LAITech2022

TAL TECH

DEPARTMENT OF LAW

COMPREHENSIBILITY AND AUTOMATION

THE PLAIN LANGUAGE PERSPECTIVE IN THE ERA OF DIGITALIZATION

István Üveges
Doctoral School of Linguistics
University of Szeged (Hungary)

01.03.2022

OVERVIEW

- Comprehensibility: intersection between law, language and informatics
- Topics discussed:
  - Historical (and contemporary) approaches to accessibility
  - Access to justice and Plain Language
  - Public Accessibility Programme of the National Tax and Customs Administration of Hungary
  - Machine Learning (ML) and automation: can accessible drafting (or some components of it) be automated?

READABILITY (FORMULAS)

- Originated at the 1920's
- Wide-spread approach e.g., in
  - journalism,
  - research,
  - health,
  - law,
  - insurance
- Coarse-grained and mechanical method to measure if one can easily understand what a given text contains

Illustration

Flesch-Kinkade Reading Easy test (resulted in a number between 0 and 100)

$$206.835 - 1.015 \cdot \left(\frac{W}{S}\right) - 84.6 \cdot \left(\frac{Syl}{W}\right)$$

Where:

- W: number Words in the document
- S: number of Sentences in the document
- Syl: number of Syllables in the document

Higher score correlates harder readability (?)

COGNITIVE PERSPECTIVE - PSYCHOLINGUISTICS

- Levels of language processing
  - **Understanding**: processing the *literal* interpretation possibilities of the content; process of the grammatical structure (or syntax), the meaning of individual words (or lexical items, phrases)
  - **Comprehension**: understanding of the information compared to our knowledge of the world, and our understanding of how things should go, and how the newly accessed information relates to this. (Taking into account global – contextual data.)
- *Analytical* vs. *Holistic* stages of interpreting information carried by natural language.

PLAIN LANGUAGE MOVEMENT

- Started in the US with civil activists, lawyers and linguists in the 1970's
- Main goal: improve the effectiveness of communication (e.g., insurance, legislation, judicial proceedings)
- Improving the overall accessibility at any professional language medium, where the addressees of the professional text are not exclusively professionals but are mostly lay people
- Including explicit reference to linguistic works in many cases (Federal Plain Language Guidelines)
- *"Clear writing from your government is a civil right."*

Albert Arnold Gore Jr., 1998, Former Vice President of the United States

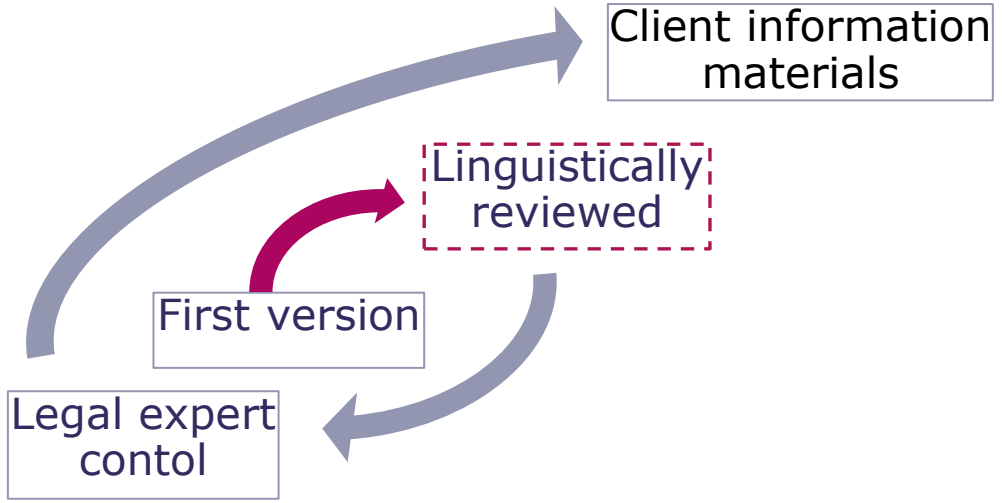
LAW AND LANGUAGE
(LANGUAGE AND LAW)

- "Law and Language" studies
  - Thoretically: Systematic efforts to use philosophical insights about language to solve problems in philosophy of law
  - Practically: investigating communicative situations where the „language of law" is used (courtroom studies, witness hearings etc.)

- Two main findings
  1. Everyone has the right to be protected from state supremacy: a **language equally understandable** to all participants should be used
  2. Lack of knowledge of a technical language (like legalese) can function as a **barrier in the way of access to justice**

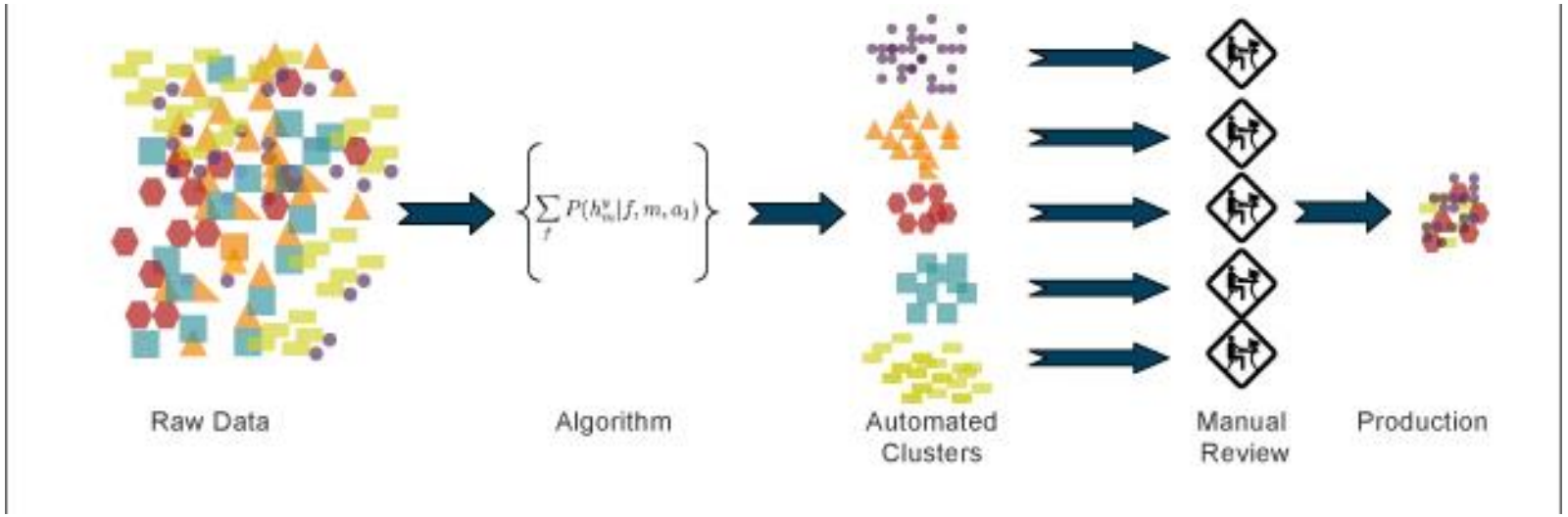
CORPUS BUILDING

- Digitalization + Automation + Plain Language = ?
- National Tax and Customs Administration of Hungary
  - public accessibility programme for about 4 years
  - 3 linguistic experts reviewing all of the Office's communication material for the lay public



MACHINE LEARNING I.

- Digitalization + Automation + Plain Language = Machine Learning (?)



MACHINE LEARNING II.

- Hungarian Tax Authority made available all the texts their linguistic experts rephrased in the past four years

Sub-corpora	„Original”	„Revised” (by linguistic experts)
Token number (and sentence number)	186.900 (5710)	87.754 (3100)
Rate	68.54%	31.45%

Table 1: *Training data*

Algorithm	Sub-corpora	Precision	Recall	F-measure
SVM	„Original”	**78%**	60%	68%
	„Revised”	48%	**68%**	56%
LR	„Original”	70%	**76%**	73%
	„Revised”	**46%**	39%	42%

Table 2: *Results after hypeparameter-tuning*

CONCLUSIONS

- What is this model can be used for?
  - Helping the work of linguistic or even legal experts when writing official material to lay people
- What are the limitations of this method?
  - Traditional ML methods, like SVM only deals with words, and NOT semantics
  - The current approach is very domain-specific
- Solution:
  - Application of more state-of-the-art methods, like neural networks (LSTM) or contextual word embeddings (BERT)

THANK YOU FOR YOUR ATTENTION!

Supported by the ÚNKP-21-3 - New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund.

TAL TECH

DEPARTMENT OF LAW

LAITech2022 STUDENT CONFERENCE

TALTECH.EE/EN/DEPARTMENT-LAW/LAITECH2022