Importance and advances in Language Data Processing in Slovakia

Radovan Garabík

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
LDS Country Workshop in Slovakia

2024-11-07, Bratislava

NLP

- start: sometime in 1950s

NLP

- start: sometime in 1950s
  - Machine Translation

NLP

- start: sometime in 1950s
  - Machine Translation
  - Information Retrieval

NLP

- start: sometime in 1950s
  - Machine Translation
  - Information Retrieval
  - ELIZA

NLP

- start: sometime in 1950s
  - Machine Translation
  - Information Retrieval
  - ELIZA
  - rule based

NLP

- start: sometime in 1950s
  - Machine Translation
  - Information Retrieval
  - ELIZA
  - rule based

NLP

- start: sometime in 1950s
  - Machine Translation
  - Information Retrieval
  - ELIZA
  - rule based
- skewed by English

NLP

- start: sometime in 1950s
  - Machine Translation
  - Information Retrieval
  - ELIZA
  - rule based
- skewed by English
- 1980s: rule based methods

NLP

- start: sometime in 1950s
  - Machine Translation
  - Information Retrieval
  - ELIZA
  - rule based
- skewed by English
- 1980s: rule based methods
- 1990s: statistical methods (golden era)

NLP

- start: sometime in 1950s
  - Machine Translation
  - Information Retrieval
  - ELIZA
  - rule based
- skewed by English
- 1980s: rule based methods
- 1990s: statistical methods (golden era)
- 2000s: machine learning, neural networks

NLP

- start: sometime in 1950s
  - Machine Translation
  - Information Retrieval
  - ELIZA
  - rule based
- skewed by English
- 1980s: rule based methods
- 1990s: statistical methods (golden era)
- 2000s: machine learning, neural networks
- 2010s: deep learning

NLP

- start: sometime in 1950s
  - Machine Translation
  - Information Retrieval
  - ELIZA
  - rule based
- skewed by English
- 1980s: rule based methods
- 1990s: statistical methods (golden era)
- 2000s: machine learning, neural networks
- 2010s: deep learning
- 2020s: transformers, LLMs

NLP

- start: sometime in 1950s
  - Machine Translation
  - Information Retrieval
  - ELIZA
  - rule based
- skewed by English
- 1980s: rule based methods
- 1990s: statistical methods (golden era)
- 2000s: machine learning, neural networks
- 2010s: deep learning
- 2020s: transformers, LLMs
- 2030s: AI will kill us all

Text Corpora

- What is a corpus?
- (a lot) of text, annotation, metadata, queries
- size matters

Text Corpora

- What is a corpus?
- (a lot) of text, annotation, metadata, queries
- size matters
- Slovak language corpora – start: 1993
- 2002 – Slovak National Corpus

Corpora sizes

- 1 million – lemmatization, morphological analysis
- 10 million – terminology extraction, named entity recognition
- 100 million – pocket dictionary (body of law, Slovak Wikipedia)
- 1 billion[1] – large dictionary, word embeddings, "classic" language models (SNK 1.6G)
- 10 billion – web corpus (HPLT+ARANEA 7G)
- 100 billion – English web corpora
- 500 billion – GPT-3
- 13 trillion[2] – GPT-4

[1] 10^9
[2] 10^{12}

Corpora sizes

- 1 million – lemmatization, morphological analysis
- 10 million – terminology extraction, named entity recognition
- 100 million – pocket dictionary (body of law, Slovak Wikipedia)
- 1 billion[1] – large dictionary, word embeddings, "classic" language models (SNK 1.6G)
- 10 billion – web corpus (HPLT+ARANEA 7G)
- 100 billion – English web corpora
- 500 billion – GPT-3
- 13 trillion[2] – GPT-4
- lifelong human exposure

[1] 10^9
[2] 10^{12}

Corpora sizes

- 1 million – lemmatization, morphological analysis
- 10 million – terminology extraction, named entity recognition
- 100 million – pocket dictionary (body of law, Slovak Wikipedia)
- 1 billion[1] – large dictionary, word embeddings, "classic" language models (SNK 1.6G)
- 10 billion – web corpus (HPLT+ARANEA 7G)
- 100 billion – English web corpora
- 500 billion – GPT-3
- 13 trillion[2] – GPT-4
- lifelong human exposure – 400 million words?

[1] 10^9
[2] 10^{12}

Use cases (corpora)

- lexicography

Use cases (corpora)

- lexicography
- all kind of linguistic research

Use cases (corpora)

- lexicography
- all kind of linguistic research
- **training data**

Use cases (corpora)

- lexicography
- all kind of linguistic research
- **training data**
- lemma, word, tag

Use cases (corpora)

- lexicography
- all kind of linguistic research
- **training data**
- lemma, word, tag
- word starting with é-

Use cases (corpora)

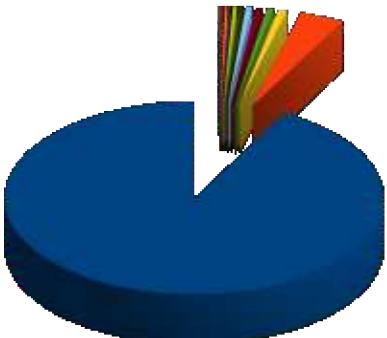
- lexicography
- all kind of linguistic research
- **training data**
- lemma, word, tag
- word starting with é-
- longest word

Use cases (corpora)

- lexicography
- all kind of linguistic research
- **training data**
- lemma, word, tag
- word starting with é-
- longest word
- tisícštyristosedemdesiatdvakorunový, devätnásťtisícšesťstoosemdesiatdva, osemnásťtisícdvestosedemdesiattri, deväťdesiatdeväťtisícdeväťstodeväťdesiatdeväť, tristoosemdesiatštyritisícštyristokilometrový

Use cases (corpora)

- lexicography
- all kind of linguistic research
- **training data**
- lemma, word, tag
- word starting with é-
- longest word
- tisícštyristosedemdesiatdvakorunový, devätnásťtisícšesťstoosemdesiatdva, osemnásťtisícdvestosedemdesiattri, deväťdesiatdeväťtisícdeväťstodeväťdesiatdeväť, tristoosemdesiatštyritisícštyristokilometrový
- hippopotomonstrosesquippedaliofóbia, hydrogénuhličitanovochloridová



slovenský
latinský
francúzsky
grécky
starohornonemecký
anglický
taliansky
český
praslovanský
stredohornonemecký
nemecký
germánsky
poľský
maďarský
španielsky
turkický
staroslovienský
ruský
turecký

Loanwords in Slovak

393475	musieť	sthnem
172916	problém	l←g
150089	škola	stlat←g
146048	otázka	čes
130222	chvíľa	sthnem
130170	skupina	čes
125669	štát	l
120238	systém	l←g
118469	situácia	f←l
117359	kniha	psl←akkad
109156	firma	t
105331	informácia	l
103613	program	l←g
94695	euro	g
94210	film	a
92650	cieľ	strhnem
91886	sociálny	l

English loanwords in Slovak

94210	film	a
81512	gól	a
63370	klub	a
62642	tím	a
55890	tréner	a←f
43347	in	10=l,9=a,1=t
35684	the	a
34992	dolár	a←n
32295	internet	l+a
31644	parlament	a←f
30633	partner	a←stfr←l
30607	televízia	g+a←l
30159	www	a+a+a
29962	of	a
27748	šport	a←strfr←f+l
26261	new	a
25058	nato	11=slk,9=a

Use cases (NLP)

- lexicography

Use cases (NLP)

- lexicography
- `https://slovnik.juls.savba.sk`

Use cases (NLP)

- lexicography
- `https://slovnik.juls.savba.sk`
- all kind of linguistic research

Use cases (NLP)

- lexicography
- `https://slovnik.juls.savba.sk`
- all kind of linguistic research
- fulltext search

Use cases (NLP)

- lexicography
- `https://slovnik.juls.savba.sk`
- all kind of linguistic research
- fulltext search
- information retrieval

Use cases (NLP)

- lexicography
- `https://slovnik.juls.savba.sk`
- all kind of linguistic research
- fulltext search
- information retrieval
- machine translation

Use cases (NLP)

- lexicography
- `https://slovnik.juls.savba.sk`
- all kind of linguistic research
- fulltext search
- information retrieval
- machine translation
- named entity recognition

Use cases (NLP)

- lexicography
- `https://slovnik.juls.savba.sk`
- all kind of linguistic research
- fulltext search
- information retrieval
- machine translation
- named entity recognition
- sentiment analysis

Use cases (NLP)

- lexicography
- `https://slovnik.juls.savba.sk`
- all kind of linguistic research
- fulltext search
- information retrieval
- machine translation
- named entity recognition
- sentiment analysis
- text classification

Use cases (NLP)

- lexicography
- `https://slovnik.juls.savba.sk`
- all kind of linguistic research
- fulltext search
- information retrieval
- machine translation
- named entity recognition
- sentiment analysis
- text classification
- TTS, speech recognition

Use cases (NLP)

- lexicography
- `https://slovnik.juls.savba.sk`
- all kind of linguistic research
- fulltext search
- information retrieval
- machine translation
- named entity recognition
- sentiment analysis
- text classification
- TTS, speech recognition
- OCR

Use cases (NLP)

- lexicography
- `https://slovnik.juls.savba.sk`
- all kind of linguistic research
- fulltext search
- information retrieval
- machine translation
- named entity recognition
- sentiment analysis
- text classification
- TTS, speech recognition
- OCR
- human-computer interatcion

Use cases (NLP)

- lexicography
- `https://slovnik.juls.savba.sk`
- all kind of linguistic research
- fulltext search
- information retrieval
- machine translation
- named entity recognition
- sentiment analysis
- text classification
- TTS, speech recognition
- OCR
- human-computer interatcion

Breaking point

- pre-LLM (“classic”)

Breaking point

- pre-LLM (“classic”)
- LLM

Corpora sizes

- $\sim$ 5 billion – teach a LLM new language
- *mistral-sk-7b-v0.1* – `https://www.juls.savba.sk/llm.html`

Sources

- Slovak National Corpus (2002 –) 1.6 G `https://korpus.sk`

Sources

- Slovak National Corpus (2002 –) 1.6 G `https://korpus.sk`
- web corpora (ARANEA) 7 G HPLT+ARANEA

Sources

- Slovak National Corpus (2002 –) 1.6 G `https://korpus.sk`
- web corpora (ARANEA) 7 G HPLT+ARANEA
- manually lemmatized & morphologically annotated corpus, 1.2 M

Sources

- Slovak National Corpus (2002 –) 1.6 G `https://korpus.sk`
- web corpora (ARANEA) 7 G HPLT+ARANEA
- manually lemmatized & morphologically annotated corpus, 1.2 M
- manually syntactically annotated corpus, 1 M tokens, 70 k sentences
  `https://korpus.sk/synt.html`

Sources

- Slovak National Corpus (2002 –) 1.6 G `https://korpus.sk`
- web corpora (ARANEA) 7 G HPLT+ARANEA
- manually lemmatized & morphologically annotated corpus, 1.2 M
- manually syntactically annotated corpus, 1 M tokens, 70 k sentences
  `https://korpus.sk/synt.html`
- specialized corpora:
  - spoken `https://korpus.juls.savba.sk/shk.html`
  - historical `https://korpus.juls.savba.sk/hist.html`
  - dialects

Sources

- Slovak National Corpus (2002 –) 1.6 G `https://korpus.sk`
- web corpora (ARANEA) 7 G HPLT+ARANEA
- manually lemmatized & morphologically annotated corpus, 1.2 M
- manually syntactically annotated corpus, 1 M tokens, 70 k sentences
  `https://korpus.sk/synt.html`
- specialized corpora:
  - spoken `https://korpus.juls.savba.sk/shk.html`
  - historical `https://korpus.juls.savba.sk/hist.html`
  - dialects
- parallel corpora – en, cs, ru, de, fr, la, hu, pl, bg, es
  `https://korpus.juls.savba.sk/par.html`

Level of support

- less resourced languages

Level of support

- less resourced languages
- Rehm, Georg, Way, Andy (eds), 2022, European Language Equality. Cognitive Technologies. Springer, Cham. https://doi.org/10.1007/978-3-031-28819-7_3

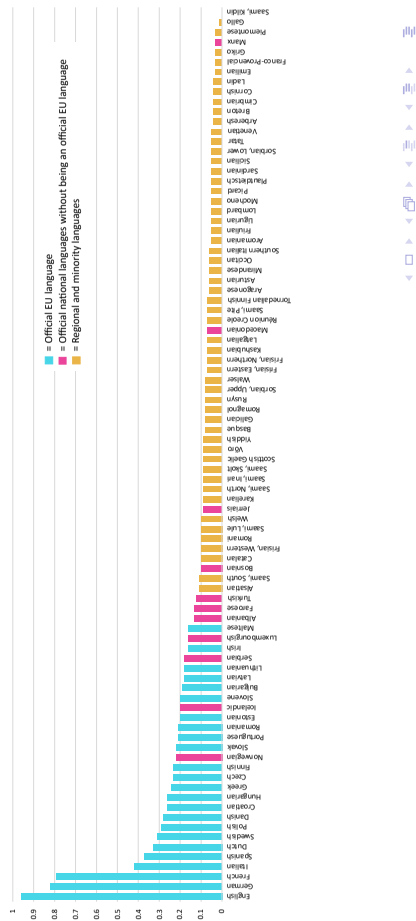


Fig. 2 Contextual Digital Language Equality scores as of late February 2023

Level of support

- LLM: the big equalizer

Level of support

- LLM: the big equalizer(?)
- multilingual models

Level of support

- LLM: the big equalizer(?)
- multilingual models
- death sentence for (even) small(er) languages?

Use case (LLM)

- Text (any) generation

Use case (LLM)

- Text (any) generation
- Spam

Use case (LLM)

- Text (any) generation
- Spam
- Fake news

Use case (LLM)

- Text (any) generation
- Spam
- Fake news
- Transition to post-truth society

Use case (LLM)

- Text (any) generation
- Spam
- Fake news
- Transition to post-truth society
- **Human-computer interaction (new level)**

Use case (LLM)

- Text (any) generation
- Spam
- Fake news
- Transition to post-truth society
- **Human-computer interaction (new level)**
- Information retrieval

Use case (LLM)

- Text (any) generation
- Spam
- Fake news
- Transition to post-truth society
- **Human-computer interaction (new level)**
- Information retrieval
- AI entertainment

Use case (LLM)

- Text (any) generation
- Spam
- Fake news
- Transition to post-truth society
- **Human-computer interaction (new level)**
- Information retrieval
- AI entertainment

Unresolved issues

- vendor lock-in

Unresolved issues

- vendor lock-in
- copyright

Unresolved issues

- vendor lock-in
- copyright
- it is distribution and derivative works that are protected

Unresolved issues

- vendor lock-in
- copyright
- it is distribution and derivative works that are protected
- “the Google way”

Unresolved issues

- vendor lock-in
- copyright
- it is distribution and derivative works that are protected
- “the Google way”
- **non-human creative work**

Unresolved issues

- vendor lock-in
- copyright
- it is distribution and derivative works that are protected
- “the Google way”
- **non-human creative work**
- precedence: juridical person

Thank you for your attention