Some steps towards the UNIDIVE BLARK: revising Latvian language resources and tools

Inguna Skadiņa, Baiba Saulīte, Laura Rituma and Lauma Pretkalniņa

aitab

How to quantify university and diversity?

To obtain a comprehensive understanding of the current situation, we propose to start with the construction of **UNIDIVE BLARK**:

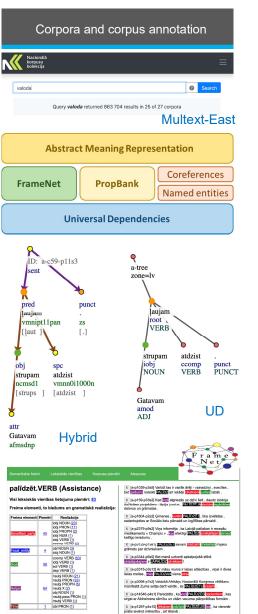
- focus of WG1-WG3
- concentrating on languages represented by the consortium
- this poster: an overview of the Latvian language technologies and tools

Results: Latvian UniDive BLARK

Corpora	FrameWork	Size
General	Morphology, MULTEXT-EAST	2,3 M tokens
Treebank	UD, Hybrid	17 K trees
Semantic	FrameNet, ProBank, AMR	13.5 K sentences
Lexical Resources		
General	TEI	388 K entries
WordNet	LMF	5,3 K synsets
FrameNet	FrameNet	2,9 K lexical units
Terminology	ТВХ	16 M entries (Eurotermbank)
Language processing tools		
Syntactic parsers	UD, hybrid	
Semantic parsers	GF	
MWE parsers	NER Bio	

Conclusion

- Latvian resources and tools are mainly developed in accordance with international standards and models (e.g. TEI, MULTEXT-EAST, LMF, TBX, UD, FrameNet, AMR, etc.)
- While MWE resources are less developed today, some recently started initiatives, e.g. LATE and DHELI projects, could allow us to fill this gap soon.



Language processing tools

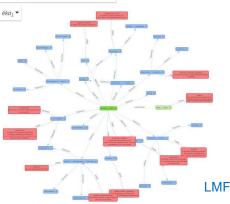


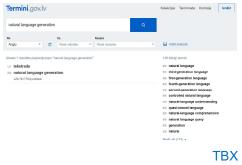
<section-header>A construction of the second construct of the second construction of the s

Latvian WordNet

IIII

WordNet graph for the synset {celtne1, ēka1}







CoNLL, Multext-East, UD, BIO

The results reported here were supported by the project "Research on Modern Latvian Language and Development of Language Technology" (LATE, project No: VPP-LETONIKA-2021/1-0006) and project "Towards Development of Open and FAIR Digital Humanities Ecosystem in Latvia" (DHELI, project No: VPP-IZM-DH-2022/1-0002)