Българският език и развитието на езиковите технологии в България

Проф. д-р Светла Коева
Институт за български език
„Проф. Любомир Андрейчин“
Българска академия на науките

Българският език

- Население на Република България: 7 245 677 (2014)
- Процент от населението в Европейския съюз: 1.4 % (2014)
- Официален език: **български**
- България е член на Европейския съюз: от 1 януари 2007
- Официален език в Европейския съюз: **български**

http://europa.eu/about-eu/countries/member-countries/bulgaria/index_en.htm

Българският език

- **Българският език** се говори от близо 9 милиона души в България и по света.
- Много българи живеят в Испания, Великобритания, Германия, Италия, Франция, Австрия, Чехия, Словакия, Унгария, САЩ.

Българският език по света

- Гърция (около 30 000 души) (1998)
- Молдова (около 54 000 души) (2009)
- Румъния (около 6 700 души) (2003)
- Сърбия (около 13 300 души) (2013)
- Турция (около 350 000 души) (2014)
- Украйна (около 234 000 души)

http://www.ethnologue.com

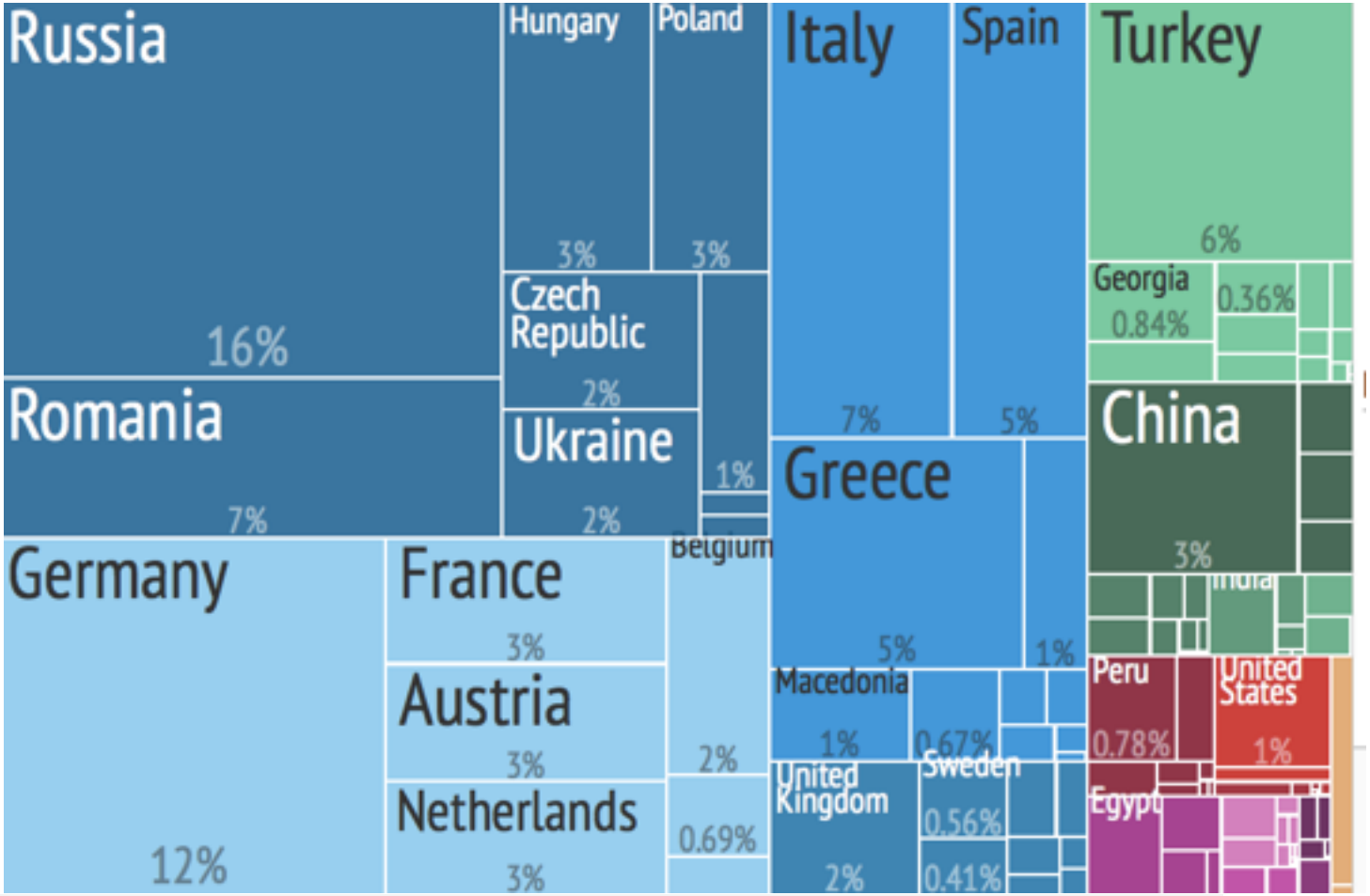
Българският език на ново място по света

- Ани Кочева-Лефеджиева, Институт за български език
- от 100 000 до 200 000 души
- от 20 000 до 50 000
- от 10 000 до 20 000

Българският език

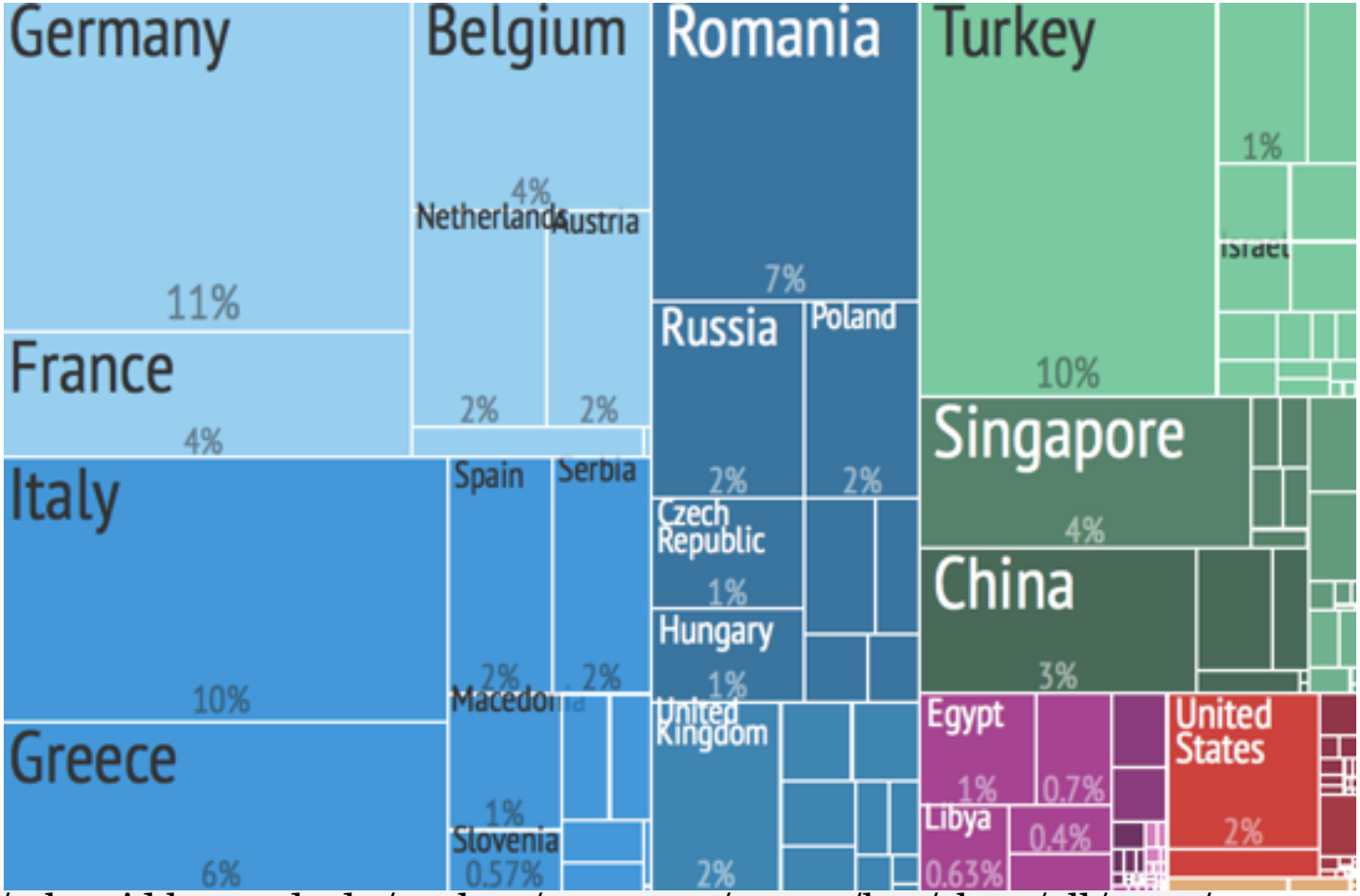
- Какво се случва в реалния живот:
  - търговия с другите европейски държави;
  - електронна търговия;
  - мобилност на пазара на труда в Европа;
  - мобилност на студентите в Европа;
  - туризъм от и в другите европейски държави;
  - миграция.

Внос на България през 2014 г.



http://atlas.cid.harvard.edu/explore/tree_map/export/bgr/show/all/2014/

Износ на България през 2014 г.



http://atlas.cid.harvard.edu/explore/tree_map/export/bgr/show/all/2014/

Гранични региони и езици

Регион	Държави	Езици
Балкански полуостров	Албания, Босна и Херцеговина, България, Гърция, Косово, Република Македония, Черна гора; част от Сърбия, Хърватия, Словения и Румъния; малка част от Турция и Италия	Населението на Балканския полуостров е разнородно. Най-многобройни са славяните – около 22 милиона, които също говорят на няколко различни езика.

Гранични региони и езици

Регион	Държави	Езици
Черно-морски регион	България, Грузия, Румъния, Русия, Турция, Украйна	български, грузински, румънски, руски, турски, украински

Гранични региони и езици

Регион	Държави	Езици
Дунавски регион	Австрия, България, Германия, Република Чехия, Румъния, Словакия, Словения, Унгария, Хърватия	български, хърватски, чешки, унгарски, словенски, словашки, немски, румънски

Многоезичието

- Основен принцип на Европейския съюз е **многоезичието.**

- Равнопоставеността на 24-те официални езика символизира равнопоставеността на страните в Европейския съюз.

Основен принцип

- След 1 януари 2007 г. българският език се използва при уреждане на отношенията на България с Европейския съюз в следните ситуации:
  - Официалният бюлетин, който съдържа правата на гражданите и текстове от европейското право, се публикува на български език.
  - Представителите на българските власти имат право да говорят на български език в Съвета на Европейския съюз.
  - **Българските граждани имат право да използват българския език в кореспонденцията си до европейските институции.**

Многоезикова Европа

- През последното десетилетие Европа успешно отстрани много от ограниченията за изграждането на Единен цифров пазар – един от ключовите европейски приоритети.
- Езиковите бариери обаче все още възпрепятстват пълноценното използване на различни възможности, предоставяни в интернет, и ограничават достъпа на потребителите до много електронни източници на информация.

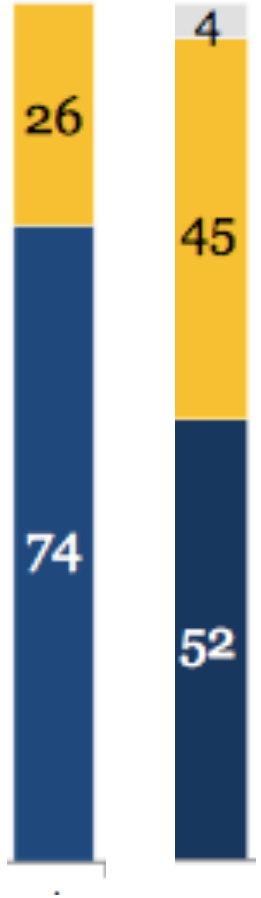
Многоезикова Европа

- Много често информацията, от която се нуждаят европейските граждани, не е достъпна на език, който разбират:
- **44% от европейците никога не са използвали език**, различен от този, който владеят и използват ежедневно, когато четат или разглеждат съдържание в интернет.
- **59% от европейците никога не са използвали език**, различен от този, който владеят и използват ежедневно, когато пишат в интернет.

http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf

Многоезикова Европа

- **26% от българите никога не са използвали друг език**, когато четат или разглеждат съдържание в интернет, а 45 % – когато пишат в интернет.



http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf

Българският език днес

- Статутът на езика зависи не само от **броя на неговите носители**, но и от **присъствието на езика в дигиталното информационно пространство** и софтуерните приложения.
  - През 2014 г. 56.7% от домакинствата в страната разполагат с интернет по данни на Националния статистически институт.
  - Към март 2016 г. Българската Уикипедия е на 36 място в света с 214 385 статии.

Алтернативи?

- Немислима алтернатива на многоезичието в Европа е да се позволи на един език да заеме доминираща позиция и да замени останалите езици.
- През вековете българският език е устоял на влиянието на думи и термини от гръцки и латински – езиците на познанието, както и на навлизането на френски и руски думи през 18-и и 19-и век и на руски думи през 20-и век.

Алтернативи?

- Един от начините за преодоляване на езиковите бариери е изучаването на чужди езици.
- Изучаването на чужди езици е полезно, но усвояването на 24-те официални езика на Европейския съюз и още около 60 други европейски езика представлява непреодолимо препятствие пред отделния човек.

Алтернативата

- Решение на проблема за отстраняването на езиковите бариери предлагат **езиковите технологии.**
- Езиковите технологии могат да подпомагат реално комуникацията между отделните хора, бизнеса и обществените институции, независимо от националните граници и езици.

Алтернативата

- Следващото поколение езикови технологии ще се усъвършенства в употребата на естествен език до такава степен, че потребителите ще общуват, използвайки собствения си език.
- Устройствата ще могат да намерят най-важната или най-релевантната информация в световното дигитално познание само с помощта на гласови команди.
- Новите езикови технологии ще предлагат по-съвършен **автоматичен превод**, адекватно резюме на текст, реч или на колекция от многоезикови документи.

Езикови технологии за български

- В областта на езиковите технологии за български съществуват редица ресурси, продукти и технологии.
- Има приложения за проверка на правописа и граматиката, за анализ и синтезиране на реч. Съществуват и програми за **автоматичен превод,** макар че не винаги се предлагат лингвистично коректни преводи, особено когато преводът е от друг език на български. Това се дължи основно на специфичните езикови характеристики на българския език.

Характеристики на българския език

- Някои от специфичните характеристики, които предопределят трудностите при компютърната обработка на българския език:
  - богата флективна система;
  - богата деривационна система;
  - видови двойки при глаголите;
  - свободно изпускане на подлога;
  - силно редуциране на падежната система;
  - наличие на задпоставен определителен член;
  - загуба на инфинитива;
  - свободен словоред.

Езикови технологии за български

- Направена е оценка на няколко области на приложение на езиковите технологии: разпознаване и синтезиране на реч, машинен превод, анализ, интерпретация и генериране на текст с помощта на 5-степенна скала по отношение на тяхното количество, достъпност, качество, покритие, развитост, устойчивост и гъвкавост.

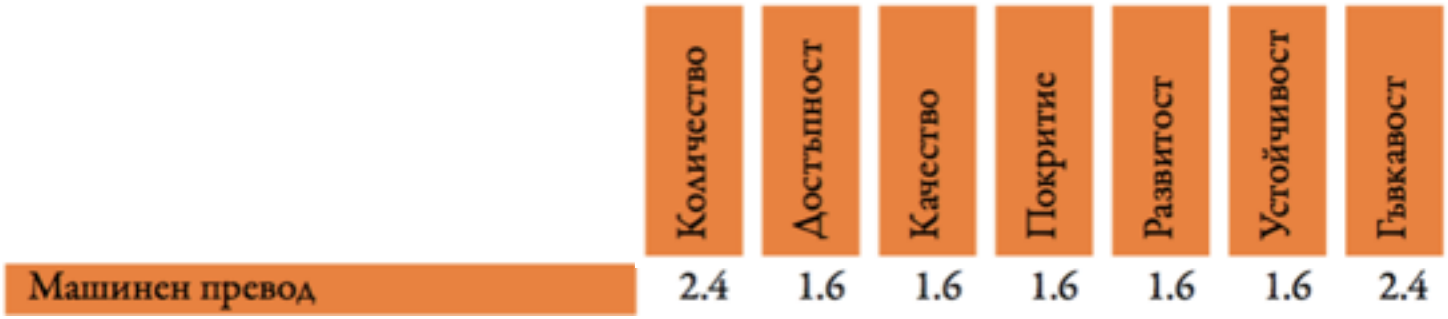
http://www.meta-net.eu/whitepapers/e-book/bulgarian.pdf

Езикови технологии за български

	Количество	Достъпност	Качество	Покритие	Развитост	Устойчивост	Гъвкавост
Езикови технологии: програми, технологии, приложения							
Разпознаване на реч	1.6	0.8	2.4	2.4	1.6	1.6	0.8
Синтезиране на реч	1.6	0.8	2.4	2.4	1.6	1.6	0.8
Анализ на текст	2.4	2	3.6	3.6	2.8	2.4	2.8
Интерпретация на текст	0.8	0.8	1.3	1.1	0.8	1.1	1.3
Генериране на текст	0.8	0.8	1.6	1.6	1.6	0.8	0.8
Машинен превод	2.4	1.6	1.6	1.6	1.6	1.6	2.4

Машинен превод

- За машинен превод са оценени качеството на съществуващите технологии, броят на езиковите двойки, покритието на езикови явления и тематични области, качеството и големината на съществуващите паралелни корпуси за обучение, броят и разнообразието на съществуващите компютърни приложения за машинен превод.



Езикови технологии за български

- Предварителна обработка: анализ или отстраняване на форматиране, разпознаване на езика на входния текст и др.
- Морфологичен анализ: разпознаване на думите, основната им форма и граматичните им характеристики; разпознаване на имена на хора, организации и географски названия.

Светла Коева (съст.). Езикови ресурси и технологии за български език. София, 2014.

Езикови технологии за български

- Граматичен анализ: разпознаване на границите на словосъчетанията и простите изречения в състава на сложното, разпознаване на вида на словосъчетанията и изреченията и синтактичните зависимости между тях.
- Семантичен анализ: отстраняване на многозначност (кое значение е правилно в даден контекст), свързване на анафорите (кое местоимение към кое съществително се отнася в дадено изречение или текст).

Светла Коева (съст.). Езикови ресурси и технологии за български език. София, 2014.

Езикови технологии за български

- Автоматична проверка на правописа и граматиката и предложения за корекция.
- Анализ на текст и реч и преобразуване на текст в реч и обратно.
- Автоматично категоризиране и клъстеризиране на (многоезикови) документи, автоматично извличане на резюме на документ или колекция от (многоезикови) документи, определяне на темата на текста.
- Търсене и извличане на информация от (многоезикови) документи, анализ на отношението на автора на текста към неговото съдържание.

Светла Коева (съст.). Езикови ресурси и технологии за български език. София, 2014.

Заключение

- Езиковите технологии в България са създадени през последните години от сравнително малко на брой хора и с много ограничени средства особено в сравнение с езици като английски, френски и немски, където различни колективи, включващи много хора, са работили дълги години, подкрепени със значително финансиране
- **Езиковите технологии за български са сравними по качество и функционалности с технологиите за останалите европейски езици.**