universität freiburg

https://github.com/HannaMahler/RWTH_Aachen_2023-11-20

Contrastive Grammar in Use

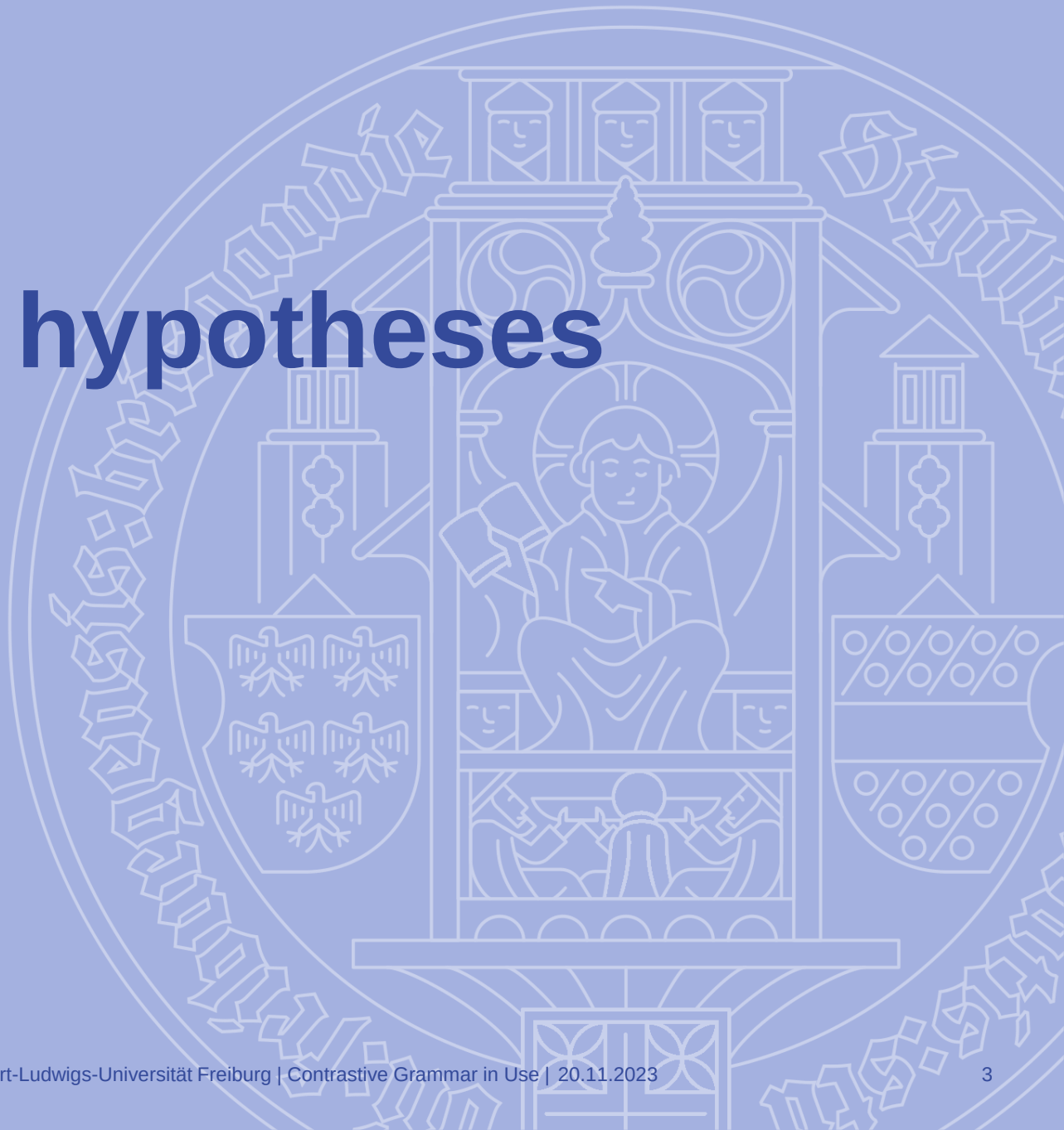
Quantitative Perspectives on the Verb Phrase in English and German

Hanna Mahler, M.A.
Albert-Ludwigs-Universität Freiburg
Aachen, 20.11.2023

Agenda

1. State of research & hypotheses
3. Methodology & data
4. Results
5. Discussion
6. Conclusion

State of research & hypotheses



Previous research on verb phrase use in English and German

- **common assertion that German is more nominal and English more verbal**
  - e.g., Kortmann & Meyer (1992: 165), Friederich (1969: 83, 88, 91), Königs (2004: 1)
  - Example: (1) “A window displaying outfits”
    (2) “Ein Schaufenster mit Kleidung”
- **state of research: mostly focus on nominal style, few corpus-based studies (using translation material), mostly use of automatic PoS frequencies**
  - Steiner (2012): CroCo corpus, PoS frequencies, more verbs in English
  - Berg et al. (2012): study on compounding, similar frequency of nouns in both languages
  - Fischer (2013): small translation corpus, manual annotation, more verb phrases in English
  - Berg (2017): type and token frequency of word classes, four comparable corpora, English more verbal
  - Neumann (2020): translation corpus, focus on nominal style, German not more nominal

Hypotheses

- Hypothesis 1: On the whole, English uses more verb phrases than German.
- Hypothesis 2: On the whole, English uses more non-finite verb phrases than German.
- Hypothesis 3: Cross-linguistically, the use of verb phrases varies by register and mode.
- Hypothesis 4: Cross-linguistically, the use of non-finite verb phrases varies by register and mode.
- Hypothesis 5: The relationship between information density and frequency of verb phrases is language-specific.
- Hypothesis 6: The relationship between information density and frequency of non-finite verb phrases is language-specific.
- Hypothesis 7: Cross-linguistically, frequency of use may differ even for parallel non-finite constructions.

Methodology & data

The corpus: GECCo
Kunz et al. (2021)

- Translation corpus and comparable corpus
- Contains spoken and written data from 14 registers
- Corpus size: around 500,000 words per language (non-translated part)
- Extraction of verb phrases via UPOS-tags, automatic + manual annotation
- Variables annotated: auxiliary or main verb, finite or non-finite, verb form, implicit or overt subject, type of dependent clause (embedded, adverbial, nominal), function of nominal clause (S, O, C_S, C_O)

GECCo
GERMAN ENGLISH CONTRASTS IN COHESION

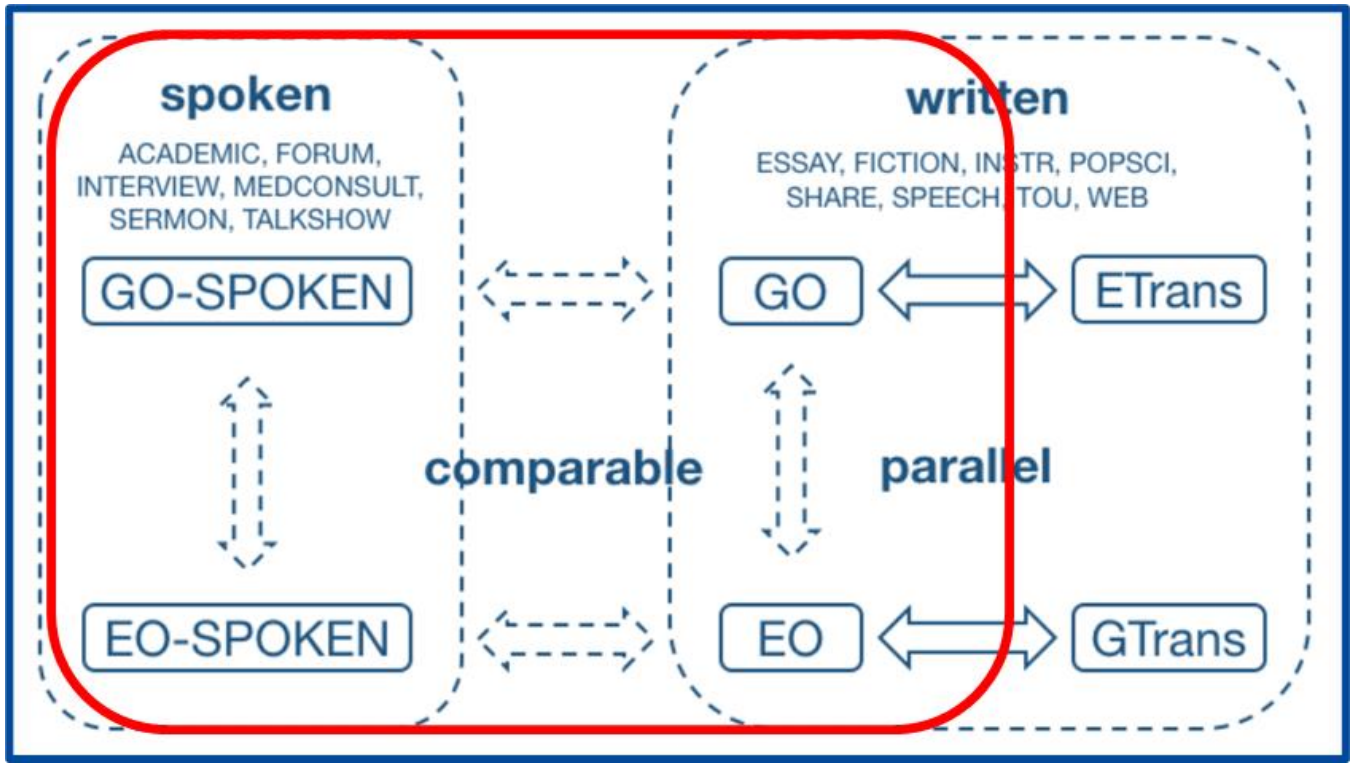


Figure 1: Composition of GECCo corpus (taken from https://fedora.clarin-d.uni-saarland.de/gecco/index.html last accessed 21.05.2023)

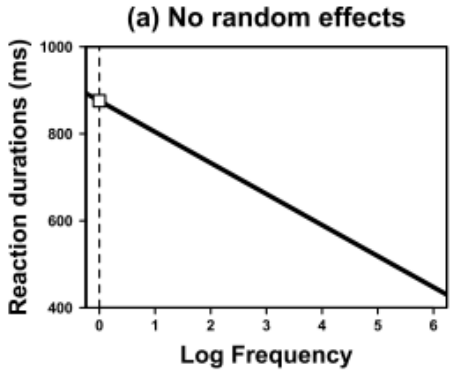
Statistical procedure

Bayesian mixed effects Poisson regression modelling

Graphs taken from Winter (2020)

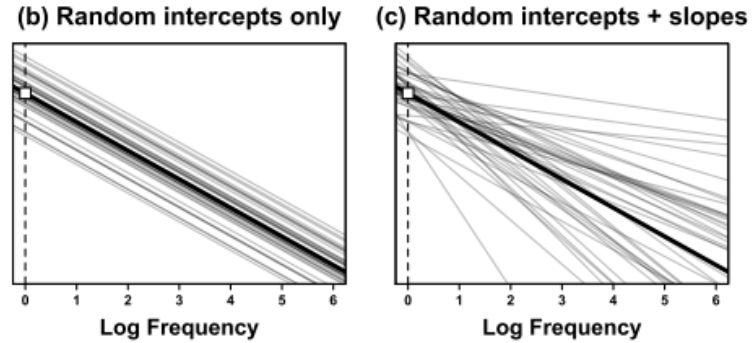
Regression modelling

$$y = b_0 + b_1 \cdot x$$

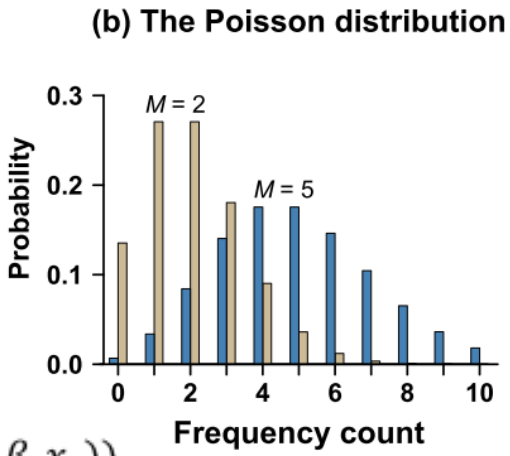


Mixed-effects regression modelling

$$y_i = \beta_{0i} + \beta_{1i}x_{1i} + \beta_2 x_2$$



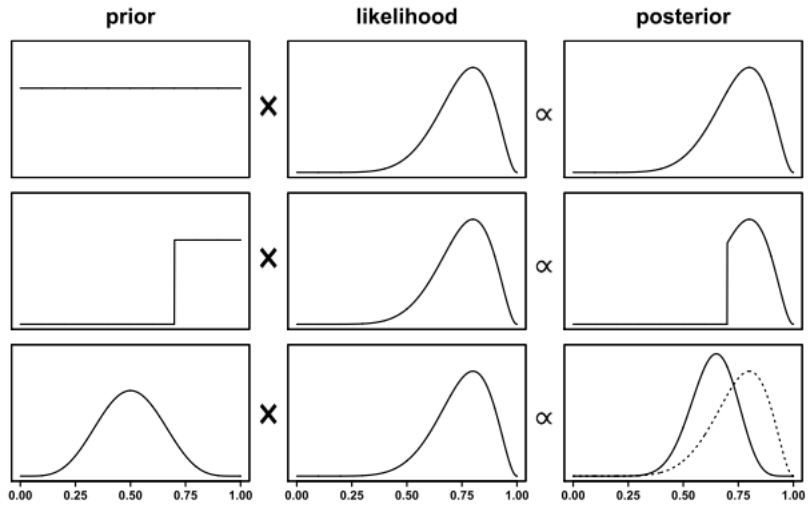
Mixed-effects Poisson regression modelling



$$y_i \sim Poisson(\exp(\beta_0 + \beta_1 x_1))$$

Bayesian Mixed-effects Poisson regression modelling

$$posterior \propto prior \cdot likelihood$$



Winter & Bürkner (2021)

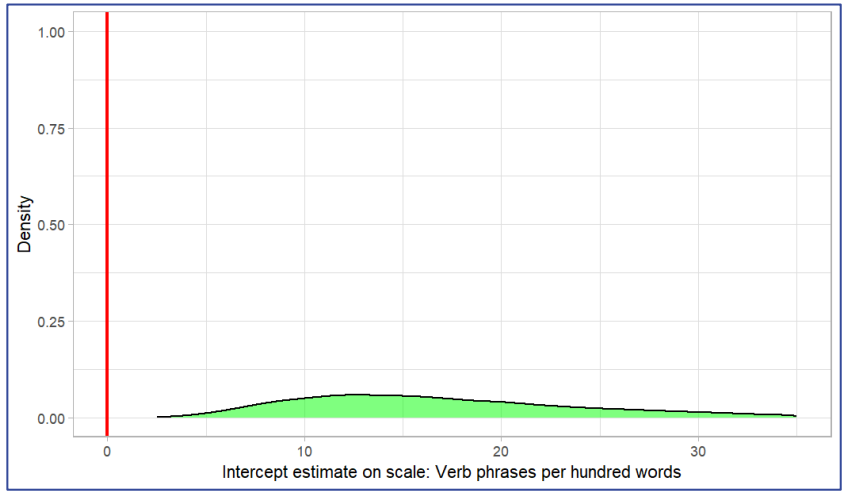
Statistical procedure

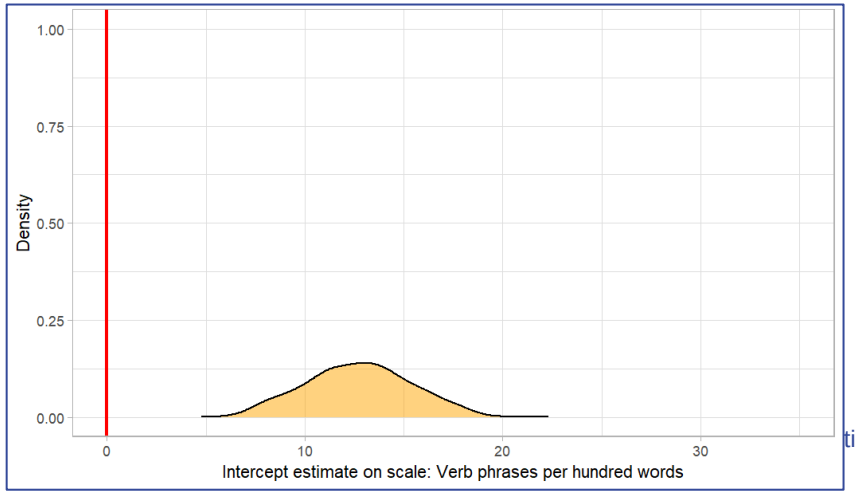
Bayesian mixed effects Poisson regression modelling

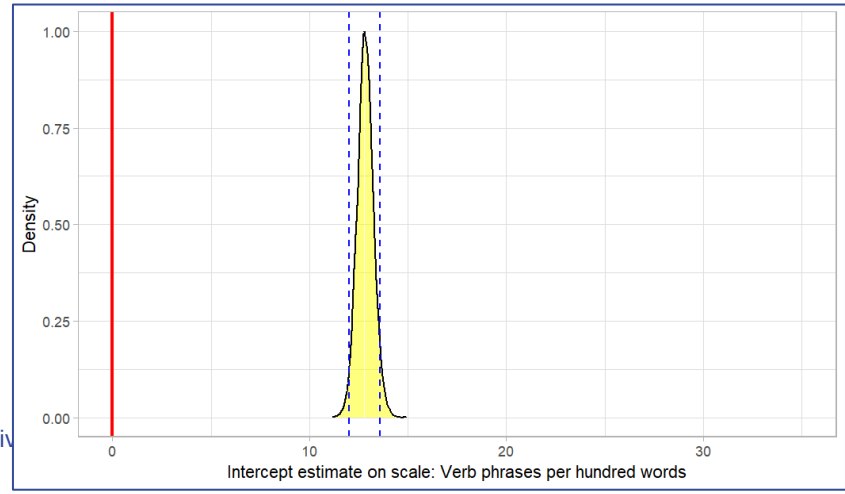
Assumptions about frequency of VPs

Observed frequency of VPs in GECCo

Prior distribution ✖ **Data** = **Posterior distribution**







Statistical procedure

Bayesian mixed effects Poisson regression modelling

What we want to predict

Count of verb phrases (for each text)

Count of non-finite verb phrases (for each text)

What we can use as predictors

- Language (binary, sum-coded, fixed effect)
- Mode (binary, sum-coded, fixed effect)
- Standardized type-token ratio (continuous, z-scored, fixed effect)
- Register (categorical, sum-coded, random effect)
- Text length (exposure variable)
- Interaction of mode and language
- Interaction of language and STTR

Statistical procedure

Bayesian mixed effects Poisson regression modelling

```
Family: poisson
Links: mu = log
Formula: NR_vp ~ 1 + Language + Mode + STTR_z + offset(log(NR_tokens_phw)) +
Language:Mode + Language:STTR_z + (1 | Register)
Data: texts (Number of observations: 372)
Draws: 3 chains, each with iter = 6000; warmup = 2000; thin = 1;
total post-warmup draws = 12000

Group-Level Effects:
~Register (Number of levels: 14)
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.12      0.02     0.08     0.17 1.00     3455     5334

Population-Level Effects:
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept             2.55      0.03     2.49     2.62 1.00     3379     4671
Language1             0.06      0.00     0.05     0.06 1.00    12256     7849
Mode1                 0.09      0.03     0.03     0.15 1.00     3325     4276
STTR_z               -0.04      0.01    -0.05    -0.03 1.00    11863     8254
Language1:Mode1       0.01      0.00     0.00     0.02 1.00    10535     8167
Language1:STTR_z     -0.01      0.00    -0.02    -0.00 1.00     9683     7379

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Figure 2: Model summary of regression model predicting the frequency of verb phrases.

Statistical procedure

Bayesian mixed effects Poisson regression modelling

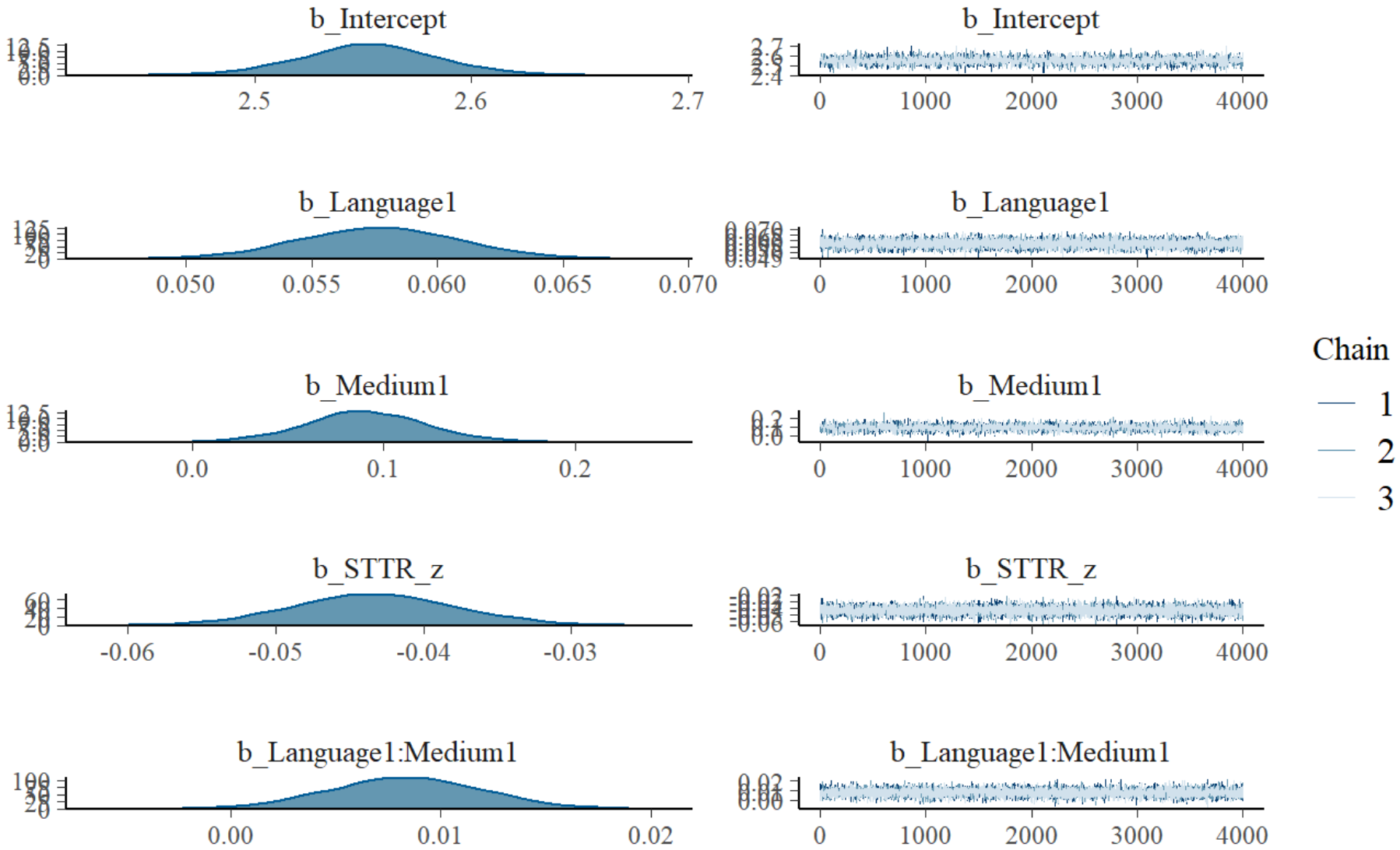


Figure 3: Visualisation of posterior estimates for regression model predicting the frequency of verb phrases.

Results

The contribution of non-finite verb phrases to the overall VP count

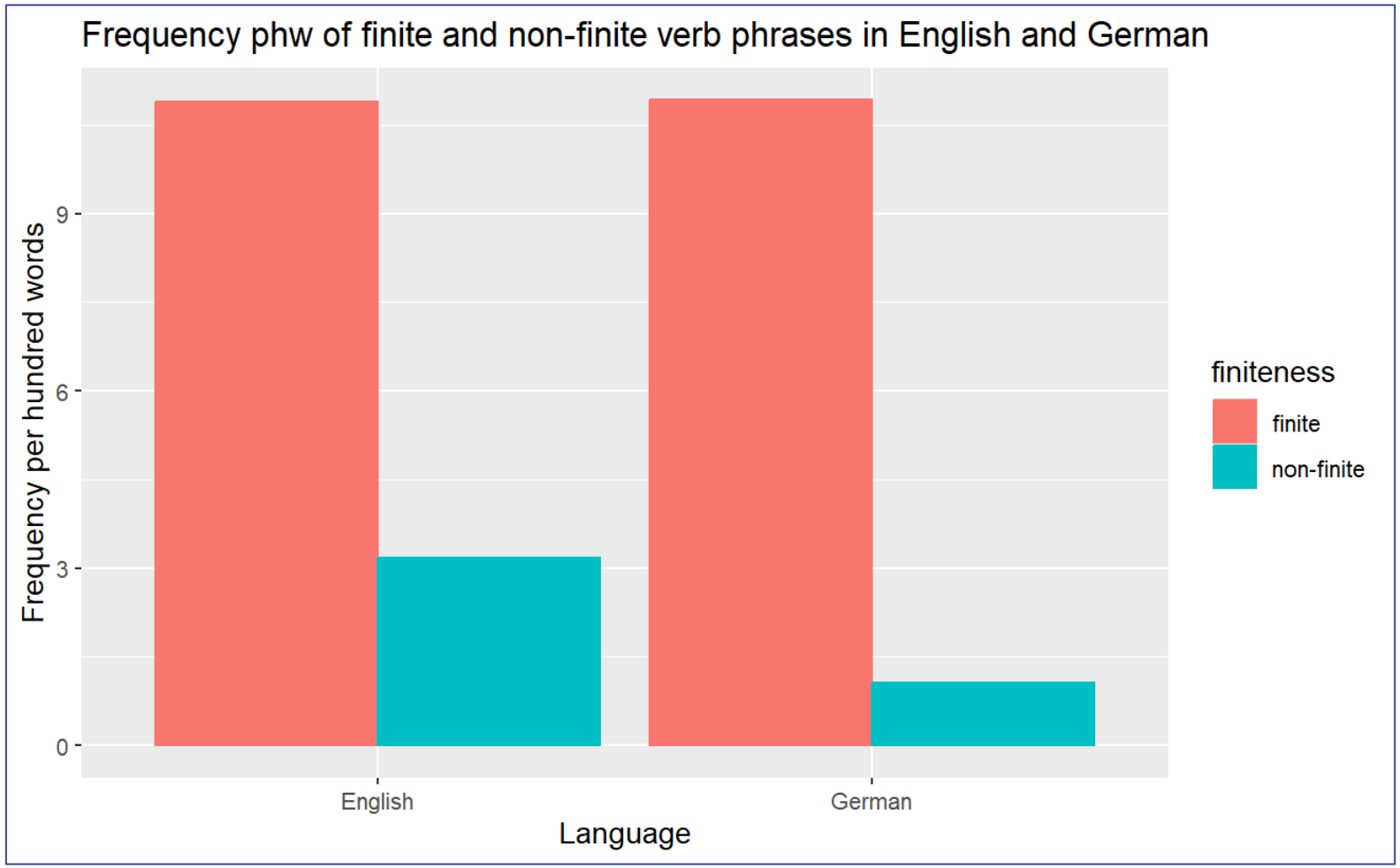


Figure 4: Frequency phw of finite and non-finite verb phrases in GECCo by language.

The most over-represented non-finite verb phrases in English

Verb form	Overt subject	Clause type and syntactic function	Frequency difference phw	Frequency phw English	Frequency phw German
to/*zu*-infinitive	no	nominal clause as object	0.49	0.71	0.22
present participle	no	embedded clause	0.44	0.48	0.04
to/*zu*-infinitive	no	adverbial clause	0.29	0.43	0.14
present participle	no	adverbial clause	0.19	0.20	0.01
to/*zu*-infinitive	no	nominal clause as subject complement	0.18	0.28	0.10
to/*zu*-infinitive	no	embedded clause	0.16	0.36	0.20
past participle	no	embedded clause	0.09	0.18	0.09

Table 1: Non-finite structures with a higher frequency phw in English compared to German.

Overall frequency of verb phrases in GECCo

Example: texts with highest ‘verbiness’

- (3) So I [don't know] whether you [want] [to go] and [see] her rather than, I [could get] a doctor [to go] and [see] her and [phone], (EO_MEDCONSULT_002)

- (4) "Danke, Juli. Und [geh] mal ins Bett, [hörst] du. [Ist] schon spät. Ich [schlaf] jetzt auch."
  (GO_FICTION_006)

Example: texts with lowest ‘verbiness’

- (5) The Einstein Planetarium [projects] images about space and astronomy onto a star-filled, domed ceiling. The Lockheed Martin IMAX Theater [shows] large-format films on a screen five stories high. (EO_WEB_008)

- (6) Mit Bus oder Bahn bequem zum Startpunkt einer Wanderung und abends wieder stressfrei zurück, das [ermöglichen] im Schwarzwald die öffentlichen Verkehrsmittel. (GO_TOU_014)

Verb phrases and mode

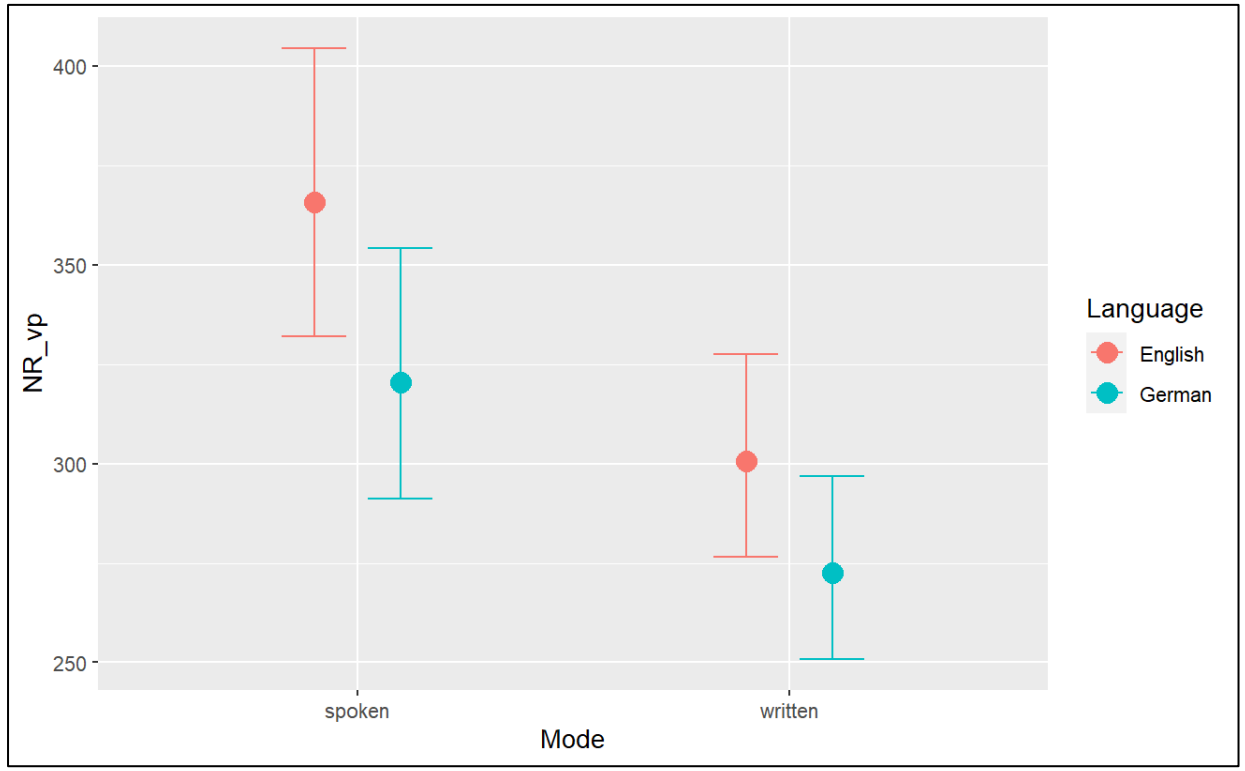


Figure 5: **Number of verb phrases** in a text of average length (2,439 words) by mode and language, as predicted by the regression model.

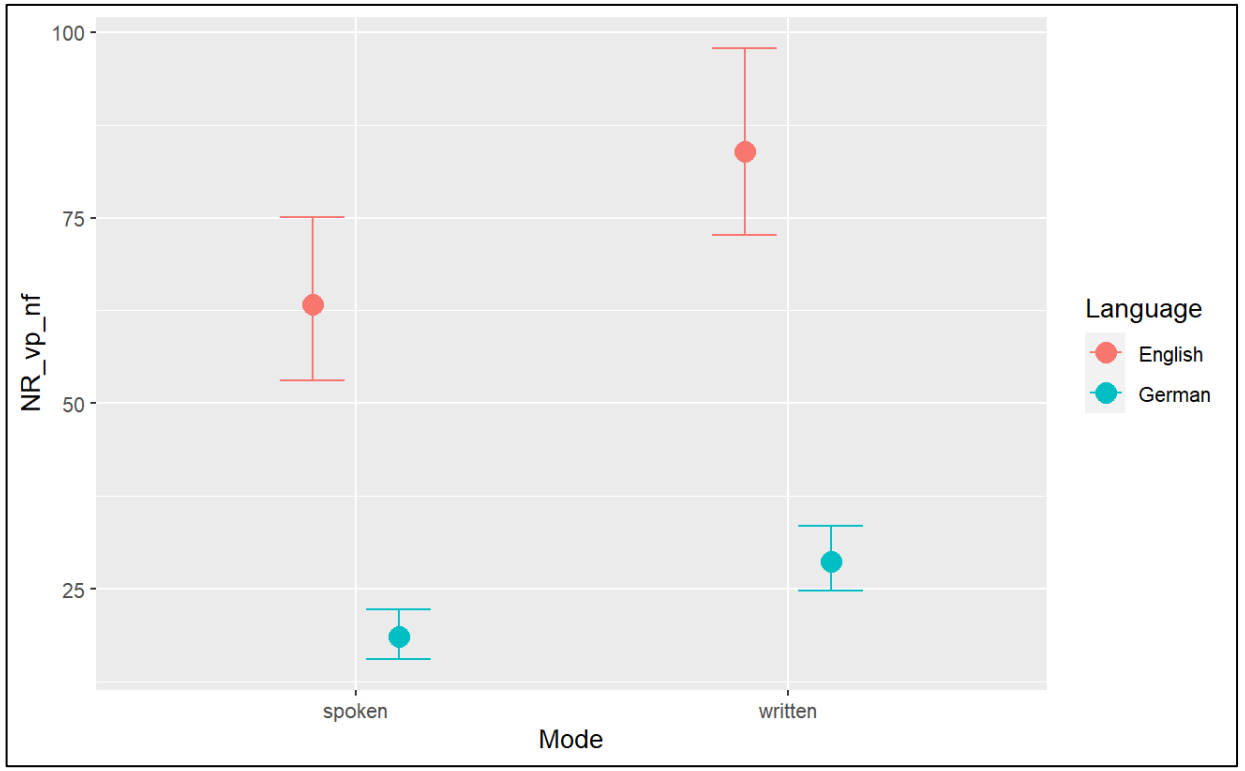


Figure 6: **Number of non-finite verb phrases** in a text of average length (2,439 words) by mode and language, as predicted by the regression model.

Verb phrases and type-token-ratio

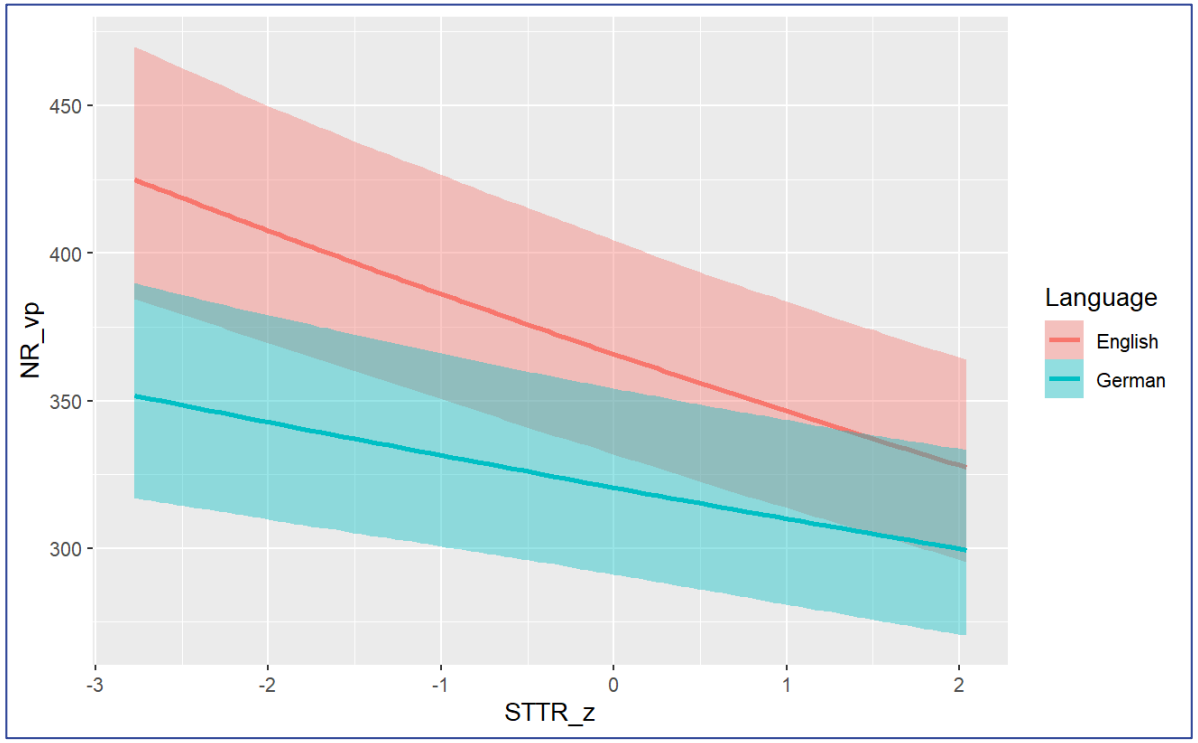


Figure 7: Number **of verb phrases** in a text of average length (2,439 words) by density (STTR_z) and language, as predicted by the regression model.

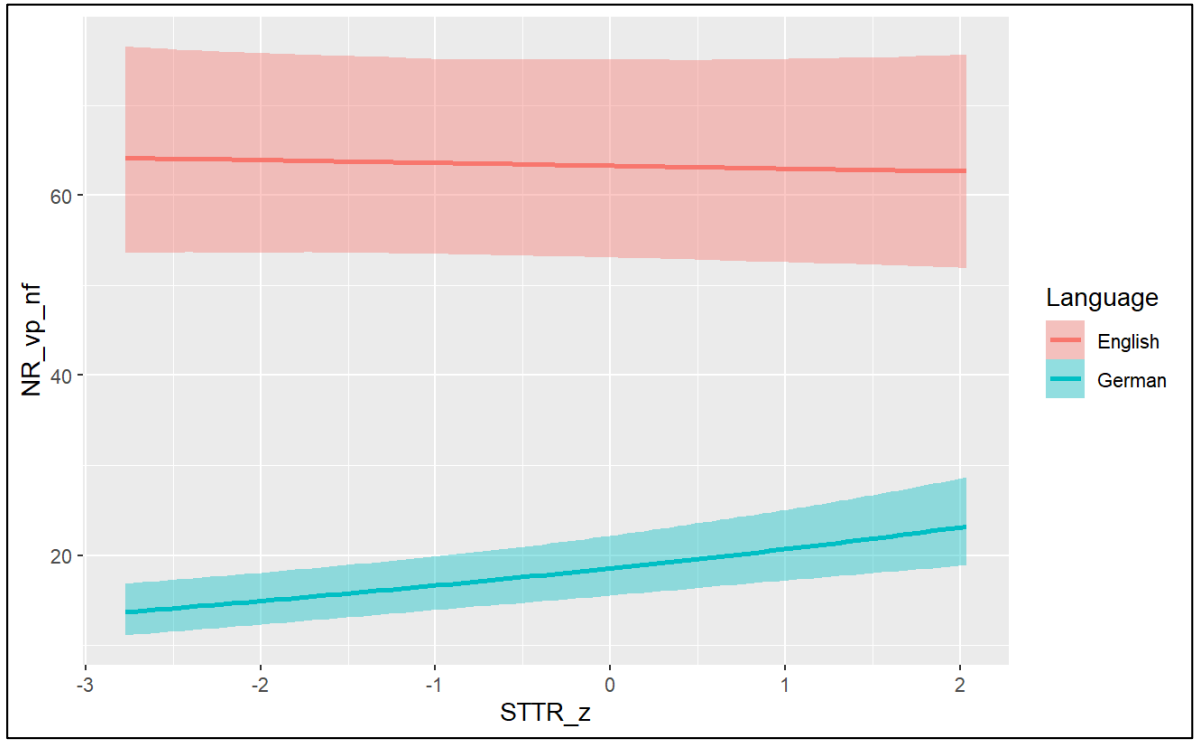


Figure 8: Number of **non-finite verb phrases** in a text of average length (2,439 words) by density (STTR_z) and language, as predicted by the regression model.

Verb phrases and type-token-ratio

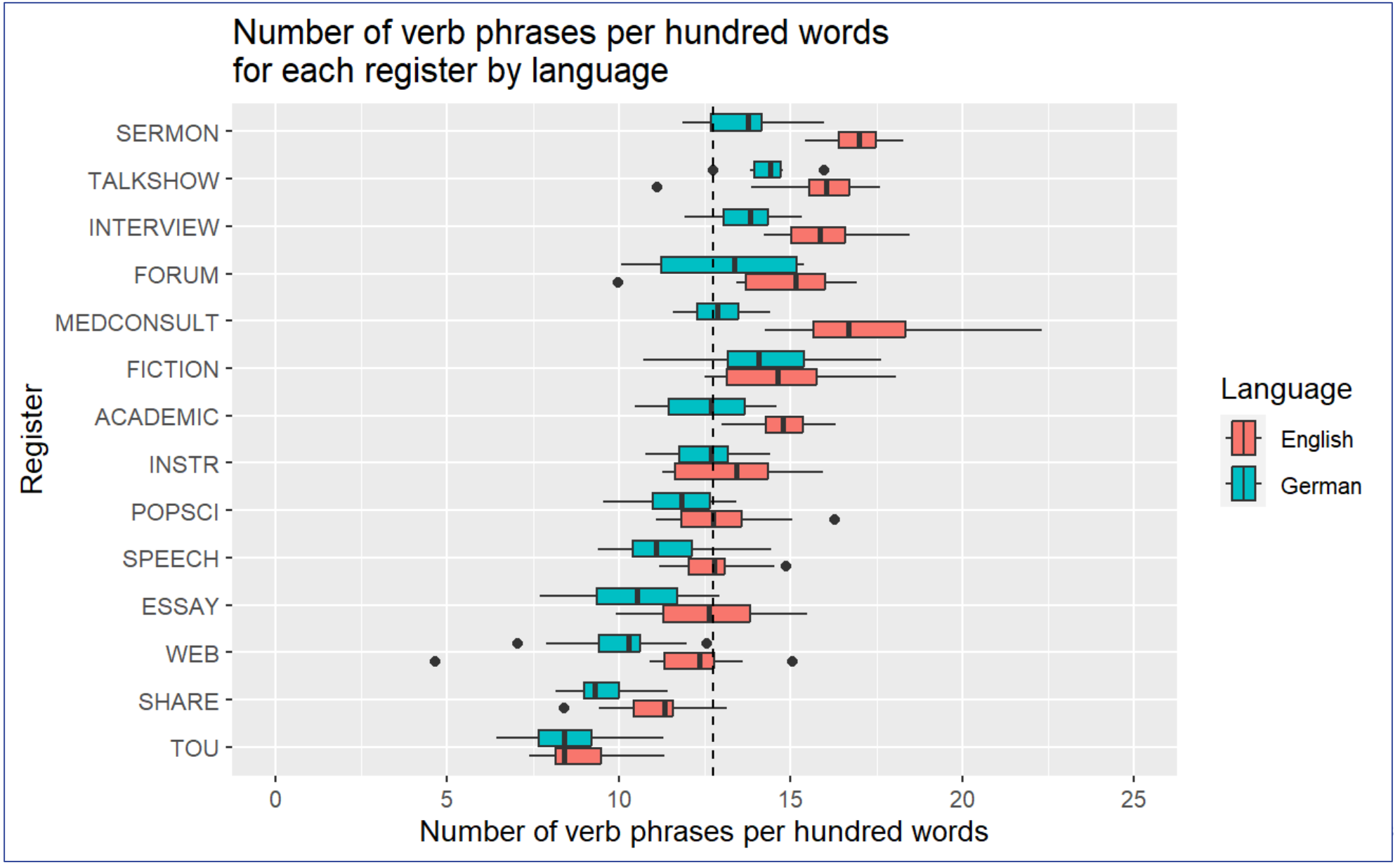


Figure 9: Relative frequency of verb phrases per register in GECCo, divided by language.

Main findings

- English uses **more verb phrases** and considerably **more non-finite verb phrases**. The non-finite verb phrases are most likely the **main reason** for the higher overall frequency of verb phrases in English.
- **Spoken language:** high frequency of verb phrases, lower frequency of non-finite verb phrases, higher frequency of the *to*/*zu*-infinitive and of nominal clauses. **Written language:** not as many verb phrases, more non-finite verb phrases, higher frequency of participles, adverbial clauses, and embedded clauses.
- English not only has more **grammatical options** for non-finite verb phrases, but also makes more frequent use of the options that are also available in German.
- Main factors impacting verbal style: language and mode. Information density only plays a minor role. Extensive variation by register.
- Correlation between **information density** and frequency of non-finite verb phrases: weak positive correlation in German, but there no apparent correlation in English.

Main findings

This all points to the conclusion that English non-finite verb phrases **penetrate all instances of language use** and could to an extent be called **grammatically determined**. In German, the use of non-finite verb phrases can be characterised as more '**multifactorial**', as they appear to be preferred in certain environments (written texts, high information density, certain registers).

Discussion

Language-specific conditions for verbal style: Differences in basic constituent order

- languages differ in their ways to achieve efficient realisations (e.g. Wurzel 2001: 385)
- Kortmann & Meyer (1992): German is (at least in parts) a verb-final language, the argument structure needs to be kept as simple as possible, but the arguments themselves can be expanded. English, as an SVO-language, can afford to introduce several argument-predicate structures (1992: 165)
- Doherty (1996: 452) calls the right-branching structure that we see in English “theoretically open-ended”. The basic constituent order in English is therefore more open to including clausal constituents (Doherty 1996: 452): “Attaching elements to the right means following the natural order of a right-branching language, like English”. In German, rendering constituents as clauses would, however, quickly over-burden the sentence bracket (Doherty 1996: 452).

Language-specific conditions for verbal style: Differences in basic constituent order

English: SVO, “right-branching”

(7) and he was thinking of these as atoms more or less and, uh *trying* to understand nature, through the possible arrangements that you can have, for spheres … (EO_ACADEMIC_001)

(8) But he does *try* to reassure them, himself as much as his two daughters. (EO_FICTION_005)

German: SOV, “left-branching”, sentence bracket

(9) die Antidiskriminierungsstelle *versucht* die Landespolitik aktiv zu beeinflussen (GO_TOU_021)

(10a) ja, und danach *hab*' ich dann *versucht*, [einen Job zu finden] (GO_INTERVIEW_007)

(10b) ? ja, und danach *hab*' ich dann [einen Job zu finden] *versucht*

But: Why non-finite subordinate clauses specifically?

So far, the argumentation would hold for finite and non-finite clauses…

Doherty (1996: 453) explains this tendency with the necessity to keep the original matrix clause “open”:

> the opening up of new clauses can be understood to signal the end of the preceding clause. Processing ease in an open-ended type of structure will, instead, promote indicators preventing the process of closing down. Thus, […] the preference for non-finite verb phrases could help to keep structures open while offering another verbal anchor for further structural extension (1996: 453).

(11) Angesichts […] war es im Interesse unserer Aktionäre nicht länger zu *akzeptieren*, weiterhin allein das gesamte Risiko zu tragen. (GO_SHARE_003)

(12) Wir mußten einen Weg finden, der es amerikanischen Richtern erlaubt zu *akzeptieren*, daß sich Regierungen […] auf eine komplexe Lösung geeinigt hatten […] (GO_SPEECH_009)

(13) or why the hunters, […], now *accept* to drag around with them over the tired land all the heavy guns […] (EO_FICTION_005)

(14) Rather, we need to *accept* that the Treaty is fundamentally in conflict with […] (EO_SPEECH_003)

“Grammatical determinism” of non-finite verb phrases in English?

Is the relationship between matrix verb and non-finite verb phrase more grammaticalized in English than in German? Are non-finite clauses in post-predicate position more grammatically determined in English?

Potential arguments:

- Vast areas of “intermediate verbs” in English, but not in German.
- No relationship between information density and non-finite verb phrases in English → not distributed stylistically
- More English matrix verbs allowing or requiring non-finite clauses (Mair 1990)
- The sheer number of *to*-infinitives serving as object clauses in English compared to German

Counterargument: this only concerns nominal clauses. Embedded clauses and adverbial clauses also play a role.

Conclusion

Conclusion

Main findings:

- Existing assumptions about use of verb phrases (and non-finite verb phrases) are overall confirmed
- English deserves its title as a language that is more verbal than German
- Basic constituent order seems like the likely reason for the higher frequency of verb phrases and non-finite verb phrases in English.
- Non-finite verb phrases are deeply entrenched in English grammar, may be called “grammatically determined”
- Language-internal variation is important to take into account
- Importance of looking at spoken material (this is where language-specific features shine)
- From “contrastive grammar” (system-based comparison) to “contrastive grammar in use”: quantitative, corpus-based contrastive linguistics (Gast 2015: 5)

Conclusion

Reasons to be sceptical:

- Comparability of corpus components
- Corpus representativeness
- Combination of automatic and manual coding does not lead to 100 percent accuracy
- Difficulties in clearly delineating verb phrases
  - Verbs vs. nouns
  - Auxiliary verbs vs. main verbs
  - Participles vs. adjectives

Reasons to be confident:

- Results are overall in line with existing research and assumptions
- Regression model is very sure of its estimates
- More methodological caution than previous studies (no translation effects, more data, not only count of PoS-tags, extensive quality control)

Thank you for your attention.

Hanna Mahler, M.A.

Albert-Ludwigs-Universität Freiburg

hanna.mahler@students.uni-freiburg.de

References

Berg, Thomas. 2017. Compounding in German and English: A Quantitative Translation study. *Languages in Contrast* 17 (1). 43–68.

Berg, Thomas; Helmer, Sabine; Neubauer, Marion; Lohmann, Arne (2012): Determinants of the Extent of Compound Use. A Contrastive Analysis. In *Linguistics* 50 (2), pp. 269–303. DOI: 10.1515/ling-2012-0010.

Doherty, Monika. 1996. Introduction. *Linguistics* 34. 441–457.

Fischer, Klaus. 2013. *Satzstrukturen im Deutschen und Englischen: Typologie und Textrealisierung* (Konvergenz und Divergenz 1). Berlin: Akademie Verlag.

Friederich, Wolf. 1969. *Technik des Übersetzens: Englisch und Deutsch*. Ismaning: Max Hueber.

Gast, Volker. 2015. On the Use of Translation Corpora in Contrastive Linguistics: A Case Study of Impersonalization in English and German. *Languages in Contrast* 15(1). 4–33.

Kortmann, Bernd & Paul Meyer. 1992. Is English Grammar more Explicit than German Grammar, after all? In Christian Mair & Manfred Markus (eds.), *New Departures in Contrastive Linguistics: Proceedings of the Conference Held at the Leopold-Franzens-Universität Innsbruck, Austria, 10-12 May 1991* (Innsbrucker Beiträge zur Kulturwissenschaft / Anglistische Reihe 4), Innsbruck: Verlag des Institutes für Sprachwissenschaft. 155–166.

Königs, Karin. 2004. *Übersetzen Englisch-Deutsch: Ein Systematischer Ansatz*, 2nd edn. (Lehr- und Handbücher zu Sprachen und Kulturen). München: Oldenbourg Wissenschaftsverlag.

Kunz, Kerstin, Ekaterina Lapshinova-Koltunski, José M. Martínez Martínez, Katrin Menzel & Erich Steiner. 2021. *GECCo - German-English Contrasts in Cohesion: Insights from Corpus-Based Studies of Languages, Registers and Modes* (Trends in linguistics. Studies and monographs 355). De Gruyter Mouton.

Mair, Christian. 1990. *Infinitival Complement Clauses in English: A Study of Syntax in Discourse*. Cambridge: Cambridge University Press (Studies in English Language).

Neumann, Stella. 2020. Is German more Nominal than English?: Evidence from a Translation Corpus. In Renata Enghels, Bart Defrancq & Marlies Jansegers (eds.), *New Approaches to Contrastive Linguistics: Empirical and Methodological Challenges*. Berlin, Boston: De Gruyter. 127–158.

Steiner, Erich. 2012. A Characterization of the Resource Based on Shallow Statistics. In Silvia Hansen-Schirra, Stella Neumann & Erich Steiner (eds.), *Cross-Linguistic Corpora for the Study of Translations: Insights from the Language Pair English-German* (Text, Translation, Computational Processing 11), 71–89. Berlin, Boston: De Gruyter Mouton.

Winter, Bodo. 2020. *Statistics for Linguists: An Introduction Using R*. New York: Routledge.

Winter, Bodo & Paul-Christian Bürkner. 2021. Poisson Regression for Linguists: A Tutorial Introduction to Modelling Count Cata with brms. *Language and Linguistics Compass* 15(11). 1–23.

Wurzel, Wolfgang U. 2001. Ökonomie. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language Typology and Language Universals: An International Handbook.* Volume 1, 384-400. Berlin, New York: Walter de Gruyter.

Delimiting verb phrases

German

- Counted as one verb phrase:
  - Auxiliary verb + non-finite verb form (*haben, sein, werden, bekommen, kommen,* …)
  - Modal verbs + non-finite verb form (*werden, dürfen, können, müssen, brauchen, …*)
- Counted as two verb phrases:
  - Full verbs + infinitive form (*gehen, fahren, lehren, lernen, hoffen, beginnen*, …)
  - Verbs of perception and causation + non-finite verb form (*sehen, hören, lassen*, …)

English

- Counted as one verb phrase:
  - Central modals + non-finite verb form (*can, may, might, should, will, would, mus*t, …)
  - Marginal modals + non-finite verb form (*dare, need, ought to, used to*)
- Counted as two verb phrases:
  - Full verb + non-finite verb form (*hope, wish, want, like, decide,* …)
  - “Intermediate verbs”
    - Modal idioms + non-finite verb form (*had better, be to, have got to, would rather*)
    - Semi-auxiliaries (*have to, be about to, be going to, be likely to,* …)
    - Catenatives + non-finite verb form (*happen to, tend to, seem to, come to,* …)

Quality of the annotation of sentence boundaries in the GECCo corpus

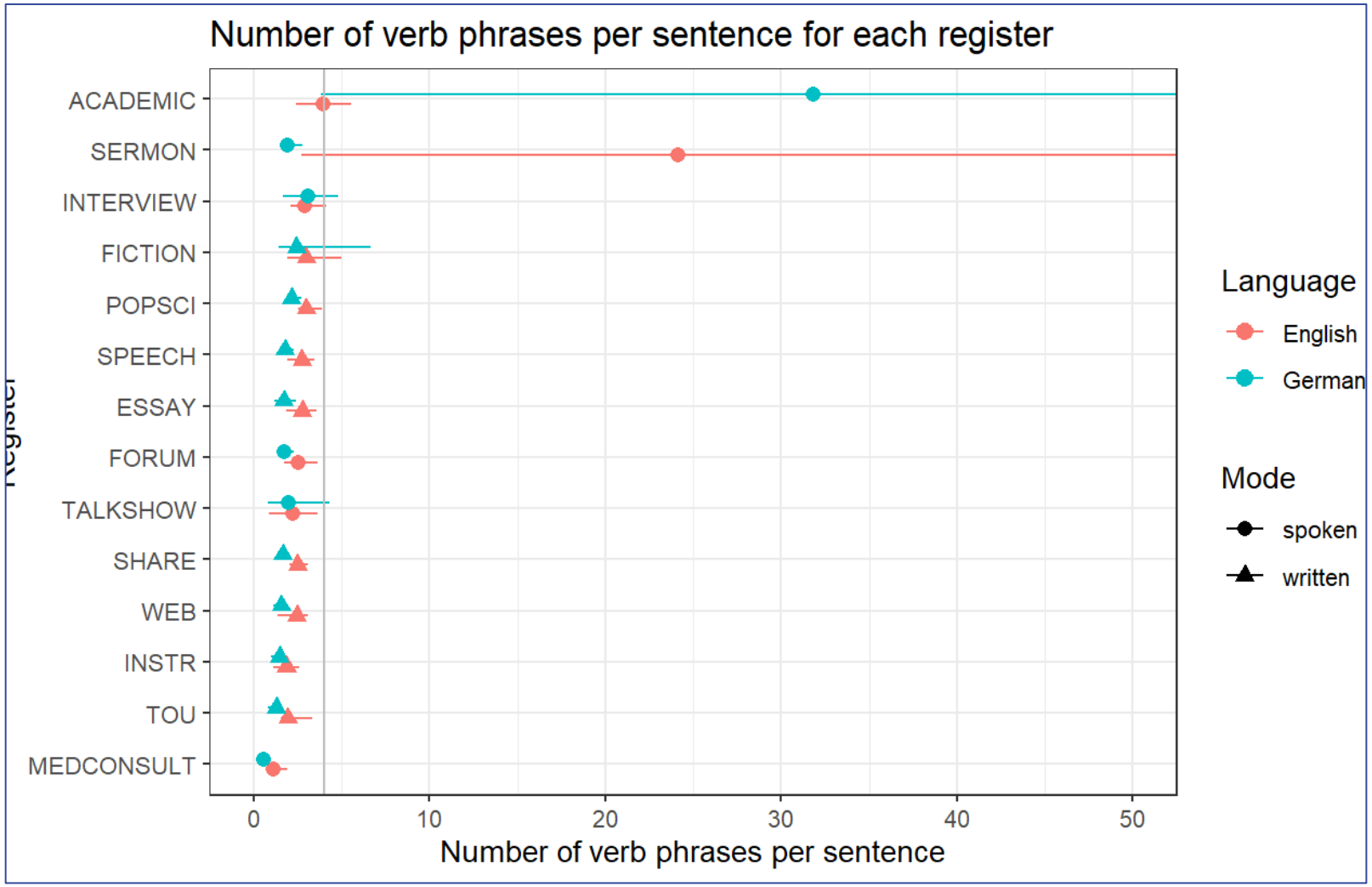


Figure 10: Frequency of verb phrases per sentence in GECCo corpus.

Clustering solution of registers in GECCo based on frequency of finite and non-finite verb phrases

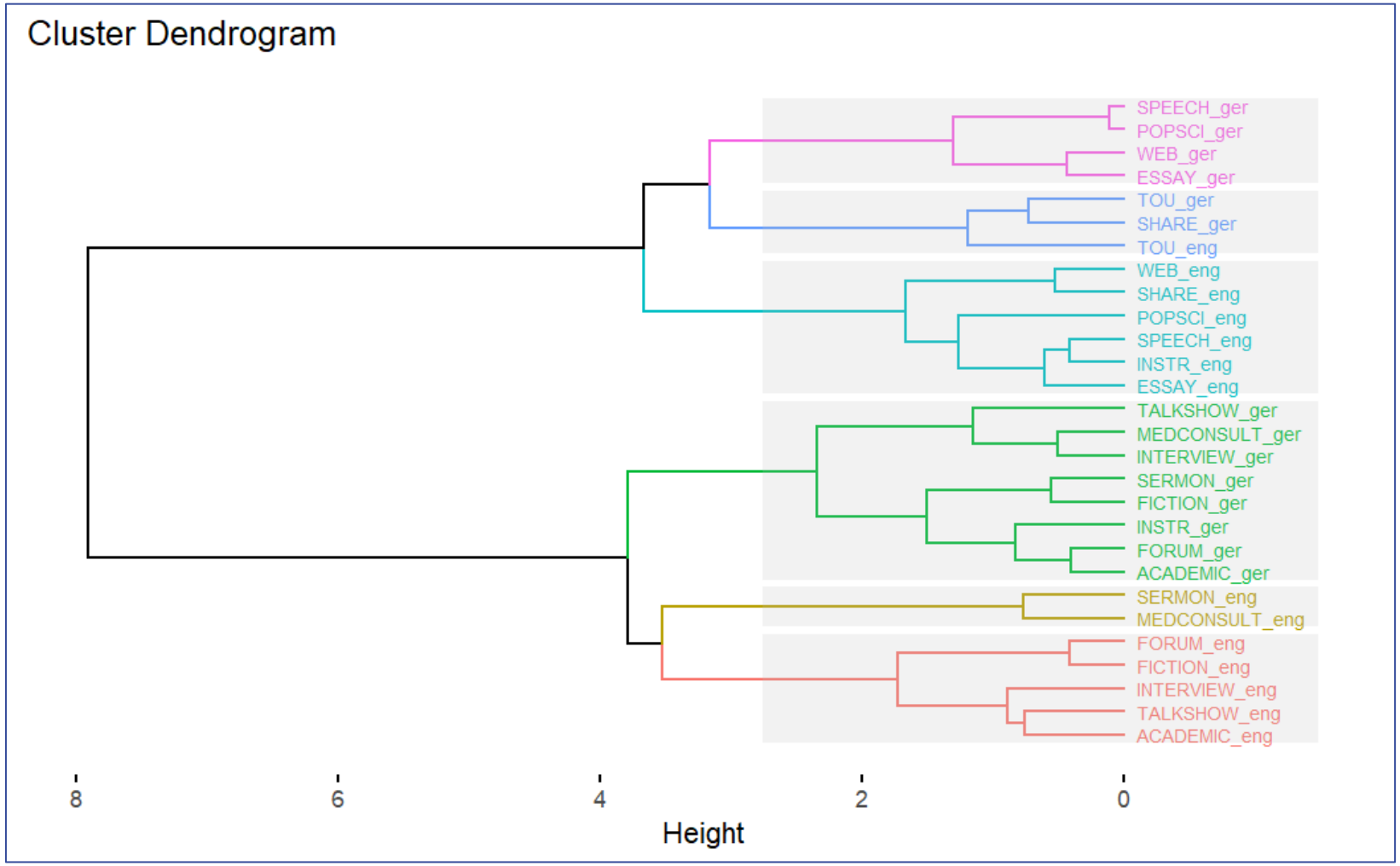


Figure 11: Cluster solution for registers in GECCo.