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

authored: st. schwarz
snc: #16142::20260330(19.24) CEST
<<< back to paper <<<

preliminary

Inspired by BSI (2025) and Yakura et al. (2025) which found evidence for GPT influenced human language after the introduction of chatGPT we tried to replicate the Yakura et al. (2025) 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 model 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.

methods / data

Our human languaga data consists of raw texts from german bundestag plenary protocols, DIP (2026).

The LLM corpus consists of model summaries of a first subset of these texts generated with the prompt (to gemini API, Wikipedia and Google (2026)) you see in C1.

We first devised model typical lemmata in the AI corpus which are distinct for that corpus using their log odds score calculated over pre/post target, cf. Figure 3.

For corpus building and evaluation scripts cf. Schwarz (2026).

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:
```

discussion / limitations

If the findings of increased AI lemma frequencies really prove manifestation of AI language depends also on influences of other factors of language variation and change which Yakura et al. (2025) already took into account. A deeper going analysis of general trends in language development and investigation on the bidirectional influences of model-training and human language could harden the results. Also dating of the adressed onset of gemini introduction and correspondingly synchronizing the target resp. the AI corpus date ranges is an issue not yet fully adressed within our pipeline. Since google gemini was introduced in different stages with changing model traing data, to exactly sync our corpora with the model output i.e. build the AI vocabulary corresponding with that timeframe is an open issue.

p2

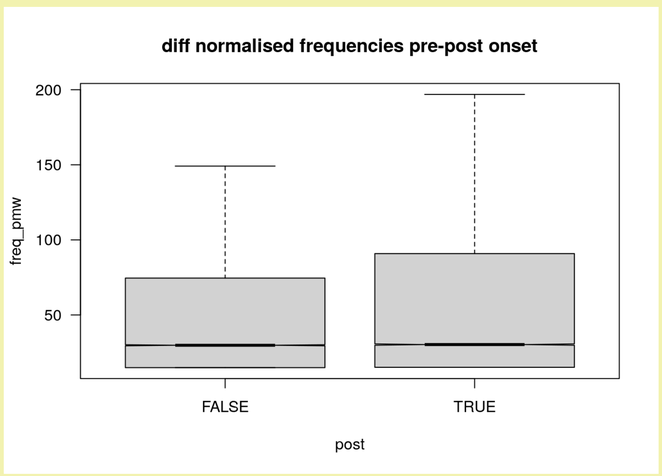

Figure 1: descriptive stats plot

p3

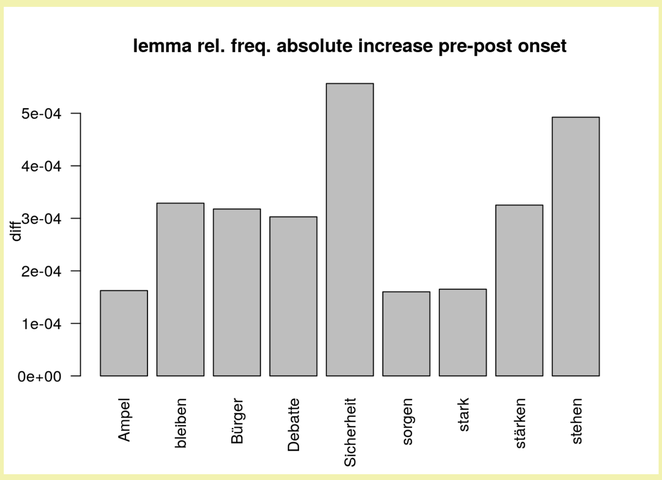

Figure 2: lemma frequency increase of relevant AI typical lemma

c2: corpus stats

stat	var	pre	post
mean	freq_rel gpt lemmas	0.000704	0.000751
mean	freq normalised pmw	29.825	30.288
tokens	n	1640446	1940385

statistics…

c3: regression simple model

```
Call:
lm(formula = freq_pmw ~ post + gp, data = df_lg.norm)

Residuals:
   Min     1Q Median     3Q    Max
 -1400   -435   -388   -243 436922

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   833.75      38.64  21.576  < 2e-16 ***
postTRUE       83.94      20.64   4.067 4.77e-05 ***
gp             68.11       5.20  13.097  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3765 on 133087 degrees of freedom
  (1393 observations deleted due to missingness)
Multiple R-squared:  0.0014,    Adjusted R-squared:  0.001385
F-statistic: 93.32 on 2 and 133087 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: freq_pmw
             Df     Sum Sq    Mean Sq F value    Pr(>F)
post          1 2.1394e+08  213944737  15.096 0.0001022 ***
gp            1 2.4310e+09 2431030842 171.537 < 2.2e-16 ***
Residuals 133087 1.8861e+12   14172027
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p4

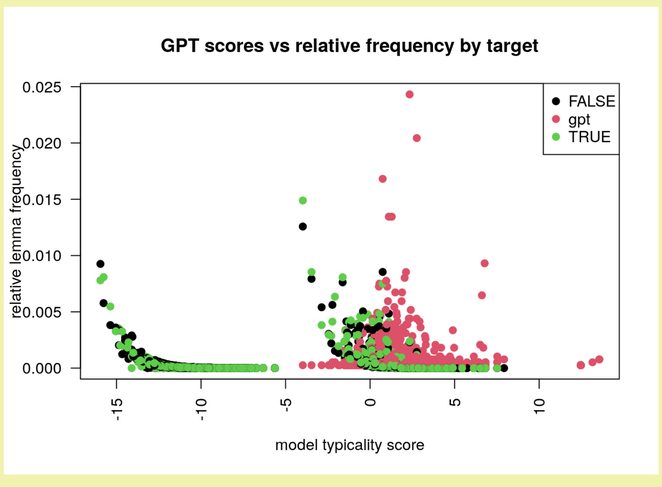

Figure 3: model typical vocab distribution

c4: regression random effects model

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: freq_pmw ~ post + gp + (1 | lemma)
   Data: df_lg.norm

REML criterion at convergence: 2490672

Scaled residuals:
    Min      1Q  Median      3Q     Max
-48.624  -0.047  -0.032   0.008  74.360

Random effects:
 Groups   Name        Variance Std.Dev.
 lemma    (Intercept) 8839356  2973
 Residual             1754704  1325
Number of obs: 133090, groups:  lemma, 99649

Fixed effects:
            Estimate Std. Error        df t value Pr(>|t|)
(Intercept) 4.344e+02  3.966e+01 1.055e+05  10.953   <2e-16 ***
postTRUE    1.246e+02  9.496e+00 5.602e+04  13.126   <2e-16 ***
gp          2.932e+01  5.723e+00 1.017e+05   5.124    3e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
        (Intr) psTRUE
postTRUE -0.110
gp        0.959  0.009

Type III Analysis of Variance Table with Satterthwaite's method
       Sum Sq   Mean Sq NumDF  DenDF F value    Pr(>F)
post 302319616 302319616     1  56015 172.291 < 2.2e-16 ***
gp    46071208  46071208     1 101716  26.256 2.996e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

model explanations

We see in the model summaries that frequency increases are significant (p<0.001) when positing a random effect of lemma (model 1, cf. C4) with an estimate increase of mean frequency (pmw) per lemma by 29.32 vs. the intercept (post=FALSE i.e. target[pre]).

In model 2, cf. C3 (simple model without lemma effect) we observe an estimate increase of mean frequency (pmw) of post=TRUE by 83.94 vs. the intercept.

In both models `gp` (independent var of above mentioned gpt score: higher score > more model typical lemma) has been included as fixed effect, but vcvs. without random lemma effect that predictor lets the model naturally collapse (p<10E-16).

results

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.

The 9 lemma where 1. the relative frequency of lemma in target=post is higher than in target=pre (n=27) and 2. that frequency exceeds the mean relative frequency is displayed in Figure 2.

We observe significant frequency increases using a linear regression model (Bates et al. (2015)), cf. C3/C4, which tested the hypothesis.

references

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.

BSI, Brand Science Institute. 2025. "Wie KI Unsere Sprache Verändert – Eine Empirische Studie." https://www.bsi.ag/cases/104-case-studie-wie-ki-unsere-sprache-veraendert---eine-empirische-studie.html.

DIP. 2026. "DIP - Bundestagsprotokolle." Docs. *DIP - API*. Berlin. https://dip.bundestag.de/%C3%BCber-dip/hilfe/api#content.

Schwarz, St. 2026. "This Papers Corpus Build & Evaluation Scripts." *GitHub/Esteeschwarz*. Berlin. https://github.com/esteeschwarz/SPUND-LX/blob/main/germanic/HA/.

Wikipedia, and Google. 2026. "Google Gemini." *Wikipedia*. https://de.wikipedia.org/w/index.php?title=Google_Gemini&oldid=263426206.

Yakura, Hiromu, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, Ivan Soraperra, and Iyad Rahwan. 2025. "Empirical Evidence of Large Language Model's Influence on Human Spoken Communication." arXiv. https://doi.org/10.48550/arXiv.2409.01754.