BRIDGING CORPUS LINGUISTICS AND RELATIONAL LEXICAL DATABASES THROUGH MACHINE LEARNING: THE SERMIS APPROACH

Dalia Jakulytė, Vitalij Denisov, Arturas Garbuz
Klaipėda University

1 Technological Foundations: XML vs. Relational Model

Sliekkas
Relies on hierarchical XML corpora and automated processing tools to manage deeply annotated historical texts [1].

SERMIS
Employs a robust MySQL and Laravel-based MVC web architecture for storing and managing complex linguistic data [3].

2 Old Lithuanian Linguistic Data Modeling Approach

```
<w ana="lexeme|POS|case|number">word</w>
```

In XML formats, grammatical categories are defined within inline annotations and notation structures [1].

→

Lexeme | **POS** | **Case** | **Number**

In the relational database, these categories are split into distinct attributes and normalized tables, enabling advanced filtering, sorting, and querying across any combination of grammatical features.

3 SERMIS Development & Training via Sliekkas

Sliekkas → **SERMIS**

The system development leverages structured datasets imported from Sliekkas.

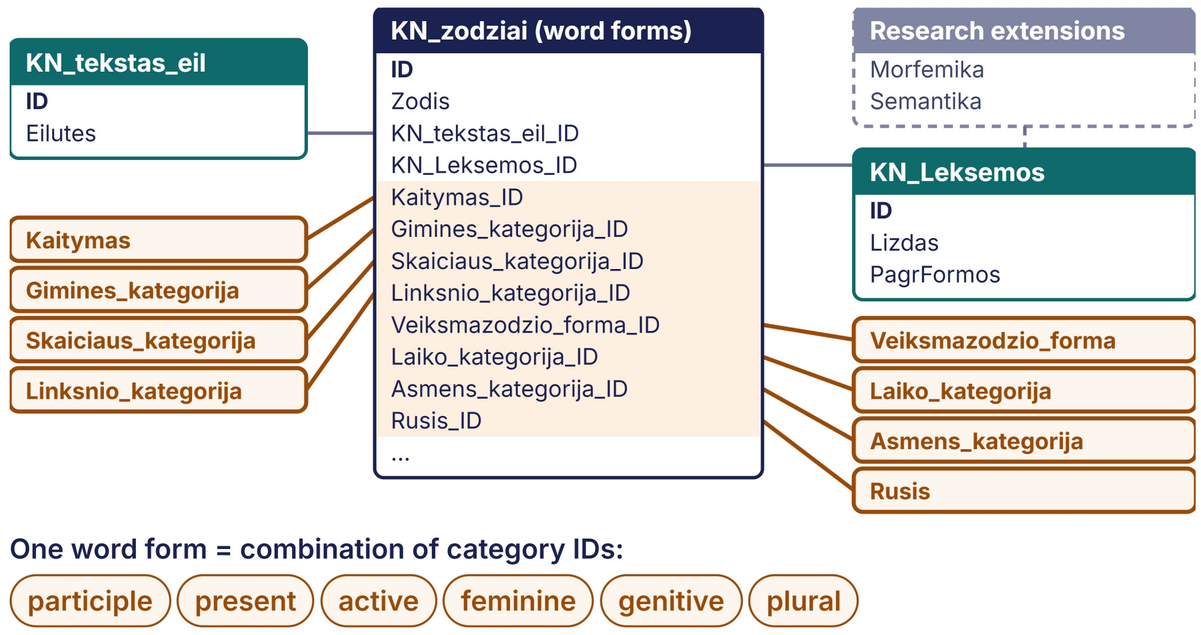
Annotated source texts are converted and fed into the database pipeline to bootstrap and train the machine learning models.

4 SERMIS Architecture & Data Evolution

Evolving from early physical card indexes and MS Access prototypes [2] into a modern SQL-based relational web architecture [3].

Past
Physical card index
→
Prototype
MS Access
→
Present
SQL-based relational architecture

The data model connects texts, lexemes, word forms, morphological features, and semantic attributes into an integrated entity network.



5 Machine Learning Integration Pipeline

1 Raw Text Corpus Processing | **2** Automated ML Feature Prediction | **3** Expert Validation (Human-in-the-Loop) | **4** Relational SQL Ingestion

Implements a human-in-the-loop workflow combining automated ML feature prediction, tokenization, and expert linguistic validation.

Validated predictions are systematically ingested and denormalized into relational tables to scale up historical text annotation.

Conclusions

- Moving from XML notation to normalized relational attributes lets researchers query historical Lithuanian texts by any combination of grammatical features.
- Annotated Sliekkas corpora provide the foundation for training machine learning models that predict morphological features for SERMIS.
- The human-in-the-loop pipeline reduces manual database entry, keeps annotation consistent across corpora, and keeps ML outputs explainable through normalized SQL tables.

References

[1] Gelumbeckaitė, J., Šinkūnas, M., & Zinkevičius, V. (2012). Old Lithuanian Reference Corpus (SLIEKKAS) and Automated Grammatical Annotation. *Journal for Language Technology and Computational Linguistics (JLCL)*, 27(2), 83–96. https://doi.org/10.21248/jlcl.27.2012.165

[2] Jakulytė, D. (2004). Knygos nobažnystės (1653) morfologijos tyrimo galimybės naudojant reliacinę duomenų bazę [Possibilities of studying the morphology of “Knyga nobažnystės” (1653) using a relational database]. *Acta Linguistica Lithuanica*, 51, 1–14. https://journals.lki.lt/actalinguisticalithuanica/article/view/1126/1206

[3] Jakulytė, D., & Pakalniškienė, D. (2018). Senųjų raštų morfologijos duomenų bazės taikymas semantikos tyrimams: veiksmažodžių semantika [Application of the morphological database of Old Writings to semantics research: Verb semantics]. *Vilnius University Open Series*, 1, 373–388. https://doi.org/10.15388/Proceedings.2018.15