FREIE UNIVERSITÄT BERLIN

AI Speech and language variation & change

A study on manifestation of AI speech in human natural language

class: [16838] germanische sprachen im vergleich
tutor: Hüning / Konvicna
term: WS25/26

snc: 2026-03-01 17:07:05.403895

Preliminary

inspired by the above paper which indeed found (evidence) for GPT influenced human language after the introduction of chatGPT we tried to replicate the pipeline of building an AI vocabulary (model preferred lemmata) and compare frequencies of model typical words across pre- and post model introduction human language corpora. The first draft essai proves their hypothesis that LLM generated language manifests within human natural language.

Questions / Hypotheses

- do humans adapt to language produced by LLM and incorporate AI speech vocabulary into their own language production?
- H1: we will find higher frequencies of LLM typical vocabulary within human natural language corpora after onset of model introduction.

p2

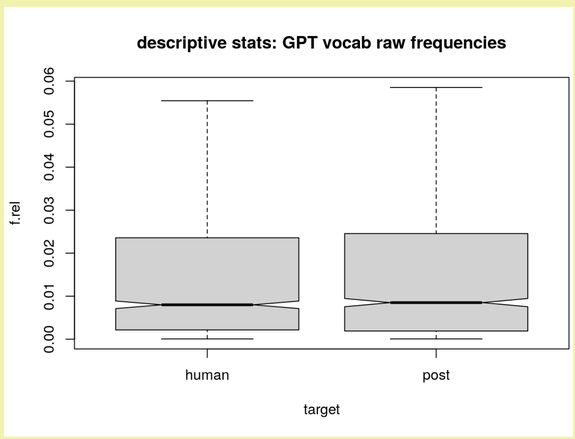

Figure 1: stats descriptive

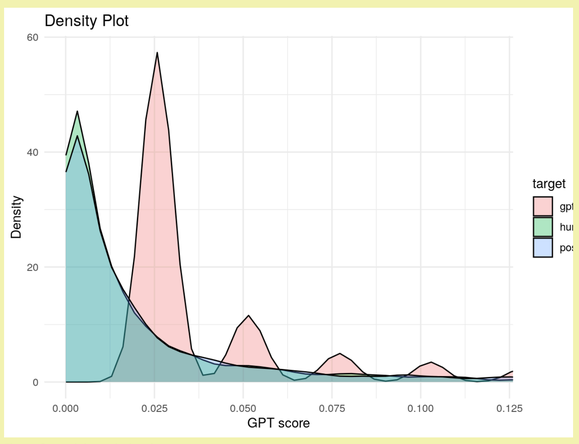

Figure 2: model vocab density

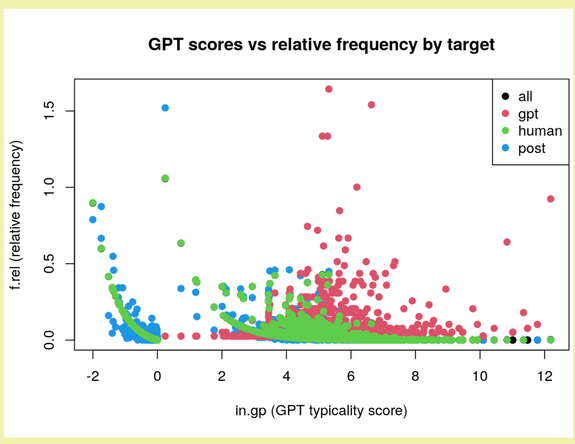

Figure 3: model vocab distribution

Methods / Data

our human langugae data consists of raw texts from german bundestag plenary protocols (). the LLM corpus consists of model summaries of a first subset of these texts generated with the prompt you see below.

we first devised AI-typical lemmata in the model corpus which are distinct for that corpus using a linear regression model that calculates a score for each lemma in the corpus, see Figure 3 and Figure 2.

c1

System prompt: You are a member of german parliament. Prepare a summary of the text provided to present at a local community meeting of your party members. Output in german language, no preamble, no extra information, just the plain text. Wordcount maximal 300 words, containing not more than 5% of the keywords of the text provided and explicitly not just a list of keywords but an entertaining text. You are supposed to interprete freely, including background insights on daily politics. Keep in mind thatthe text will be used as is as keynotes to the talk being held to the locals. Text:

Results

to gather an insight, yet with simple descriptive stats comparing the raw frequencies of gpt-preferred lemmas in pre- and post-gemini onset we find that in the target corpus the occurences of these lemma increase, see Figure 1.

References

Gilquin (2008) Mehl (2021) UCSB Santa Barbara Corpus

```
human"]
/SPUND-
trace =
################################################################
API_KEY<-api_key<-key
# range<-c("2021-04-01","2021-12-31")
gg.json
# levels(factor(lmdf.c$target))
get.q<-function(baseurl,qdf,d.start,d.end){
# gpt_typical  <- re_lemma[order(-re_lemma$res), ]
tokens.3<-tokens$t2
  head(tokens.2$word,20)
    human_count = human,
#fg1$gus<-fg1$FREQ*fg1$gu
  cat("processing: ",k,"\r")
  lt<-list(p=1)

library(dplyr)
mh1<-lmdf.c$gus.c==1&lmdf.c$target=="0-human"
load(paste0(Sys.getenv("GIT_TOP"),"/SPUND-LX/germanic/HA/lm-base_lmdf.c.RData"))
#1z1[1,]<-c("intercept",1,"0-intercept",1)
### now with lemma, udpipe pipeline...


lhp<-lhp[!is.na(lhp)]
stops.u<-lapply(stops.j, function(x){
  fj11$fg[m]<-fg$Var1[k]
)
  pivot_wider(lmdf.c,
# library(clipr)
url<-'https://search.dip.bundestag.de/api/v1/plenarprotokoll-text?f.aktualisiert.start=2022-12-06T10%3A00%3A00&f.aktualisiert.end=2022-12-06T20%3A00%3A
})
summary(m1)
m3 <- glmer(
#r2<-GET(turl)
################################################################
# t1<-table(tok.r2$lemma)
p61<-round(s2$coefficients[1,1],6)
gg.js<-gg.json
f1p1$gus<-0
  #  mode(tdb6$size)<-"double"
tokens.r<-tokens.r[!m,]
    values_from = freq,
```