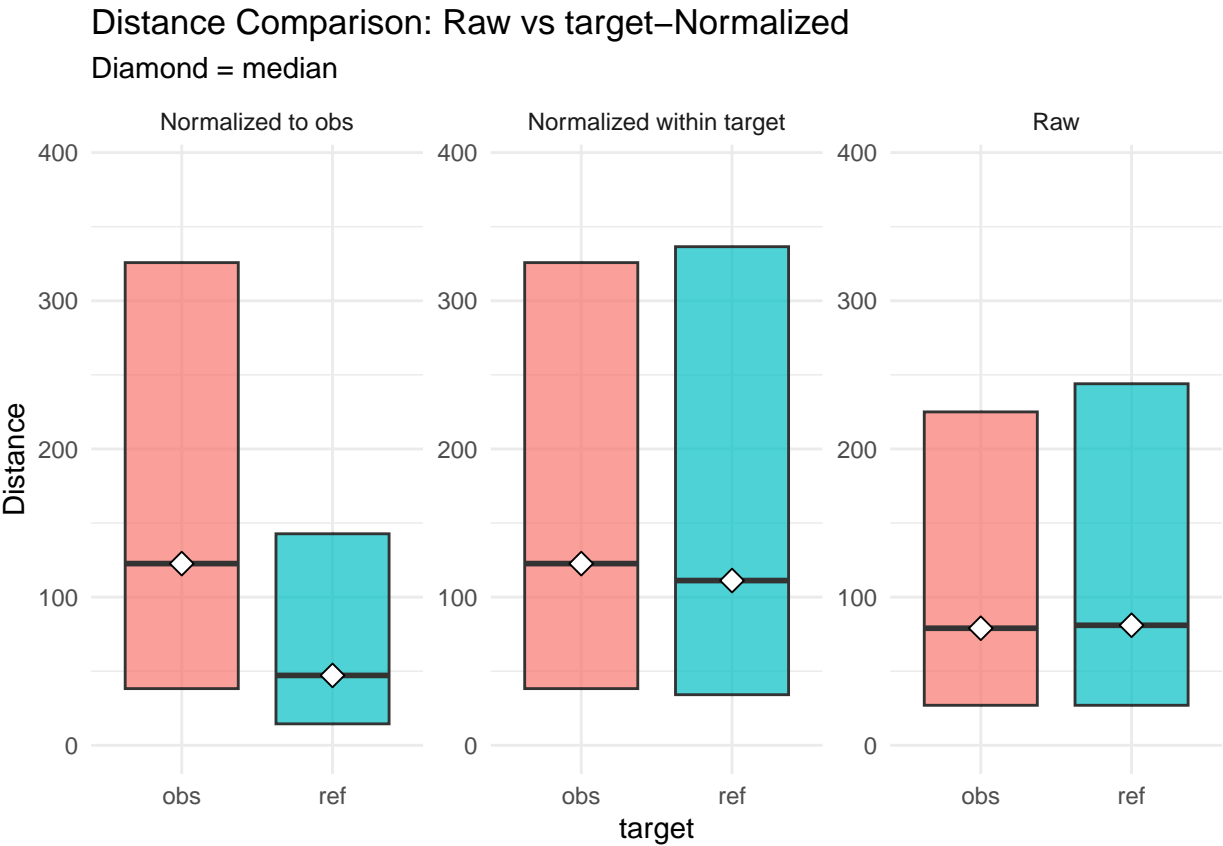


Figure 5: distances normalised vs. raw

evaluation model: 2

meta

eval output data: 13, not normalised, distance ceiling =outliers not removed

parameter setting

```
##          value
## norm_target
## det.t     TRUE
## limit     FALSE
## author    TRUE
## url       TRUE
## embed1    TRUE
## embed2    f
```

```
## range1      TRUE
## range2      f
## rel         FALSE
## lme         FALSE
## lemma       FALSE
```

anova analysis

anova plain

```
formula: [dist ~ target*q*det]
```

```
##              Df      Sum Sq    Mean Sq  F value  Pr(>F)
## target          1 1.1152e+11 1.1152e+11 268.8154 < 2e-16 ***
## q                5 9.8792e+08 1.9758e+08  0.4763 0.79425
## det              1 4.1537e+08 4.1537e+08  1.0012 0.31702
## target:q         5 2.3050e+09 4.6101e+08  1.1112 0.35184
## target:det        1 2.7199e+09 2.7199e+09  6.5561 0.01045 *
## q:det             2 2.4028e+08 1.2014e+08  0.2896 0.74857
## target:q:det      1 7.0024e+06 7.0024e+06  0.0169 0.89663
## Residuals    142304 5.9037e+13 4.1487e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

anova of linear regression model

```
[anova(summary(lmer))]
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##              Sum Sq    Mean Sq NumDF  DenDF  F value  Pr(>F)
## target      1.2717e+09 1.2717e+09     1   3751   5.5781 0.01824 *
## q            6.3534e+08 1.2707e+08     5 137654   0.5574 0.73281
## det          7.3359e+05 7.3359e+05     1 133172   0.0032 0.95476
## range        2.8637e+07 2.8637e+07     1   2113   0.1256 0.72307
## embed.score  2.7199e+10 2.7199e+10     1 141732 119.3005 < 2e-16 ***
## target:q      3.0753e+09 6.1507e+08     5 138840   2.6979 0.01920 *
## target:det    8.1028e+08 8.1028e+08     1 138434   3.5541 0.05940 .
## q:det         4.8717e+08 2.4358e+08     2 135770   1.0684 0.34355
## target:q:det  2.4585e+06 2.4585e+06     1 138496   0.0108 0.91729
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

linear regression coefficients

```
formula: [dist ~ target*q*det+(1|aut_id)+range+(embed.score)+(1|url_id)]
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
```

```

## Formula: eval(expr(lmeform))
## Data: dfa
##
## REML criterion at convergence: 3153653
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -23.760  -0.034  -0.006   0.025  55.672
##
## Random effects:
## Groups Name Variance Std.Dev.
## aut_id (Intercept) 28986087 5384
## url_id (Intercept) 98382082 9919
## Residual 227983587 15099
## Number of obs: 142321, groups: aut_id, 8395; url_id, 2145
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  2.873e+03  4.211e+02  8.594e+03  6.823 9.53e-12 ***
## targetref    1.341e+03  6.536e+02  2.412e+03  2.051  0.0404 *
## qb           6.895e+01  1.008e+03  1.363e+05  0.068  0.9454
## qc          -6.307e+02  3.622e+02  1.372e+05 -1.741  0.0816 .
## qd          -1.993e+03  1.522e+04  1.332e+05 -0.131  0.8958
## qe          -1.006e+02  2.520e+02  1.385e+05 -0.399  0.6899
## qf          -1.355e+02  3.218e+02  1.384e+05 -0.421  0.6737
## det           7.031e+02  3.145e+02  1.375e+05  2.236  0.0254 *
## range         6.798e-02  1.918e-01  2.113e+03  0.354  0.7231
## embed.score  -5.793e+01  5.304e+00  1.417e+05 -10.922 < 2e-16 ***
## targetref:qb  6.675e+02  1.124e+03  1.371e+05  0.594  0.5527
## targetref:qc  3.752e+01  8.128e+02  1.395e+05  0.046  0.9632
## targetref:qd  2.022e+03  7.989e+02  1.395e+05  2.531  0.0114 *
## targetref:qe  2.269e+02  6.042e+02  1.395e+05  0.376  0.7073
## targetref:qf  3.210e+02  7.643e+02  1.393e+05  0.420  0.6745
## targetref:det -1.416e+03  6.890e+02  1.397e+05 -2.055  0.0398 *
## qb:det        -1.077e+03  1.107e+03  1.364e+05 -0.973  0.3304
## qd:det         1.039e+03  1.521e+04  1.332e+05  0.068  0.9456
## targetref:qb:det -1.651e+02  1.590e+03  1.385e+05 -0.104  0.9173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 7 columns / coefficients
## Some predictor variables are on very different scales: consider rescaling

```

plots

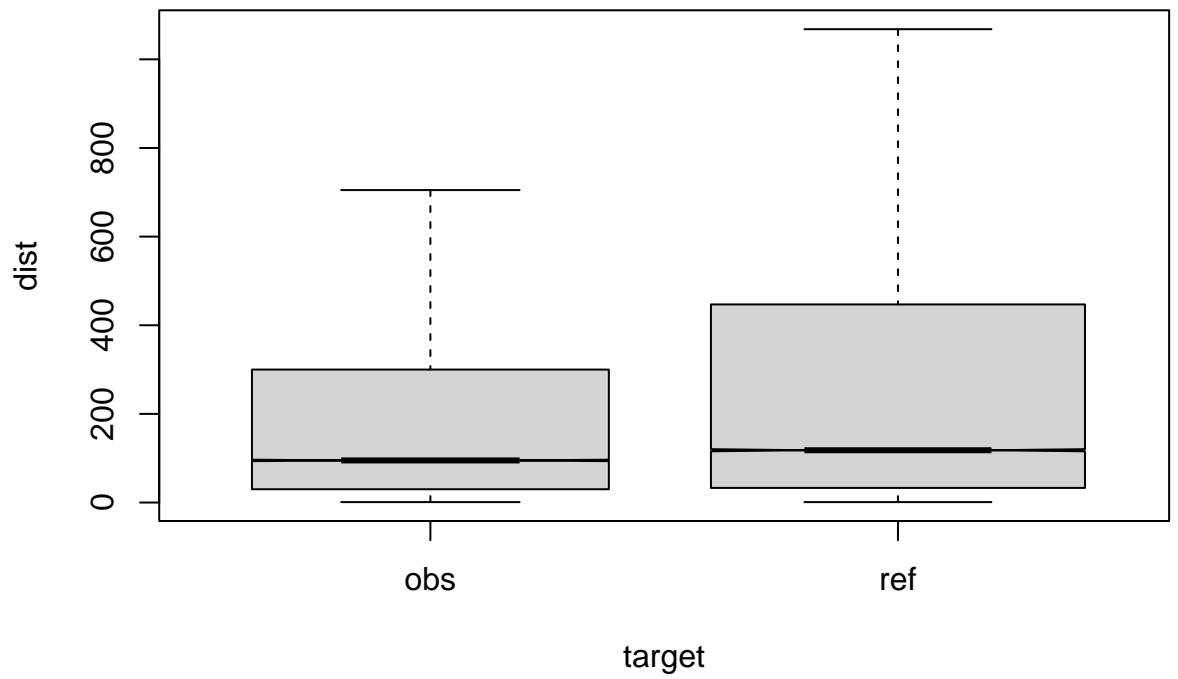


Figure 6: compare distances by corpus, not normalised, distance ceiling =outliers not removed

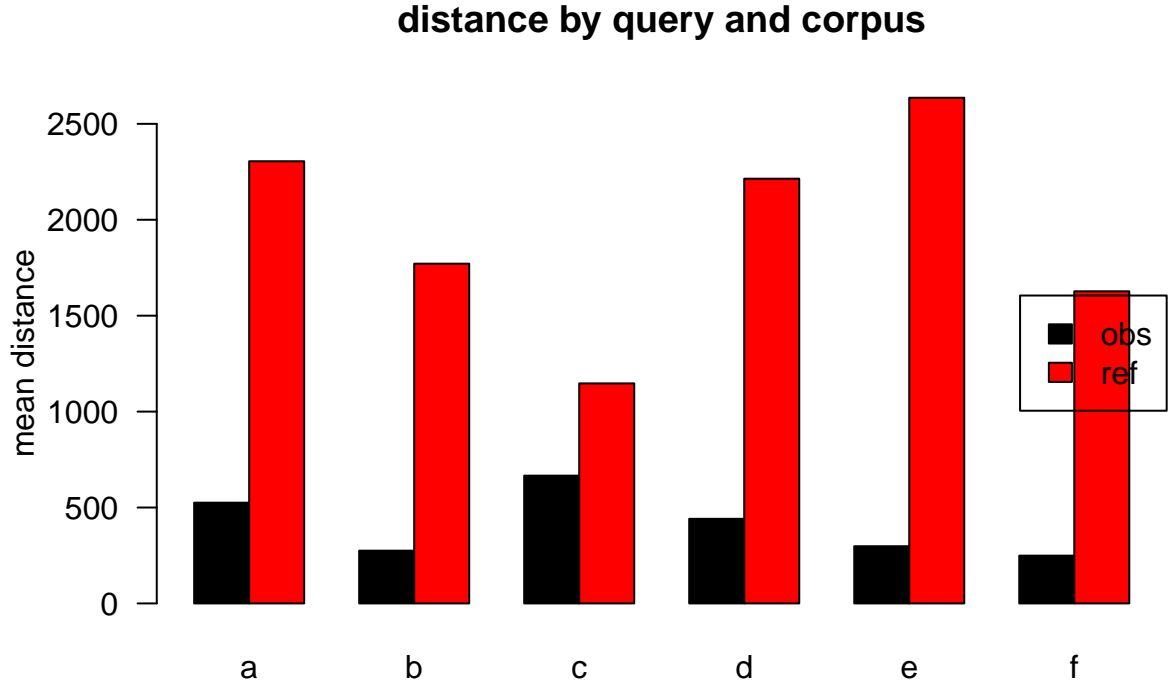


Figure 7: mean distances over query/corpus, not normalised, distance ceiling =outliers not removed

Table 5: mean/median table for model: 2

target	q	n	mean	median
obs	a	46318	525	92
ref	a	68618	2305	118
obs	b	2287	275	109
ref	b	1315	1771	111
obs	c	6253	666	89
ref	c	1504	1147	119
obs	d	6171	441	105
ref	d	1765	2214	124
obs	e	4278	298	109
ref	e	795	2636	116
obs	f	2520	249	77
ref	f	497	1627	124

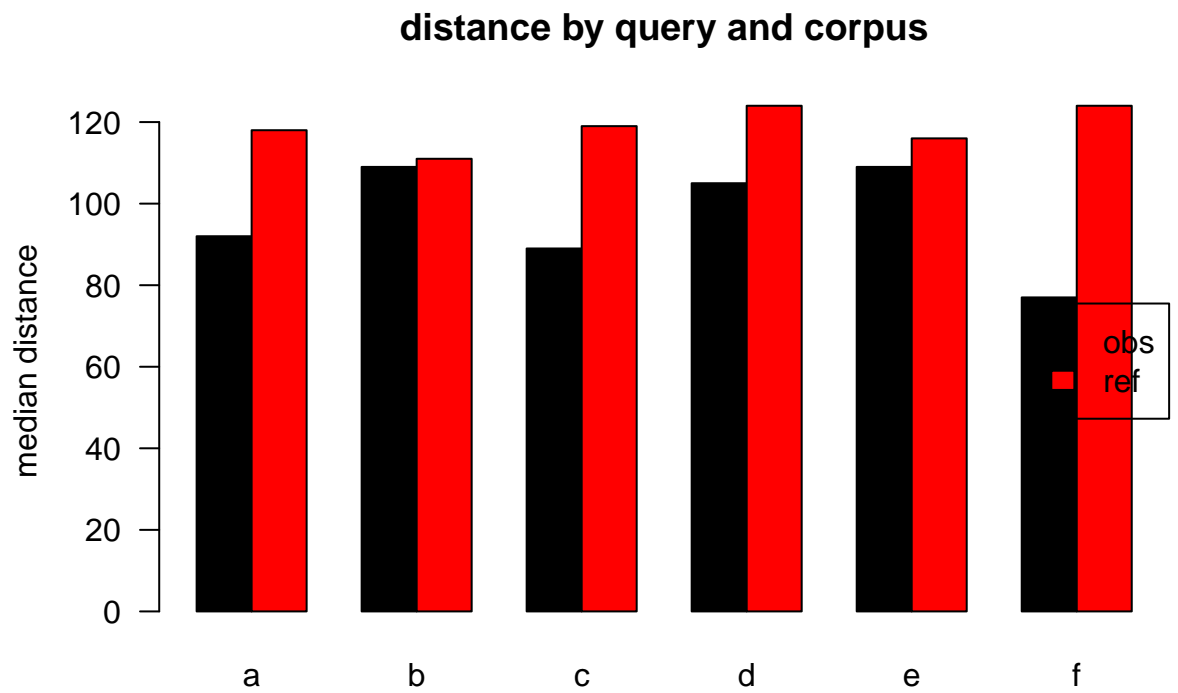


Figure 8: median distances over query/corpus, not normalised, distance ceiling =outliers not removed

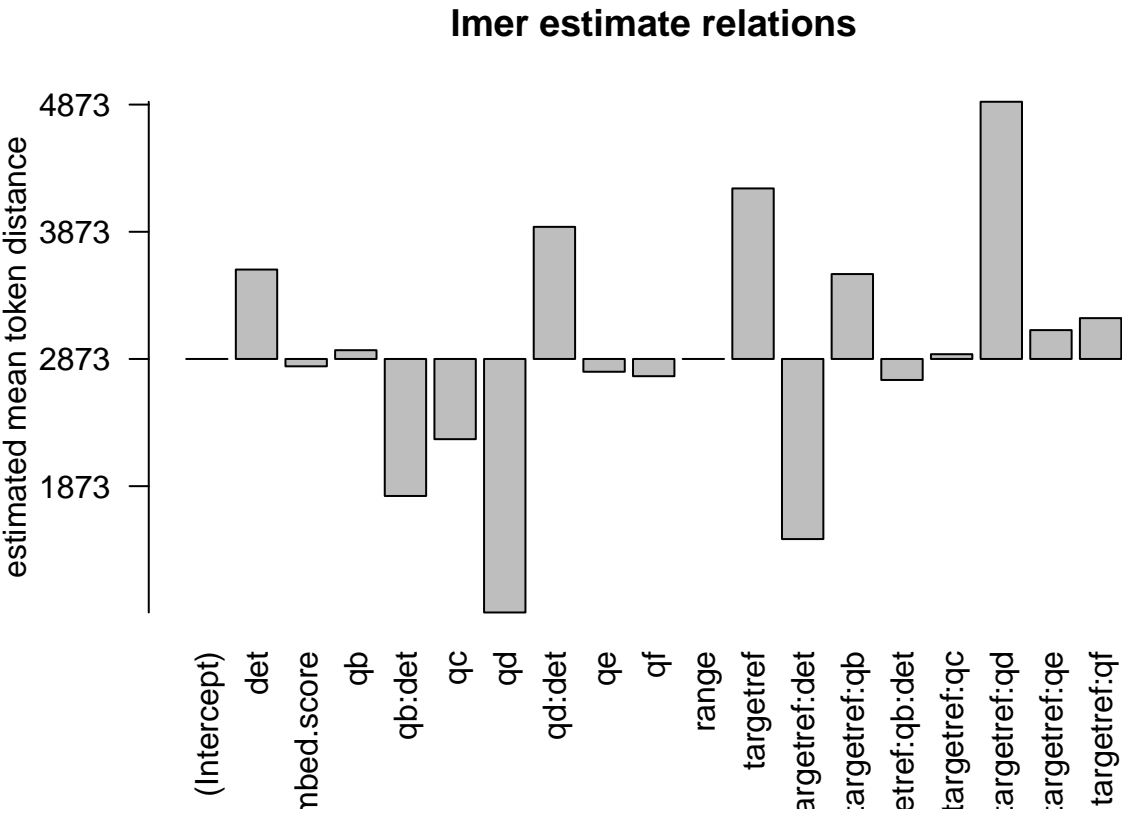


Figure 9: distances relation, not normalised, distance ceiling =outliers not removed

Distance Comparison: Raw vs target-Normalized

Diamond = median

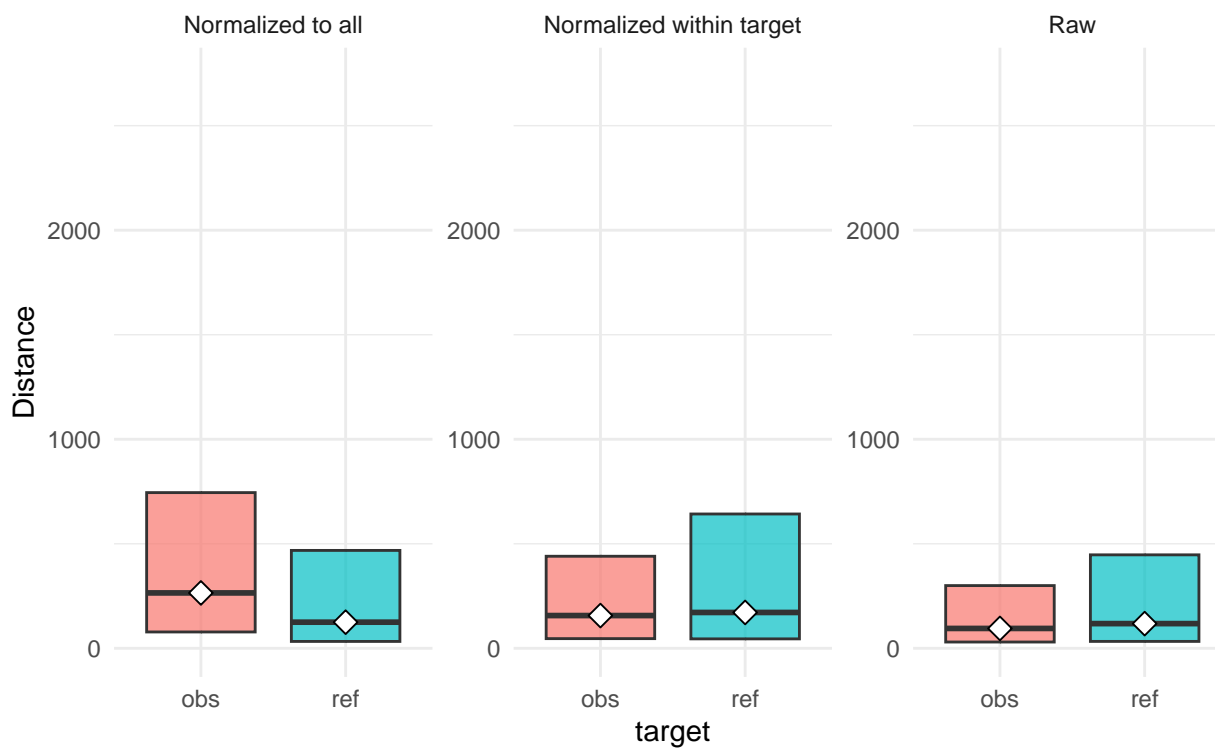


Figure 10: distances normalised vs. raw

evaluation model: 3

meta

eval output data: 13, normalised to all, distance ceiling = outliers removed

parameter setting

```
##          value
## norm_target _rel_all
## det.t      TRUE
## limit      TRUE
## author     TRUE
## url        TRUE
## embed1     TRUE
## embed2     f
```



```
## range1      TRUE
## range2      f
## rel         TRUE
## lme         FALSE
## lemma      FALSE
```

anova analysis

anova plain

```
formula: [dist_rel_all ~ target*q*det]
```

```
##           Df      Sum Sq   Mean Sq  F value    Pr(>F)
## target      1 1.2830e+09 1283010757 7336.4625 < 2.2e-16 ***
## q           5 3.4949e+07   6989793   39.9688 < 2.2e-16 ***
## det         1 4.6410e+06   4641007   26.5380 2.588e-07 ***
## target:q     5 7.7932e+06   1558646    8.9126 1.786e-08 ***
## target:det   1 7.1283e+05    712833    4.0761 0.043496 *
## q:det        2 2.5680e+06   1283981    7.3420 0.000648 ***
## target:q:det 1 2.0345e+06   2034482   11.6335 0.000648 ***
## Residuals 126209 2.2072e+10    174881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

anova of linear regression model

```
[anova(summary(lmer))]
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##           Sum Sq   Mean Sq NumDF  DenDF   F value    Pr(>F)
## target      3245706   3245706     1    3519   23.4567 1.333e-06 ***
## q           2091953    418391     5 122421    3.0237 0.0098706 **
## det          34508     34508     1 118425    0.2494 0.6175055
## range     142964302 142964302     1    1025 1033.2042 < 2.2e-16 ***
## embed.score 71204325 71204325     1 122690 514.5942 < 2.2e-16 ***
## target:q     2202162    440432     5 123486    3.1830 0.0070933 **
## target:det   1534830   1534830     1 123325   11.0922 0.0008672 ***
## q:det        1019818    509909     2 120804    3.6851 0.0250971 *
## target:q:det  623611    623611     1 123315    4.5068 0.0337615 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

linear regression coefficients

```
formula: [dist_rel_all ~ target*q*det+(1|aut_id)+range+(embed.score)+(1|url_id)]
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
```

```

## Formula: eval(expr(lmeform))
## Data: dfa
##
## REML criterion at convergence: 1859233
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.8643 -0.5282 -0.1721  0.2469  6.9244
##
## Random effects:
## Groups Name Variance Std.Dev.
## aut_id (Intercept) 8101 90.01
## url_id (Intercept) 23223 152.39
## Residual 138370 371.98
## Number of obs: 126226, groups: aut_id, 8238; url_id, 2145
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  7.789e+02  8.688e+00  8.969e+03  89.651 < 2e-16 ***
## targetref    -7.312e+01  1.061e+01  1.300e+03  -6.893 8.50e-12 ***
## qb           -3.390e+01  2.572e+01  1.218e+05  -1.318 0.187483
## qc           -3.717e+01  9.261e+00  1.226e+05  -4.014 5.98e-05 ***
## qd           -5.353e+01  3.748e+02  1.184e+05  -0.143 0.886426
## qe           4.198e+01  6.460e+00  1.247e+05   6.498 8.14e-11 ***
## qf           -3.185e+01  8.240e+00  1.244e+05  -3.866 0.000111 ***
## det          2.144e+01  8.041e+00  1.229e+05   2.667 0.007662 **
## range        -9.786e-02  3.044e-03  1.025e+03 -32.143 < 2e-16 ***
## embed.score  -3.080e+00  1.358e-01  1.227e+05 -22.685 < 2e-16 ***
## targetref:qb  3.136e+01  2.894e+01  1.225e+05   1.083 0.278599
## targetref:qc  3.842e+01  2.154e+01  1.237e+05   1.784 0.074435 .
## targetref:qd  7.432e-01  2.113e+01  1.238e+05   0.035 0.971935
## targetref:qe -3.910e+01  1.602e+01  1.239e+05  -2.441 0.014662 *
## targetref:qf  3.033e+01  2.039e+01  1.238e+05   1.488 0.136766
## targetref:det -2.490e+01  1.826e+01  1.239e+05  -1.363 0.172784
## qb:det       9.962e+01  2.826e+01  1.219e+05   3.526 0.000423 ***
## qd:det       6.144e+01  3.747e+02  1.184e+05   0.164 0.869736
## targetref:qb:det -8.754e+01  4.124e+01  1.233e+05  -2.123 0.033761 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 7 columns / coefficients
## Some predictor variables are on very different scales: consider rescaling

```

plots

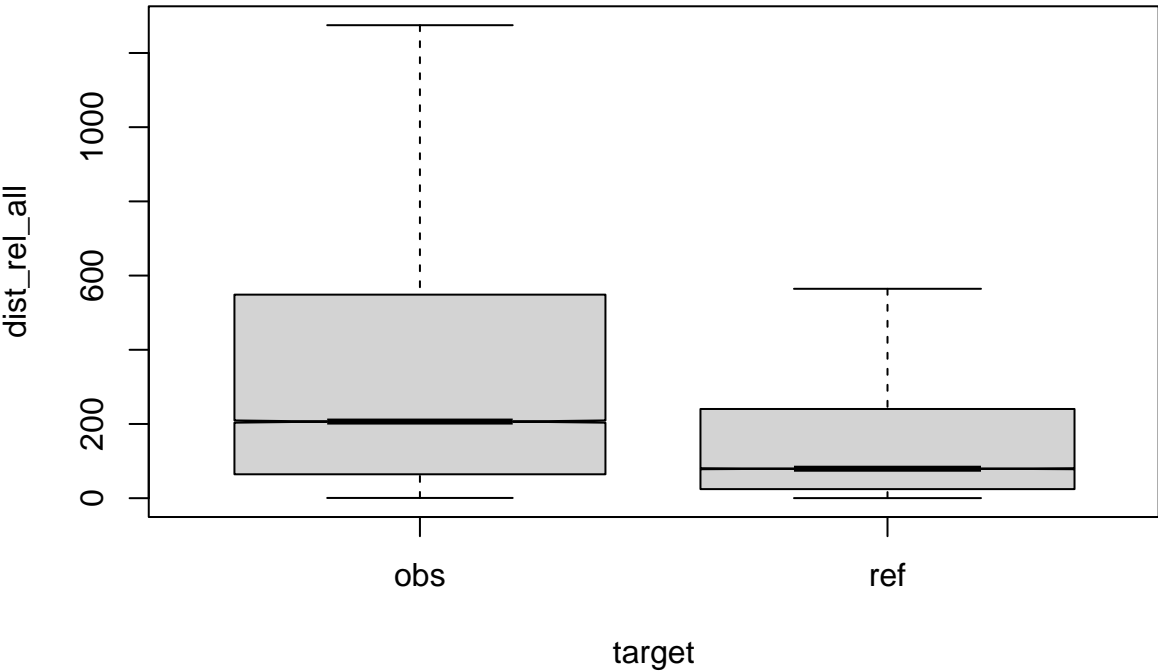


Figure 11: compare distances by corpus, normalised to all, distance ceiling = outliers removed

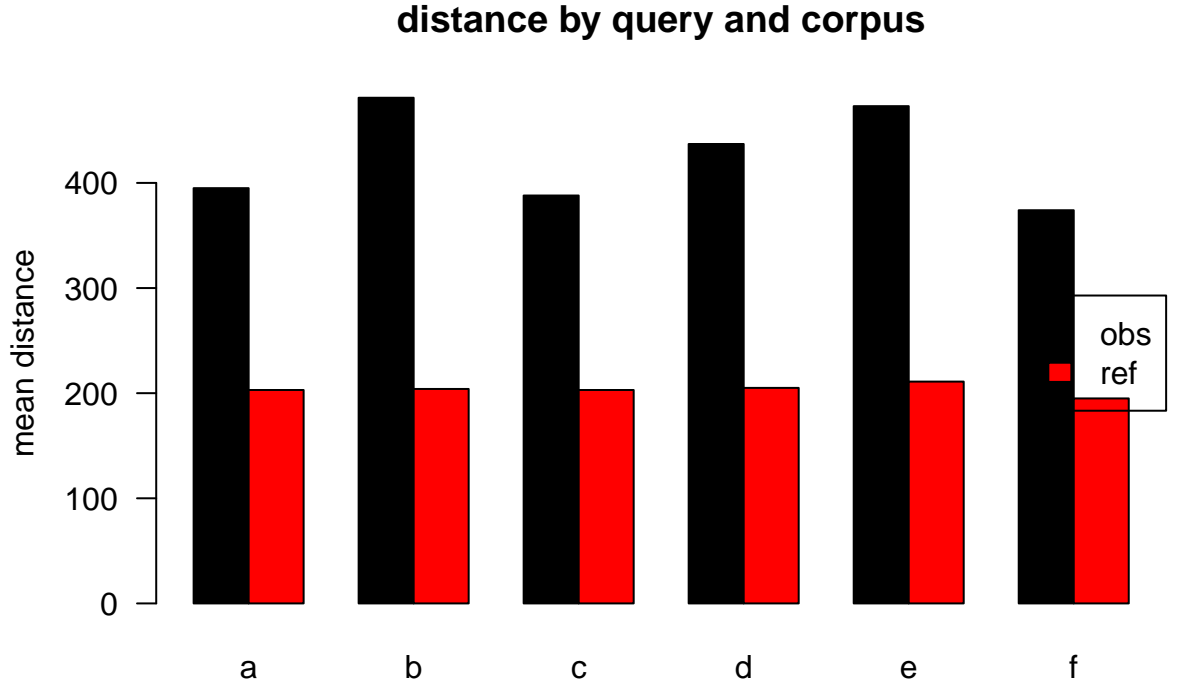


Figure 12: mean distances over query/corpus, normalised to all, distance ceiling = outliers removed

Table 6: mean/median table for model: 3

target	q	n	mean	median
obs	a	42836	395	196
ref	a	58615	203	79
obs	b	2116	481	279
ref	b	1130	204	75
obs	c	5770	388	191
ref	c	1274	203	80
obs	d	5654	437	243
ref	d	1525	205	83
obs	e	3911	473	248
ref	e	671	211	75
obs	f	2311	374	224
ref	f	413	195	79

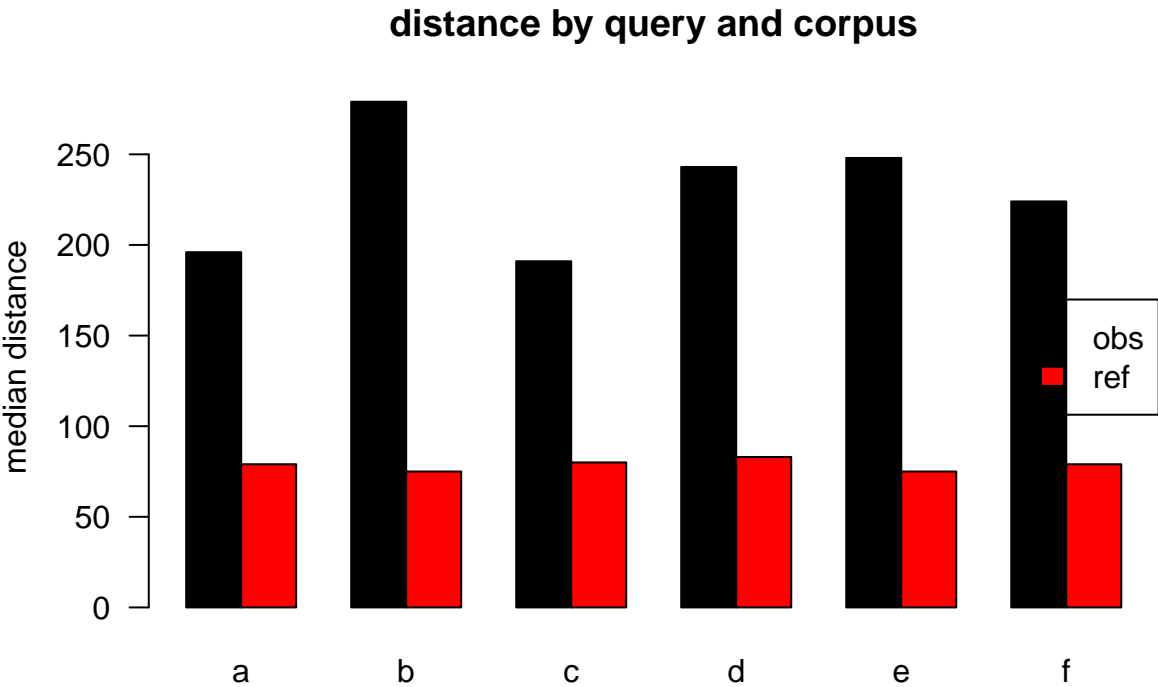


Figure 13: median distances over query/corpus, normalised to all, distance ceiling = outliers removed

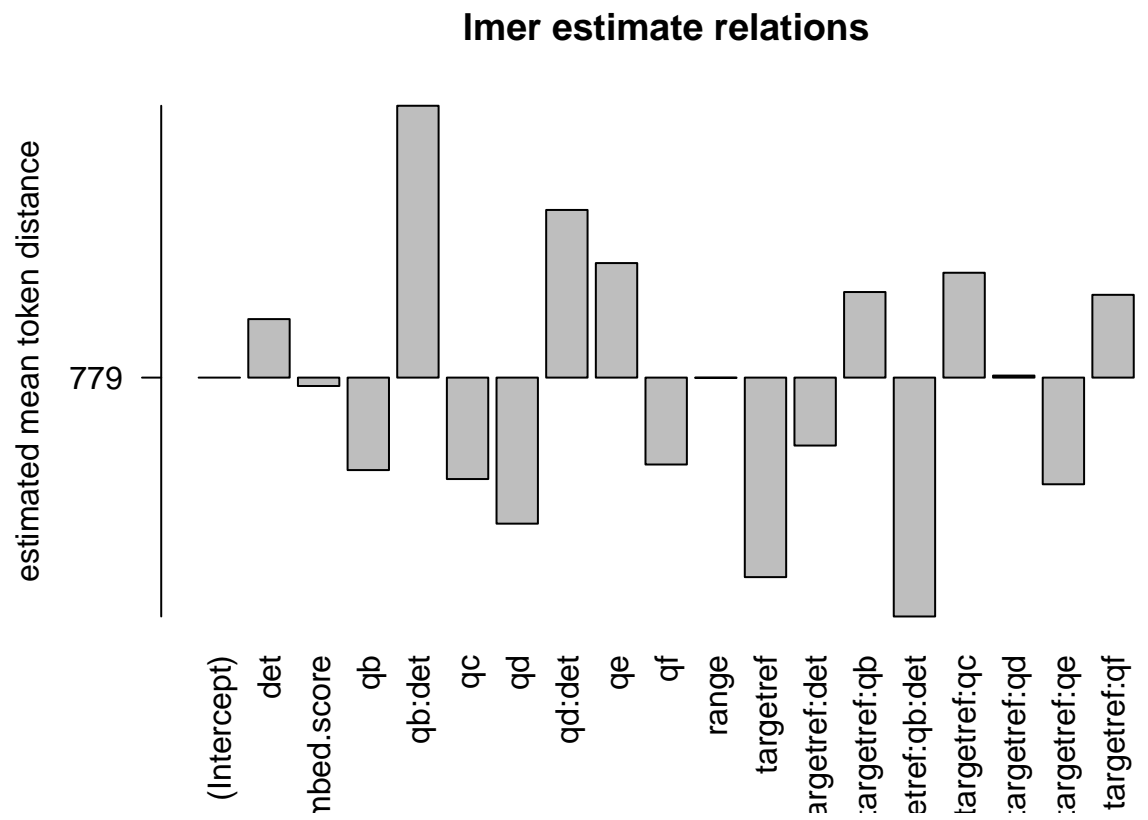


Figure 14: distances relation, normalised to all, distance ceiling = outliers removed

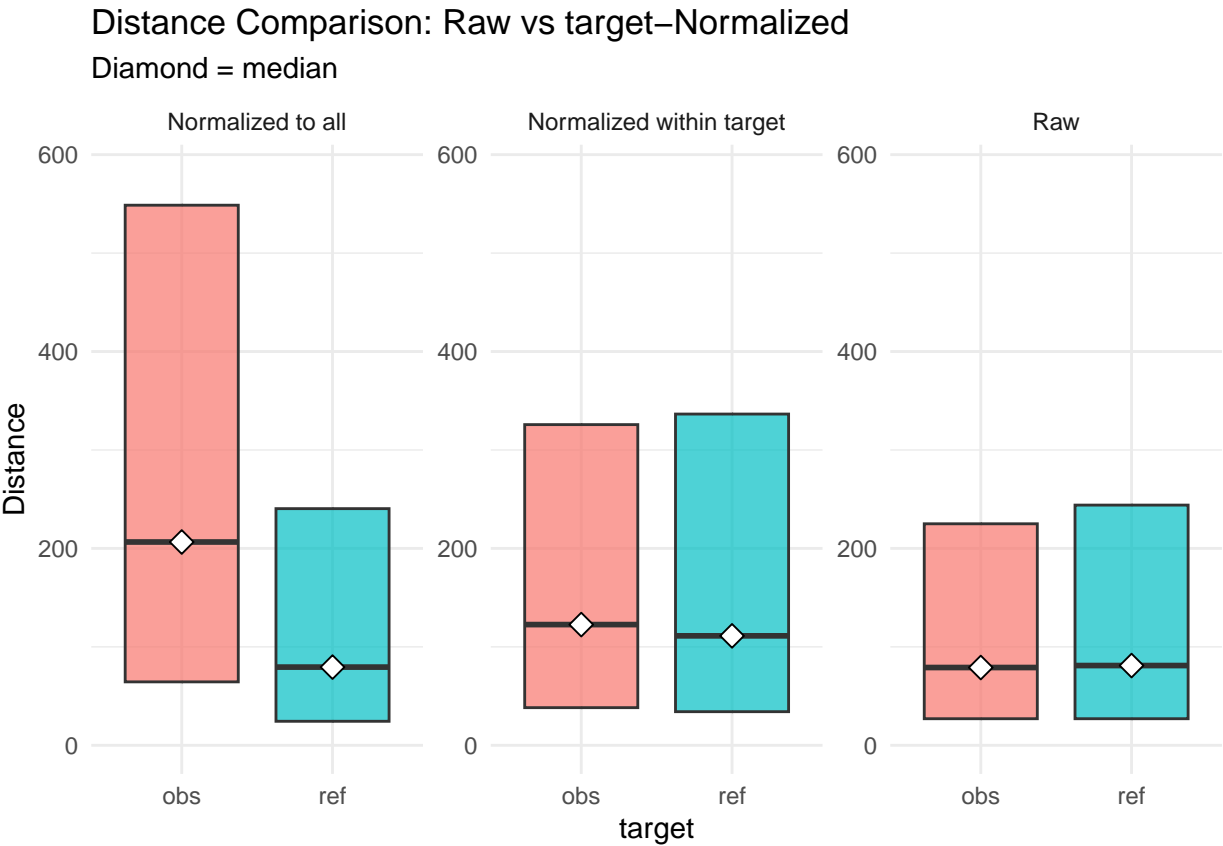


Figure 15: distances normalised vs. raw

evaluation model: 4

meta

eval output data: 13, normalised to ref, distance ceiling = outliers removed

parameter setting

```
##           value
## norm_target _rel_ref
## det.t      TRUE
## limit      TRUE
## author     TRUE
## url        TRUE
## embed1     TRUE
## embed2     f
```

```
## range1      TRUE
## range2      f
## rel         TRUE
## lme         FALSE
## lemma       FALSE
```

anova analysis

anova plain

```
formula: [dist_rel_ref ~ target*q*det]
```

```
##              Df      Sum Sq   Mean Sq   F value    Pr(>F)
## target          1 2.5135e+09 2513546743 7336.4625 < 2.2e-16 ***
## q                5 6.8469e+07  13693706   39.9688 < 2.2e-16 ***
## det              1 9.0922e+06   9092198   26.5380 2.588e-07 ***
## target:q         5 1.5268e+07   3053543    8.9126 1.786e-08 ***
## target:det        1 1.3965e+06   1396511    4.0761 0.043496 *
## q:det             2 5.0309e+06   2515448    7.3420 0.000648 ***
## target:q:det      1 3.9858e+06   3985754   11.6335 0.000648 ***
## Residuals    126209 4.3240e+10    342610
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

anova of linear regression model

```
[anova(summary(lmer))]
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##              Sum Sq   Mean Sq NumDF   DenDF   F value    Pr(>F)
## target          6358663   6358663     1    3519   23.4567 1.333e-06 ***
## q                4098347    819669     5  122421    3.0237 0.0098706 **
## det              67605      67605     1  118425    0.2494 0.6175055
## range          280081406 280081406     1    1025 1033.2042 < 2.2e-16 ***
## embed.score    139496414 139496414     1  122690   514.5942 < 2.2e-16 ***
## target:q         4314256    862851     5  123486    3.1830 0.0070933 **
## target:det       3006886    3006886     1  123325   11.0922 0.0008672 ***
## q:det            1997926    998963     2  120804    3.6851 0.0250971 *
## target:q:det     1221717    1221717     1  123315    4.5068 0.0337615 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

linear regression coefficients

```
formula: [dist_rel_ref ~ target*q*det+(1|aut_id)+range+(embed.score)+(1|url_id)]
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
```



```

## Formula: eval(expr(lmeform))
## Data: dfa
##
## REML criterion at convergence: 1944105
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.8643 -0.5282 -0.1721  0.2469  6.9244
##
## Random effects:
## Groups Name Variance Std.Dev.
## aut_id (Intercept) 15871 126.0
## url_id (Intercept) 45496 213.3
## Residual 271080 520.7
## Number of obs: 126226, groups: aut_id, 8238; url_id, 2145
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  1.090e+03  1.216e+01  8.969e+03  89.651 < 2e-16 ***
## targetref    -1.024e+02  1.485e+01  1.300e+03  -6.893 8.50e-12 ***
## qb           -4.744e+01  3.600e+01  1.218e+05  -1.318 0.187483
## qc           -5.203e+01  1.296e+01  1.226e+05  -4.014 5.98e-05 ***
## qd           -7.492e+01  5.246e+02  1.184e+05  -0.143 0.886426
## qe           5.876e+01  9.042e+00  1.247e+05   6.498 8.14e-11 ***
## qf           -4.458e+01  1.153e+01  1.244e+05  -3.866 0.000111 ***
## det          3.001e+01  1.125e+01  1.229e+05   2.667 0.007662 **
## range       -1.370e-01  4.261e-03  1.025e+03 -32.143 < 2e-16 ***
## embed.score  -4.311e+00  1.900e-01  1.227e+05 -22.685 < 2e-16 ***
## targetref:qb  4.389e+01  4.051e+01  1.225e+05   1.083 0.278599
## targetref:qc  5.378e+01  3.015e+01  1.237e+05   1.784 0.074435 .
## targetref:qd  1.040e+00  2.957e+01  1.238e+05   0.035 0.971935
## targetref:qe -5.472e+01  2.242e+01  1.239e+05  -2.441 0.014662 *
## targetref:qf  4.246e+01  2.853e+01  1.238e+05   1.488 0.136766
## targetref:det -3.485e+01  2.556e+01  1.239e+05  -1.363 0.172784
## qb:det       1.394e+02  3.955e+01  1.219e+05   3.526 0.000423 ***
## qd:det       8.600e+01  5.244e+02  1.184e+05   0.164 0.869736
## targetref:qb:det -1.225e+02  5.772e+01  1.233e+05  -2.123 0.033761 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 7 columns / coefficients
## Some predictor variables are on very different scales: consider rescaling

```

plots

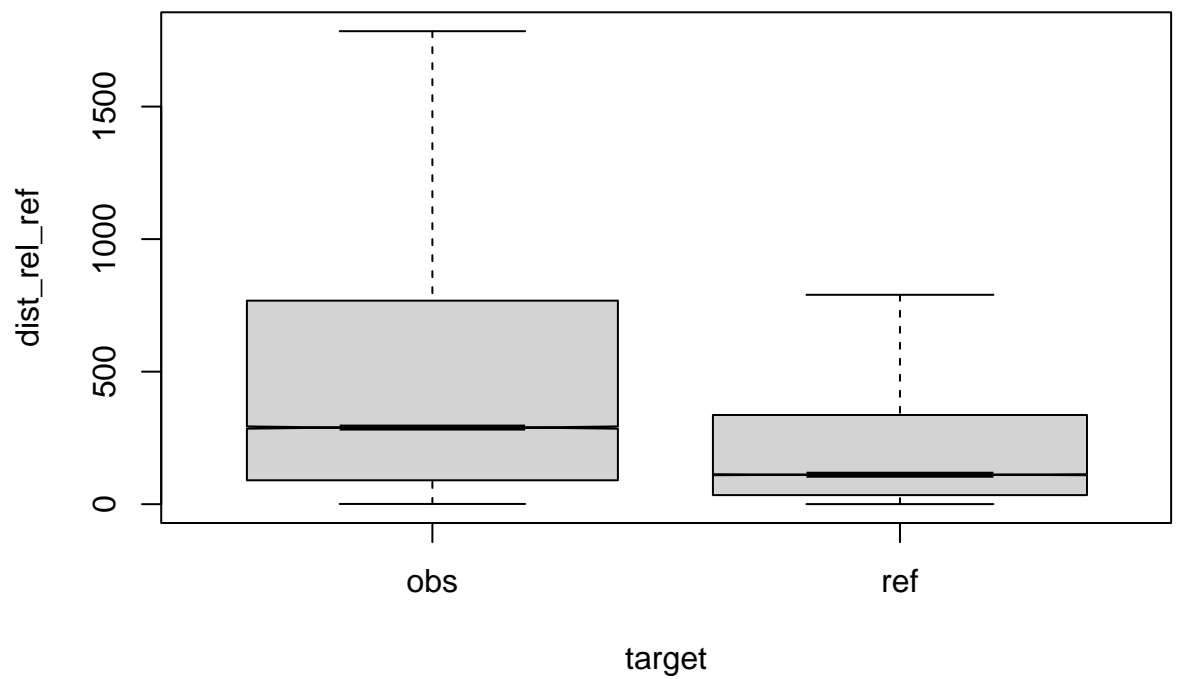


Figure 16: compare distances by corpus, normalised to ref, distance ceiling = outliers removed

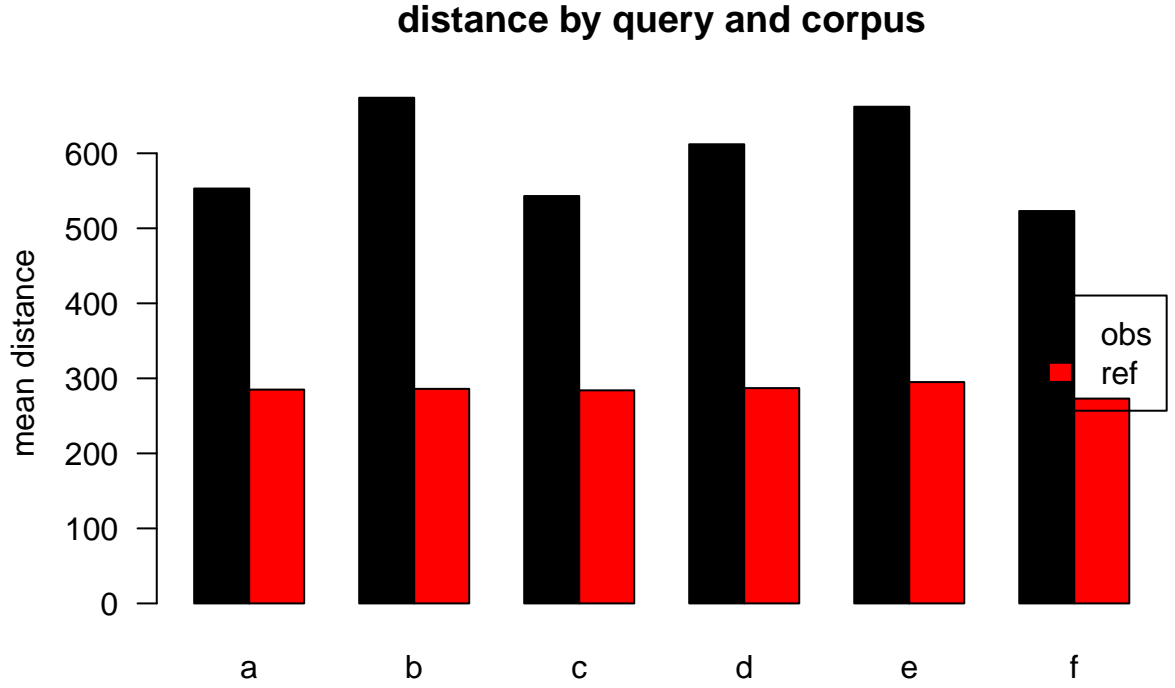


Figure 17: mean distances over query/corpus, normalised to ref, distance ceiling = outliers removed

Table 7: mean/median table for model: 4

target	q	n	mean	median
obs	a	42836	553	275
ref	a	58615	285	111
obs	b	2116	674	390
ref	b	1130	286	104
obs	c	5770	543	268
ref	c	1274	284	112
obs	d	5654	612	340
ref	d	1525	287	116
obs	e	3911	662	347
ref	e	671	295	105
obs	f	2311	523	313
ref	f	413	273	111

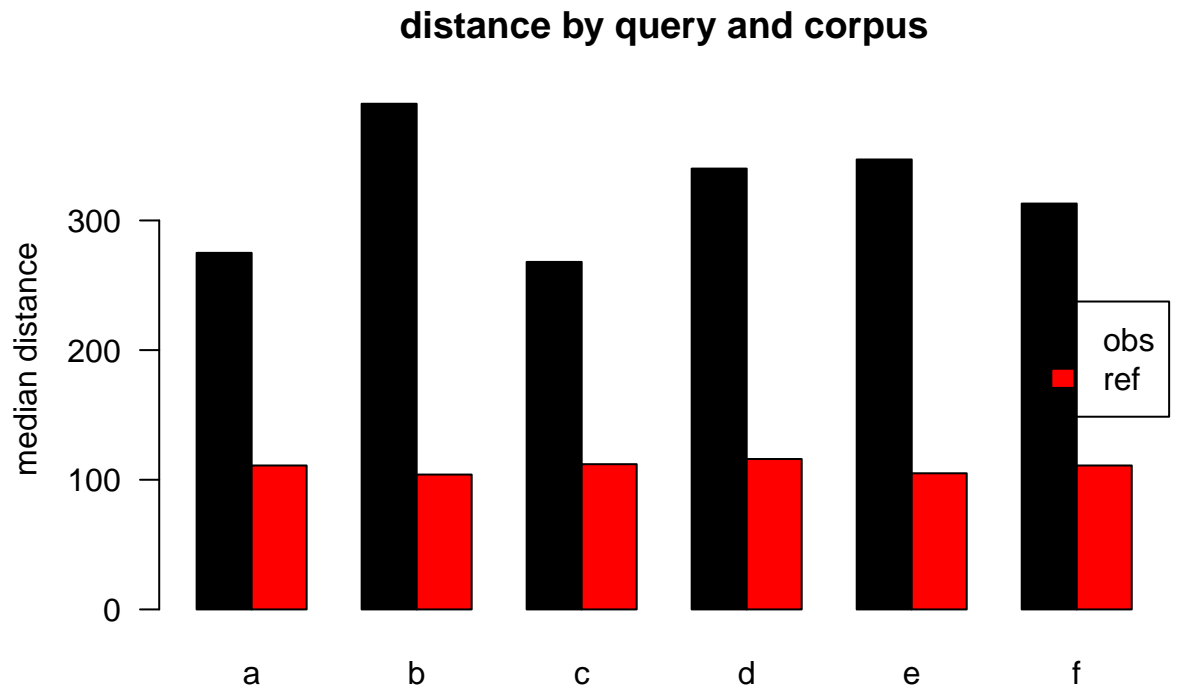


Figure 18: median distances over query/corpus, normalised to ref, distance ceiling = outliers removed

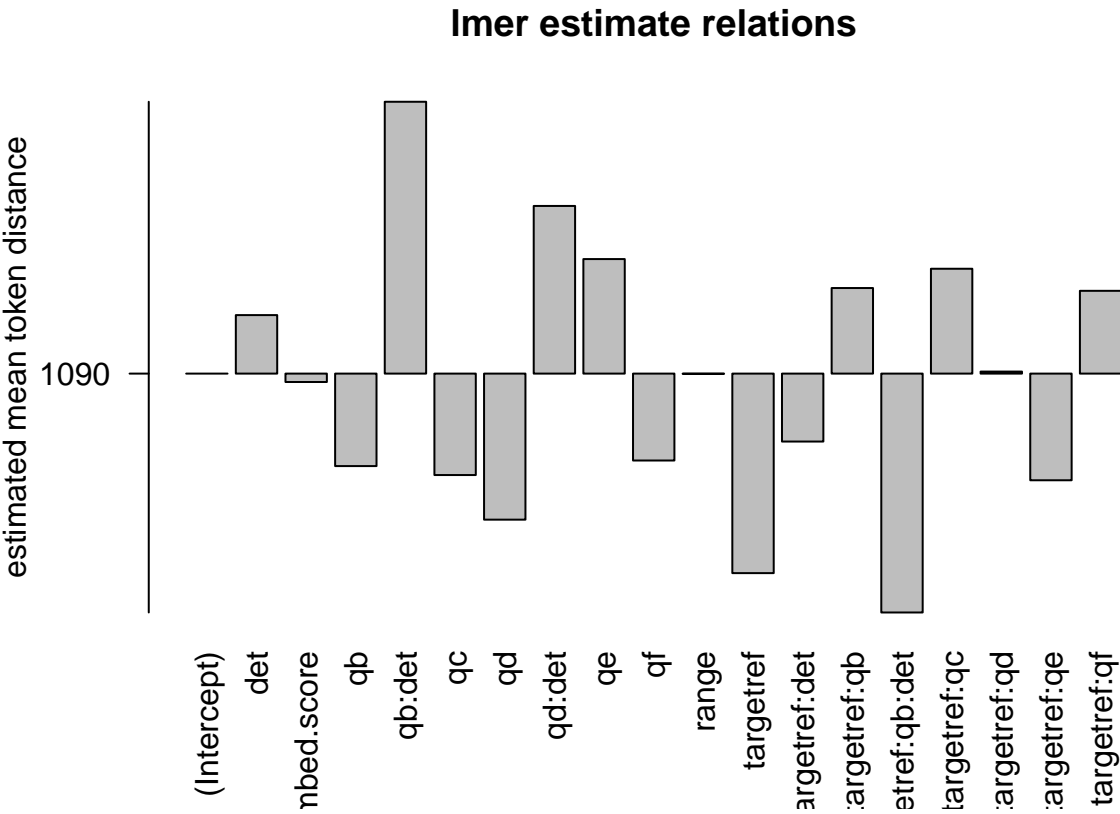


Figure 19: distances relation, normalised to ref, distance ceiling = outliers removed

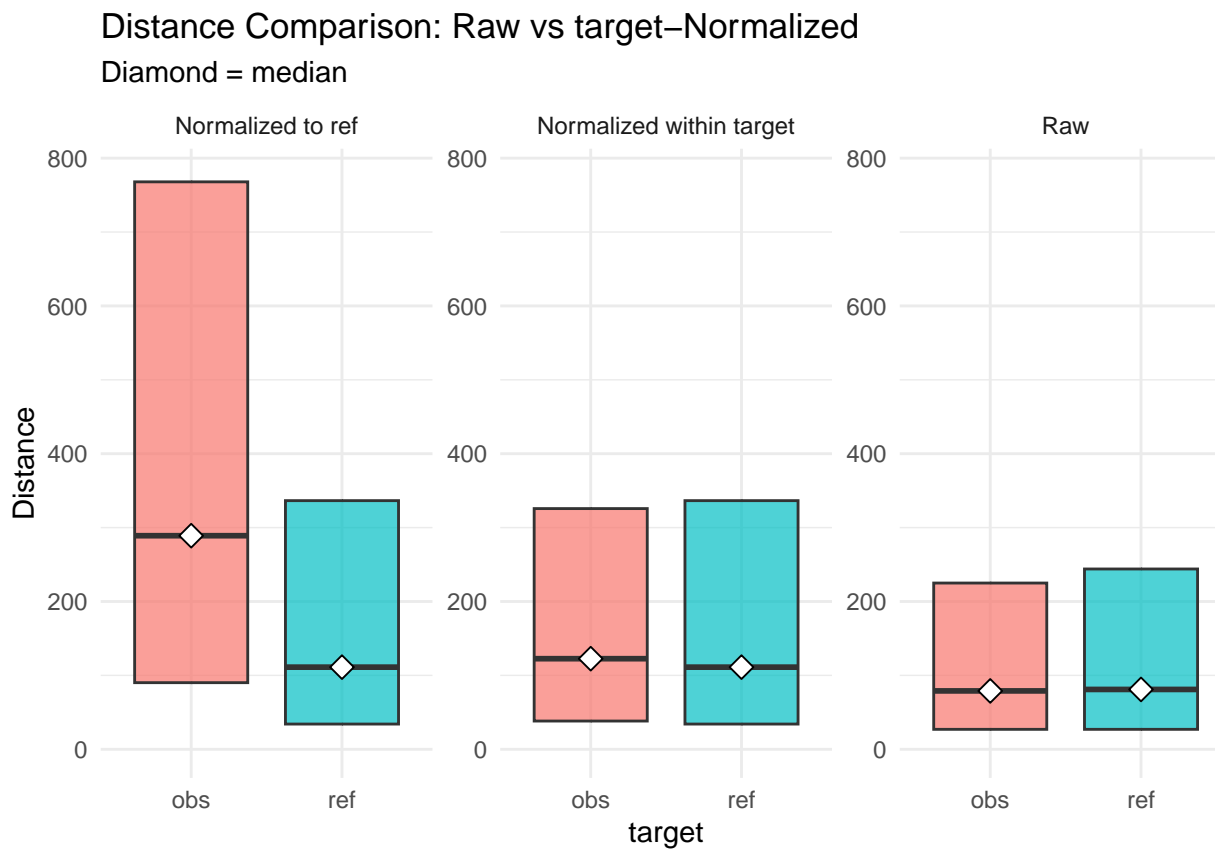


Figure 20: distances normalised vs. raw

evaluation model: 6

meta

eval output data: 13, not normalised, distance ceiling =outliers removed

parameter setting

```
##          value
## norm_target
## det.t      TRUE
## limit      TRUE
## author     TRUE
## url        TRUE
## embed1     TRUE
## embed2     f
```

```
## range1      TRUE
## range2      f
## rel         FALSE
## lme         FALSE
## lemma       FALSE
```

anova analysis

anova plain

```
formula: [dist ~ target*q*det]
```

```
##              Df      Sum Sq Mean Sq F value    Pr(>F)
## target          1    3284330 3284330 84.1223 < 2.2e-16 ***
## q                5    1633205  326641  8.3663  6.39e-08 ***
## det              1    431404   431404 11.0496 0.0008873 ***
## target:q         5    441118    88224  2.2597 0.0457798 *
## target:det        1     16732    16732  0.4286 0.5126999
## q:det             2     25549    12774  0.3272 0.7209470
## target:q:det      1        6009     6009  0.1539 0.6948226
## Residuals    126209 4927490433   39042
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

anova of linear regression model

```
[anova(summary(lmer))]
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##              Sum Sq  Mean Sq NumDF  DenDF    F value Pr(>F)
## target          218        218     1   17034    0.0061 0.9377
## q              109358      21872     5  124317    0.6129 0.6901
## det             20678      20678     1  121247    0.5794 0.4465
## range          15332432 15332432     1     912  429.6377 <2e-16 ***
## embed.score    77286240 77286240     1  105351 2165.6761 <2e-16 ***
## target:q        304923     60985     5  125126    1.7089 0.1287
## target:det      17833     17833     1  124982    0.4997 0.4796
## q:det           37151     18576     2  123066    0.5205 0.5942
## target:q:det    23985     23985     1  124972    0.6721 0.4123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

linear regression coefficients

```
formula: [dist ~ target*q*det+(1|aut_id)+range+(embed.score)+(1|url_id)]
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
```

```

## Formula: eval(expr(lmeform))
## Data: dfa
##
## REML criterion at convergence: 1685342
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.0402 -0.6622 -0.3317  0.3419  4.1697
##
## Random effects:
## Groups Name Variance Std.Dev.
## aut_id (Intercept) 1394 37.34
## url_id (Intercept) 1072 32.74
## Residual 35687 188.91
## Number of obs: 126226, groups: aut_id, 8238; url_id, 2145
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  2.533e+02  3.618e+00 1.966e+04  70.000 < 2e-16 ***
## targetref    1.326e+00  2.954e+00 1.890e+03   0.449  0.65362
## qb          -8.195e+00  1.300e+01 1.239e+05  -0.630  0.52845
## qc          -8.144e+00  4.675e+00 1.243e+05  -1.742  0.08150 .
## qd          -1.117e+02  1.902e+02 1.212e+05  -0.587  0.55726
## qe           1.392e+01  3.248e+00 1.256e+05   4.285 1.83e-05 ***
## qf          -6.628e+00  4.145e+00 1.253e+05  -1.599  0.10981
## det           3.793e+00  4.058e+00 1.245e+05   0.935  0.35005
## range        1.535e-02  7.406e-04 9.124e+02  20.728 < 2e-16 ***
## embed.score  -3.110e+00  6.682e-02 1.054e+05 -46.537 < 2e-16 ***
## targetref:qb  4.017e+00  1.464e+01 1.244e+05   0.274  0.78373
## targetref:qc  4.577e+00  1.089e+01 1.253e+05   0.420  0.67442
## targetref:qd -2.061e+00  1.069e+01 1.253e+05  -0.193  0.84707
## targetref:qe -2.134e+01  8.099e+00 1.255e+05  -2.635  0.00841 **
## targetref:qf  8.889e+00  1.031e+01 1.254e+05   0.862  0.38849
## targetref:det 1.178e+00  9.236e+00 1.253e+05   0.127  0.89855
## qb:det       1.714e+01  1.428e+01 1.239e+05   1.200  0.23002
## qd:det       1.126e+02  1.902e+02 1.212e+05   0.592  0.55380
## targetref:qb:det -1.710e+01  2.086e+01 1.250e+05  -0.820  0.41233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 7 columns / coefficients
## Some predictor variables are on very different scales: consider rescaling

```


plots

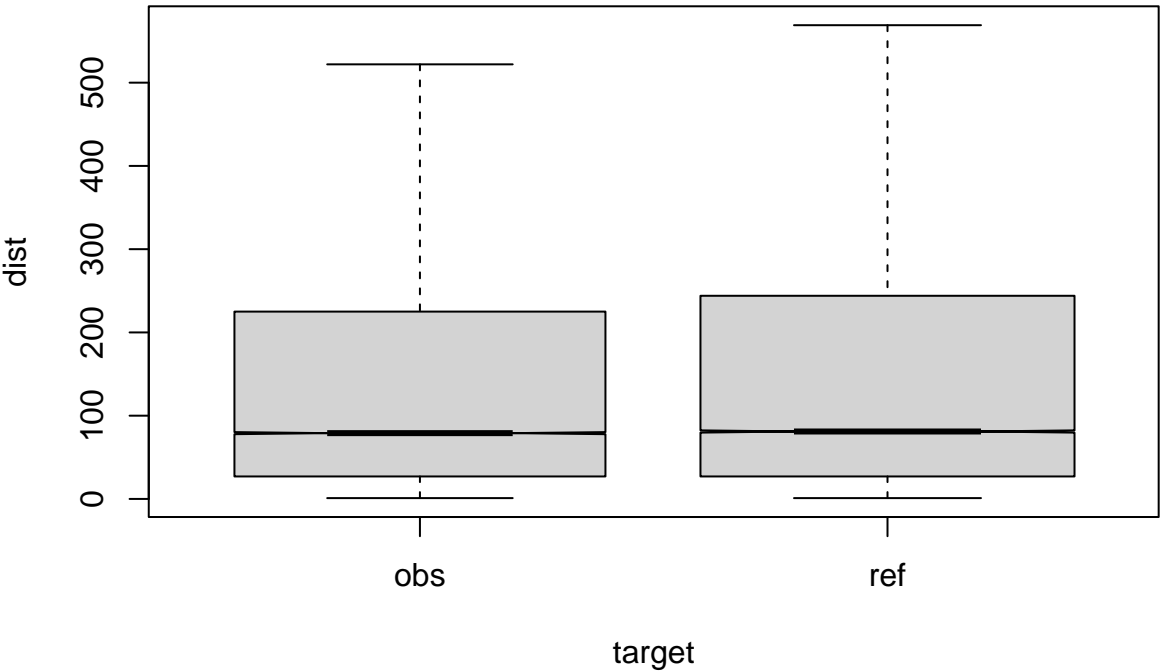


Figure 21: compare distances by corpus, not normalised, distance ceiling =outliers removed

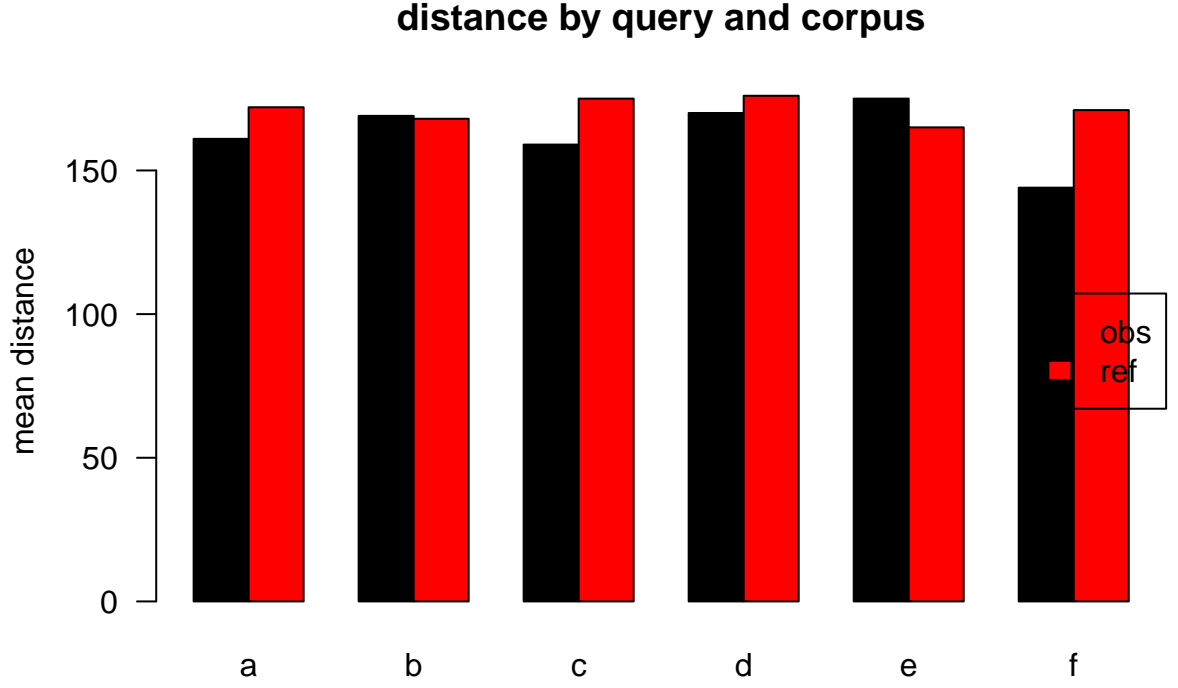


Figure 22: mean distances over query/corpus, not normalised, distance ceiling =outliers removed

Table 8: mean/median table for model: 6

target	q	n	mean	median
obs	a	42836	161	77
ref	a	58615	172	81
obs	b	2116	169	109
ref	b	1130	168	78
obs	c	5770	159	75
ref	c	1274	175	84
obs	d	5654	170	86
ref	d	1525	176	83
obs	e	3911	175	92
ref	e	671	165	71
obs	f	2311	144	62
ref	f	413	171	82

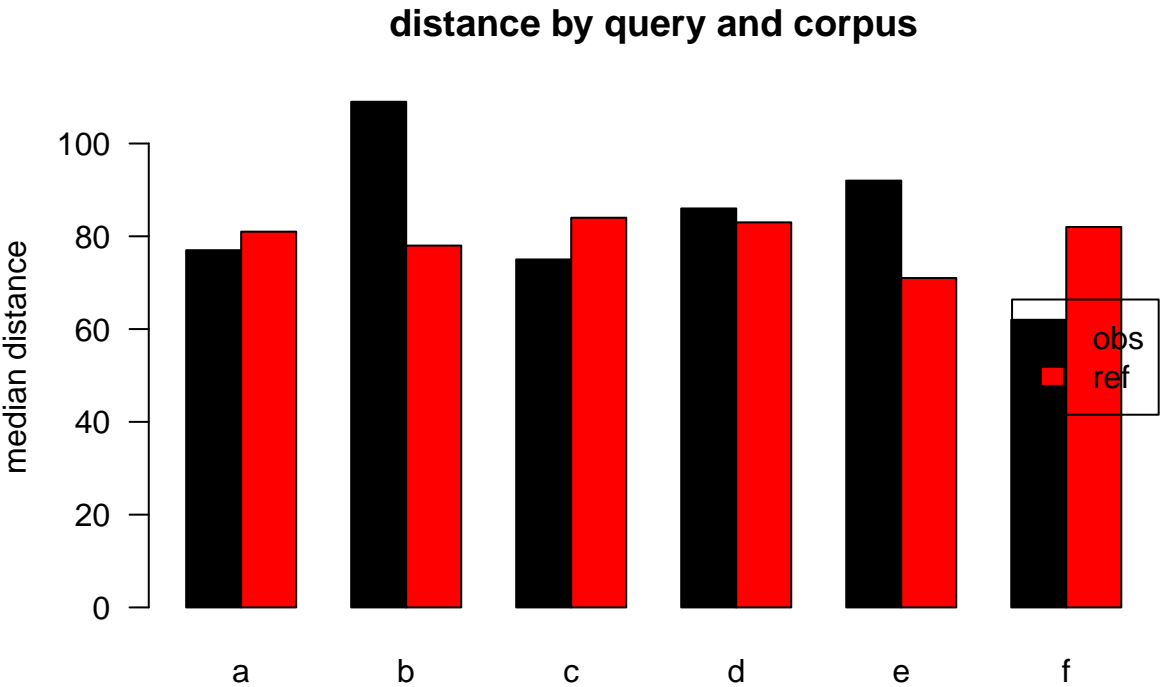


Figure 23: median distances over query/corpus, not normalised, distance ceiling =outliers removed

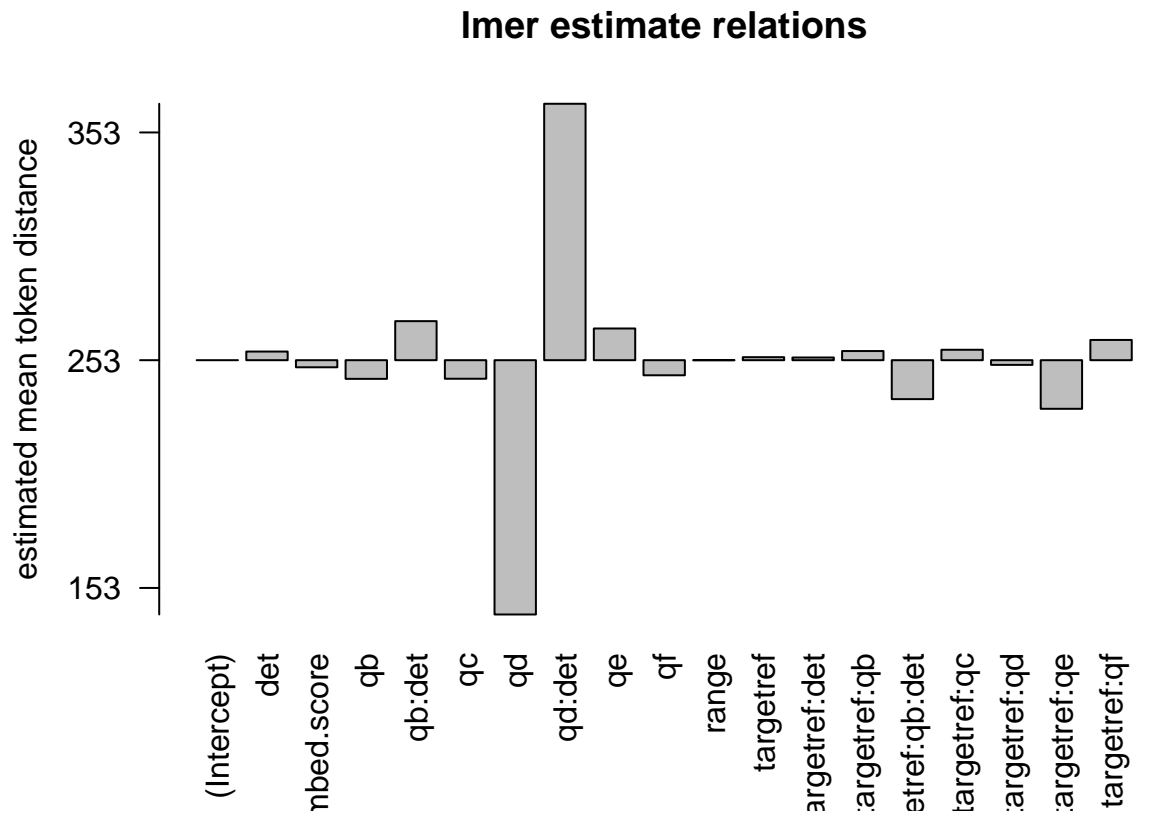


Figure 24: distances relation, not normalised, distance ceiling =outliers removed

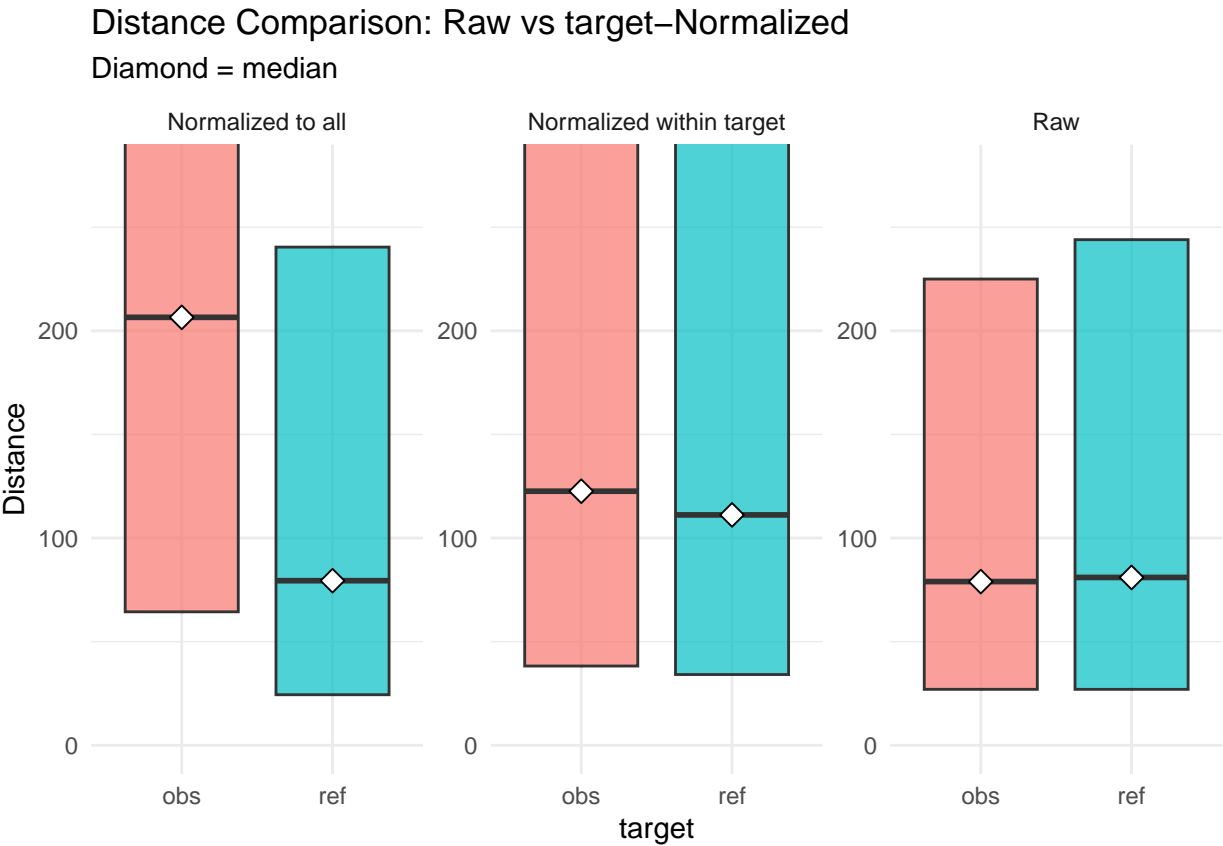


Figure 25: distances normalised vs. raw

Selbständigkeit: benutzte Hilfestellung

In der vorliegenden Arbeit wurden keinerlei nicht erlaubte Hilfsmittel zur Erstellung von Inhalten verwendet. Die Benutzung von KI beschränkt sich auf (Tabelle):

Table 9: verwendete Hilfsmittel

Hilfsmittel	Verwendung
github copilot	Hilfe bei der Skripterstellung (R, Python) zur Programmierung der Distanzenberechnung, semantic embeddings und statistischen Auswertung
chatgpt.com	dito
claude.ai	dito
deepseek.com	dito
nomic-embed-text (model)	calculate semantic embeddings

references

literature used and alii. . .

- Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software* 67 (1): 1–48. <https://doi.org/10.18637/jss.v067.i01>.
- De Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. “Universal Dependencies.” *Computational Linguistics*, May, 1–54. https://doi.org/10.1162/coli_a_00402.
- HuggingFace. 2025. “All-MiniLM-L6-V2 · Hugging Face.” *Sentence Transformers*. <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>.
- Kjell, Oscar, Salvatore Giorgi, and H. Andrew Schwartz. 2023. “The Text-Package: An R-Package for Analyzing and Visualizing Human Language Using Natural Language Processing and Deep Learning.” *Psychological Methods*. <https://doi.org/10.1037/met0000542>.
- Kuperberg, Gina R. 2010. “Language in Schizophrenia Part 2: What Can Psycholinguistics Bring to the Study of Schizophrenia... and Vice Versa?” *Language and Linguistics Compass* 4 (8): 590–604. <https://doi.org/10.1111/j.1749-818X.2010.00217.x>.
- Lee, Kenton, Luheng He, and Luke Zettlemoyer. 2018. “Higher-Order Coreference Resolution with Coarse-to-Fine Inference.” arXiv. <https://doi.org/10.48550/arXiv.1804.05392>.
- Mishara, Aaron L. 2010. “Klaus Conrad (1905–1961): Delusional Mood, Psychosis, and Beginning Schizophrenia.” *Schizophrenia Bulletin* 36 (1): 9–13. <https://doi.org/10.1093/schbul/sbp144>.
- Nenchev, Ivan, Tatjana Scheffler, Marie de la Fuente, Heiner Stuke, Benjamin Wilck, Sandra Anna Just, and Christiane Montag. 2024. “Linguistic Markers of Schizophrenia: A Case Study of Robert Walser.” In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, edited by Andrew Yates, Bart Desmet, Emily Prud’hommeaux, Ayah Zirikly, Steven Bedrick, Sean MacAvaney, Kfir Bar, Molly Ireland, and Yaakov Ophir, 41–60. St. Julians, Malta: Association for Computational Linguistics. <https://aclanthology.org/2024.clpsych-1.4/>.
- “Nomic-Ai/Nomic-Embed-Text-V1.5 · Hugging Face.” 2024. <https://huggingface.co/nomic-ai/nomic-embed-text-v1.5>.
- “Nomic-Embed-Text.” n.d. Accessed October 6, 2025. <https://ollama.com/nomic-embed-text>.
- Nussbaum, Zach, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. “Nomic Embed: Training a Reproducible Long Context Text Embedder.” <https://huggingface.co/nomic-ai/nomic-embed-text-v1.5>.
- ottiram. 2025. “Ottiram/MMAX2.” <https://github.com/ottiram/MMAX2>.
- Poesio, Massimo, Artstein, Ron, Uryupina, Olga, Rodriguez, Kepa, Delogu, Francesca, Bristot, Antonella, and Hitzeman, Janet. 2013. “The ARRAU Corpus of Anaphoric Information.” Linguistic Data Consortium. <https://doi.org/10.35111/Y3MR-HE10>.

- Prince, Ellen F. 1981. “Toward a Taxonomy of Given-New Information.” In *Syntax and Semantics: Vol. 14. Radical Pragmatics*, edited by P. Cole, 223–55. New York: Academic Press.
- Rivera, Ivan. 2023. “RedditExtractoR: Reddit Data Extraction Toolkit.” <https://CRAN.R-project.org/package=RedditExtractoR>.
- Schwarz, St. 2025. “Poster Appendix: This Papers Scripts for Corpus Build and Statistics on Github.” <https://github.com/esteschwarz/SPUND-LX/tree/main/psych/HA>.
- Wijffels, Jan. 2023. *Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. <https://CRAN.R-project.org/package=udpipe>.
- Zimmerer, Vitor C., Stuart Watson, Douglas Turkington, I. Nicol Ferrier, and Wolfram Hinzen. 2017. “Deictic and Propositional Meaning—New Perspectives on Language in Schizophrenia.” *Frontiers in Psychiatry* 8 (February). <https://doi.org/10.3389/fpsy.2017.00017>.