ParHistVis: Visualization of Parallel Multilingual Historical Data

Aikaterini-Lida Kalouli, Rebecca Kehlbeck, Rita Sevastjanova,
Katharina Kaiser, Georg A. Kaiser, Miriam Butt
Universität Konstanz
firstname.lastname@uni-konstanz.de

Funded by the German Research Foundation (DFG)
Research Unit 2111 "Questions at the Interfaces"

Motivation & Challenges

Parallel corpora for the analysis of historical linguistic data facilitate:

- **direct comparability** of concrete examples across time periods
- **selective investigation** of passages with potentially relevant structures
- analysis of **languages not spoken** by the researcher, based on the known languages

However, such data are often (cf. [1]):

- too **sparse** for state-of-the-art statistical methods
- too large and **high-dimensional** (time, domain, language) for manual inspection
- unsuitable for learning methods which necessarily **reduce dimensionality**

Parallel and Aggregated Analysis of Linguistic Change

- **preserves dimensionality**: investigation of the data in a parallel manner (Fig. 3)
- **avoids overwhelming**: different features encoded with different colors (Fig. 3)
- allows **detailed view**: user selects subset to investigate, subset gets highlighted (Fig. 3)

Use Case: Romance interrogatives

- **Data:** 3 French and 3 Spanish Bible translations of the 12th, 16th and 20th centuries
- **Features:** a) word order in interrogatives
  b) interrogative pronouns and verbs of speaking introducing questions
  c) particles used with interrogatives
- **Goal:** investigation of the strict word order in Old Romance vs. the greater word order variation in Modern Romance
- **Observations:** a) emergence of complex inversion in Modern French [2]: the orange stream (whSNPVSCl) first appears in Middle French and increases its frequency in Modern French (Fig. 1)
  b) diachronic non-adjacency of the wh-element and the verb when a particle is present: the blue (whPtcVS) stream stays stable over time (Fig. 1)
  c) some interrogative pronouns allow more variation in the sentence structure [3]: *why* allows for more frequent use of the particle *pues* and *donc* in Spanish and French, respectively, than other pronouns (Fig. 2)

ParHistVis

Function: interactive **visualization tool** for **parallel**, multilingual data of a) the same time period across languages; b) of different periods of the same language; c) across languages.
Input: tabulated file with aligned data over time or/and language, annotated with features
Output: color-encoded matrix view of the data
URL: `typo.uni-konstanz.de/parhistvis/`

Pattern Recognition & Interaction

User selects features for comparison:

- **Streamgraphs:** represent selected features as streams (Fig. 1)
- **Sankey diagrams:** represent selected features as nodes and their interaction as a flow between them (Fig. 2)

⇒ **at-a-glance view of patterns and interactions** across time, language and data course

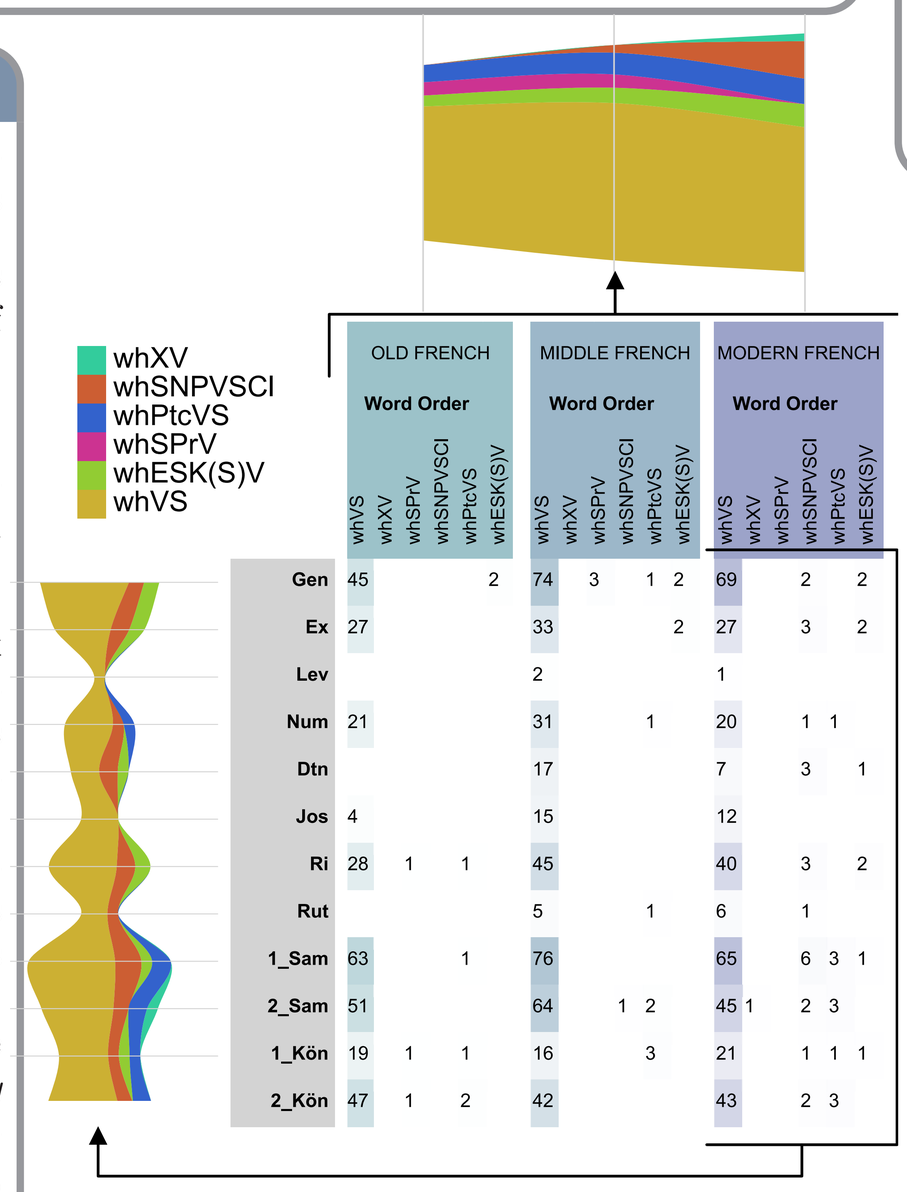

Figure 1: Streamgraphs: Word order in French across time and in Modern French across the aggregated Bible

Figure 2: Sankey Diagrams: interactions between particles and interrogative pronouns in French and Spanish

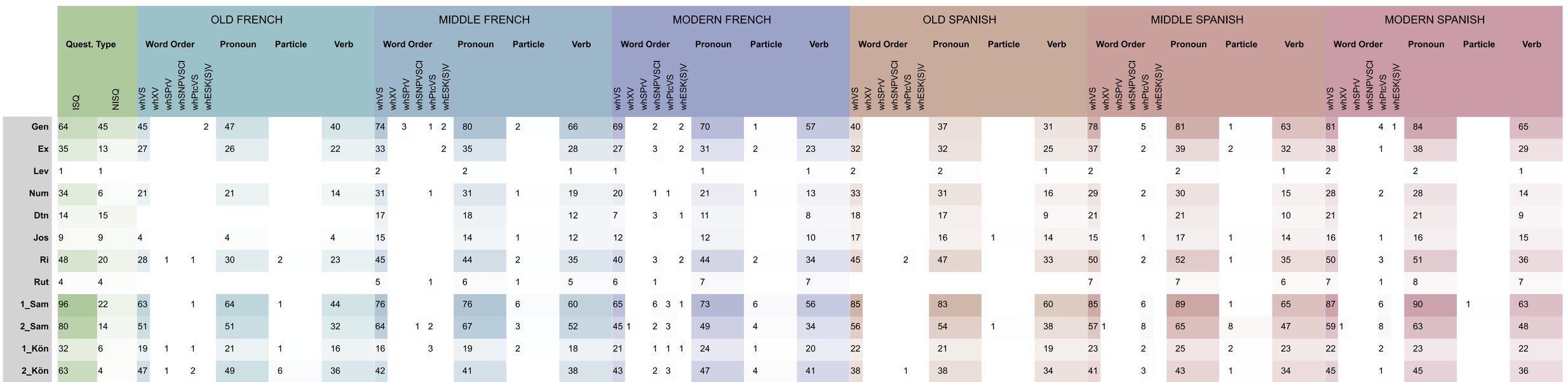

Figure 3: Aggregated matrix view of the books of the Old Testament across time periods and languages.

References

[1] Qihong Gan, Min Zhu, Mingzhao Li, Ting Liang, Yu Cao, and Baoyao Zhou. Document visualization: an overview of current research. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(1):19–36, 2014.

[2] Ian Roberts. *Verbs and Diachronic Syntax. A Comparative History of English and French.* Kluwer, Dordrecht, 1993.

[3] Francisco Ordóñez. *Word order and clause structure in Spanish and other Romance languages.* PhD thesis, The City University of New York, 1997.