The extent of legal control over language data: the case of language technologies

The presenters:

Aleksei Kelli

Penny Labropoulou
The authors:

Aleksei Kelli, Arvi Tavast, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värv
The aims of the research:

• To determine the extent of impact on the development of language technologies (LTs);
• To define a favourable regulatory framework for the development of LTs.
From Language Data to LTs:

raw data → datasets → annotated data

end-user applications → "models"
Building models process:

dataset (raw/annotated) → training → test → evaluation

human-driven → auto processing

select dataset + algorithm → extract/select features → train → test

final model → play with parameters
Language data & copyright:

- Press publications
- Written text
- Videos
- Oral text
- Performances
- Copyrighted databases
- *Sui generis* databases
Language data & personal data:

- Personal data
  - Types of personal data
    - General personal data (e.g., name, drawing, photo, etc.)
    - Special categories of personal data
      - Genetic data
        - Data concerning health
        - Biometric data
        - ....
  - Publicly available personal data (C-73/07)
  - Pseudonymous & anonymous data
Use of LD & copyright protection:

- Quotation right (InfoSoc Art. 5 (3))
- Right holder's consent
- Text and data mining exception (Digital Copyright Directive Art. 3, 4)
- Legal bases for using data-sets
- Temporary acts of reproduction exempted from the reproduction right (InfoSoc Art. 5 (1))
- Private use exception (InfoSoc Art. 5 (2))
- Research exception (InfoSoc Art. 5 (3))
Use of LD & personal data protection:

- Performance of a task in the public interest or in the exercise of official authority
- Legitimate interests
- Protection of the vital interests of the data subject or of another natural person

Legal grounds for lawful processing of personal data (GDPR Art. 6):

- Performance of a contract
- Consent
- Compliance with a legal obligation
Research use & copyright:

- Covers works, objects of related rights
- Restriction of technological protection measures and contractual provisions
- Right to use copies with an appropriate level of security for the purposes of scientific research (e.g., verification)

Text and data mining as any automated analytical technique (Digital Copyright Directive Art. 3)

- Right to make reproductions and extractions for research purposes
- Covers research organisations & cultural heritage institutions
- Lawful access as a precondition
Research use & personal data:

- Derogation from the right to be forgotten and the right to be informed (GDPR Art. 17, 14)
- Consent
- Public interest
- Legitimate interest

Legal grounds for processing (GDPR Art. 6)

- Technological development and demonstration
- Fundamental & applied research
- Privately funded research

Scientific research interpreted in a broad manner (Recital 159)

Processing for research purposes (GDPR Art. 89)

- Appropriate safeguards
  - Technical and organisational measures
  - Pseudonymisation

- Derogations from the rights (optional exceptions)
  - Right of access (Art 15)
  - Right to rectification (Art. 16)
  - Right to restriction of processing (Art. 18)
  - Right to object (Art. 21)
Models:

- Whether copyrighted material is used
- Legal ground for the use of copyrighted material
- Legal status of the model
Models:

- Recreation of personal data/copyrighted work from models
- Model as a derivative work?
- Substantial parts of the input data set remaining in the model

- Choice/creation of a data set
- Choice/creation of the programme used for the training of the model
- Various cycles of testing and validation by tuning the parameters of the training programme

Model as a result of human intellect investment
Models:

- record
- raw data
- Dataset
- Annotated dataset
- Model
- develop
- Semi-finished product
- develop
- use
- Finished product

copyrighted works and personal data
Personal data problems concerning voice:

- Argue that voice without any identifying information is not personal data
- Ask for consent for commercial use/public dissemination
- Argue that the use of voice in the model is based on the legitimate interest

Scenarios to overcome problems relating to human voice in a model
Conclusions:

• Models are not necessarily subject to the same copyright and personal data restrictions as language data used as input;

• The limited impact of copyright and personal data restrictions on models potentially allows to:
  – Rely on favourable copyright/personal data provisions supporting research;
  – Make models publicly available;
  – Commercial use of models;
  – Personal data problems with voice need to be addressed.
Thank you