EUROPEAN LANGUAGE EQUALITY

META-NET

META-FORUM 2023

# Artificial Intelligence Data Kit 2030

Svetla Koeva (Institut for Bulgarian Language Prof. Lyubomir Andreychin, Bulgaria)
svetla@dcl.bas.bg

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu

# Artificial Intelligence Data Kit 2030

EUROPEAN LANGUAGE EQUALITY

- The specification of a **data kit** designed for artificial intelligence is based on:
  - **Analysis** of prominent large language models, and datasets
  - **Study** of trustworthy research, starting with the European Language Equality reports and findings and expanding to scientific publications, surveys and artificial intelligence strategies
  - **Interactions** with representatives from Europe‘s language technology and artificial intelligence companies
- **Large language models** are trained on large datasets, which contain billions of words. The concept of modern large language datasets is based on several criteria:
  - **Quantity**: massive amounts of data that cover a variety of languages, media types, styles, domains, genres, and topics and are dynamically updated.
  - **Diversity**: wide range of sources: web pages, books, news, patents, code, images, video, and speech records
  - **Quality**: cleaned and filtered data with no re-duplication or subsumption of samples and no toxic or biased. content.
  - **Structure**: data and metadata organised in a conceptual graph.

EUROPEAN LANGUAGE EQUALITY

# Thank you!

The European Language Equality project has received funding from the European Union under grant agreements № LC-01641480 – 101018166 (ELE) and № LC-01884166 – 101075356 (ELE2).

Svetla Koeva (Institut for Bulgarian Language Prof. Lyubomir Andreychin, Bulgaria)
svetla@dcl.bas.bg

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu