Register variation in Latin

Hanna-Mari Kupari (hmknie@utu.fi), School of Languages and Translation Studies

University of Turku, Finland

Aim and research question

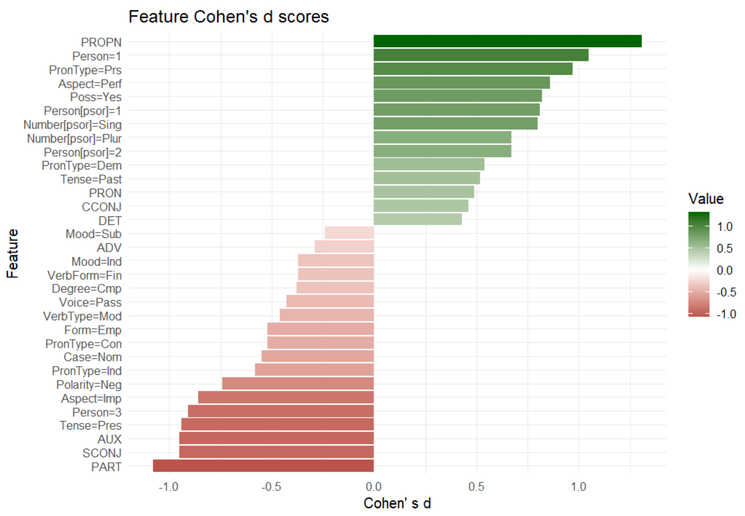
- Grouping texts by different context traced back to Aristotle dividing rhetoric into deliberative, forensic, and epideictic* (*Rh* 1.3)
- Register, as defined by Biber and Conrad (2019), is a variety of text associated with a particular situation of use:

- Previous studies on situational context, e.g. philological and historical research or studies of isolated grammatical features
- Using the Key Feature Analysis, KFA, (Egbert & Biber, 2023) method to find distinct features
- How does the statistical and computational register perspective confirm and add to previous research?

Register classes

Register	Example works and auctors	Books	Tokens
Charter	Late Latin Charter Treebank	21	242 449
Christian Biblical	Jerome's *Vulgate*	27	109 677
Christian Philosophy	Thomas of Aquinas texts	2	451 923
Fable	Phaedrus' *Fabulae*	1	4 150
History	E.g. Caesar's *Bello Gallico*	5	41 546
Letter	E.g. Cicero's *Letters to Atticus*	9	51 529
Monumental Inscription	*Res Gestae* propaganda text	1	713
Philosophy	Cicero's *De Officis*	1	10 098
Play	E.g. Tragedy, Seneca's *Oedipus*	3	19 741
Poem	Including epic, e.g. *The Aeneid*	4	9 042
Prose	Dante's *De Vulgari Eloquentia*	1	13 502
Satire	Petrinius' *Satyricon*	1	6 010
Speech	Courtroom speeches, e.g. *In Catilinam*	1	1 933
Treatise	Dante's *Monarchia*	3	39 885

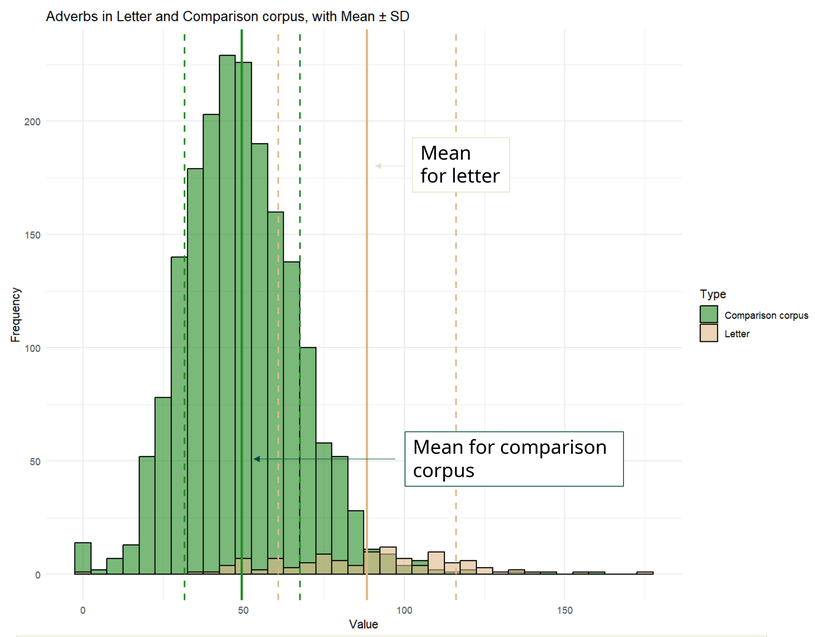
Prominent *Charter* features



Data & Methods

- Universal Dependencies (UD) provides consistent grammatical annotation (e.g. POS, morphology) across all natural languages
- The six UD Latin treebanks
- The KFA method identifies features that are statistically over- or underrepresented in each register
- KFA applies Cohen's *d* to detect feature differences between registers
- Cohen's *d* measures effect size using a comparison of standard deviation and mean,

formula: $Cohen's\ d = \frac{M_2 - M_1}{SD_{pooled}}$



A comparison in occurrences of adverbs - letter vs. comparison corpus

Counts normalized to occurrences of 1 000 tokens
Chunked into 500 tokens
Extreme outliers removed

Treebanks: "Latin Morphology through the Centuries: Ensuring Consistency for Better Language Processing" (Gamba & Zeman, 2023)

Register text examples

- *Letter*: frequent use of **adverbs** and **first-person verbs**

  dixi *hanc legem Publium Clodium* ***iam ante*** *servasse*
  I said that Publius Clodius had **already previously** complied with this law
  quod ***ego*** *non* ***credo***
  what ***I*** don't **believe** (Cic. *Att*. 1.16)

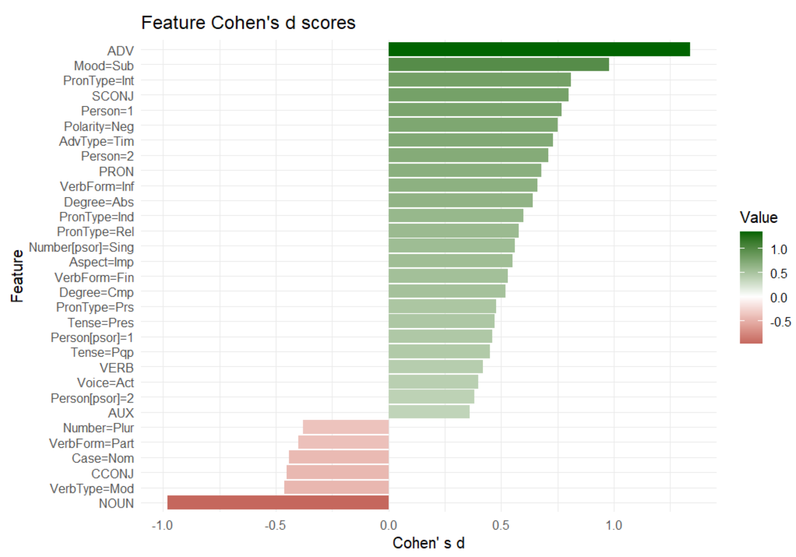
- Logical in **personal correspondence**, mimicking conversational **casual** speech

- *Charter*: overrepresentation of **proper nouns** and **first-person pronouns**

  constat ***me Sanitulum*** *filium quondam* ***Cicchi*** *de loco* ***Brancalo***
  It is established that **I**, **Sanitulus**, son of the late **Cicchus** of the place **Brancale**
  Gheipertum *clericum scribere* ***rogavi***
  I asked the clerk **Gheipertus** to write (LLCT, doc. 36)

- Suitable for **legal texts** reporting the selling of personal property, precisely **identifying actors**

Prominent *Letter* features



Discussion & Results

- Close reading of the results aligns with established knowledge
- KFA highlights the use of adverbs in *Letter*, as is common knowledge for Cicero's letters, pronouns are typical features of conversation (Biber and Conrad, 2019)
- In charters, proper nouns are prominently emphasized, consistent with findings from previous studies (Korkiakangas, 2020)

Future work

- Expand method to include lesser-known text collections, e.g. *Corpus Corporum*
- Extending the analysis to syntactic analysis
- As Hudspeth et al. (2024) note, distinctions between Classical, Medieval, and Neo-Latin should also be considered when examining variation
- How stable are register features in texts from different time contexts?

Notes & References

* Deliberative (συμβουλευτικόν), forensic (δικανικόν) and epideictic (ἐπιδεικτικόν)

The Universal Dependencies framework: https://universaldependencies.org/

Biber & Conrad. *Register, Genre, and Style*. CUP, 2019.

Egbert & Biber. "Key feature analysis: a simple, yet powerful method for comparing text varieties", *Corpora*, 2023.

Korkiakangas. "Theoretical and pragmatic considerations on the lemmatization of non-standard Early Medieval Latin charters". *Studi e Saggi Linguistici*. 2020

Hudspeth et al., "Latin Treebanks in Review: An Evaluation of Morphological Tagging Across Time". ML4AL 2024.

Thank you!

Emil Aaltonen foundation

Travel grant: COST Action CA21167: Universality, diversity and idiosyncrasy in language technology (UniDive): Young Researcher and Innovator Conference

Veronika Laippala and Timo Korkiakangas

I wish to acknowledge *CSC – IT Center for Science, Finland*, for computational resources

Virtual poster in GitHub with additional information

UNIVERSITY OF TURKU

TURKUNLP .ORG

EMIL AALTOSEN SÄÄTIÖ