Езикови технологии за български

Светла Коева, Институт за български език, Българска академия на науките

24 ноември 2022

Трети национален семинар за споделяне на езикови ресурси

- Налице ли е готовност за компютърна обработка на българския език в ерата на изкуствения интелект?
- Съществуват ли (съвременни) езикови технологии за български език?
- Кои са езиковите ресурси и технологии за български, които могат да подобрят цифровите услуги?

- Налице ли е готовност за компютърна обработка на българския език в ерата на изкуствения интелект?

НЕ

- Съществуват ли (съвременни) езикови технологии за български език?

ДОНЯКЪДЕ

- Кои са езиковите ресурси и технологии за български, които могат да подобрят цифровите услуги?

ДА ВИДИМ

Езикови технологии

- Езиковите технологии обхващат широка интердисциплинарна научна област, която се занимава с разработването на системи, способни да обработват, анализират, възпроизвеждат и „разбират“ човешките езици, независимо дали са в писмена, или в устна форма.

Езикови технологии

- Разчленяване и категоризиране на езикови единици
- Търсене и извличане на информация
- Анализ текст и реч
- Генериране на текст и реч
- Преобрзуване на текст и реч
- Обработка на езика в реално време
- Обработка на текст, реч, изображения и
  едновременно
- Многоезикова обработка

Приложения на езиковите технологии

- Откриване на правописни и граматични грешки и предложения за корекция
- Автоматично отговаряне на въпроси или изпълнение на гласови команди
- Откриване на дезинформация, на авторство, медиен мониторинг
- Анализ и предсказване на тенденции, пазарни проучвания, анализ на отношението към стоки и услуги

Приложения на езиковите технологии

- Компютърноподпомогнато обучение, опростяване на структурата на текстове с цел правилното им разбиране
- Автоматичен превод и компютърноподпомогнат превод
- Автоматично субтитриране, включително комбинирано с автоматичен превод в реално време

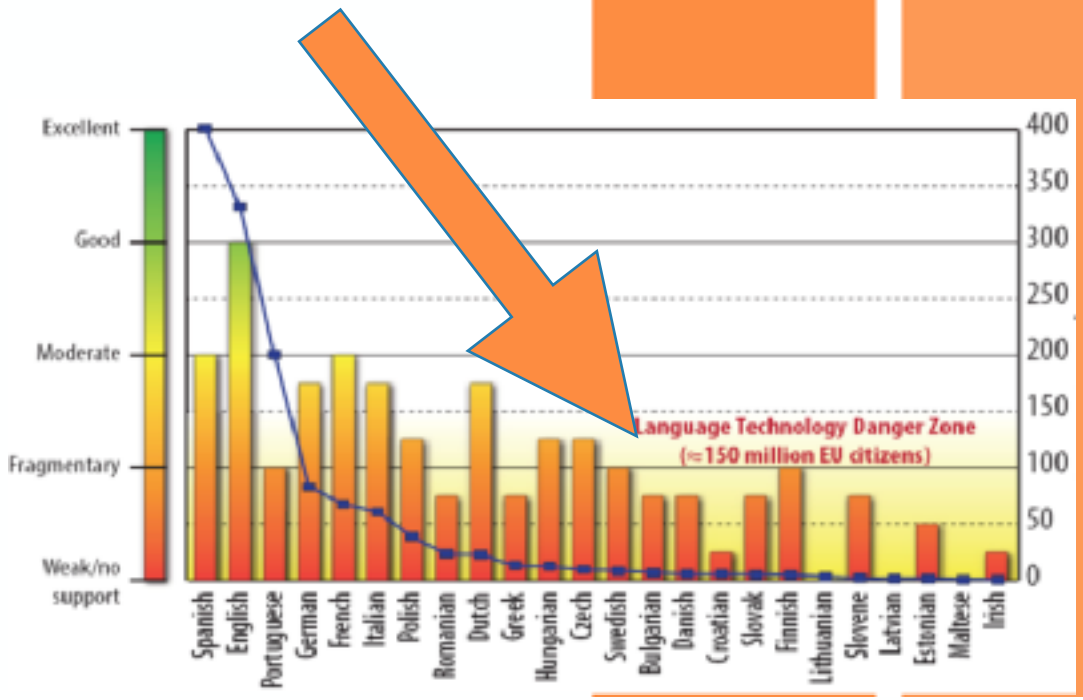
Бели книги на езиците в Европа

2012

- Езикови ресурси и технологии
- Проучване през 2012 г. на META-NET (Многоезикова европейска технологична асоциация)
- 31 европейски езика

Езиковите технологии ще дадат по-големи възможности за междуезикова комуникация и сътрудничество; ще осигурят еквивалентен достъп на носителите на различни езици до информация и познание (особено в условията на единен цифров пазар).

Бели книги на езиците в Европа



Отлично развити	Добре развити	Средно развити	Фрагментарно развити	Слабо/не-развити
	английски	испански италиански немски полски унгарски френски холандски чешки шведски	баски **български** галски гръцки датски естонски каталонски норвежки португалски румънски словашки словенски сръбски фински хърватски	ирландски исландски латвийски литовски малтийски

Българският език е застрашен от дигитална смърт

Българският е един от европейските езици, които са застрашени от дигитална смърт. Това сочат резултатите от ново изследване, проведено от Европейската мрежа за върхови постижения META-NET, в

Българският език заплашен "дигитална смърт"

Подобна е съдбата на 21 езика на Стария континент, с

Българският език – дигитална смърт

За 21 европейски езици има реален р езикови технологии

- 2012 г.: За момента българският език … не е застрашен, но ситуацията би могла да се промени значително с навлизането на ново поколение технологии.

Европейско езиково равенство

2022

- Технологична поддръжка за европейските езици в началото на 2022 година
- Европейско езиково равенство
- Повече от 40 изследователски институции за повече от 60 европейски езика[1] в началото на 2022 година

EUROPEAN LANGUAGE GR

HUGGING FACE

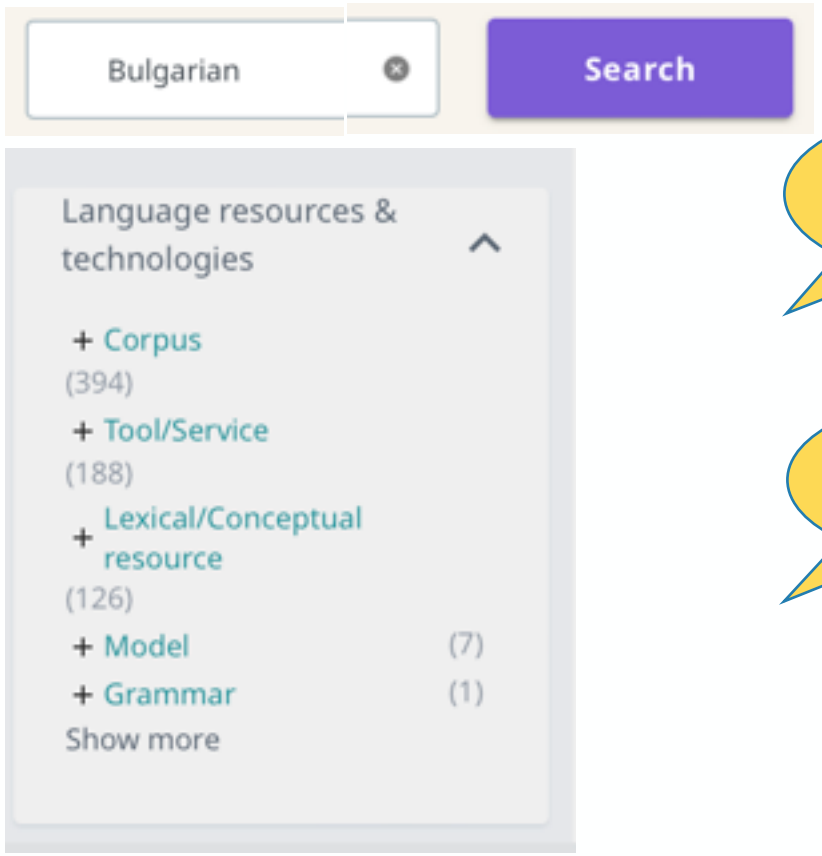
European Language Resource Coordination
Connecting Europe Facility
ELRC-SHARE

GitHub

META-NET

CLARIN

EUROPEAN LANGUAGE GRID



394 корпуса

126 ресурса

188 програми

Езикови технологии за български

- Налице са множество от инструменти за предварителна обработка на текста: **токънизация** (разделяне текста на последователности от символи); **разделяне на изречения**; откриване на **границите на абзаци**; **проверка на правописа**; **тагиране** (разпознаване на граматичните характеристики на думите или морфологичен анализ); **лематизация** (извеждане на основната форма на думите); **разпознаване на имена** на лица, организации и географски обекти; **синтактичен анализ, семантичен анализ** и др.

UDPipe NLP Cube

Езикови технологии за български

- Текстовият анализ все още доминира в областта на езиковите технологии за български език, а мултимодални данни (текст, изображения, аудио и видео) се обработват рядко едновременно, въпреки че прогнозите сочат, че видеосъдържанието скоро ще доминира в интернет.

Езикови технологии за български

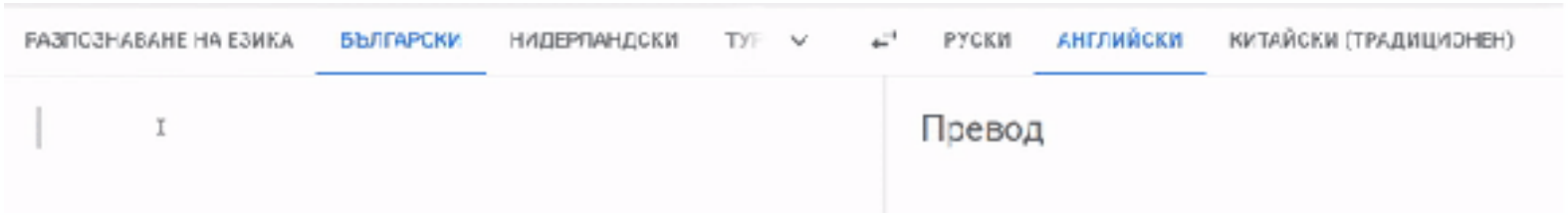
- Наличието на езикови модели е съществена предпоставка за развитието на приложения за компютърна обработка. Обучението на подобни модели отнема много време и изисква голямо количество подходящи ресурси.

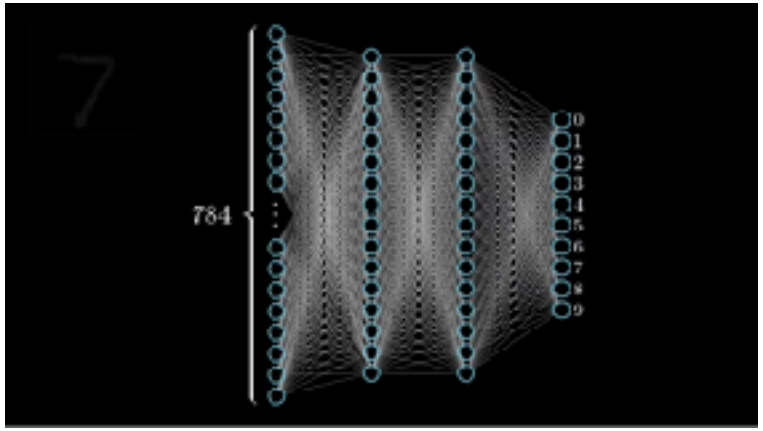
BERT RoBERTa slavic-bert XLM-R

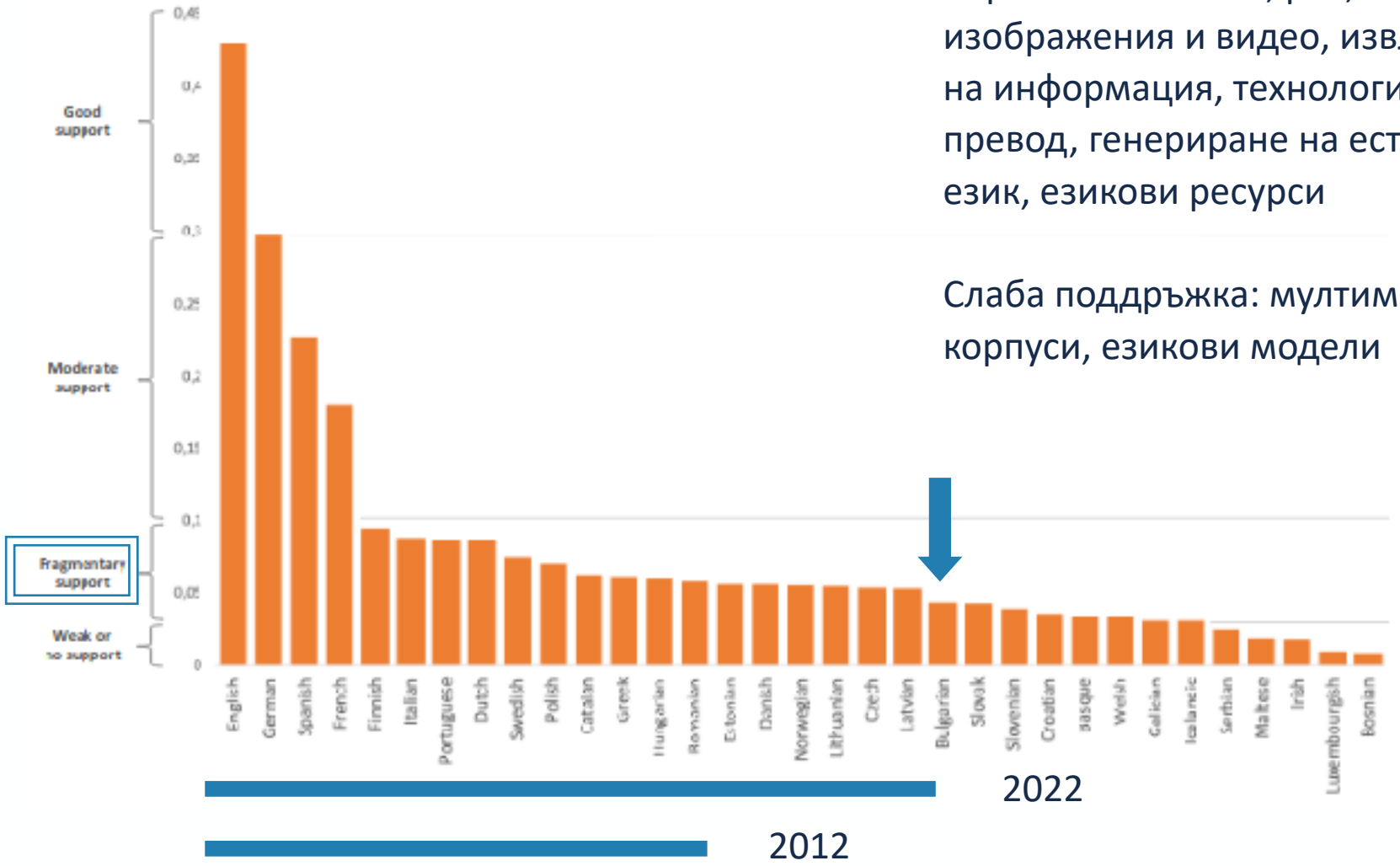
Иван застана ____ Мария.

Езикови технологии за български

- Съществуват редица системи за машинен превод от и на български език, базирани на технологията за невронен машинен превод.
- Оценката на качеството на съществуващите услуги за машинен превод, броят на езиковите двойки и обхватът на тематичните области все още определят технологиите за машинен превод за български език като недостатъчно развити.







Фрагментарна поддръжка: обработка на текст, реч, изображения и видео, извличане на информация, технологии за превод, генериране на естествен език, езикови ресурси

Слаба поддръжка: мултимодални корпуси, езикови модели

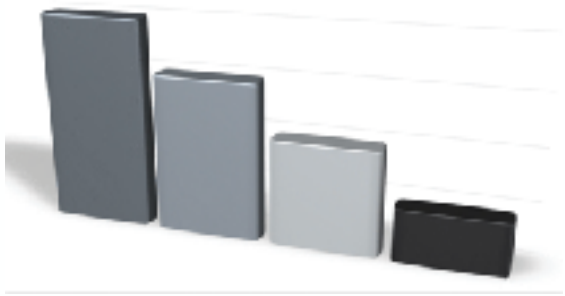
Много технологии все още не са налични (взаимодействие човек-компютър, мултимодална обработка, генериране на език и др.)

Няма готови за използване полезни приложения (автоматично резюмиране, отговаряне на въпроси и др.)

Технологиите за машинен превод на български са все още недостатъчно развити

Няма достъпни и надеждни системи за преобразуване на реч в текст на български език, особено работещи в реално време

Езикови технологии за български



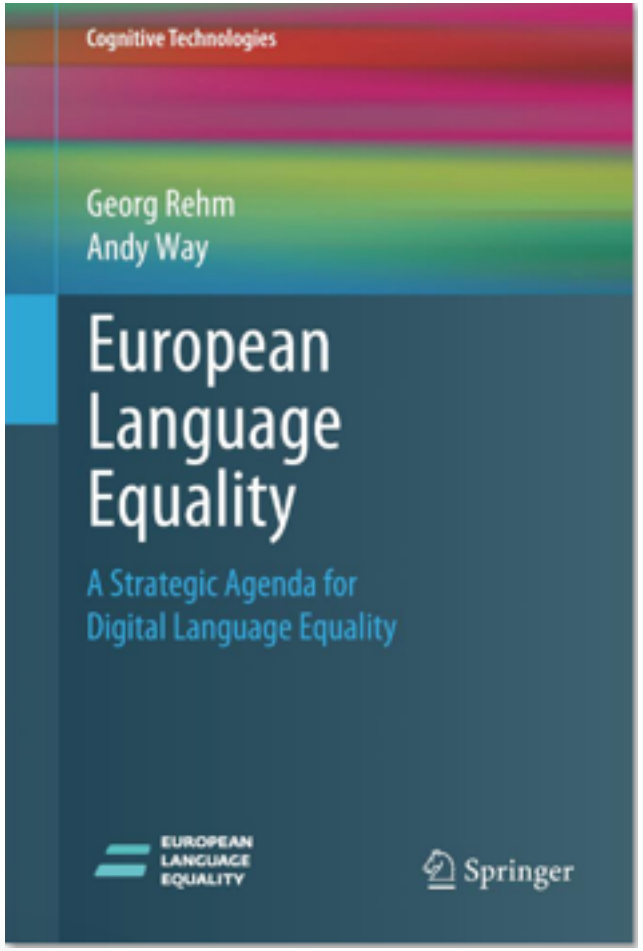
- Има нужда от отворени услуги в реално време за машинен превод от и на български език, комбиниращи текст и реч, които отчитат контекста, комуникативната ситуация и средата
- Технологиите за реч и текст за български трябва да се комбинират с технологии за други модалности: обработка на изображения и видео в реално време, работеща в многоезикова среда
- Разбирането на естествения език и генерирането на български трябва да стане част от многоезиковаа и мултимодална обработка

Езиковите технологии в помощ на публичната администрация

- През 2021 г. повече от 204 милиона страници са преведени с eTranslation – което е над двойно повече от предишния рекорд от почти 95 милиона страници, преведени през 2019 г.
- През последните години не само наборът от езици и услуги, поддържани от eTranslation, се разширява, но също така са предоставени редица инструменти,
  като преобразуване на реч в текст,
  разпознаване на имена на хора, организации и локации, класификация и автоматично анонимизиране на текст.

Езикови технологии за български

- Големи и разнообразни данни от публичния сектор, радио- и телевизионните оператори, социалните медии, издателите и др.
- Гъвкав достъп до високопроизводителна изчислителна техника, базирана на графични процесори
- Квалифицирани експерти в академичните изследователски центрове и изследователските звена на компаниите
- Цифрово езиково равенство за всички езици на Европейския съюз до 2030 година



Изкуственият интелект в полза на многоезикова Европа

- Необходимост от разбиране на значимостта на езиковите данни
  - Актуализация на Директивата за отворени данни (2019/1024/ЕС), която трябва да определи езиковите данни като данни с висока стойност
  - Подкрепа за споделянето на езикови данни и езикови технологии

ELRC White Paper

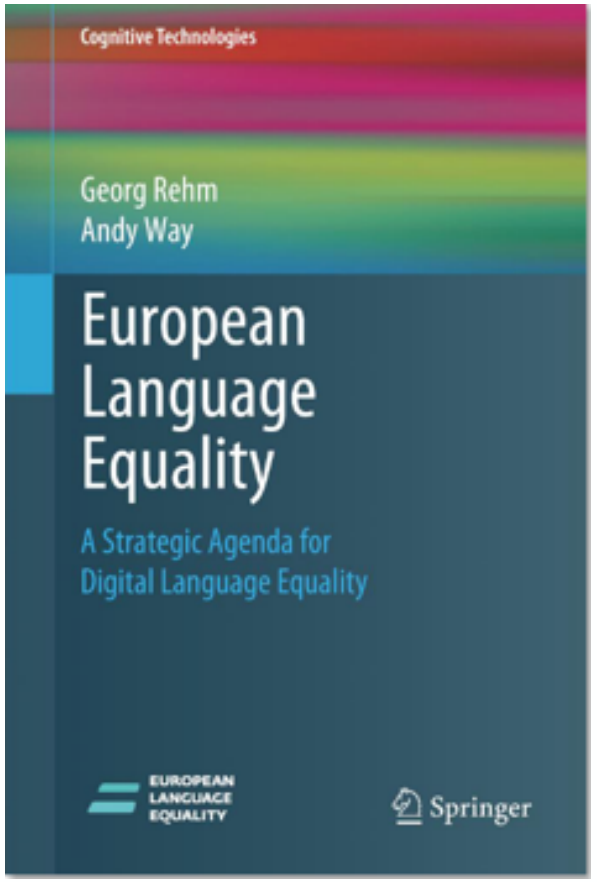
Изкуственият интелект в полза на многоезикова Европа

- Използване на компютърноподпомогнат превод и усвояване на дигитални умения
- Автоматизация на процеса на създаване на езикови данни и на техния превод, както и на работните процеси, свързани с тези дейности

ELRC White Paper

2012

2022



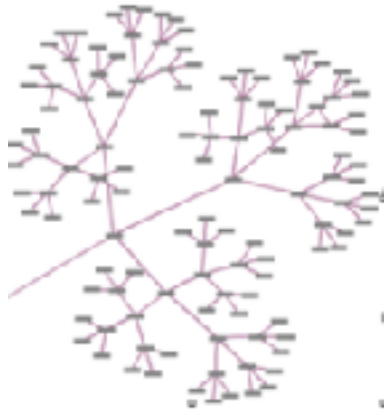
2022

ELRC White Paper

2030 → 2050 → 2100 →

- Важно е да се научим да се изненадваме от простите неща – например от факта, че телата падат надолу, а не нагоре. Началото на всяка наука се състои в осъзнаването, че най-простите явления от всекидневния живот повдигат много сериозни проблеми: защо явленията са такива, каквито са, а не други.

Ноам Чомски



Благодаря Ви за вниманието!

svetla@dcl.bas.bg