universität freiburg

https://github.com/HannaMahler/ICLC2023_Mahler_material

The Use of Verb Phrases in English and German

A Quantitative Case Study Using Comparable Corpus Data

Hanna Mahler, M.A.
Albert-Ludwigs-Universität Freiburg
Mannheim, 19.07.2023

Agenda

https://github.com/HannaMahler/ICLC2023_Mahler_material

1. State of research on verb phrase use
2. Hypotheses
3. Methodology & data
4. Results
5. Discussion
6. Conclusion

State of research on verb phrase use

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)**
  - Steiner (2012): written part of GECCo 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 more nominal

Hypotheses

Hypotheses

- **Hypothesis 1: On the whole, English uses a more verbal style than German.**
- **Hypothesis 2: The degree of verbal style within each language varies by register and mode.**
- **Hypothesis 3: Cross-linguistically, there is a correlation between verbal style and information density within both languages.**

Verbal style is operationalised as number of verb phrases per one hundred words
Information density is operationalised as standardised type-token ratio.

Methodology & data

The corpus: German-English Contrasts in Cohesion (GECCo)
Kunz et al. (2021)

- Translation corpus and comparable corpus
- Queries via CQP, no public full-text access
- 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

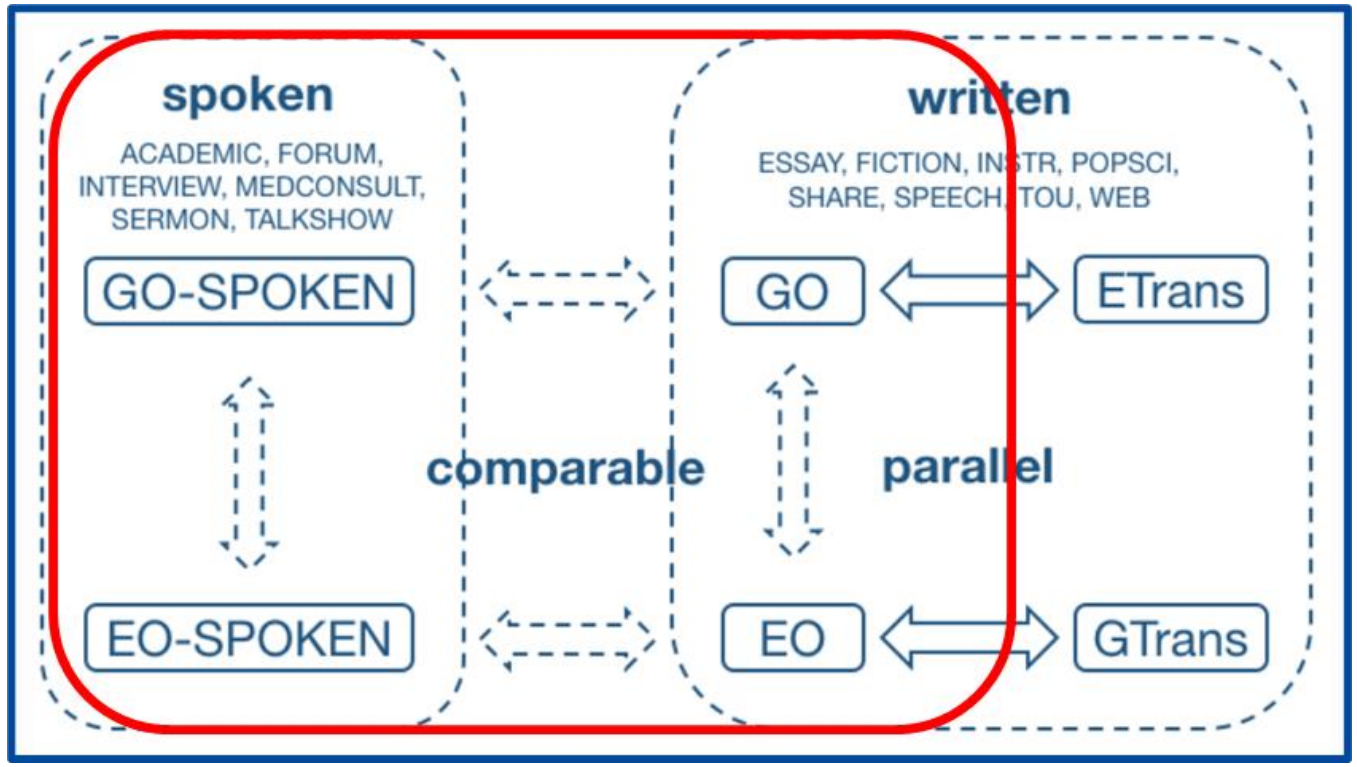


Figure 1: Composition of GECCo corpus (taken from https://fedora.clarin-d.uni-saarland.de/gecco/index.html last accessed 21.05.2023)

Statistical procedure

Bayesian mixed effects Poisson regression modelling

What we want to predict

Frequency of 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

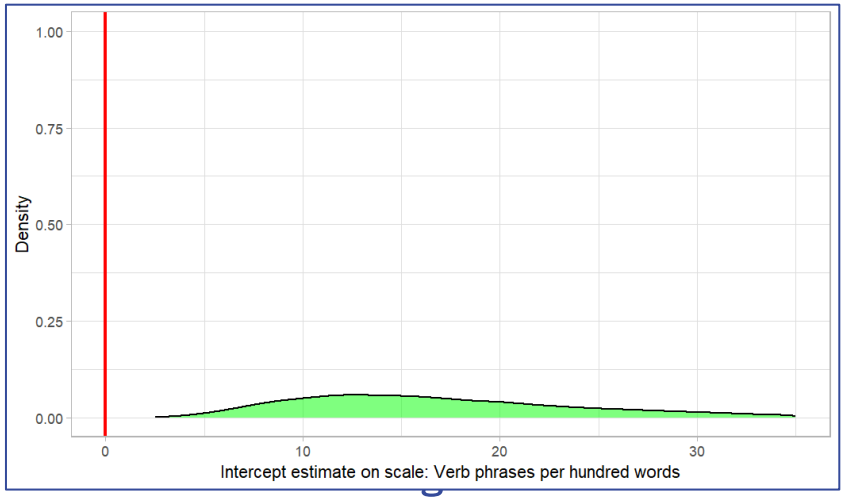
Statistical procedure

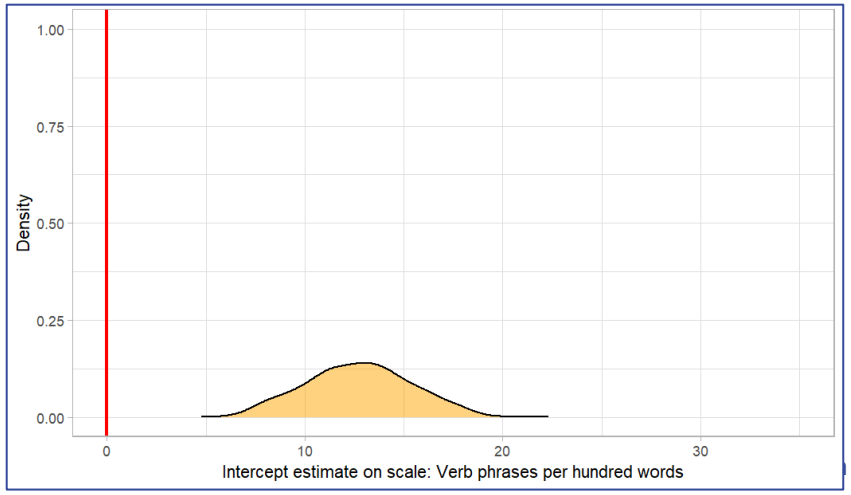
Bayesian mixed effects Poisson regression modelling

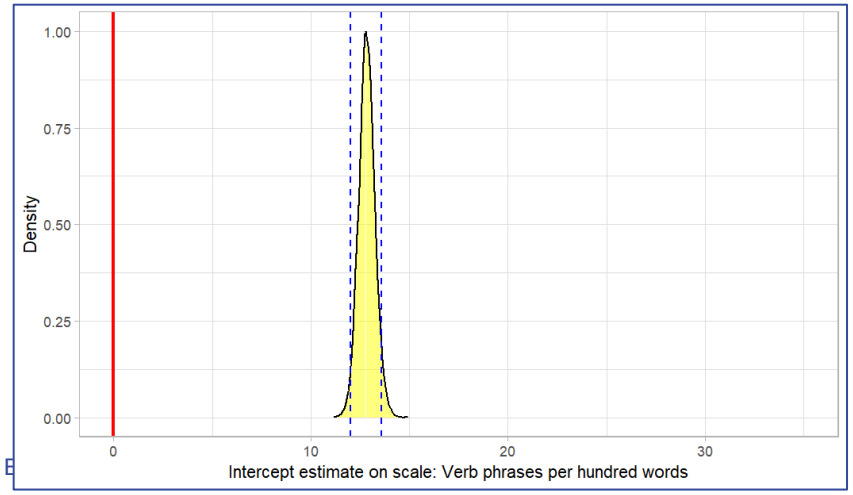
Assumptions about frequency of VPs

Observed frequency of VPs in GECCo

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







Methodological framework:
Quantitative, corpus-based contrastive linguistics

- Lack of quantitative methods mentioned by Gast (2013: 5), see also comments in Gries et al. (2020)
- Examples of recent work:
  - Neumann (2020): mixed-effects regression
  - Gries et al. (2020): inference trees, random forests, clustering, network analysis
  - Gast & Levshina (2014): correspondence analysis
  - Gast (2015): regression

The contributions of statistical methods to contrastive linguistics:

- Focus on relative preferences of languages for certain construction
- Shed light on varying importance of factors influencing the choice of construction in language A and B
- Move from "**contrastive grammar**" (comparison of decontextualised system of choices) to "**contrastive grammar in use**" (choices made in textual contexts)

Results

Overall frequency of verb phrases in GECCo

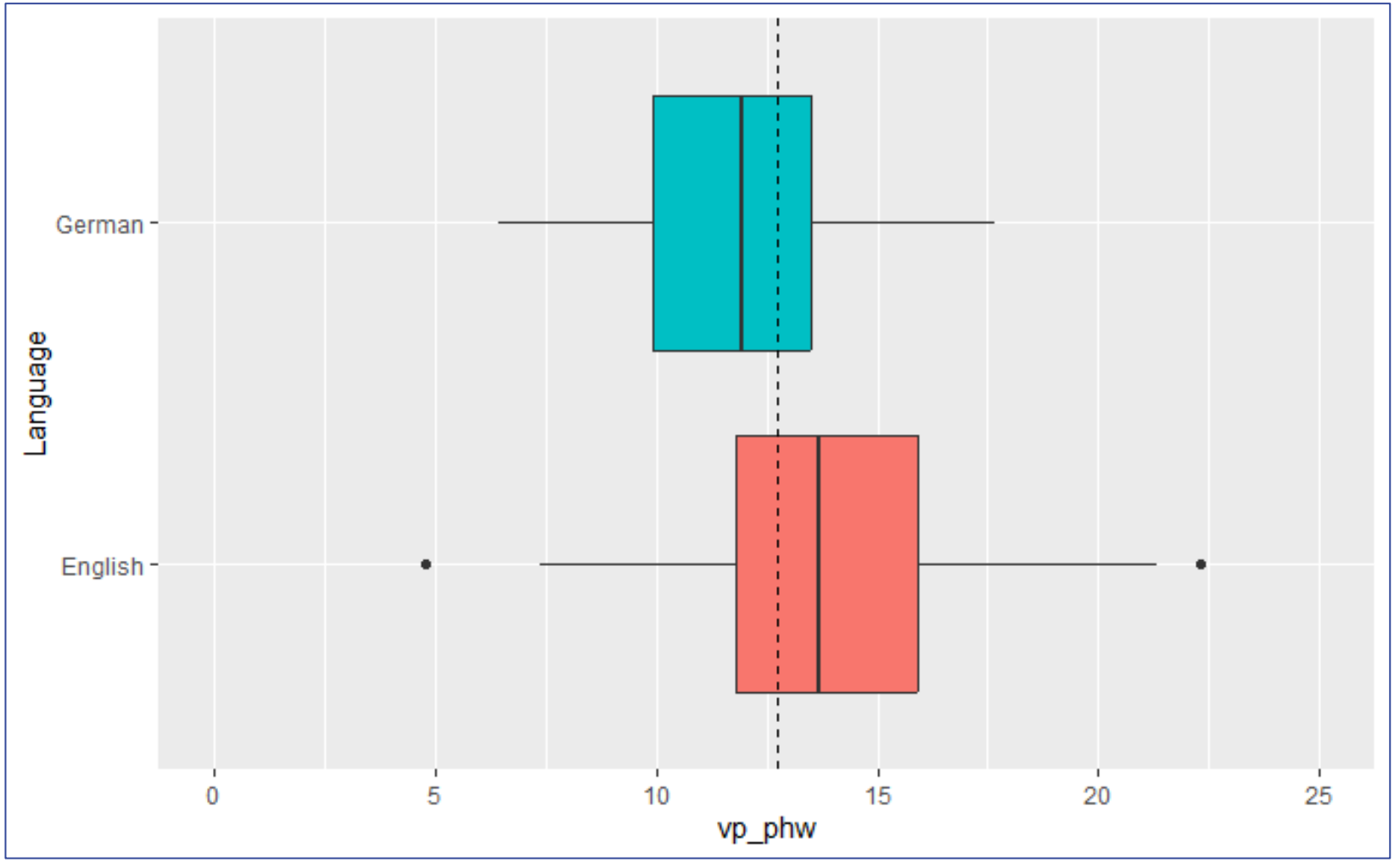


Figure 2: Verb phrase frequency phw in GECCo corpus by language.

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)

Distribution of verb phrases by mode

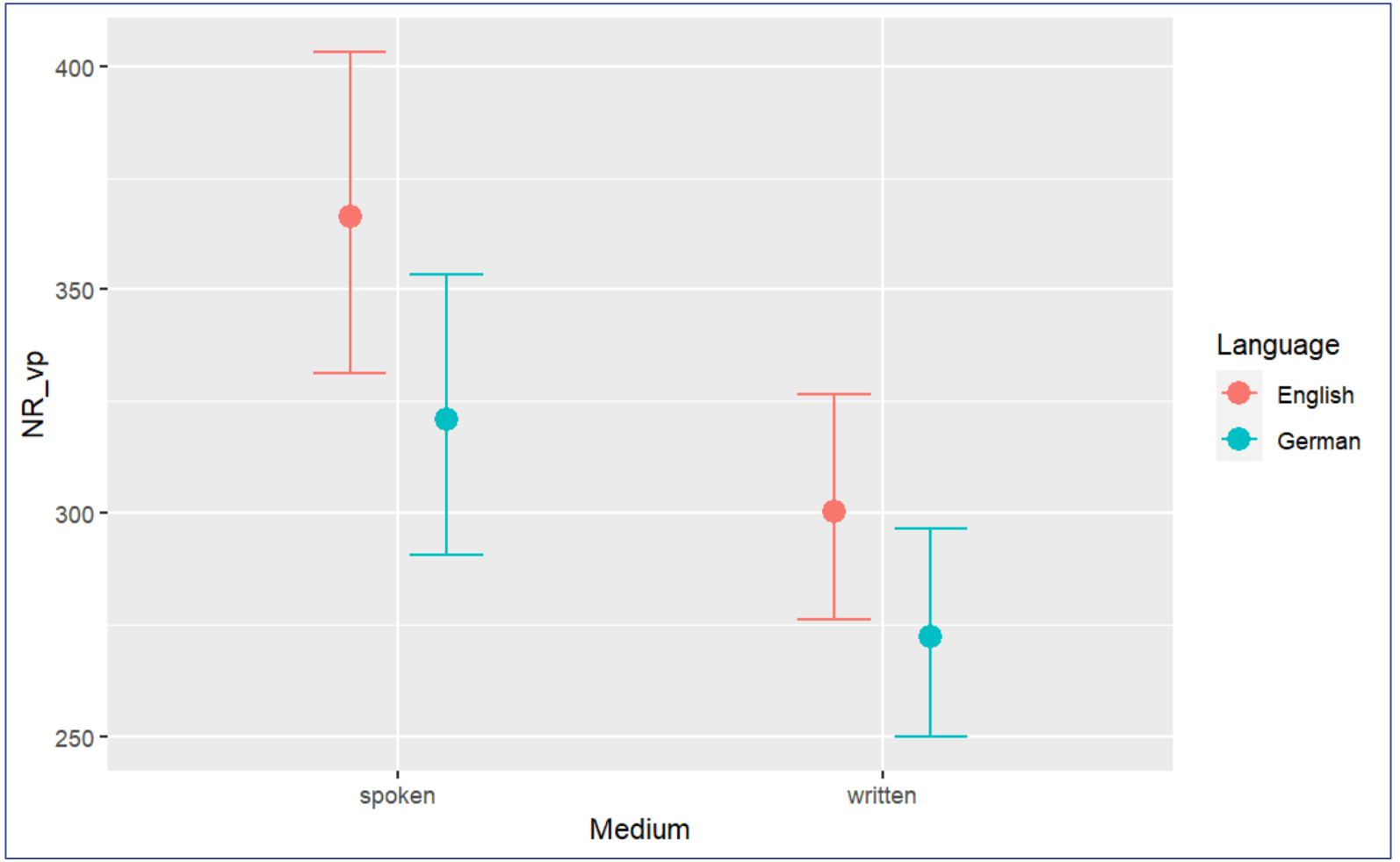


Figure 3: Number of verb phrases in a text of average length (2,439 words) by mode and language, as predicted by the regression model.

Verb phrases and type-token-ratio

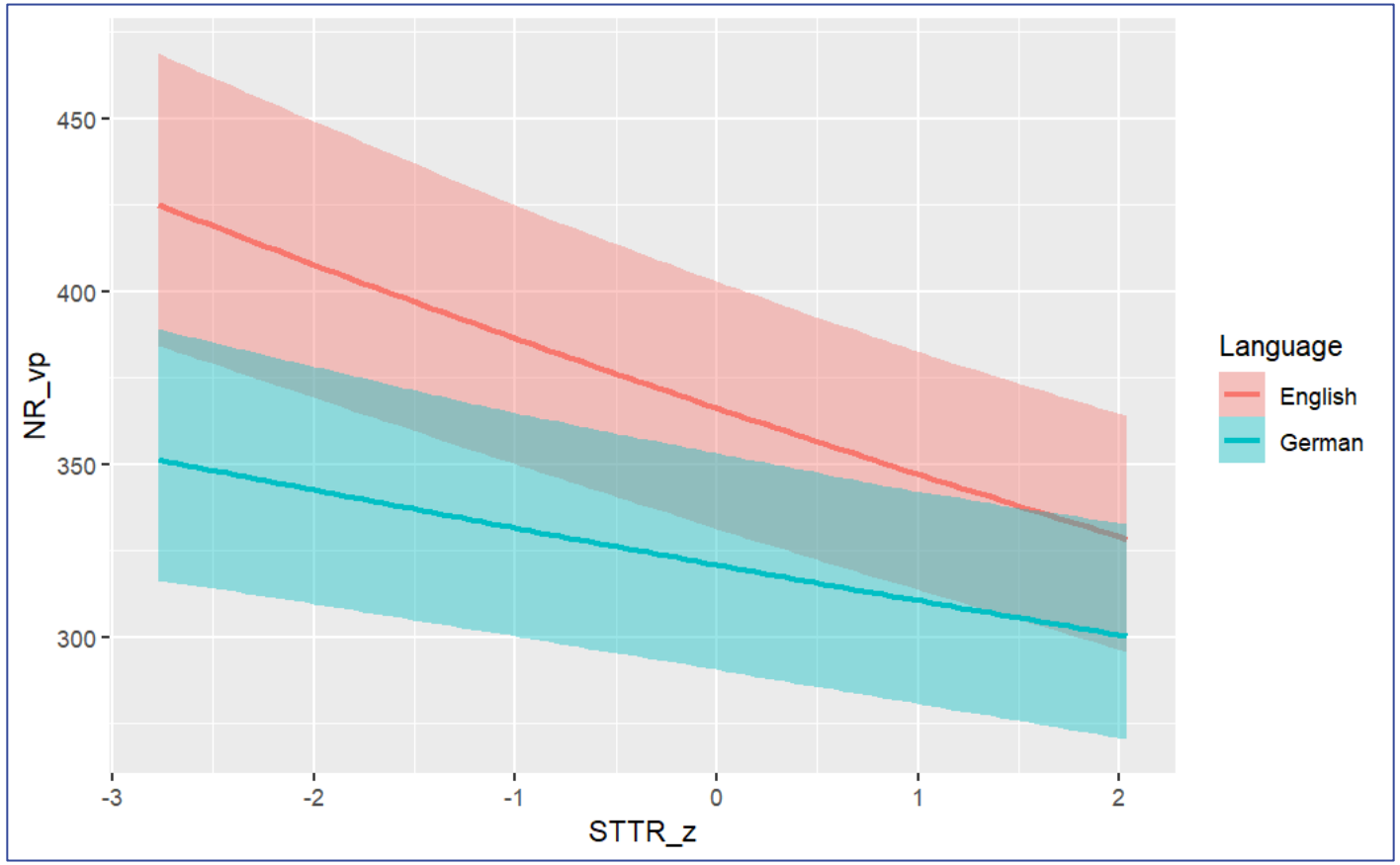


Figure 4: Number of verb phrases in a text of average length (2,439 words) by density (STTR_z) and language, as predicted by the regression model.

Verb phrases and type-token-ratio

Example of a text with high information density and low ‘verbiness’

(7) Allein 17 von 20 Qualitäts-Signets "Familienferien" in Baden-Württemberg [schmücken] Orte im Schwarzwald: Baiersbronn, Feldberg, Lenzkirch, Lossburg, Ottenhöfen, Sasbachwalden, Seebach, Schluchsee, Schonach, Schönwald, Titisee-Neustadt, Todmoos [sic], Todtnauberg, Unterkirnach. Ein weiteres den Europapark Rust. (GO_TOU_013)

Example of a text with low information density and high ‘verbiness’

(8) Oh look Natalie.

Behave, come on.

Natalie look, look.

Oh Look, it's just a wee light.

Look at the wee light.

Look it's just a wee light.

(EO_MEDCONSULT_014)

Distribution of verb phrases by register

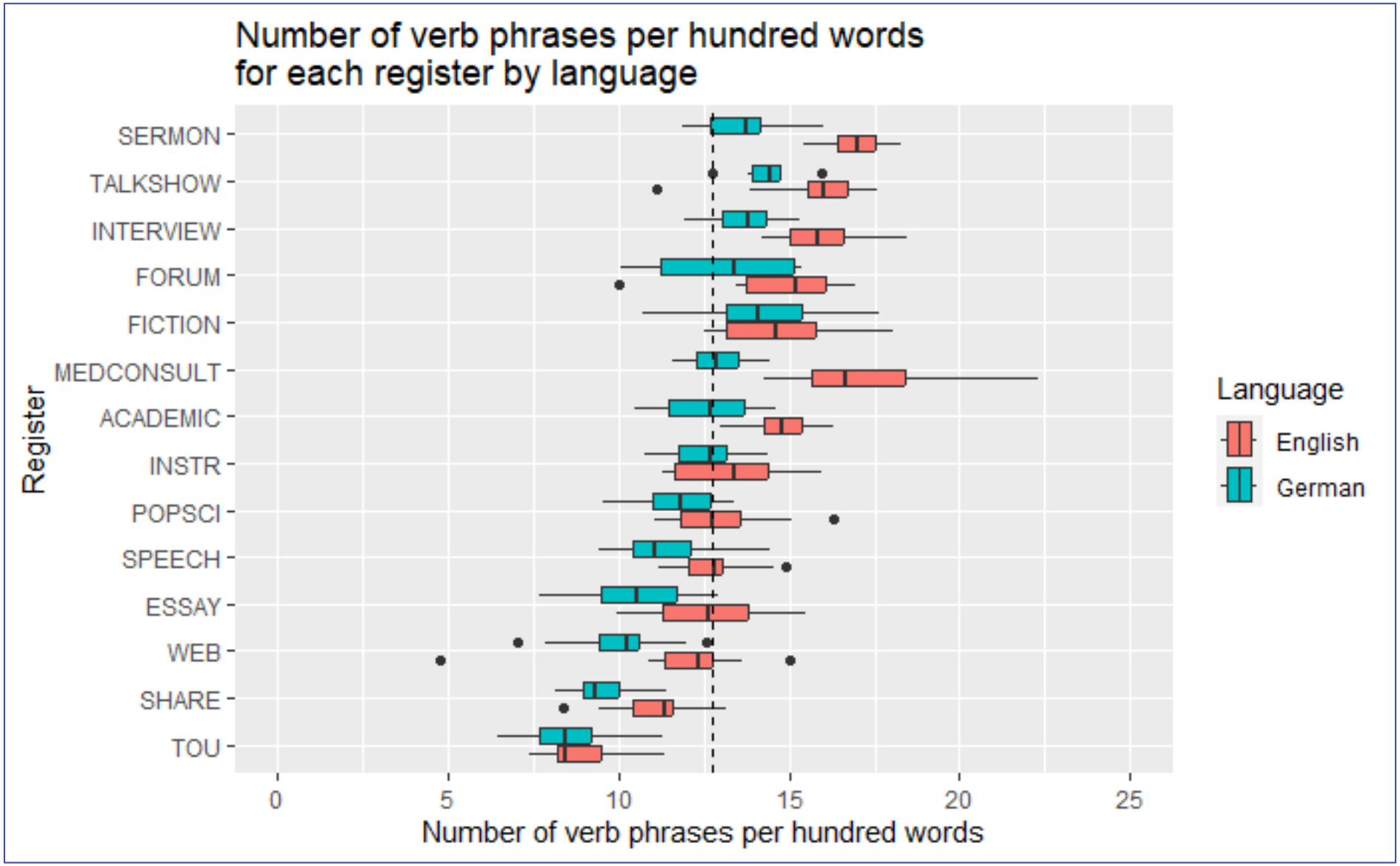


Figure 5: Number of verb phrases phw by language and register in GECCo.

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

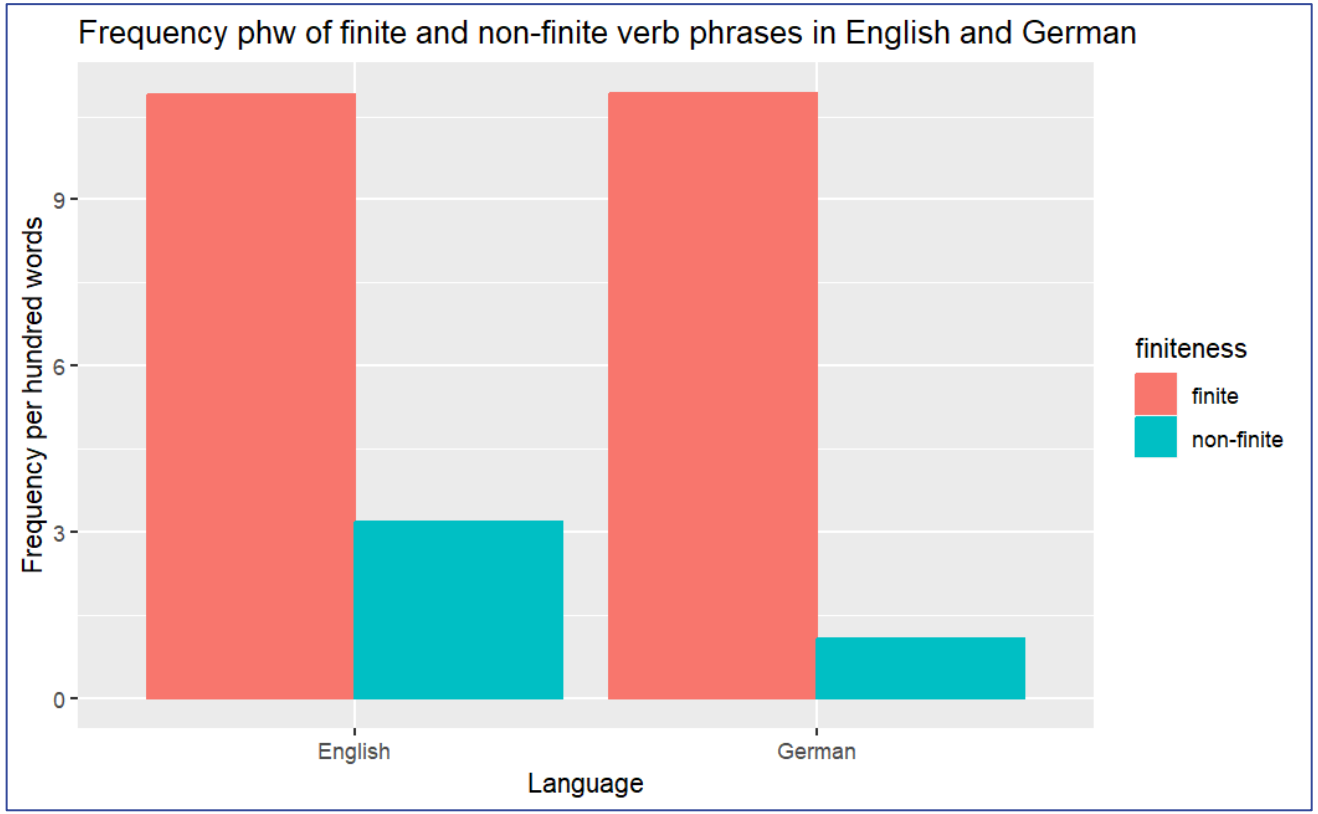


Figure 7: Frequency phw of finite and non-finite verb phrases in GECCo by language.

		English	German
All verb phrases	Total	68,658	50,289
	phw	13.74	11.72
Finite verb phrases	Total	53,150	45,827
	Percentage	77.4%	91.1%
Non-finite verb phrases	Total	15,508	4,462
	Percentage	22.6%	8.9%

Table 4: Total and relative verb phrase frequency in GECCo corpus by language, finite and non-finite.

Summary

Hypothesis 1: On the whole, English uses a more verbal style than German.

Yes. English consistently has a higher relative frequency of verb phrases, across mode and register.

Hypothesis 2: The degree of verbal style within each language varies by register and mode.

Yes. Spoken language uses more verb phrases compared to written language, and there is extensive variation by register in the GECCo corpus.

Variation by language is more pronounced in the spoken data.

The model expects language and mode to have a comparable effect on the frequency of verb phrases.

Hypothesis 3: Cross-linguistically, there is a correlation between verbal style and information density.

Yes. In both languages texts with higher information density use fewer verb phrases.

In English this effect is, however, slightly stronger.

Discussion

Relevant concepts

- Hawkins (2019): **word-external (in English) vs. word-internal properties (in German)**
  - Non-finite verb phrases, as a prime example of a structure that relies more on surrounding words, blends into the system well in English, but not in German.
- Kortmann & Meyer (1992): **English as a verb-oriented language (vs. German noun-oriented)**
  - 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 (“NP-expansion”)
  - English, as an SVO-language, can afford to introduce several argument-predicate structures (“VP-maximalization”) (1992: 165)
- Levshina (2022): **communicative efficiency**
  - Why does the drive for communicative efficiency (that English and German share) lead to systems and usage patterns that are so different?

Conclusion

Conclusion

Goals of the project:

- provide insights on the cross-linguistic relevance of mode, information density, and register for the use of verb phrases
- provide empirically grounded advice for translation and language acquisition material
- provide empirical evidence for common claims about English-German contrasts

Findings so far:

- Existing assumptions about use of verb phrases (and also non-finite verb phrases) are overall confirmed
- English deserves its title as a language that is more verbal than German
- Information density appears to be cross-linguistically correlated to verb phrase use
- Importance of looking at spoken material (this is where language-specific features shine)

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)

Thank you to:

Prof. Dr. Christian Mair

Albert-Ludwigs-Universität Freiburg

First supervisor

Prof. Dr. Ekaterina Lapshinova-Koltunski

Universität Hildesheim

GECCo Team

Thank you for your attention.

Hanna Mahler, M.A.

Albert-Ludwigs-Universität Freiburg

hanna.mahler@students.uni-freiburg.de

www.linkedin.com/in/hanna-mahler-2b4697193

Conference participation funded by:

- Internationale Graduiertenakademie Freiburg
- Wissenschaftliche Gesellschaft Freiburg

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.

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 (2013): Contrastive Analysis. In : *The Routledge Encyclopedia of Language Teaching and Learning*. London: Routledge, pp. 153–158.

Gast, Volker (2015): On the Use of Translation Corpora in Contrastive Linguistics: A Case Study of Impersonalization in English and German. In *Languages in Contrast*, 15(1), pp. 4–33. DOI: 10.1075/lic.15.1.02gas.

Gast, Volker; Levshina, Natalia (2014): Motivating w(h)-Clefts in English and German. A Hypothesis-Driven Parallel Corpus Study. In Anna-Maria Cesare (Ed.): *Frequency, Forms and Functions of Cleft Constructions in Romance and Germanic*. Berlin, München, Boston: De Gruyter, pp. 377–414.

Gries, Stefan; Jansegers, Marlies; Miglio, Viola (2020): Quantitative Methods for Corpus-Based Contrastive Linguistics. In Renata Enghels, Bart Defrancq, Marlies Jansegers (Eds.): *New Approaches to Contrastive Linguistics. Empirical and Methodological Challenges*. Berlin, Boston: De Gruyter, pp. 53–84.

Hawkins, John (2019): Word-External Properties in a Typology of Modern English. A Comparison with German. In *English Language and Linguistics* 23 (3), pp. 701–727. DOI: 10.1017/S1360674318000060.

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 edition. München: Oldenburg Wissenschaftsverlag (Lehr- und Handbücher zu Sprachen und Kulturen).

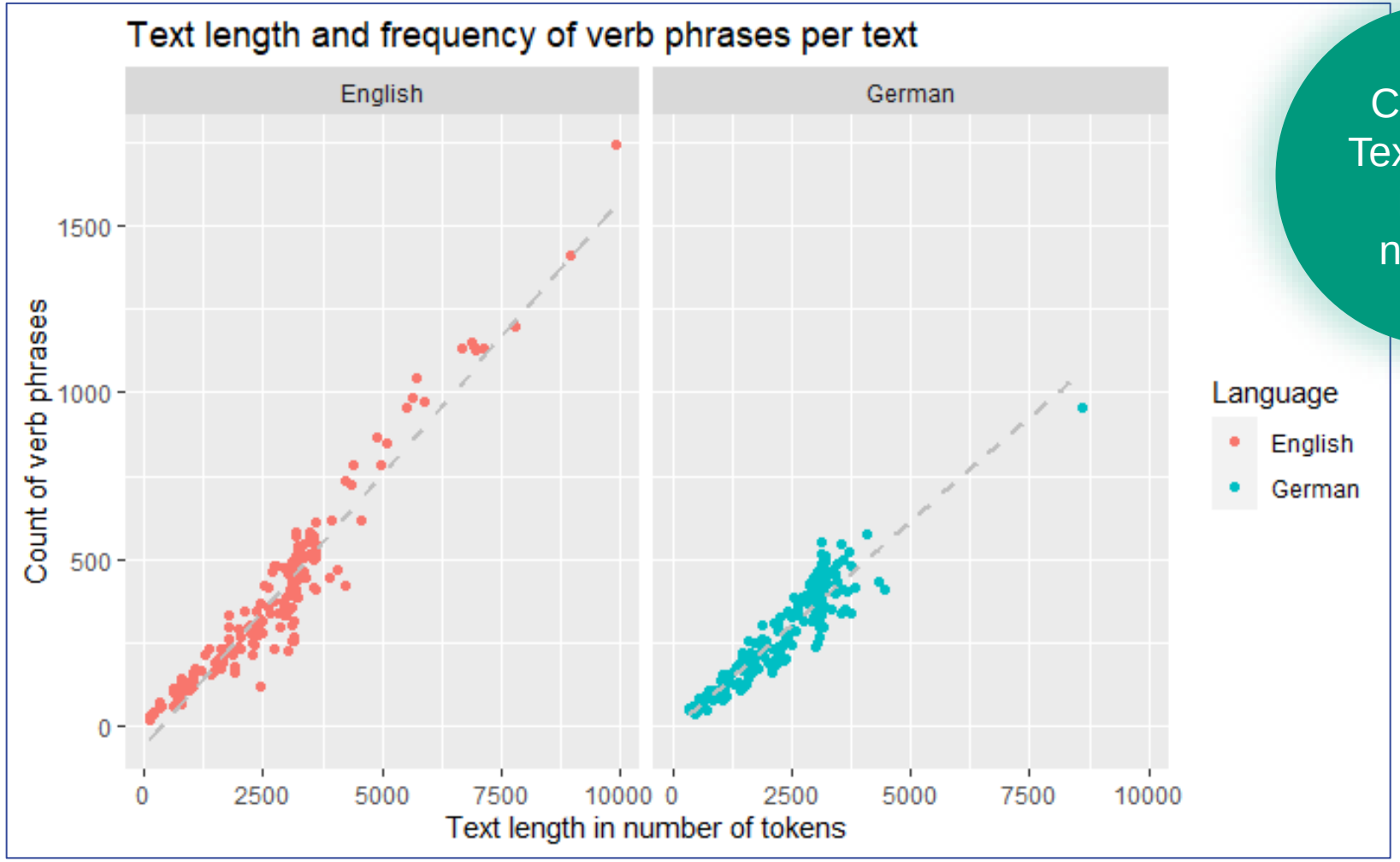
Kunz, Kerstin/Lapshinova-Koltunski, Ekaterina/Martínez Martínez, José M./Menzel, Katrin/Steiner, Erich (2021): *GECCo - German-English Contrasts in Cohesion: Insights from Corpus-Based Studies of Languages, Registers and Modes* (Trends in Linguistics. Studies and Monographs 355). De Gruyter Mouton.

Levshina, Natalia (2022): *Communicative Efficiency. Language Structure and Use*. Cambridge: Cambridge University Press.

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). Berlin, Boston: De Gruyter Mouton. 71–89.

Distribution of verb phrases by text length



Caution: Text length is not natural!

Figure 6: Verb phrase count plotted against text length in GECCo by language.

Delineating verb phrases

German

- Counted as two verb phrases:
  - Verbs of perception and causation + non-finite verb form (*sehen, hören, lassen*, …)
  - Full verbs + infinitive form (*gehen, fahren, lehren, lernen, hoffen, beginnen*, …)
- 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, …*)

English

- 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,* …)
- 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*)

Quality of the annotation of sentence boundaries in the GECCo corpus

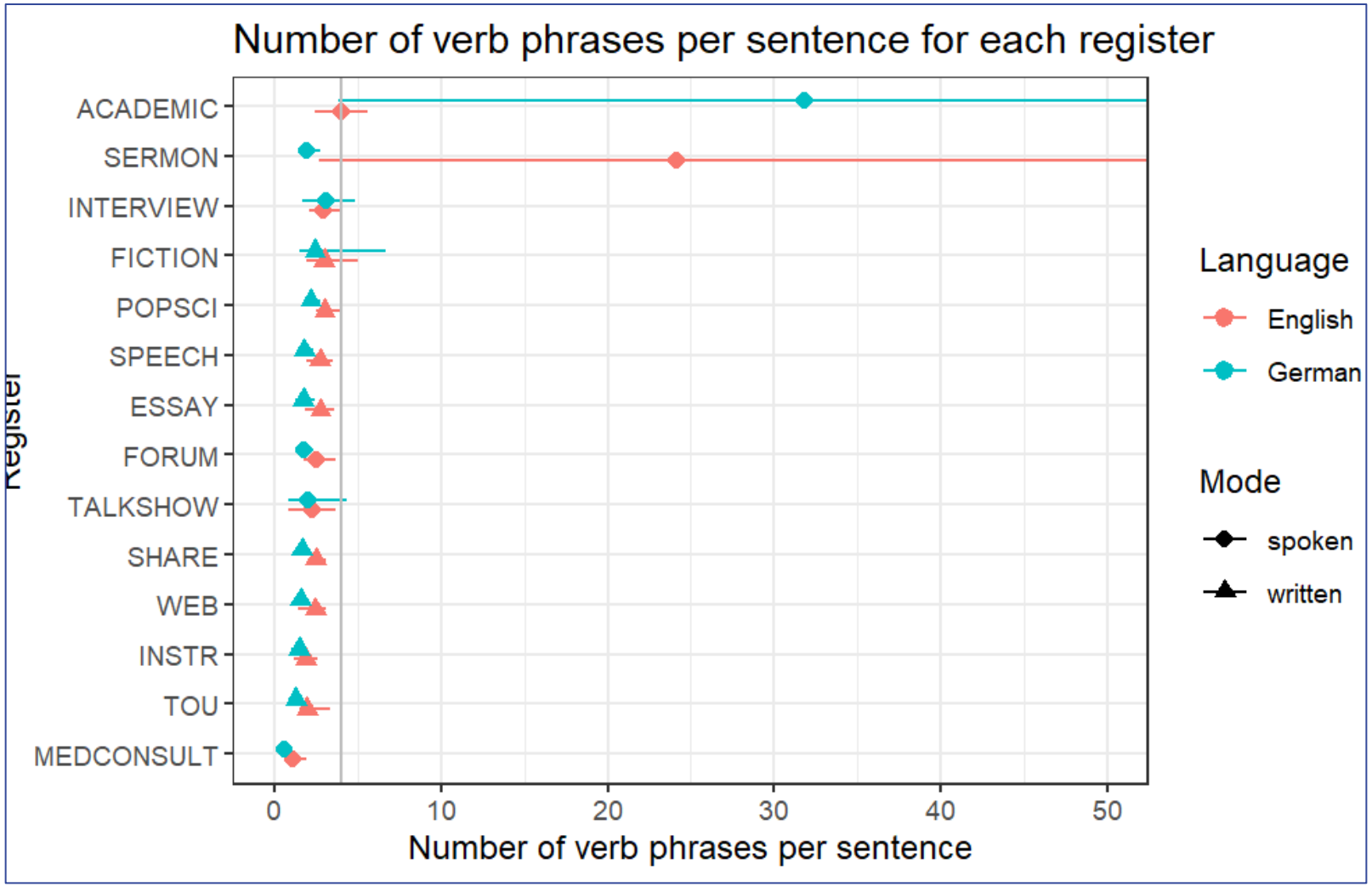


Figure 7: Frequency of verb phrases per sentence in GECCo corpus.

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

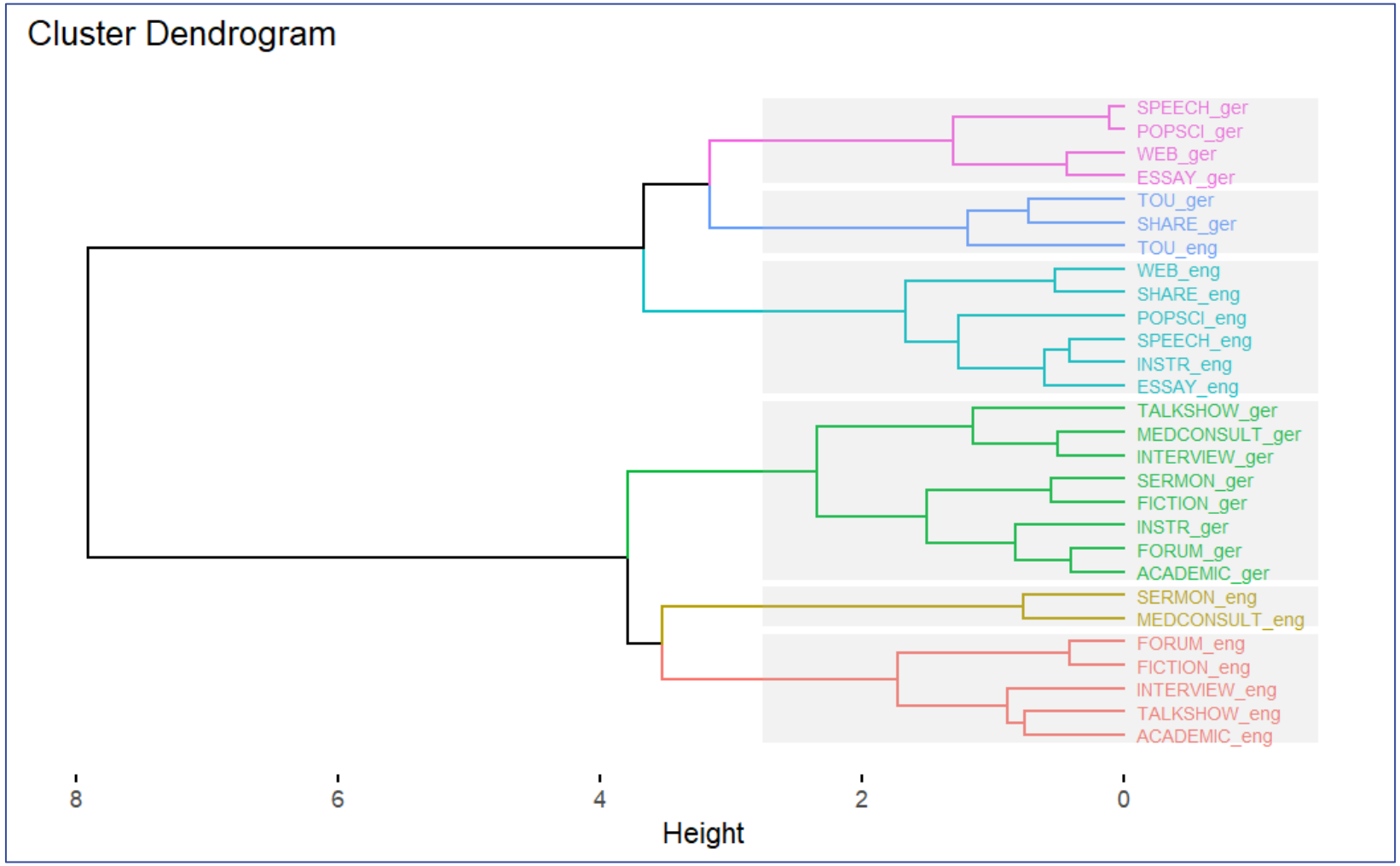


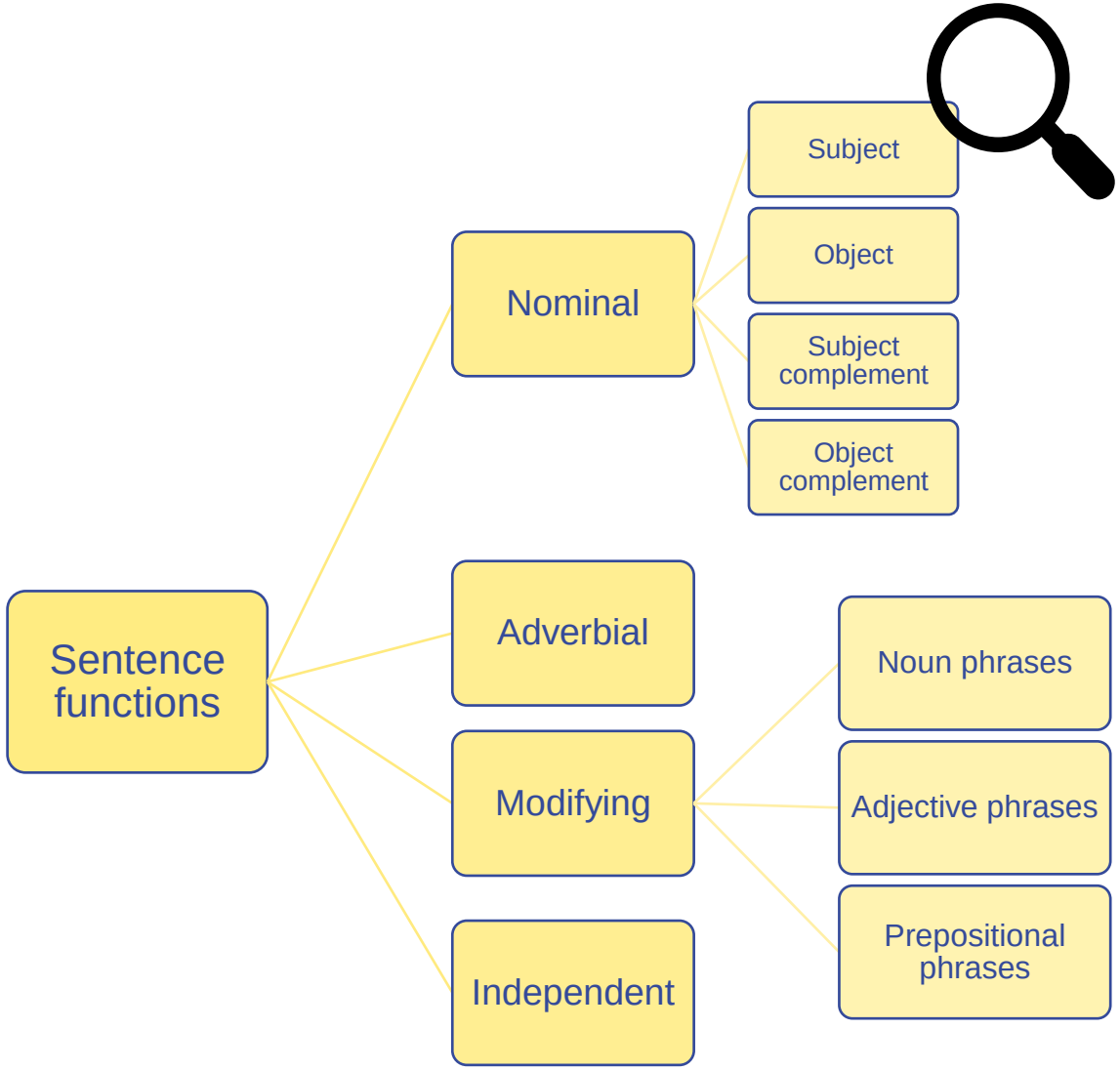
Figure 8: Cluster solution for registers in GECCo.

Digression: non-finite verb phrases

Inventory of non-finite verb phrases

English	German
to-infinitive	*zu*-infinitive
Bare infinitive	Bare infinitive
Present participle	Present participle
Past participle	Past participle

Table 1: Available non-finite verb forms in English and German.



Grammatical possibilities vs. Usage frequencies

Example: non-finite verb phrases functioning as subject clauses

		Subject clause	
		Explicit subject	Implicit subject
English	*to*-infinitive	Possible	Possible
	Bare infinitive	Not possible	Marginal
	Present participle	Possible	Possible
	Past participle	Not possible	Not possible
German	*zu*-infinitive	Not possible	Possible
	Bare infinitive	Not possible	Marginal
	Present participle	Not possible	Not possible
	Past participle	Not possible	Marginal

Table 2: Grammatical possibilities for forming non-finite subject clauses in English and German.

		Subject clause	
		Explicit subject	Implicit subject
English	*to*-infinitive	0.01 phw	0.08 phw
	Bare infinitive	0	0
	Present participle	0.002 phw	0.07 phw
	Past participle	0	0
German	*zu*-infinitive	0	0.08 phw
	Bare infinitive	0	0.03 phw
	Present participle	0	0
	Past participle	0	0

Table 3: Frequency phw for non-finite subject clauses in English and German in GECCo corpus.