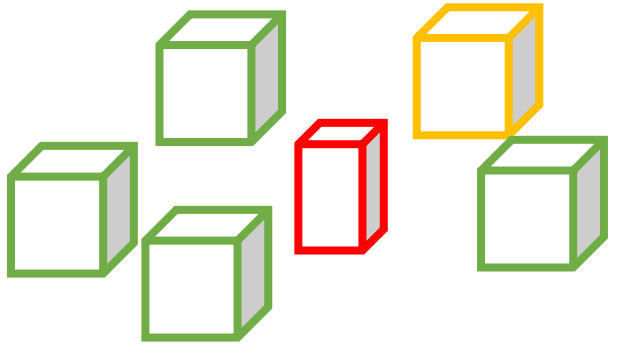


Recent Developments of Language Technologies in Lithuania

Andrius Utka

Vytautas Magnus University, Kaunas LT-44243, Lithuania

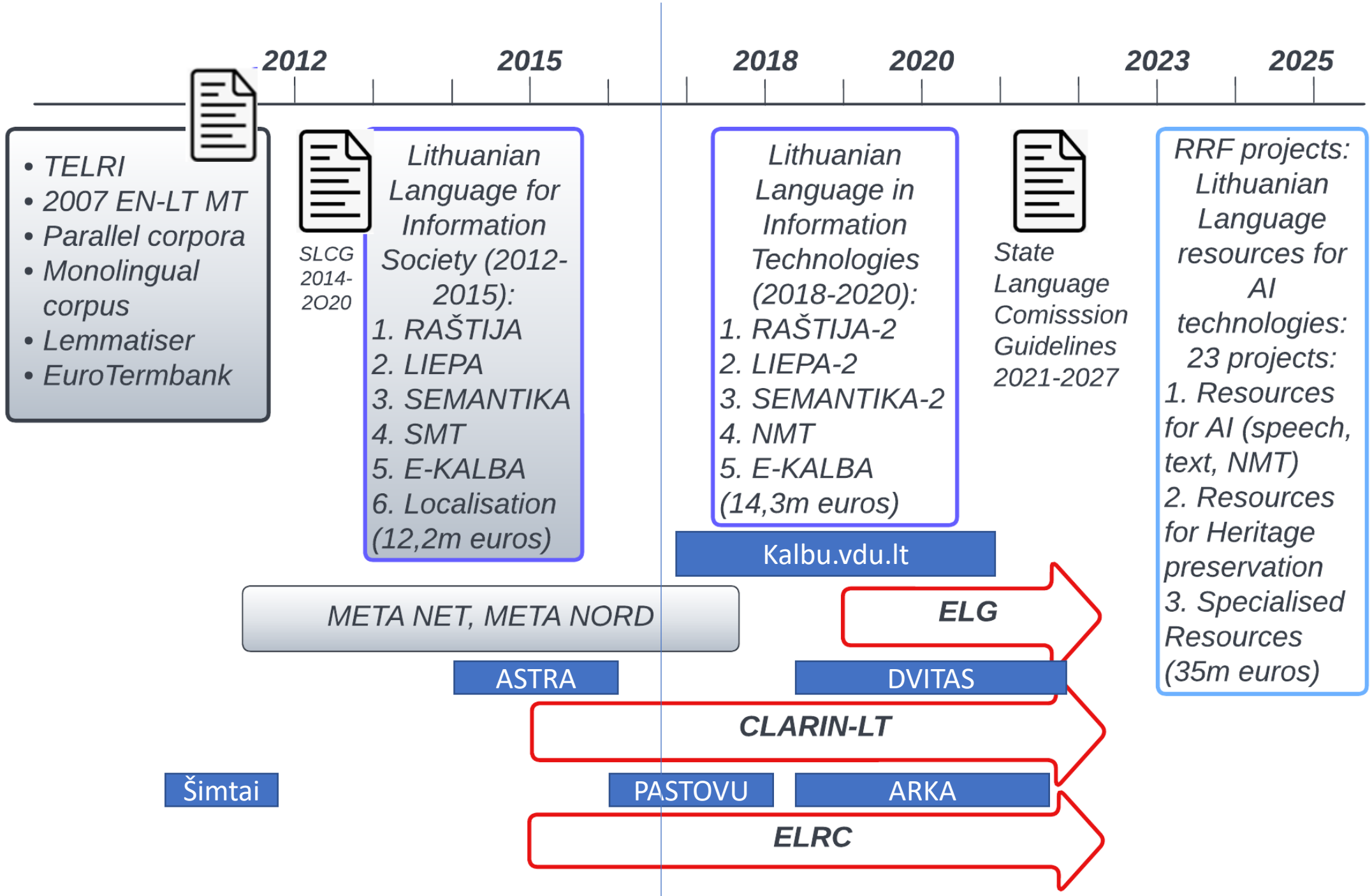
Human Language Technologies – The Baltic Perspective

Riga, Latvia

October 6-7, 2022

Plan of the presentation

1. Timeline
2. Infrastructures / Platforms / Resources
3. Resources for machine learning
4. Basic NLP
5. Speech Recognition
6. Speech Synthesis
7. Machine translation and localization
8. Conclusions


2012
2015
2018
2020
2023
2025
• TELRI
• 2007 EN-LT MT
• Parallel corpora
• Monolingual corpus
• Lemmatiser
• EuroTermbank
SLCG 2014-2020
Lithuanian Language for Information Society (2012-2015):
1. RAŠTIJA
2. LIEPA
3. SEMANTIKA
4. SMT
5. E-KALBA
6. Localisation
(12,2m euros)
Lithuanian Language in Information Technologies (2018-2020):
1. RAŠTIJA-2
2. LIEPA-2
3. SEMANTIKA-2
4. NMT
5. E-KALBA
(14,3m euros)
State Language Comisssion Guidelines 2021-2027
RRF projects: Lithuanian Language resources for AI technologies: 23 projects:
1. Resources for AI (speech, text, NMT)
2. Resources for Heritage preservation
3. Specialised Resources
(35m euros)
Kalbu.vdu.lt
META NET, META NORD
ELG
ASTRA
DVITAS
CLARIN-LT
Šimtai
PASTOVU
ARKA
ELRC

Infrastructures / Platforms / Services

- CLARIN-LT.LT – research infrastructure (RI), a Lithuanian centre of CLARIN ERIC
- KLC.VDU.LT – resources from the CCL
- RAŠTIJA.LT – integrated infrastructure of EU-funded projects
- VERTIMAS.VU.LT – infrastructure for MT services
- SEMANTIKA.LT – Infrastructure for EU-funded project SEMANTIKA-2
- EKALBA.LT – information system of Lithuanian language resources
- KALBU.VDU.LT – infrastructure of Lithuanian language teaching resources

INSTITUTIONAL

EU-FUNDED

EU NEW

Resources for machine learning

- Not many of these nice platforms have downloadable and usable resources for machine learning;
- Except for CLARIN-LT and other CLARIN centres

Assessment of resources for ML

✓ Very limited number of resources available for ML

Basic NLP for Lithuanian

Tokenisation

Morphology

lemmatiser, morphological analyser, part of speech tagger, spelling checker,

Syntax

parser

Summarization

extractive summarization

MWE detection, terminology extraction

Colloc tool

Basic NLP assessment

- ✓ Lithuanian is well supported in terms of basic NLP.
- ✓ Freely usable systems are available.
- ✓ Necessary improvement: quality, expansion of domains

Speech Recognition (STT)

Free (EU-funded) applications

- SEMANTIKA-2: VMU Transcription service (semantika.vdu.lt)
- LIEPA-2: VU Applications for computer control
- VU: within MT application

Commercial applications

- GoogleDocs Transription
- Microsoft Azure
- Tilde demo transcription service
- Intelektika service

Speech recognition assessment

- ✓ Lithuanian is well supported in terms of speech recognition.
- ✓ Both free (EU-funded) and commercial systems are available.
- ✓ Necessary improvement: quality (especially of noisy signals), expansion of domains

Speech Synthesis (TTS)

Free (EU-funded) applications

- LIEPA-2: Tartuvas (dictionary-based speech synthesis), SAPI5 (for blind people)

Commercial applications

- Tilde Text-To-Speech (TTS), Lithuanian Language, Man's Voice
- Intelektika speech synthesis service
- Microsoft Azure

Use cases

- lrt.lt and lrs.lt by Intelektika

Speech synthesis assessment

- ✓ The support of Lithuanian speech synthesis is increasing.
- ✓ Both EU-funded and Commercial systems are available.
- ✓ Necessary improvement: quality, expansion of voices, disambiguation, reading of numbers

Machine translation and localisation

Free (EU funded) MT systems

- Vilnius university translation system (LT ⇔ EN, RU, DE, FR, PL)
- eTranslation (for translating documents) (LT ⇔ EU languages)

Commercial MT systems

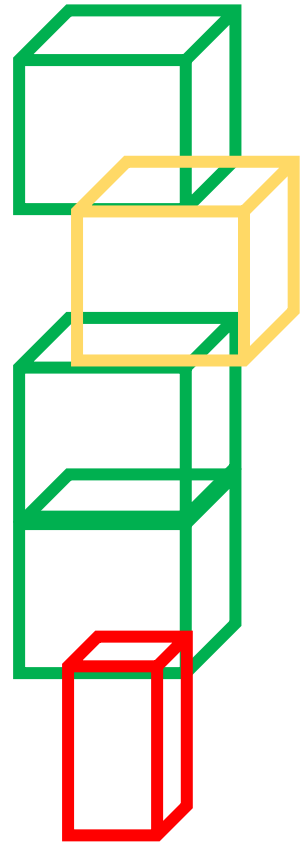
- Tilde translation system
- Google Translate
- Microsoft Bing Translator
- DeepL

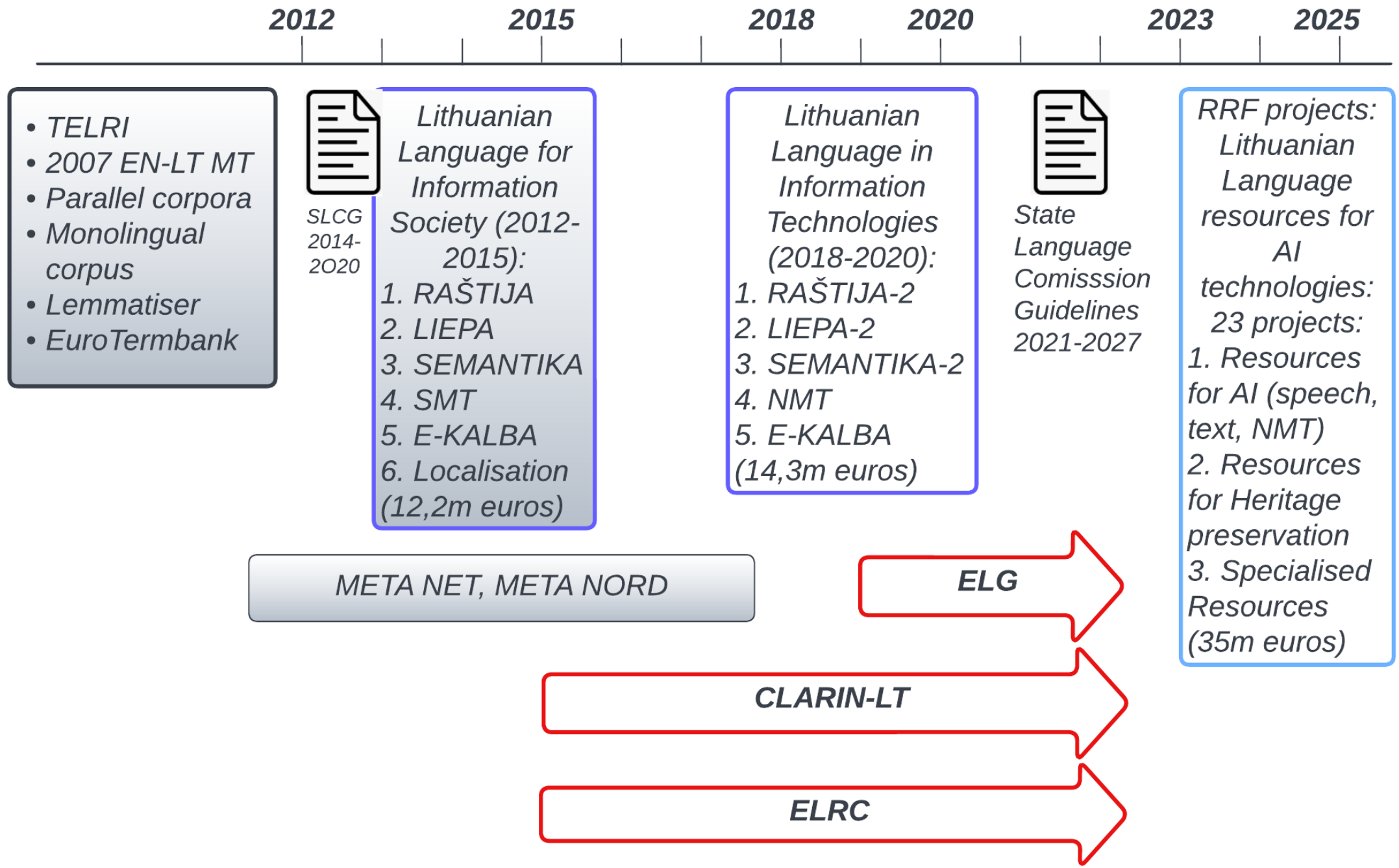
MT assessment

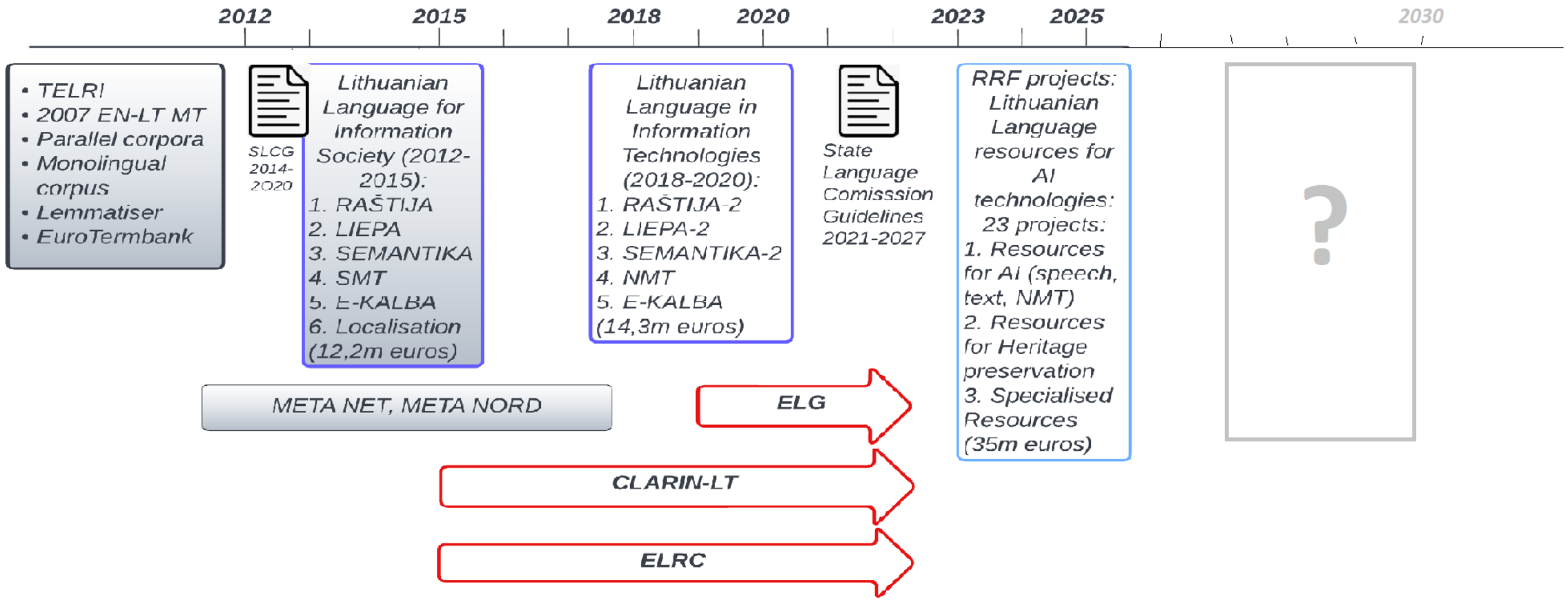
- ✓ Lithuanian is well supported in terms of MT.
- ✓ Both EU-funded and Commercial systems are available.
- ✓ Necessary improvement: quality, expansion of domains, NER

Conclusions

- Rather well supported areas for Lithuanian are:
  - Basic NLP
  - Machine translation
  - Speech recognition
- Some areas require more attention:
  - Speech synthesis
- Some areas are still largely untouched:
  - Natural language understanding
  - Natural language generation
- NOW we believe that production of freely available data may produce further improvements in all areas, especially in AI.




2012
2015
2018
2020
2023
2025
• TELRI
• 2007 EN-LT MT
• Parallel corpora
• Monolingual corpus
• Lemmatiser
• EuroTermbank
SLCG 2014-2020
Lithuanian Language for Information Society (2012-2015):
1. RAŠTIJA
2. LIEPA
3. SEMANTIKA
4. SMT
5. E-KALBA
6. Localisation
(12,2m euros)
Lithuanian Language in Information Technologies (2018-2020):
1. RAŠTIJA-2
2. LIEPA-2
3. SEMANTIKA-2
4. NMT
5. E-KALBA
(14,3m euros)
State Language Comisssion Guidelines 2021-2027
RRF projects: Lithuanian Language resources for AI technologies: 23 projects:
1. Resources for AI (speech, text, NMT)
2. Resources for Heritage preservation
3. Specialised Resources
(35m euros)
META NET, META NORD
ELG
CLARIN-LT
ELRC


2012
2015
2018
2020
2023
2025
2030
• TELRI
• 2007 EN-LT MT
• Parallel corpora
• Monolingual corpus
• Lemmatiser
• EuroTermbank
SLCG 2014-2020
Lithuanian Language for Information Society (2012-2015):
1. RAŠTIJA
2. LIEPA
3. SEMANTIKA
4. SMT
5. E-KALBA
6. Localisation
(12,2m euros)
Lithuanian Language in Information Technologies (2018-2020):
1. RAŠTIJA-2
2. LIEPA-2
3. SEMANTIKA-2
4. NMT
5. E-KALBA
(14,3m euros)
State Language Comisssion Guidelines 2021-2027
RRF projects: Lithuanian Language resources for AI technologies: 23 projects:
1. Resources for AI (speech, text, NMT)
2. Resources for Heritage preservation
3. Specialised Resources
(35m euros)
?
META NET, META NORD
ELG
CLARIN-LT
ELRC

Thank you!