EUROPEAN LANGUAGE EQUALITY

META-NET
META-FORUM 2022

EUROPEAN LANGUAGE GRID

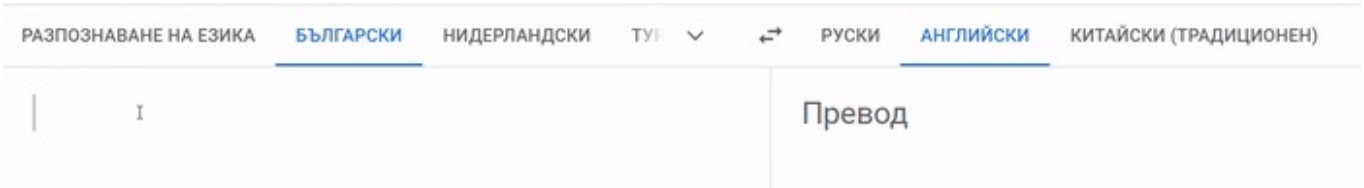
The Bulgarian Language in the Digital Age

Svetla Koeva (Institute for Bulgarian Language, Bulgarian Academy of Sciences)
svetla@dcl.bas.bg

08/09-06-2022 META-FORUM 2022 – Joining the European Language Grid (hybrid conference)
http://www.european-language-grid.eu – http://european-language-equality.eu

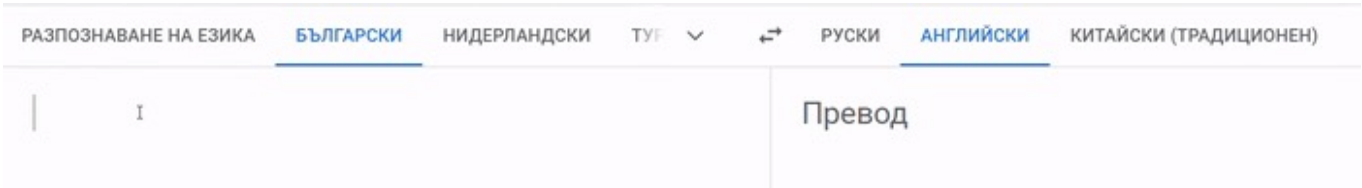
Technologies and Resources for Bulgarian

- **Machine translation technologies** for Bulgarian are still underdeveloped.
  - Bulgarian is present in several monolingual and multilingual corpora (PARSEME, UD treebank), and some of the multilingual corpora are sentence-aligned.
  - CEF-funded data-collection projects (CEF Automated Translation, MARCELL, CURLICAT).



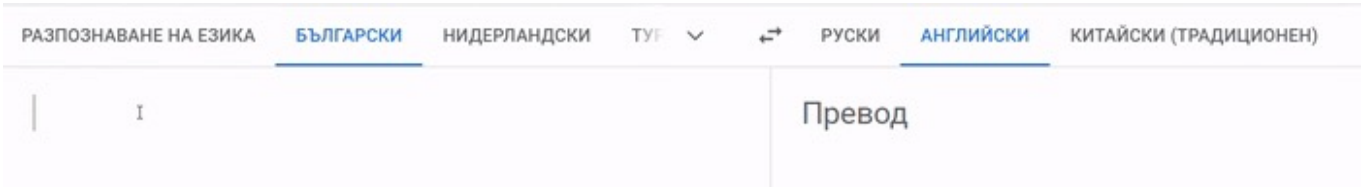
Technologies and Resources for Bulgarian

- **Machine translation technologies** for Bulgarian are still underdeveloped.
  - Bulgarian is present in several monolingual and multilingual corpora (PARSEME, UD treebank), and some of the multilingual corpora are sentence-aligned.
  - CEF-funded data-collection projects (CEF Automated Translation, MARCELL, CURLICAT).
- The quality of **speech technology** is not yet satisfactory for a low-resourced language such as Bulgarian.
  - Bulgarian speech corpora (BGSpeech, BulPhonC) have been sporadically developed.
  - There are a few systems for speech synthesis (SpeechLab 2.0), specifically designed for Bulgarian.
  - No accessible and reliable speech to text systems for Bulgarian, especially working in real time.



Technologies and Resources for Bulgarian

- **Machine translation technologies** for Bulgarian are still underdeveloped.
  - Bulgarian is present in several monolingual and multilingual corpora (PARSEME, UD treebank), and some of the multilingual corpora are sentence-aligned.
  - CEF-funded data-collection projects (CEF Automated Translation, MARCELL, CURLICAT).
- The quality of **speech technology** is not yet satisfactory for a low-resourced language such as Bulgarian.
  - Bulgarian speech corpora (BGSpeech, BulPhonC) have been sporadically developed.
  - There are a few systems for speech synthesis (SpeechLab 2.0), specifically designed for Bulgarian.
- Recently, there has been serious advances in research based on **information extraction** for Bulgarian: event extraction, sentiment analysis, fake news detection, fact-checking.



Strengths

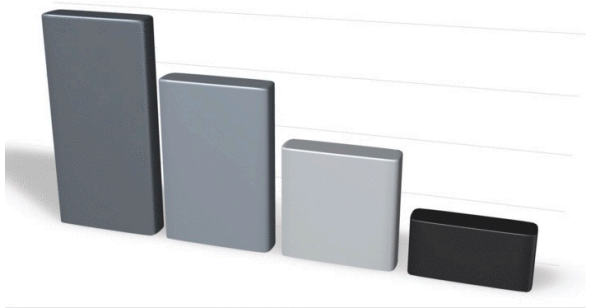
- **Strengths** of the Language Technologies for Bulgarian in 2022 include:
  - The availability of some LTs for text analysis (UDpipe, NLP-Cube),
  - Number of available language models for Bulgarian is growing (RoBERTa-base, XLM-R);
  - 220 high-schools with ICT focused curricula;
  - A number of very successful high-tech companies (around 50 AI companies).

Strengths and Weaknesses

- **Strengths** of the Language Technologies for Bulgarian in 2022 include:
  - The availability of some LTs for text analysis (UDpipe, NLP-Cube),
  - Number of available language models for Bulgarian is growing (RoBERTa-base, XLM-R);
  - 220 high-schools with ICT focused curricula;
  - A number of very successful high-tech companies (around 50 AI companies).
- **Weaknesses** of the Language Technologies for Bulgarian in 2022 include:
  - Many technologies are still not available (human-computer interaction, multimodal processing, language generation, etc.) ;
  - There are no available ready to use applications (summarisation, question answering, speech recognition, etc.);
  - Copyright issues are still the major barrier to the access and re-use of the available language resources;
  - There are not in general adequate LT funding policies.

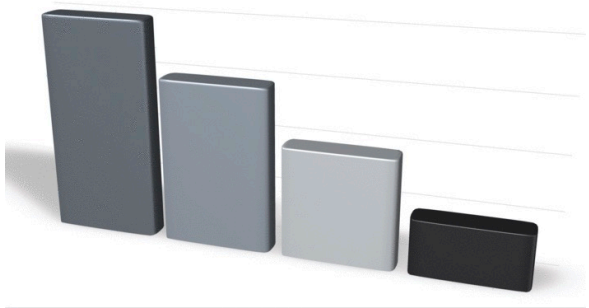
Recommendations and Next Steps

- There is a need for open real time services for machine translation from and to Bulgarian combining text and speech, **taking into account context, communicative purposes and different environments**.
- Speech and text technologies for Bulgarian have to be combined with technologies for other modalities: real time image and video processing **working simultaneously in a multilingual environment**.
- **Natural language understanding and generation** of Bulgarian have to become part of multilingual and multimodal processing.



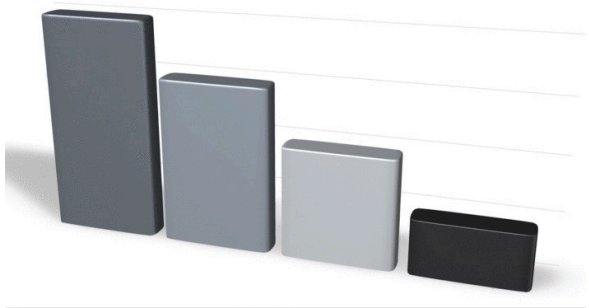
Recommendations and Next Steps

- Digital Bulgarian needs a **large-scale, long-term support**, harmonised with the support for all European languages.
- **BLARK-like minimum set of language resources and technologies** for all European languages should be developed and maintained.
- A **convenient and well-regulated access to data** is needed for the development of new products, applications and services.
- **Dedicated education and training programmes** in the field of LT and AI are essential for the advance.
- **The ELG** – the European hub and repository for ready-to-use (open-source) datasets, models, tools and services, has to be strengthen and further developed.



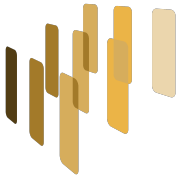
Recommendations and Next Steps

- **The vision for the future is high-quality LT for all European languages that supports political and economic unity through cultural diversity.**



European Language Grid

European Language Equality



EUROPEAN LANGUAGE GRID

EUROPEAN LANGUAGE EQUALITY

Thank you!

The European Language Grid has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 825627 (ELG).
The European Language Equality project has received funding from the European Union under grant agreement № LC-01641480 – 101018166 (ELE),

Svetla Koeva (Institute for Bulgarian Language, Bulgarian Academy of Sciences)
svetla@dcl.bas.bg

08/09-06-2022 META-FORUM 2022 – Joining the European Language Grid (hybrid conference)
http://www.european-language-grid.eu – http://european-language-equality.eu