META-NET White Paper Series

# Езиците в европейското езиково информационно общество

## Том X – Български език

# Предговор

Настоящата разработка (наречена още “Бяла книга”) е част от серия документи, които представят развитието в областта на езиковите технологии и потенциала им при обработката на съответните езици. Целевите групи на тези документи са учители и преподаватели, журналисти, политици, езикови общности и други.

Достъпът до езикови технологии за различните езици, които се говорят в Европа, е много различен. Затова и действията, които се изискват за подкрепа на изследванията и развитието на езиковите технологии, също са различни. Належащите мерки зависят от много фактори, като например сложността на дадения език и размера и устойчивостта на общността на носителите на езика.

META-NET е мрежа за високи постижения, изградена с подкрепата на Европейската комисия и целяща анализ на съществуващите на този етап езикови ресурси и технологии. Анализирани са 23-те официални европейски езика, както и редица други по-важни национални и регионални езици в Европа. Резултатите подсказват, че все още липсват някои или повечето от необходими технологии за почти всички езици. Един по-детайлен и експертен анализ и оценка на актуалната ситуация ще помогне за увеличаване на ефекта от допълнителните изследвания и минимизиране на риска от пропуски.

В META-NET участват 47 изследователски центрове от 31 страни, които работят съвместно с бизнеса, правителствени агенции, индустрия, изследователски организации, софтуерни компании, доставчици на технологични услуги и продукти и европейски университети. Тази мрежа създава обща технологична визия и разработва програма за стратегически проучвания, която да предложи стратегии, чрез които езиково-технологичните приложения да попълнят липсите сред известните ни научно-изследователските продукти до 2020 г.

## Отпечатък

Автори:

доц. д-р Диана Благоева, Институт за български език към БАН
проф. д-р Светла Коева, Институт за български език към БАН
проф. дфн Владко Мурдаров, Институт за български език към БАН

# СЪДЪРЖАНИЕ

# Резюме

Много европейски езици са заплашени от изчезване в дигиталната епоха, тъй като не са достатъчно добре представени в интернет пространството. Широките възможности на регионалните пазари остават неизползвани заради езиковите бариери. Ако не се действа своевременно и достатъчно бързо, много европейски граждани ще се окажат в социална и икономическа изолация само защото говорят предимно родния си език.

Иновативните езикови технологии са посредник, който ще помогне на всички европейци да участват в едно равноправно и икономически успешно общество, основано на знанието и информацията. Мултиезиковите технологии могат да помогнат за бързата, евтина и лесна комуникация и взаимодействие между хората и организациите отвъд езиковите граници.

Днес услуги, базирани на езикови технологии, се предлагат предимно от търговски компании от САЩ. Една от тях е Google Translate, която предоставя свободно онлайн превод от и на различни езици. Успехът на Watson, компютърната система на IBM, която спечели епизод от играта Jeopardy в съревнование с хора, илюстрира огромния потенциал на езиковите технологии. Като европейци трябва да си зададем следните важни въпроси:

- Трябва ли общуването ни и инфраструктурата за достъп до знания да зависят от компании, които имат ефективен монопол на съответния пазар?
- Можем ли наистина да разчитаме на езикови услуги, които във всеки момент могат да бъдат прекратени, независимо от нашата воля?
- Активни участници ли сме в глобалния пазар за изследвания и развитие на езиковите технологии?
- Дали трети страни от други държави имат желание да работят за разрешаване на проблема с автоматичния превод, както и за другите предизвикателства пред европейското многоезиково общество?
- Може ли европейският културен опит да помогне за израстване на общество, основано на знание, като осигури достъп до по-добри, по-сигурни, по-прецизни, иновативни и стабилни висококачествени езикови технологии?
- Бялата книга за българския език показва, че България разполага с жизнена индустрия за езикови технологии и подходяща научна среда. Макар да съществуват редица технологии и езикови ресурси за български език, все още за него няма толкова много и качествени технологии и ресурси, колкото за английски.

Анализът, представен в настоящия документ, налага извода, че са необходими незабавни действия, за да бъде отчетен значим подем в изграждането на езикови технологии за български.

META-NET допринася за изграждането на силно, многоезиково европейско дигитално информационно пространство. С постигането на тази цел мултикултурният съюз от нации може да просперира и да се превърне в модел за мирно и равнопоставено сътрудничество. Ако целта не бъде постигната, Европа ще трябва да избира между отказ от културната си идентичност и икономическия провал.

# Заплаха за езиците и предизвикателството пред езиковите технологии

*В момента сме свидетели на дигитална революция, която може да се сравни с изобретяването на печетарската преса от Гутенберг.*

През последните десетилетия сме свидетели на дигитална революция, която промени драматично комуникацията и обществото като цяло. Последното развитие на дигитализацията и технологиите за мрежови комуникации понякога се сравняват с откриването на печатарската преса от Гутенберг. Как тази аналогия ни позволява да погледнем в бъдещето на европейското информационно общество и нашите езици?

След откритието на Гутенберг истинските революции в комуникацията и обмена на знание са плод на инициатива например като превода на говорим език, който Мартин Лутер прави на Библията. В следващите столетия тези техники претърпяват сериозно развитие с цел компютърната обработка на естествения език и обмена на информация:

- стандартизацията на правописа и граматиката помогат за по-бързото разпространение на новите научни и интелектуални идеи;
- появата на официалните книжовни езици позволява на гражданите да общуват в рамките на определени (често политически) граници;
- обучението по и преводът от различни езици улесняват междуезиковия обмен на информация;
- журналистическите и библиографските документи, издадени на стандартен писмен език, осигуряват качество и по-лесен достъп до публикуваните материали;
- съществуването на различни медии, сред които вестници, радио, телевизия, книги и документи в различни формати, задоволява нуждите от различни видове комуникация.

В последните 20 години информационните технологии позволяват автоматизирането на процеса за по-голяма ефективност:

- Софтуерът за електронно публикуване замества пишещите и печатните машини;
- Microsoft PowerPoint (и други програми) заменят слайдовете за прожектиране;
- Документите се изпращат и се получават по-бързо по имейл, отколкото по факс.
- Skype (и други програми) позволяват телефонни разговори по интернет и организиране на виртуални срещи и семинари.
- Форматите за кодиране на аудио и видео съдържание помагат за обмена на мултимедийни документи.

- Търсачките осигуряват достъп до уеб страници по ключови думи.
- Онлайн услугите като Google Translate осигуряват бърз и приблизителен превод.
- Социалните медии улесняват общуването и споделянето на информация.

Макар подобни инструменти и приложения да са в помощ, те не могат да предоставят достатъчно ефективна подкрепа за мултиезиковото европейско информационно общество, което осигурява равни възможности и в което информацията и стоките могат да се придвижват свободно.

## Езиковите граници като пречка пред европейското информационно общество

Не можем да знаем с точност как ще изглежда информационното общество на бъдещето. Когато става дума за общата европейска стратегия в областта на енергетиката или външната политика, се вслушваме в думите на външните министри на европейските страни, произнесени на родния им език. Може да се създаде система, чрез която хората, говорещи различни езици и с различно ниво на езикова компетентност, ще могат да обсъждат специални теми, докато технологията автоматично събира мненията им и генерира кратки резюмета. Или да се говори с обслужващия клон на здравноосигурителна компания, която се намира в чужда страна.

*В глобалната икономика и информационно пространство се срещаме с все повече езици, с носителите им и със съдържание.*

Ясно е, че днес нуждите от общуване са много по-различни от преди няколко години. В глобалната икономика и информационно пространство се срещаме с все повече езици, с носителите им и със съдържание, които изискват от нас бързо общуване чрез новите видове медии. Повишената популярност на социалните медии (Wikidepia, Facebook, Twitter, YouTube) е само върхът на айсберга.

Днес прехвърляме гигабайтове текст от една точка на света до друга за няколко секунди, преди да разберем, че е написан на език, който не познаваме. Според скорошно проучване, изготвено по поръчка на Европейската комисия, 57% от потребителите на интернет в Европа купуват стоки и услуги на езици, които не са им родни (най-разпространеният е английският, следван от френския, немския и испанския). 55% от потребителите четат съдържание на чужди езици, докато само 35% използват чужд език, когато пишат имейли или публикуват съобщения и коментари онлайн. Преди две години английският е бил lingua franca на уеб мрежата, защото голяма част от съдържанието в интернет е било на английски. Сега обаче ситуацията е коренно променена след експлозивното нарастване на

съдържанието на други езици (особено такива, които се говорят в Азия и в арабския свят).

Дигиталната пропаст между езиковите граници изненадващо не привлича внимание в публичното пространство. И все пак именно тя налага логичния въпрос: "Кои европейски езици ще се развият и ще останат в информационното и познавателното общество на световната мрежа?"

## Езици в риск

Печатарската преса засилва обмена на информация в Европа, но води и до отмиране на много европейски езици. Текстове на регионални и малцинствени езици рядко се появяват в печатна форма. Като резултат много езици, например корнуолския или далматинския, често са ограничени до говоримите си форми, което намалява възможностите за тяхното усвояване, разпространение и употреба.

Близо 60-те езика, които се говорят в Европа, са сред най-богатите и важни културни активи на Стария континент. Езиковото богатство в Европа е и важен компонент от социалния успех [i]. Популярни езици като английски или испански със сигурност ще продължат да съществуват в развиващото се дигитално общество, както и на дигиталния пазар, но много европейски езици ще бъдат изключени от дигиталната комуникация и няма да имат голямо значение в интернет обществото. Това не е особено желателно. От една страна, ще изгубим стратегически възможности, което ще подкопае глобалните позиции на Европа. От друга страна, подобни загуби противоречат на целта за равно участие на всеки европейски гражданин, независимо от езика, на който говори. Според доклад за многоезиковото общество на ЮНЕСКО езиците са важен посредник за осигуряване на достъп до основни човешки и потребителски права, като правото на изразяване на политическите убеждения, образованието и участието в обществото[ii].

*Голямото разнообразие от езици в Европа е една от най-важните ценности - от съществено значение за успеха на Европа.*

## Езиковите технологии – ключ към уменията

В миналото усилията на инвеститорите са били насочени към езиковото обучение и превода. Според някои изчисления през 2008 г. европейският пазар за писмен и устен превод, локализация на софтуер и глобализация на уеб сайтове се е равнявал на 8,4 млрд. евро и се е очаквало да нараства с 10% на година[iii]. Дори и този актуален капацитет обаче не е достатъчен, за да се отговори на днешните и бъдещите нужди на потребителите.

Езиковите технологии са ключови помощни технологии, които могат да предпазят и подкрепят европейските езици. Те помагат на хората да си сътрудничат, да търгуват, да споделят знание и да участват в социалния и

политическия живот независимо от езиковите бариери и компютърните си умения. Езиковите технологии ни помагат в ежедневните дейности като писане на имейли, търсене на информация, стоки и услуги в интернет или резервиране на полети. Ние се възползваме от езиковите технологии, когато:

*Езиковите технологии помагат на хората да си сътрудничат, да правят бизнес, да споделят знание и да участват в социалния и политическия живот*

- търсим информация в интернет с помощта на търсачка;
- проверяваме правописа на написания от нас текст в текстообработваща програма;
- разглеждаме предложенията в онлайн магазин;
- слушаме гласови указания на навигационна система;
- превеждаме уеб страници онлайн.

Езиковите технологии, представени в настоящия документ, са важна част от бъдещите иновативни приложения. Те са помощни технологии в рамките на една по-широка система от приложения, например навигационни системи или търсачки. Документите, наречени "бяла книга", анализират възможностите на основните езикови ресурси и технологии в даден сектор за работа с всеки един от европейските езиците.

В близко бъдеще ще се нуждаем от езикови технологии за всички европейски езици. Тези технологии трябва да са налични, работещи, достъпни от финансова гледна точка и тясно интегрирани в обща софтуерна среда. Работата на потребителя в интерактивна, мултимедийна и многоезикова среда е невъзможна без посредничеството на езиковите технологии.

*Многоезичието е правило, не изключение.*

## Възможности за езиковите технологии

Езиковите технологии участват при изграждането на приложенията за автоматичен превод, при създаването на съдържание и при управлението на многоезикова информация и знания на всички европейски езици. Тези технологии ще подпомогнат развитието на интуитивни езиково базирани интерфейси за домашна електротехника, машини и машинни съоръжения, автомобили, компютри и роботи. Макар че прототипите за някои от тях вече съществуват, те все още са в началото на развитието си. В последните години обаче се откриха много възможности. Машинният превод например вече е достигнал задоволителна точност в специализирани области, а експерименталните приложения осигуряват възможности за управление на многоезикова информация и знание, както и за създаване на съдържание на много европейски езици.

*Сегашното развитие на технологичния прогрес е твърде бавно, за са осигури през следващите десет - двайсет години значими софтуерни продукти.*

Приложенията, като езиковите и гласовите потребителски интерфейси и диалогови системи, се използват само в специализирани сфери и са с ограничени възможности. Активно се разработват технологии за подкрепа при спасителни операции в

райони, засегнати от природни бедствия. В подобна рискова среда точността на превода е въпрос на живот и смърт. Същото важи за езиковите технологии в здравеопазването. Интелигентните роботи, които могат да предават информация на много езици, спасяват човешки живот.

Съществуват значителни пазарни ниши в образователната и развлекателната индустрия, където езиковите приложения могат да бъдат интегрирани в игри, образователно-развлекателни програми, симулации на среда и тренировъчни системи.

Услугите за мобилна информация, приложенията за компютърно подпомогнато езиково обучение и за електронно обучение, както и софтуерът за разпознаване на плагиатство и за самооценка, са още една от сферите за внедряване на тези технологии. Социалните медии като Twitter и Facebook са друго поле за работа на усъвършенствани езикови технологии за наблюдение, обобщение, разпознаване на мнения и емоции, както и за идентифициране на нарушени авторски права или злоупотреба с новите комуникационни канали.

Езиковите технологии предлагат невероятни икономически и културни възможности за Европейския съюз. Мултикултурализмът вече се е наложил в Европа. Европейският бизнес, организациите и училищата също са мултинационални и залагат на разнообразието. При общуването си гражданите искат да преодолеят езиковите граници, които все още съществуват на единния европейски пазар. Езиковите технологии могат да ни помогнат да преодолеем останалите бариери, подкрепяйки свободната употреба на езика. Иновативните, многоезикови технологии за европейските езици могат да ни помогнат и при контакти с глобални ни партньори и съответните им мултиезикови общества. Езиковите технологии подкрепят разширяването на международните икономически възможности.

## Предизвикателства пред езиковите технологии

Макар че езиковите технологии отбелязаха значителен напредък в последните години, настоящият ход на технологичния прогрес и иновативното производство е твърде бавен. Не можем да чакаме 10 или 20 години за съществен подем в развитието на комуникациите и производителността в многоезикова среда.

По-широко използваните технологии, като приложенията за корекция на правописни и граматични грешки в съвременните системи за текстообработка, или са едноезикови, или са достъпни само на определени езици. Приложенията за многоезикова комуникация,

например за машинен превод, изискват много по-интелигентни решения. Услугите за онлайн превод като Google Translate или Bing Translator са полезни, ако искаме да придобием най-обща идея за съдържанието на даден чуждоезиков текст. Но тези услуги, както и приложенията за професионален машинен превод, често се затрудняват, особено при нужда от точен и пълен превод. Забавните грешки в превода (като дословните преводи на имена като 'Bush' или 'Kohl') са сред най-известните примери на трудностите, които стоят пред езиковите технологии.

## Как се учи език?

За да разберем как компютрите обработват езиковата информация и защо се затрудняват с тази задача, нека хвърлим поглед към начина, по който човешк научава родния си език и как се учи втори език, а след това накратко ще представим и начина, по който работят системите за машинен превод. Не бива да се забравя, че езиковите технологии са тясно свързани с усилията за създаване на изкуствен интелект.

Човек усвоява езика по два различни механизма. Детето научава родния си език, докато слуша общуването между носителите на езика. Излагането на конкретни примери от езика, произнесени от родители, братя, сестри и други членове на семейството, помага на децата след двегодишна възраст да формулират първите си думи и кратки фрази. Детето е в състояние да прави това, тъй като наследява специална генетична структура, която му позволява да учи езици.

*Хората придобиват езикови умения по два различни начина: посредством примери или езикови правила.*

Усвояването на втори език обикновено изисква доста повече усилия, ако детето не живее в обществото на носителите на този език. В училищна възраст чуждите езици най-често се усвояват чрез заучаване на граматичната структура на съответния език, както и на речника и ортографията му с помощта на учебници и учебни пособия, които описват лингвистичното знание под формата на абстрактни правила, парадигми и примерни текстове. Усвояването на чужд език отнема време и усилия и е все по-трудно с напредване на възрастта.

Двата основни вида системи за езикови технологии усвояват езика по начин, подобен на този при човешките същества. Статистическите приложения придобиват лингвистични знания от големи корпуси с определени текстове, които са написани на един или на два и повече езика и са част от т. нар. паралелни корпуси. Алгоритмите за машинно обучение могат, поне до известна степен, да учат моделите, по които думи, кратки фрази или цели изречения са използвани в даден език, или как се превеждат от един на друг език. За тази цел са нужни огромен брой изречения. Качеството на

*Двата основни типа системи, базирани на езикови технологии, усвояват езика по подобен на хората начин.*

резултатите се повишава заедно с броя на анализираните изречения. Не е рядкост подобни системи да се обучават въз основа на корпуси от милиони изречения. Това е и една от причините, поради които компаниите, разработващи търсачки, се стремят да съберат колкото може повече писмени езикови данни. Корекцията на правописа в софтуера за текстообработка, достъпната онлайн информация, услугите за превод като Google Search и Google Translate работят чрез тези статистически (базирани на данните) методи.

Езиковите системи, основани на формални правила, са вторият вид езикови технологии. Лингвисти, компютърни лингвисти и програмисти кодират граматични правила (правила за превод) и изграждат речникови списъци (лексикони). Изграждането на системи, основани на правила, струва много време и човешки труд. Нужна е и интензивната намеса на висококвалифицирани експерти. Някои от водещите системи за машинен превод, основани на правила, се изграждат в продължение на над 20 години. Предимство на системите, базирани на правила, е, че експертите могат да контролират обработката на езиковата информация. Те могат да коригират грешки на софтуера, както и да отговарят на въпроси и искания на потребителите, особено в случаи, когато системите, основани на правила, се използват в езиковото обучение. Заради финансовите ограничения езиковите технологии, основани на правила, са достъпни и ефективни само за най-разпространените езици.

# Българският език в европейското информационно общество

## Общи данни

Българският език е официалният език на Република България. Говори се от близо 10 млн. души предимно в България, но също и в Гърция, Македония, Румъния, Сърбия, Турция (европейската част), Украйна, Австралия, Канада, САЩ, Германия и Испания. По-големи общности от говорещи български са регистрирани и в Хърватия, Чешката република, Унгария, Израел, Либия, Молдова, Руската федерация (европейската част), Словакия[iv]. Агенцията за българите в чужбина към Министерския съвет отговаря за българските общности в съседните на България страни и имигрантските общности в различни държави по света[v].

Предварителните данни на Националния статистически институт[vi] след преброяването на населението на България показват, че към 1 февруари 2011 г. населението на страната е било 7 351 633 души. Колко българи живеят в страни от Европейския съюз ще стане ясно през 2012 г., когато ще бъдат публикувани данните от преброяването и в тези държави.

Официалната азбука за писане на български език е кирилицата. Българският е първият славянски език, който разполага със собствена писмена система, датираща от IX в. на н. е. През 886 г. България приема глаголическата азбука, създадена от равноапостолите св. св. Кирил и Методий, обявени за покровители на Европа. Глаголицата обаче постепенно е заменена от кирилицата – азбука, създадена и наложена от Охридската и Преславската книжовна школа в началото на X век. На 1 януари 2007 г., когато България е приета за пълноправен член на Европейския съюз, кирилицата става третата официална азбука на Европейския съюз след латинската и гръцката.

Българските диалекти са регионалните говорими варианти на българския език, които са разпространени както в границите на страната, така и извън нея. Българските диалекти са разделени на западни и източни от т.нар. “ятова граница”, която разграничава две по-големи диалектни области въз основа на различните звукови варианти на старобългарската гласна “ят” (ѣ), произнасяна в определени условия като [’а] или [е] на изток от границата (например в ед. ч. [б'ал], но в мн. ч. [бели]) или само като [е] на запад от ятовата граница (ед. ч. [бел] и мн. ч. [бели]) .

## Особености на българския език

Българският език принадлежи към групата на южнославянските езици и едновременно с това е част от

т. нар. Балкански езиков съюз (Balkan Sprachbund), поради което има характеристики и на двете езикови групи. Като славянски език българският притежава богата флективна и деривационна система, глаголите са разделени по двойки въз основа на техния аспект. В същото време езиците в Балканския езиков съюз са си повлияли взаимно и българският загубва падежите при имената (с изключение на звателната форма) и инфинитива при глагола.

Освен общата лексика, някои от най-важните граматични характеристики на българския език, които го отличават като славянски език (някои от които споделя и с други балкански езици), са:

- Богата флективна система. Богатата флективна система до известна степен затруднява работата в областта на езиковите технологии. При лематизацията например могат да възникнат проблеми при разпознаване на определени флективни видове, които принадлежат към различни части на речта. Думата “бели” например е омограф и може да означава:

  форма за мн. ч. на съществителното “беля”;

  форма за мн. ч. на прилагателното “бял”;

  форма за 3 лице, ед. ч., сегашно време на глагола “беля”;

  форма за 2 или 3 лице, ед. ч., минало свършено време на глагола “беля”;

  форма за 2 лице, ед. ч., повелително наклонение на глагола “беля”.

- Богата деривационна система: умалителни и увеличителни съществителни: стол – столче; суфикси за женски род на съществителните за лица (професии): учител – учителка; и други.
- Видови двойки при глаголите: родя – раждам;
- Свободно изпускане на подлога (име или местоимение), защото глаголното окончание носи информация за вършителя на действието: Аз чета книга. - Чета книга.

В хода на историческото развитие на българския език и контактите му със съседните неславянски езици на Балканския полуостров в структурата му настъпват значителни промени в сравнение с останалите славянски езици. Сред типичните характеристики на езиците от Balkan Sprachbund са:

- Опростяване на именното склонение: наблюдават се само остатъци от винителен и дателен падеж при местоименията и звателен падеж при някои съществителни от мъжки и женски род – одушевени или лица.

- Наличие на задпоставен определителен член при съществителните имена в български (жена – жената), албански, арумънски и румънски. Това явление възниква вероятно под влияние на старобългарския език, защото отсъства в новогръцки, където определителният член е предпоставен.
- Загуба на инфинитива и замяната му с лични глаголни форми в български и новогръцки.

Българската писмена система е много по-лесно разбираема от английската например. За произношението на дадена дума се съди по изписването ѝ. Въпреки това комбинацията от правописни и правоговорни принципи (асимилация при звучните и беззвучните съгласни, редукция на гласните и други) може да доведе до поява на омонимни форми, например: *кос – коз*. Така различни графични форми се произнасят по един и същи начин. Явлението предизвиква различни грешки, които програмите за корекция на правописа и на граматиката са в състояние да разпознаят и коригират.

Главните букви в книжовния писмен български език се използват доста по-рядко в сравнение с други езици. Най-общо с главни букви започват само първата дума от заглавието на книга, филм, песен, както и първата дума в изречението; личните имена и географските названия (*Стара Загора*), имена на компании, правителствени институции и други; някои абревиатури. Названията на националности и езици не се изписват с главна буква, нито пък названията на месеците и на дните от седмицата.

Българският език има характеристики, които могат сериозно да затруднят автоматичното описание на синтактичните структури. Словоредът например е сравнително свободен – прилагателните винаги са предпоставени модификатори на съществителните, а наречията са предпоставени модификатори на прилагателните и други наречия, но подлогът и допълнението на глагола могат да сменят местата си сравнително свободно. В изречение с три конституента (подлог, сказуемо/глагол, допълнение) например са възможни всичките шест словоредни комбинации, а изречение с четири конституента (подлог, сказуемо/глагол, пряко допълнение, непряко допълнение) може да се изрази в 24 словоредни комбинации; и пр. Например в изречението *"Дивата котка гони злото куче цяла сутрин"* не става ясно, извън по-широкия контекст, коя от именните групи е подлогът (*дивата котка* или *злото куче*). За разлика от други езици със сравнително свободен словоред (например останалите славянски езици), българският не използва падежни окончания при имената, за да отбелязва синтактичните отношения между думите и само

използването на определителния член при имената от мъжки род отчасти може да помогне да бъдат разкрити граматичните зависимости.

Друга особеност на българския, която затруднява синтактичния анализ, е свободното изпускане на подлога. Това явление в контекста на разместването на позициите на подлога и допълнението прави задачата още по-сложна. Да разгледаме например следното изречение: *"* Извика Анна"*. Значението може да е, че Анна е извикала някого, т.е. Анна е подлогът, или че Анна е била извикана от някого, т.е. Анна е допълнението. Липсва морфологичен маркер, който да разграничи подлога от допълнението, освен в случаите на графично представяне на определителния член при имената от мъжки род в позицията на подлог. Например в изречението *"Попита учителят"*, учителят е попитал нещо, а в изречението *"Попита учителя"* – учителят е бил запитан за нещо.

Сравнително свободният словоред, съчетан с липсата на морфологично разграничение на падежните форми при имената и вероятността подлогът да е изпуснат, затрудняват създаването на компютърни програми за обработка на българския език.

## Актуално

От 1990 г. насам българският телевизионен и филмов пазар е доминиран от американски филми и сериали. Чуждоезиковите филми и сериали или са дублирани (ако се излъчват по националната телевизия или друга голяма телевизия с национално покритие), или са съпроводени със субтитри (предимно при излъчване по по-малките частни телевизии). Независимо как са преведени обаче, присъствието на чужд начин на живот в медиите определено влияе върху българската култура и език. Поради продължаващата доминация на английската и американската музика от 60-те години на миналия век насам (която се засилва значително през 90-те години на миналия век насам) българите непрекъснато са обкръжени от английския език. Английският бързо се сдобива със статут на "куул/хип" език в определени общности и сфери.

През последните 20 години се наблюдава отчетливо интернационализиране на българската лексика, най-вече под влияние на английския език. Навлизат нови думи, значения, съчетания, които произхождат от английски (т. нар. англицизми), но немалка част от тях са заети и от други европейски и неевропейски езици.

В „Речник на новите думи в българския език" (2010)[vii], който отразява навлезлите нови думи, фразеологизми и терминологични съчетания през последните 20 години, са включени около 4300 нови лексикални единици,

близо една четвърт от които (около 1020) са заемки от английски[viii]. В редица терминологични области специализираната лексика се изгражда основно под влияние на английски – компютърни технологии и интернет (*файл, сайт*), финанси, икономика и бизнес (*дилър, брокер*), съвременна музика (*диджей, техно, клип*), спорт (*джогинг, бодибилдинг*). Навлизането на английски заемки се наблюдава и при общоупотребимата лексика, например *тостер, стикер, бодигард*, както и в младежкия жаргон.

В речника са отразени и над 700 нововъзникнали значения на познати думи, немалка част от които се появяват под влияние на английски, т. е. представляват семантични калки, например компютърните термини *мишка, папка, гласова поща* и пр., нови значения на думите *интелигентен, агресивен* и други. Влиянието на английския език тук остава до голяма степен скрито, тъй като засяга обикновено семантичната страна, но не и формалния състав на думите. Немалко от новите английски заемки създават трудности. Някои са трудно произносими – *блокбъстър, мърчандайзинг*, други имат затруднена морфологична адаптация, например има колебания при формите за мн. ч. – *бодигарди* или *бодигардове*, *чипсети* или *чипсетове, билборди* или *билбордове*. Хората от по-възрастните поколения, които не владеят английски, срещат затруднения с разбирането на тези заемки.

Специфичен проблем при новата българска лексика, непознат в периода преди 90-те години на XX в., е изписването с латинска графика на част от чуждите думи, особено на англицизмите. Става дума както за абревиатури (например *CV, CD-ROM, PR*), така и за широко използвани думи като internet или за предпоставени съставни части като *e-* за electronic (*е-търговия*). Използването на латинска графика се среща особено често в някои специални сфери – например компютърната терминология, при изписване имена на компании или търговски марки, а също и в общоупотребимия език например в езика на рекламата. Някои хора се тревожат, че използването на латински букви при писане на български (т.нар. „шльокавица“ [shl'okavitsa][ix], която се използва предимно в неофициално общуване – текстови съобщения и други) ще повлияе негативно на качеството на говоримия и писмения български език и може да предизвика отказване от характерната кирилица.

Примерите, дадени по-горе, показват значението на усилията за повишаване информираността за процесите в обществото, при които има опасност големи социални групи да бъдат изключени от участие в информационното общество, особено ако не владеят английски.

## Езикови политики в България

Българският език е официалният език в Република България и този му статут е уреден от Конституцията. С решение на Министерския съвет Институтът за български език към Българската академия на науките е официалната институция, която следи за промените в езика и определя книжовната норма, като отразява промените в правописа и правоговора. Основната му дейност е свързана с изследвания в областта на българското езикознание – общата, теоретичната, приложната и компютърната лингвистика. Институтът издава Речник на българския език и поддържа архивни материали за развоя на езика. Занимава се с изследване на българските диалекти на и извън територията на България, както и с публикуване на документи, свързани с езиковата политика в рамките на европейската интеграция. Сред останалите задачи на Института са създаването на езикови корпуси и бази данни и изграждането на лингвистична основа за изработване на различни компютърни приложения. В изпълнение на тези си функции Институтът вече 60 години издава списание "Български език", поддържа предаването "Език мой" по Българската национална телевизия (БНТ) и осигурява справки и консултации чрез службата "Езикови справки". Излъчват се и редовни рубрики към Българското националното радио (БНР).

Паралелно, в много от националните и провинциални ежедневници се поддържат рубрики по езикови въпроси, например рубриката във вестник "Труд", водена от професори от Софийския университет. От 2002 г. насам Софийският университет издава списание "Родна реч", чиято задача е да отговаря на различни актуални въпроси и да служи за повишаване езиковата култура на обществото. Първият академичен правописен речник, който отразява нормативността по отношение на правописа и правоговора, излиза през 1983 г. През 2002 г. излиза нов правописен речник, в който се добавят нови думи, а други се променят в съответствие с развитието на езика, а през 2011 г. завършва работата по най-новия правописен речник, който отразява последните промени в речниковия състав на езика. Освен нормативните речници се издават и други правописни и правоговорни речници, които следват нормативната уредба. Поддържат се и редица поредици, свързани с проблемите на правописа и правоговора.

Повечето от тези дейности отразяват чисто академични интереси и не са достатъчно популярни, особено сред по-младите поколения. Медиите използват по-разчупен език, с цел да привлекат и забавляват аудиторията си, но същевременно този език често се отклонява от нормативните предписания. Няма официални квоти за музиката на български език, която би трябвало да се излъчва по медиите (няма такава уредба дори за БНР и

БНТ), за разлика от регулациите във Франция, Унгария, Словения например. Въпреки това в Закона за радио и телевизия е посочено, че БНР трябва да отделя за създаване и изпълнение на българска музика и радио театър не по-малко от 5% от субсидията си от държавния бюджет, а БНТ – не по-малко от 10% от своята субсидия за продуциране на български телевизионни филми.

Последната правописна реформа е проведена през 1945 г. с цел осъвременяване на българския правопис. Тя премахва някои букви, остатъци от историческия правопис. На няколко пъти депутати от Народното събрание внасят проектопредложения за Закон за българския език с цел да бъдат ограничени чуждите влияния в него. Тези предложения са съпроводени от горещи и емоционални дебати. През 2007 г. е приет Закон за транслитерацията, чиято цел е да внесе унификация в разнообразната практика за предаване на думи (собствени имена) с латински букви.

Българският език се намира в по-незавидна ситуация в сравнение с езици като френския, които имат силна подкрепа от цялата френскоговоряща общност, от т. нар. Франкофонски съюз. Широкото навлизане на езиковите технологии може да помогне за напредък в тази посока, чрез предлагане на инструменти за езикови услуги при работата на медиите, интернет и мобилните комуникации, които включват програми за проверка на правописа и на граматиката, програми за стилистична корекция, изграждане на инструменти за справка в синонимни речници и за правилното произношение на думите.

## Езикът в образованието

След XIX в. българският език и литература са важна част от обучението. Според законодателството в България обучението като част от актуалната държавно одобрена учебна програма, от предучилищно до университетско ниво, трябва да се провежда на български език. Специални клаузи уреждат обучението на децата, чийто майчин език не е българският. Изучаването на българския език е задължително в основното и гимназиалното образователно ниво.

В изследването PISA'2009[x] (за уменията за използване на научно знание, разпознаване на проблеми и привеждане на доказателства) България е на 46-о място от 65 страни. Според PISA'2006 България е 43-та от 55 страни, а според PISA'2000 – 33-та от 41 страни.

Според проучването PIRLS'2006[xi] (грамотност при четене, оценена като способност да се разбират и използват формите на писмения език, които са част от обществения живот) България заема 14-о място с 547 точки. Резултатите отново са по-ниски, отколкото в PIRLS'2001 – 4-о място с 550 точки.

Резултатите за България от проучванията PISA и PIRLS могат да послужат като индикатор (Международен референтен стандарт), който да определи до каква степен националната учебна програма на училищно ниво покрива изискванията на международните образователни стандарти. През 2009 г. българските ученици показват основна грамотност, но почти половината се затрудняват да тълкуват или анализират текст.
Недостатъците на обучението по български език, например в гимназията, могат да се обобщят по следните показатели: отделя се недостатъчно време в програмата за обучение по български – 36 часа (урока) годишно; преподаването и усвояването не следват принципите на комуникативния процес; предлага се неадекватно съдържание.

Според последните Държавни образователни изисквания обучението по български език се провежда в рамките на културно-образователната област "Български език и литература". Този предмет е традиционен за българските училища и университетите обучават специалисти – преподаватели по български език за основното и за средното училище. Един от начините за подобряване ефективността на обучението по български език е свързването му със специфична и значима научна сфера – езикознанието. Макар по традиция да се класифицира като хуманитарна дисциплина, езикознанието се занимава с формулиране на правила, по които елементите на езика се съчетават, и в този смисъл е по-близо до точните дисциплини, за разлика от литературознанието.

На университетско ниво се забелязва липса на курсове по български (в някои университети), които да помогнат на бъдещите експерти за успешната им професионална комуникация и практическата им грамотност.
Езиковите умения са сред най-важните изисквания в процеса на обучение, както и при личната и професионалната комуникация. Увеличаването на броя на часовете по български език в училище е една от възможните мерки за подобряване на езиковите умения на учениците, необходими им за активно участие в обществото. Езиковите технологии също играят роля в този процес, например чрез създаване на т. нар. системи за компютърно подпомогнато езиково обучение (CALL). Те позволяват на учащите се да обогатяват своя езиков опит на игрови принцип; например чрез свързване на лексикалните единици в електронен текст с разбираеми дефиниции или в аудио и видео файлове, които осигуряват допълнителна информация, например за произношението на думата.

## Международният статут на българския език

Използването и влиянието на българския език извън границите на България и неговите носители е ограничено. От дълги години Министерството на образованието, младежта и науката организира конкурси за изпращане на български университетски преподаватели като лектори по български език, литература и култура в множество чуждестранни университети, където се изучава български език. Националният фонд "Култура" към Министерството на културата периодично организира конкурси за превод на българска художествена литература на чужди езици. Разбира се, творбите на български автори - класици на българската литература, като Христо Ботев и Иван Вазов, отдавна са преведени на почти всички европейски езици.

България се гордее със своите световноизвестни певци (Николай Гяуров, Гена Димитрова, Борис Христов, Валя Балканска), автори и артисти (Юлия Кръстева, Цветан Тодоров, Христо Явашев, Доньо Донев) и други. Поетът Пенчо Славейков (1866 – 1912) е централна фигура в българската литература и единственият българин, предложен за Нобелова награда. Извън тесните специализирани кръгове в чужбина, в които е обект на изучаване или изследване, българският език е непознат и екзотичен.

Както навсякъде в научния свят, така и от българските учени се изисква да публикуват в достъпни и престижни (най-вече международни) списания, където повечето материали са на английски език. В света на бизнеса и икономиката ситуацията е подобна. В много големи международни компании английският е lingua franca в писмената и в устната комуникация. В същото време 44% от възрастните българи не говорят чужд език по данни от изследване, публикувано от Европейската статистическа служба Евростат през 2009 г.[xii].
Българският език получава статут на официален административен език за Европейския съюз, редом до английския, немския или френския, тъй като Европейският съюз е изграден върху принципите на солидарност и равноправие между членовете си. 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.

Именно членството на България в Европейския съюз и идеята за единство и многообразие в контекста на глобализацията и запазването на националната идентичност осигурява реална възможност за равноправно използване на българския език наред с големите европейски езици.

Езиковите технологии могат да помогнат за разрешаването и на проблемите, възникващи в тази ситуация, но от различна перспектива, тъй като предлагат услуги като машинен превод или извличане на информация на различни езици от текстове, които са написани на друг език. Това ще улесни преодоляването на социалното и икономическото неравенство и проблемното включване, с което се сблъскват гражданите, чийто роден език не е английски.

## Българският език в Интернет

Потребителите на интернет в България през 2009 г. са се увеличили с 31% спрямо 2007 г. и вече са 46% от цялото население. България е сред страните с най-висок процент проникване на интернет според проучване на gemiusAudience (gemius.com)[xiii], публикувано в доклада Do you CEE? (internetcee.com)[xiv].
По данни на internetworldstats.com[xv] в България има около 3,5 млн. потребители на интернет, а според статистиката, публикувана от Gemius, ръстът на аудиторията на следените от анализаторите сайтове е почти 10,7 % на годишна база. През 2010 г. потребителите са се увеличили с още 5 %.
Като изключим най-разпространените и посещавани международни уеб сайтове, най-популярните сайтове в българското интернет пространство са предлагащите новини на български език (dir.bg, gbg.bg news.bg и други). Българската Уикипедия е важен източник на данни за изграждане на програми за обработка на естествения език със своите близо 117 000 статии, макар и доста по-малко от най-големите портали на Уикипедия за английски, немски и френски. В същото време по брой на статиите българската Уикипедия е на 34-а позиция сред 270-те Уикипедии.[xvi] в света.
Често се посочва, че английският език е доминиращ в света на компютрите и в интернет и желаещите да ги ползват първо трябва да научат английски. Това схващане бе валидно в ранните години от развитието на технологиите, но незнанието на английски вече не представлява толкова голяма пречка, колкото тогава. Явлението интернет, което започва в среда за англоговорящи, бързо се превърна в многоезиково. Създава се софтуер, който може да разчита и визуализира различна писменост. Много корпоративни

уеб сайтове използват многоезикови платформи и предлагат езика, предпочитан от потребителя. Машинният превод на уеб съдържание може да се получи само с няколко натискания на мишката. Увеличаващото се значение на интернет е важно за езиковите технологии в два основни аспекта. От една страна, огромното количество дигитално достъпни езикови данни предоставя големи възможности за анализ на употребите на естествения език, от особено значение е количеството на данните, което прави възможно получаването на достоверна статистическа информация. От друга страна, интернет предлага множество сфери на приложение на езиковите технологии.

Най-широко използваното уеб приложение са уеб търсачките, при които езикът се обработва автоматично на няколко равнища. Този процес ще бъде изяснен по-подробно във втората част на този документ. Използват се съвременни езикови технологии, които са различни за различните езици. За български например това означава да се разпознаят 52 различни форми на основната форма на преходен глагол от несвършен вид в ед. ч. Потребителите на интернет и авторите на уеб съдържание може да се възползват от езиковите технологии по не толкова очевидни начини, например при автоматичен превод на уеб съдържание от един език на друг. Поради високата цена на преводаческите услуги е изненадващо колко малко ефективна езикова технология е създадена спрямо съществуващите нужди от нея.

Всичко това обаче е не чак толкова изненадващо, когато си дадем сметка за сложността на българския език и количеството софтуер, който се използва в типичните приложения за езикови технологии. В следващата част на този документ ще представим накратко езиковите технологии и основните им сфери на приложение, както и ще дадем оценка на актуалната ситуация в областта на езиковите технологии за български език.

## Избрана литература

Съвременен български език Фонетика. Лексикология. Словообразуване. Йордан Пенчев, Иван Куцаров, Тодор Бояджиев, Издателска къща Петър Берон, София, 1999, 654 с.

Речник на новите думи в българския език. Емилия Пернишка, Диана Благоева, Сия Колковска, Наука и изкуство, София, 2010, 515 с.

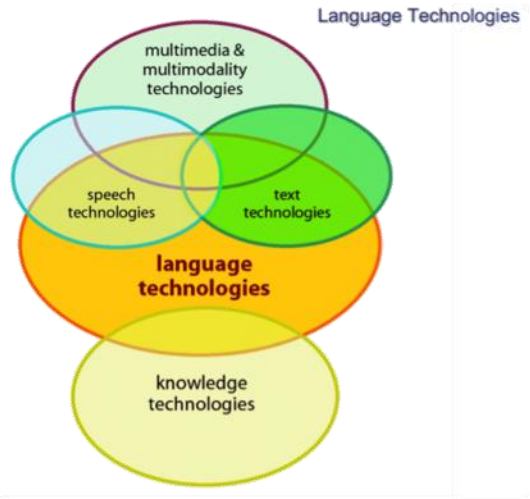
Български диалектен атлас. Обобщаващ том. Ч. I-III. Фонетика. Акцентология. Лексика. (авторски колектив). Книгоиздателска къща Труд. София, 2001, 538 с.

Българите, книжовността, езикът XIX – XX век. (авторски колектив) състав. В. Мурдаров. София, ЕМАС, 2008, 389 с.

# Езикови и технологични приложения за български език

## Езикови технологии

Езиковите технологии са информационни технологии, специално насочени към обработката на естествения език. Поради това те често се класифицират като подвид на технологиите за естествен език. Естественият език се използва от хората в говорима и в писмена форма. Макар речта да е най-старата и естествена форма за езикова комуникация, по-сложната информация и голяма част от познанието се записват и разпространяват писмено. Речевите и текстовите технологии анализират и генерират език в тези две форми. Езикът има аспекти, които са общи и за двете форми, като речниковата информация, голяма част от граматиката и значението на изреченията. Затова редица езикови технологии не могат да бъдат класифицирани само като речеви или само като текстови технологии. Сред тях са технологиите, които свързват езика с когнитивните модели. На фигурата вдясно е представена схема на езиковите технологии. В комуникационния процес ние смесваме езика с други модуси на комуникация и други информационни медии. Смесваме речта с жестовете и израженията на лицето. Дигиталните текстове се съчетават с картини и звуци. Филмите съдържат език в говорима и писмена форма. Така речевите и текстовите технологии се съчетават и взаимодействат с други технологии, които подкрепят обработването на мултимодусни комуникативни и мултимедийни документи.




## Архитектура за езиково-технологични приложения

Типичните софтуерни приложения за езикова обработка се състоят от няколко компонента, които отразяват различни аспекти от езика. Фигурата вдясно представя силно опростена архитектура на система за текстообработка. Първите три модула отговарят за структурата и значението на входния текст:

- Предварителна обработка: изчистване на данните, премахване на форматиране, разпознаване на езика на входния текст и др.
- Граматичен анализ: разпознаване на глагола и допълненията, определенията и други; разпознаване на изреченска структура.
- Семантичен анализ: снемане на многозначност (кое значение на “apple/Apple” е правилно в дадения контекст?); разрешаване на анафора и референтност в изрази като “тя”, “колата” и други; представяне на значението на изречението във вид, който да е читаем от машина.

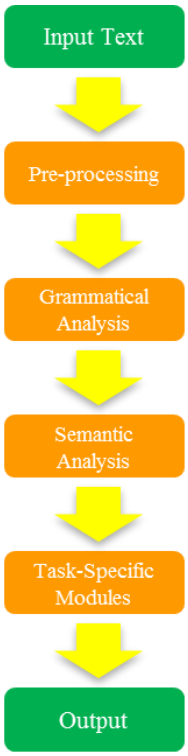



Figure 2: A Typical Text Processing Application Architecture

Модулите, отговарящи за специфичните задачи, изпълняват различни операции, като осъществяване на автоматично резюме на входния текст, сверяване с базата данни и много други. По-долу ще дадем примери за най-важните сфери на приложение и разпознаване на основните модули. Архитектурите на приложенията са опростени и идеализирани, но илюстрират сложността на приложенията за езикови технологии в разбираем вид.

След сферите на приложение ще ви запознаем накратко със ситуацията в изследователската дейност и обучението с помощта на езиковите технологии и ще завършим с преглед на изпълнени и действащи изследователски програми. В края ще представим експертна оценка на наличните основни инструменти и езикови ресурси по критерии като достъпност, фаза на разработка и качество. Таблицата дава представа за ситуацията в сектора за езикови технологии за български език.

Най-важните инструменти и ресурси, които участват в процеса, са подчертани в текста и фигурират в таблицата, приложена в края. Отделните части разглеждат най-важните сфери на приложение заедно с преглед на фирмите, работещи в този сектор в България.

## Ключови сфери на приложение

### Разпознаване на език

Всеки, който е използвал програма за текстообработка като Microsoft Word, се е сблъсквал с приложението за коригиране на правописа (spell checker), което разпознава правописни грешки и дава предложения за поправка. 40 години след първата програма за коригиране на правописа, създадена от Ралф Горин, програмите за корекция на правописни грешки не просто сравняват срещаните в текста думи спрямо списък от правилно изписани думи, но и стават все по-усъвършенствани. Освен че ползват езиково зависими алгоритми за разпознаване на морфология (например множествено число), някои разпознават грешки в синтаксиса, като липсващ глагол или глаголна форма, която не се съгласува с подлога по лице и число, като “Тя *написахме писмото.” Най-разпространените приложения за корекция на правописа (включително вграденото в Microsoft Word) обаче няма да открият грешки в първия стих на следната поема от Джерълд Х. Зар (1992):

*Eye have a spelling chequer,*
*It came with my Pea Sea.*
*It plane lee marks four my revue*
*Miss Steaks I can knot sea.*

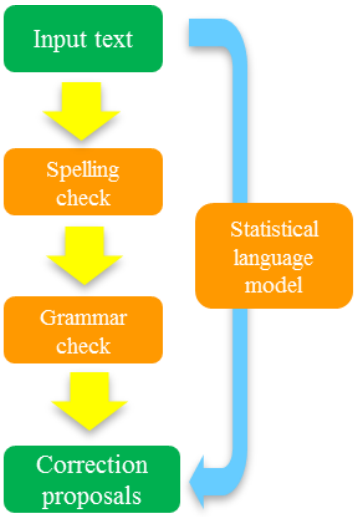


Figure 3: Language Checking (left: rule-based; right: statistical)

За разпознаване на този тип грешки е необходим анализ на контекста. В други случаи трябва да проверим в контекста дали една дума започва с главна буква, или не, като например в:

Тя живее в Стара Загора.
Тя е стара жена.

Тези случаи изискват както формулиране на езиково специфични граматични правила, т. е. висока степен на експертно познание, или използване на т. нар. статистически езикови модели, които изчисляват вероятността определена дума да се появи в специфичен контекст (на предхождащи и следхождащи думи). Статистическият езиков модел може да бъде автоматично конструиран въз основа на информация от голямо количество стандартни езикови данни (т. нар. корпус). Досега тези подходи са били изграждани и оценявани върху данни от английски. Резултатите невинаги са лесно приложими за български, който е с по-свободен словоред, с незадължително изразен подлог (subject pro-drop) и по-богати флективни модели.

Употребата на програми за корекция на езика не е ограничена до инструментите за езикова обработка, те се прилагат и в помощни системи, които се вграждат в редица продукти с техническа документация, която бързо нараства в последните десетилетия. Поради опасенията си от оплаквания от клиенти за неправилна употреба и нанесени щети заради неточни или неправилно разбрани инструкции компаниите отделят все повече ресурси за качеството на техническата документация, особено когато се стремят към международния пазар. Напредъкът в сферата на обработката на естествения език доведе до разработване на помощен софтуер, който да съдейства на пишещите техническа документация с речник и изреченски структури, съобразени с определени правила и (корпоративни) терминологични ограничения.

Освен програмите за корекция на правопис и помощните програми, приложенията за езикова корекция се използват и при компютърно асистираното езиково обучение и за автоматичното коригиране на заявки в уеб търсачките, например предложенията ‘Did you mean…’ на Google.

**Уеб търсене**

Търсенето в мрежата, в интранет пространства или в дигитални библиотеки е вероятно най-широко използваната, но и най-недостатъчно развитата езикова технология днес. В момента търсачката Google, която се появява през 1998 г., се използва за около 80% от всички

търсения на информация в мрежата във всички точки на света.[xvii] Нито търсещият интерфейс, нито формата на намерените резултати са претърпели значителни промени от първата версия досега. В настоящата си версия Google предлага корекция на словоформите, ако разпознае думите като сгрешени, а от 2009 г. вгражда и основни възможности за семантично търсене чрез търсещия си алгоритъм, които могат да подобрят точността на търсенето чрез анализ на значението на термините в контекст.[xviii] Историята на успеха на Google показва, че при много налични данни и ефективни техники за индексиране статистически базираният подход може да постигне задоволителни резултати.

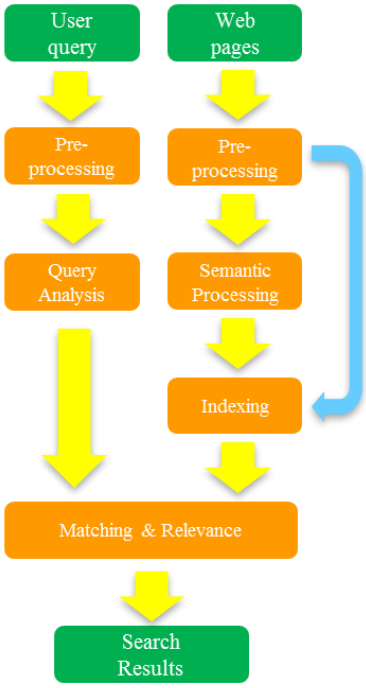


Figure 4: Web Search Architecture

За по-детайлно търсене на информация обаче е нужно да се интегрира и значително лингвистично знание. В изследователските лаборатории експериментите, които използват електронни тезауруси и онтологични езикови ресурси като WordNet (или еквивалента му за български BulNet), са усъвършенствани, за да могат да открият уеб страница чрез сверка със синонимите на търсените думи/изрази, като например атомна енергия и ядрена енергия, или дори по-далечно свързани думи и изрази.

Следващото поколение търсачки трябва да се базира на по-съвременни езикови технологии. Ако търсенето се състои от въпрос или друг вид изречение вместо от списък от ключови думи, в резултатите се включват подходящи отговори на търсенето след анализ на изречението на синтактично и семантично равнище и индекс, който позволява бързото намиране на релевантни документи. Да вземем за пример потребител, който търси: “Искам списъка на всички компании, които са били придобити от други компании в последните пет години.” Изречението трябва да се раздели синтактично, за да се анализира граматичната му структура и да се определи дали потребителят търси компании, които са били придобити, или компании, които са придобили други компании. Изразът “последните пет години” трябва да бъде обработен, за да се уточни към кои точно пет години се отнася.
Обработеното търсене трябва да бъде съпоставено с голямо количество неструктурирани данни, за да се открие онова късче или късчета информация, които потребителят търси. Този процес обичайно се нарича извличане на информация и включва търсене за и класифициране на референтни документи. За генерирането на списък от компании трябва да извлечем информация, според която определена поредица от думи в документа реферира към име на компания. До този тип информация се достига чрез т. нар. идентификатори на именувани обекти.

Изискванията са дори още по-големи при опитите за разпознаване на търсения термин в документи на различни езици. За извличане на информация от такива

документи трябва автоматично да се преведе търсенето на всички възможни изходни езици и да се прехвърли получената информация отново на изходния език. Данните в нетекстови формати стават все повече и се увеличава търсенето на услуги, улесняващи извличането на мултимедийна информация, т.е. информация под формата на образи, аудио и видео данни. При аудио и видео файловете се използва механизъм за разпознаване на реч, конвертиращ съдържанието в текст или фонетична репрезентация, спрямо която да бъдат съпоставени търсенията.

Jabse.com е българска търсачка (акроним на Just Another Bulgarian Search Engine), която индексира сайтове от българското интернет пространство, изградени с отворен код. Jabse разполага със своя собствена система за оценка, която определя важността на страниците и термините според набор от критерии. Определени български портали разполагат с crawler софтуер (т. нар. "паяк"), подобен на използвания от глобалните търсачки, който индексира сайтовете, класифицирани в категории. Например Dir.bg е един от първите и най-големи уеб портали в България, който стартира автономна услуга – Diri.bg. Тази услуга е пряк конкурент на Jabse и все още не е ясно коя от двете услуги ще се развие достатъчно, за да надделее в българското интернет пространство до такава степен, че да успее да конкурира Google. Технологиите с отворен код като Lucene и SOL често се използват от компании, разработващи софтуер за търсене, като базисна инфраструктура за търсачките. Други компании, работещи в областта, разчитат на международни технологии за търсене като FAST или Exalead.

За да постигнат ръст, тези компании предлагат допълнителни приложения и софтуер за търсене от последно поколение за портали, към които съществува специален интерес. Те основават разработваните технологии на семантичен анализ на езиковите явления, срещащи се в съответната област. Заради високите изисквания към процесорите подобни търсещи машини са икономически оправдани само при обработване на сравнително малки корпуси. Времето за обработка спокойно може да надмине нужното на една традиционна статистическа търсеща програма като тази на Google. Тези търсещи машини се предпочитат все по-често за моделиране на области със специфична тематика, но не са особено рентабилни в уеб мащаб.

**Гласова комуникация**

Технологията за гласова комуникация създава интерфейси, които позволяват на потребителя да взаимодейства с машините чрез реч, а не чрез графични картини, клавиатура и мишка. Днес подобен гласов потребителски интерфейс (VUI) се използва в приложения, осигуряващи частични или пълни

автоматични гласови услуги, които компаниите предлагат на своите клиенти, служители или партньори по телефона. VUI е най-разпространен в банкирането, логистиката, обществения транспорт и телекомуникациите. Технологиите за гласова комуникация се вграждат и в интерфейси на определени устройства, например вградената в колата система за навигация и реч като алтернатива на входните/изходните данни на графичния потребителски интерфейс в т.нар. смартфони.

Технологиите за гласова комуникация включват четири различни модула:

- Автоматичното разпознаване на реч (ASR) идентифицира думите, които действително са изречени зад поредицата от звукове, произнесени от потребителя.
- Синтактичната и семантичната интерпретация анализират синтактичната структура на изказването на потребителя и го интерпретират според целта на съответната система.
- Нужен е и диалогов мениджмънт за наблюдение на системите, с които си взаимодейства потребителят, както и за преценка какво действие да бъде предприето въз основа на входните данни и функциите на системата.

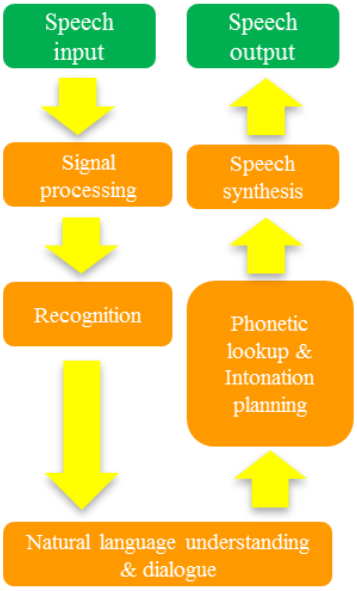



Figure 5: Simple Speech-based Dialogue Architecture

Технологията за синтез на реч (превръщане на текст в реч или TTS) трансформира елементите от изказването в звукове, които ще бъдат насочени като изходни данни към потребителя.

Едно от основните предизвикателства е разработването на ASR система, която да разпознава произнесените думи колкото е възможно по-точно. Това изисква или ограничение на обхвата на възможните потребителски произношения до определен набор от ключови думи, или ръчно създаване на езикови модели, които покриват голям обхват от произношения на носителите на езика. Това предполага сравнително трудна и негъвкава употреба на VUI и затруднява възприемането от потребителя, усложнява създаването и поддържането на езиковите модели, а и значително увеличава разходите. VUI обаче се базират на езикови модели и позволяват на потребителя да се изразява достатъчно свободно, например “Как мога да Ви помогна?”. Тези технологии са по-автоматизирани и по-приемливи за употреба в сравнение със заучените фрази.

Често компаниите използват за изходни данни на VUI предварително записани изказвания на хора със съответната професионална и (в идеалния случай)

корпоративна подготовка. Нужен е богат потребителски опит за заучените изказвания, чиито произношение и структура не зависят от определен контекст на употреба или от особеностите на дадения потребител. В колкото по-динамично съдържание се внедри произнесеното, толкова по-трудно ще е възприемането от потребителя заради лошото качество на прозодиката след наслагване на единични аудио файлове. Днешните TTS системи са доста по-добри (макар че и от тях има какво да се желае) по отношение на прозодичната естественост на динамичните изказвания.

През последното десетилетие пазарът за технологии за гласова комуникация се радва на висока стандартизация на интерфейсите на различните технологични компоненти и на поява на стандарти за софтуер за определени приложения. Освен това станахме свидетели на значима пазарна консолидация, особено при ASR и TTS. Националните пазари на икономически силните и гъсто населени страни от Г-20 са доминирани от по-малко от пет компании в този сегмент, а Nuance и Loquendo са водещи в Европа.

На българския TTS пазар се предлагат няколко български системи за синтезиране на текст в реч. SpeechLab 2.0 се разпространява свободно за потребители със зрителни затруднения, а няколко компании, като например издателят на юридическа литература Ciela, са разработили свои системи за разпознаване на българска реч. При технологиите за диалогово управление и гласова комуникация все още няма развит пазар за лингвистични технологии за синтактичен и семантичен анализ.

Търсенето на VUI в България се увеличи значително в последните пет години. За това изигра роля повишеното търсене от крайните потребители на системи за самообслужване, както и значителната оптимизация на разходите при автоматизираните телефонни услуги и засилената роля на устната реч при контакта човек-машина.

В бъдеще вероятно ще настъпят значителни промени в резултат на разпространението на смартфоните като нова платформа за управление на контакти между потребителите (освен телефоните, интернет и имейла). Тази тенденция ще се отрази и при технологиите за гласова комуникация. От една страна, търсенето на телефонно базирани VUI ще намалее в дългосрочен план. От друга страна, употребата на говорим език под формата на входни данни за смартфоните ще придобива все по-голямо значение. Тази тенденция се подкрепя и от забележимото подобрение на точността при разпознаването на реч, независимо от особеностите на говорещия, особено в услугите за гласови указания, които вече се предлагат от централите за обслужване на

смартфон потребители. Предвид отделянето на задачата по разпознаването на реч и прехвърлянето ѝ към инфраструктурата на някои ежедневни приложения, в бъдеще лингвистичните технологии все по-често ще се приемат като част от по-специфични приложения.

**Машинен превод**

Идеята за използване на дигитални компютри за превод на естествен език се появява още през 1946 г. в резултат на изследванията на А. Д. Бут. През 50-те години на миналия век задачата получава значително финансиране, а интересът се възражда през 80-те години. Машинният превод (МТ) обаче все още не успява да оправдае надеждите, съпътстващи го от раждането му.

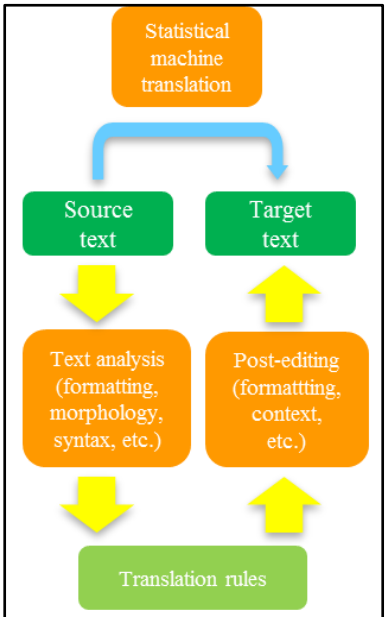


Figure 6: Machine translation (top: statistical; bottom: rule-based)

В първата си фаза технологиите за МТ просто заменят думи от един естествен език с думи от друг. Това може да е приемливо в определени жанрове с клиширан език, изпълнен с формули, например прогнозите за времето. За добър превод на по-слабо стандартизирани текстове обаче трябва да се приспособят по-големи текстови сегменти (фрази, изречения и дори цели параграфи) към най-близките им паралелни отрязъци текст на другия език. Основната трудност се крие във факта, че човешкият език се характеризира с многозначност на всички равнища, например отстраняване на многозначност на лексикално ниво (Ягуар – животно или модел кола) или позиция на предлозите на синтактично ниво, като в:

*Полицаят наблюдава престъпника с телескопа.*
*Полицаят наблюдава престъпника с пистолета.*

Един от подходите е основан на лингвистични правила. При превод между близки езици прекият превод е очакван, като в примера по-горе. Често обаче системи, базирани на правила (или основани на знания), анализират входния текст и създават преходна, символна репрезентация, от която се генерира текстът в производния език. Успехът на този метод зависи от достъпа до богати речници с морфологична, синтактична и семантична информация, както и от наличието на богат набор от граматични правила, изградени от професионален лингвист.

От края на 80-те години на миналия век компютрите стават все по-достъпни и по-мощни, а и по-евтини, затова интересът към статистическите модели за машинен превод се засилва. Параметрите им се извличат след анализ на двуезични текстови корпуси като паралелния корпус Europarl със стенограмите на Европейския парламент на 11 европейски езика. При наличието на много данни статистическият машинен превод е достатъчно добър, за да извлече приблизителното значение на езиковите сегменти в

чуждоезичния текст. За разлика от базираните на предварително знание системи обаче статистическият (или основан на данни) машинен превод често генерира неграматичен изходен текст. От друга страна, въпреки предимствата на спестения човешки труд статистическият машинен превод може да покрие особеностите на езика, които липсват при основаните на знания системи, например при идиомите.

Силните и слабите места на двата метода се компенсират взаимно, така че изследователите се насочват към хибридни подходи, които комбинират решения и от двата модела по няколко начина. Първо, могат да се използват едновременно системи, основани на данни, и такива, основани на правила, и да се изработи модул за избор, който да преценява кой е най-добрият изходен резултат за всяко изречение. При по-дългите изречения резултатът никога не е идеален. По-добро решение е да се комбинират най-добре преведените части от всяко изречение от множеството различни изходни резултати. Задачата е доста сложна, тъй като съответстващите си части в множеството невинаги са очевидни и трябва да бъдат съпоставени и сравнени.

Машинният превод за българския език е още по-голямо предизвикателство. Проблемите са свързани с феномени като липсата на падежно окончание при имената, сравнително свободния словоред и незадължително изразения подлог. Богатата морфология при глаголите е допълнително затруднение.

Един от успешните примери е WebTrance на SkayCode – система за машинен превод, която автоматично превежда текстове, помощни файлове, менюта, прозорци и интернет страници от английски, немски, френски, испански, италиански и турски на и от български. Превод, който следва значението и не е дума по дума, е предизвикателство дори за хората, които все още учат чужд език, така че целта на WebTrance е да даде смислен превод на текстовете. Ако се внедри и програма, която да се съобрази със специфичната терминология, използвана от потребителя, и да осигури интеграция в работната схема, използването на машинния превод може значително да увеличи продуктивността.
Има още доста какво да се желае за качеството на системите за машинен превод. Трудностите са свързани с адаптацията на езиковите ресурси към определена тематична или потребителска област и интеграцията в съществуващите работни схеми със запомняне на терминологични и преводни варианти. Повечето познати системи са изградени около и/или са насочени към английския език и поддържат превод от и на малко езици спрямо български. Липсва унификация, така че се налага потребителите на машинен превод да усвоят различни методи за кодиране на лексиконите за различните системи.

Инициативите за оценка на тези приложения позволяват сравнение на качеството на работа на системите за машинен превод, ефективността на различните подходи и статуса на системите за машинен превод за различните езици. В таблица 1, разработена по проекта EC Euromatrix+, са представени резултатите при машинен превод между два езика за 22-та официални езика на ЕС (без ирландски), изчислени по точковата система BLEU[xix].

Най-добри резултати (зелено и синьо) дават езици, които са били обект на значителни изследователски усилия, били са част от координирани програми и са представени в паралелни корпуси (английски, френски, холандски, испански и немски). С най-слаби резултати (в червено) са езици, които не са били обект на проекти и проучвания или такива, които се различават значително от останалите езици (унгарски, малтийски, фински).

Target Language

| | en | bg | de | cs | da | el | es | et | fi | fr | hu | it | lt | lv | mt | nl | pl | pt | ro | sk | sl | sv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| en | – | 40.5 | 46.8 | 52.6 | 50.0 | 41.0 | 55.2 | 34.8 | 38.6 | 50.1 | 37.2 | 50.4 | 39.6 | 43.4 | 39.8 | 52.3 | 49.2 | 55.0 | 49.0 | 44.7 | 50.7 | 52.0 |
| bg | 61.3 | – | 38.7 | 39.4 | 39.6 | 34.5 | 46.9 | 25.5 | 26.7 | 42.4 | 22.0 | 43.5 | 29.3 | 29.1 | 25.9 | 44.9 | 35.1 | 45.9 | 36.8 | 34.1 | 34.1 | 39.9 |
| de | 53.6 | 26.3 | – | 35.4 | 43.1 | 32.8 | 47.1 | 26.7 | 29.5 | 39.4 | 27.6 | 42.7 | 27.6 | 30.3 | 19.8 | 50.2 | 30.2 | 44.1 | 30.7 | 29.4 | 31.4 | 41.2 |
| cs | 58.4 | 32.0 | 42.6 | – | 43.6 | 34.6 | 48.9 | 30.7 | 30.5 | 41.6 | 27.4 | 44.3 | 34.5 | 35.8 | 26.3 | 46.5 | 39.2 | 45.7 | 36.5 | 43.6 | 41.3 | 42.9 |
| da | 57.6 | 28.7 | 44.1 | 35.7 | – | 34.3 | 47.5 | 27.8 | 31.6 | 41.3 | 24.2 | 43.8 | 29.7 | 32.9 | 21.1 | 48.5 | 34.3 | 45.4 | 33.9 | 33.0 | 36.2 | 47.2 |
| el | 59.5 | 32.4 | 43.1 | 37.7 | 44.5 | – | 54.0 | 26.5 | 29.0 | 48.3 | 23.7 | 49.6 | 29.0 | 32.6 | 23.8 | 48.9 | 34.2 | 52.5 | 37.2 | 33.1 | 36.3 | 43.3 |
| es | 60.0 | 31.1 | 42.7 | 37.5 | 44.4 | 39.4 | – | 25.4 | 28.5 | 51.3 | 24.0 | 51.7 | 26.8 | 30.5 | 24.6 | 48.8 | 33.9 | 57.3 | 38.1 | 31.7 | 33.9 | 43.7 |
| et | 52.0 | 24.6 | 37.3 | 35.2 | 37.8 | 28.2 | 40.4 | – | 37.7 | 33.4 | 30.9 | 37.0 | 35.0 | 36.9 | 20.5 | 41.3 | 32.0 | 37.8 | 28.0 | 30.6 | 32.9 | 37.3 |
| fi | 49.3 | 23.2 | 36.0 | 32.0 | 37.9 | 27.2 | 39.7 | 34.9 | – | 29.5 | 27.2 | 36.6 | 30.5 | 32.5 | 19.4 | 40.6 | 28.8 | 37.5 | 26.5 | 27.3 | 28.2 | 37.6 |
| fr | 64.0 | 34.5 | 45.1 | 39.5 | 47.4 | 42.8 | 60.9 | 26.7 | 30.0 | – | 25.5 | 56.1 | 28.3 | 31.9 | 25.3 | 51.6 | 35.7 | 61.0 | 43.8 | 33.1 | 35.6 | 45.8 |
| hu | 48.0 | 24.7 | 34.3 | 30.0 | 33.0 | 25.5 | 34.1 | 29.6 | 29.4 | 30.7 | – | 33.5 | 29.6 | 31.9 | 18.1 | 36.1 | 29.8 | 34.2 | 25.7 | 25.6 | 28.2 | 30.5 |
| it | 61.0 | 32.1 | 44.3 | 38.9 | 45.8 | 40.6 | 26.9 | 25.0 | 29.7 | 52.7 | 24.2 | – | 29.4 | 32.6 | 24.6 | 50.5 | 35.2 | 56.5 | 39.3 | 32.5 | 34.7 | 44.3 |
| lt | 51.8 | 27.6 | 33.9 | 37.0 | 36.8 | 26.5 | 21.1 | 34.2 | 32.0 | 34.4 | 28.5 | 36.8 | – | 40.1 | 22.2 | 38.1 | 31.6 | 31.6 | 29.3 | 31.8 | 35.3 | 35.3 |
| lv | 54.0 | 29.1 | 35.0 | 37.8 | 38.5 | 29.7 | 8.0 | 34.2 | 32.4 | 35.6 | 29.3 | 38.9 | 38.4 | – | 23.3 | 41.5 | 34.4 | 39.6 | 31.0 | 33.3 | 37.1 | 38.0 |
| mt | 72.1 | 32.2 | 37.2 | 37.9 | 38.9 | 33.7 | 48.7 | 26.9 | 25.8 | 42.4 | 22.4 | 43.7 | 30.2 | 33.2 | – | 44.0 | 37.1 | 45.9 | 38.9 | 35.8 | 40.0 | 41.6 |
| nl | 56.9 | 29.3 | 46.9 | 37.0 | 45.4 | 35.3 | 49.7 | 27.5 | 29.8 | 43.4 | 25.3 | 44.5 | 28.6 | 31.7 | 22.0 | – | 32.0 | 47.7 | 33.0 | 30.1 | 34.6 | 43.6 |
| pl | 60.8 | 31.5 | 40.2 | 44.2 | 42.1 | 34.2 | 46.2 | 29.2 | 29.0 | 40.0 | 24.5 | 43.2 | 33.2 | 35.6 | 27.9 | 44.8 | – | 44.1 | 38.2 | 38.2 | 39.8 | 42.1 |
| pt | 60.7 | 31.4 | 42.9 | 38.4 | 42.8 | 40.2 | 60.7 | 26.4 | 29.2 | 53.2 | 23.8 | 52.8 | 28.0 | 31.5 | 24.8 | 49.3 | 34.5 | – | 39.4 | 32.1 | 34.4 | 43.9 |
| ro | 60.8 | 33.1 | 38.5 | 37.8 | 40.3 | 35.6 | 50.4 | 24.6 | 26.2 | 46.5 | 25.0 | 44.8 | 28.4 | 29.9 | 28.7 | 43.0 | 35.8 | 48.5 | – | 31.5 | 35.1 | 39.4 |
| sk | 60.8 | 32.6 | 39.4 | 48.1 | 41.0 | 33.3 | 46.2 | 29.8 | 28.4 | 39.4 | 27.4 | 41.8 | 33.8 | 36.7 | 28.5 | 44.4 | 39.0 | 43.3 | 35.3 | – | 42.6 | 41.8 |
| sl | 61.0 | 33.1 | 37.9 | 43.5 | 42.6 | 34.0 | 47.0 | 31.1 | 28.8 | 38.2 | 25.7 | 42.3 | 34.6 | 37.3 | 30.0 | 45.9 | 38.2 | 44.1 | 35.8 | 38.9 | – | 42.7 |
| sv | 58.5 | 26.9 | 41.0 | 35.6 | 46.6 | 33.3 | 46.6 | 27.4 | 30.9 | 38.9 | 22.7 | 42.0 | 28.2 | 31.0 | 23.7 | 45.6 | 32.2 | 44.2 | 32.7 | 31.3 | 33.5 | – |

Таблица 1: Резултати от работата на приложенията за машинен превод между 22-та езика на ЕС (по езикови двойки; източник: Euromatrix+)

## Зад кулисите на езиковите технологии

Изграждането на приложения за езикови технологии включва редица подфази, които не винаги са видими при взаимодействието с потребителя, но предлагат значителни предимства. Те са обект на изследване от отделни поддисциплини на компютърната лингвистика.

Отговорът на въпроси (QA) е област, на която се отделя все повече внимание и за която се изграждат анотирани корпуси и се организират изследователски състезания. Идеята е да се премине от търсене по ключови думи (на което се реагира с цяла колекция от потенциално подходящи документи) към сценарий, при който потребителят задава конкретен въпрос и системата дава един-единствен отговор, например: “На каква възраст е бил Нийл Армстронг, когато е стъпил на Луната?” - “38.” Тази задача очевидно е свързана с методите за уеб търсене, но автоматичното отговаряне на въпроси е по-общ термин за цяла област, която се занимава с множество проблеми, например: какви са видовете въпроси и как трябва да се подхожда при всеки от тях; как се изгражда набор от документи, потенциално

съдържащи отговора, който трябва да бъде анализиран и сравнен (дали отговорите не си противоречат?); как специфичната информация (отговорът) може да бъде достоверно експерпирана от документа, без да се пренебрегва контекстът.

Това, на свой ред, е свързано със задачата за извличане на информация (IR), която бе доста популярна в периода на "статистическата революция" в компютърната лингвистика в началото на 90-те години. Извличането на информация има за цел да разпознае специфични късове информация в специфични класове документи, например разпознаване на ключови играчи при придобиване на компании според материали в медиите. Друг сценарий е свързан с новините за терористични инциденти, в които задачата е да се съпостави текстът с модела, в който участват извършителят, целта, времето и мястото на инцидента, както и резултатът. Попълването на модела, специфично за всяка тематична област, е най-важният етап при извличането на информация – още един пример за добре разработена изследователска област за практически цели, която впоследствие трябва да бъде внедрена в подходящи приложения.
Две гранични области, които понякога се определят като автономни приложения, а понякога и като помощна, вградена услуга, са резюмирането на текст и генерирането на текст. При резюмирането се създава кратък текст от по-дълъг текст. Приложение за резюмиране е вградено и в MS Word. Тази задача се основава предимно на статистически данни, като първоначално идентифицира "важните" думи в даден текст (тоест, думите, които са най-често срещани, но все пак по-рядко, отколкото в текстове с обща употреба) и след това определя онези изречения, които съдържат голямо количество важни думи. Тези изречения се маркират в документа или се извличат от него и влизат в резюмето. Засега тази стратегия е най-популярната и в нея резюмирането е равносилно на извличане на изречение: текстът се свежда до набор от изречения. Всички търговски продукти, които използват резюмиране, използват този принцип. Алтернативният подход включва действително синтезиране на нови изречения, които не се появяват във формата си от изходния текст. Този подход изисква добро разбиране на текста и е доста по-бавен и по-несигурен. Генераторът на текст в повечето случаи не е автономно приложение, а е интегрирано в по-голяма софтуерна среда, например в медицинските информационни системи, където се събират, съхраняват и обработват данни за пациентите, а генерирането на информационна справка е само една от многото изходни данни.

Във всички тези технологии българският език е много по-слабо представен от английския, при който отговорът на въпроси, извличането на информация и

резюмирането на текст са били обект на множество конкурентни проекти и най-вече на инициативи, организирани от DAPRA/INST в САЩ от 90-те години насам. Те подобряват значително развитието в сектора, но фокусът винаги е бил върху английски. Някои проекти са включвали повече езици, но българският никога не е бил сред основните. Затова и няма анотирани корпуси и други ресурси за тези цели. Системите за резюмиране, основани на чисто статистически методи, често са до известна степен езиково независими, а някои изследователски прототипи се предлагат и на пазара. За генериране на текст повторно използваемите компоненти обичайно са ограничени до модули за повърхнинно генериране (т. нар. генериращи граматики). И тук по-голяма част от наличния софтуер е предназначен за английски език.

## Езиковите технологии в образованието

Езиковите технологии са интердисциплинарна област, която разчита на висококвалифицирания труд на лингвисти, програмисти, математици, философи, психолингвисти и невролози. Необходимостта от експертно знание и ресурси е и една от пречките пред утвърждаването им в задължителните образователни програми в България. Българските университети предлагат ограничени възможности за специализирано обучение в областта на компютърната лингвистика. Предлагат се курсове и програми, но те са насочени или към студенти в областта на хуманитарните науки, или към математици, но не покриват нуждите и на двете групи специалисти. Преди две години Пловдивският университет започна да предлага бакалавърска програма по лингвистика и информационни технологии. През учебната 2009/10 г. за тази програма се записаха 30 студенти, а през следващата броят на желаещите се удвои. На студентите се предлага широк набор от курсове по основи на лингвистиката, математика и програмиране, както и възможности за запознаване и работа с приложенията за езикови технологии. Бакалавърската програма по информатика в Пловдивския университет по традиция предлага лекционен курс по компютърна лингвистика.

От 2004 г. Факултетът по математика и информатика към Софийския университет предлага магистърска програма по изкуствен интелект. Софийският университет предлага и магистърска програма по компютърна лингвистика, в която са включени курсове в областта на математика, логиката, програмирането и теоретичната лингвистика. Предлагат се и въвеждащи курсове в сферата на компютърната лингвистика, като например статистически методи за обработка на езика, машинен превод, семантични мрежи и онтологии и други. Дипломантите на тези програми получават добро обучение за започване на научна и академична кариера,

както и по-обща теоретична и компютърна грамотност, която им позволява да работят по изграждането на неконвенционални практически приложения. За успеха на програмите може да се съди по професионалното развитие на завършилите ги, които работят във водещи технологични и компютърни компании, национални и специализирани медии и научни и академични институции.

## Програми за езикови технологии

Първите проекти за създаване на необходимите езикови технологии с международно и национално финансиране се появяват в самото начало на 90-те години на миналия век. За кратък период от време редица научни проекти получават финансиране от европейските институции. Сред тях са: LaTeSLav (1991 – 1994), който има за цел създаване на прототип на програма за граматична корекция; BILEDITA (1996 – 1998) – за създаване на двуезични електронни речници; GLOSSER (1996 – 1998) – в подкрепа на обучението по чужди езици и други образователни цели. Multext-East (1995 – 1997) е продължение на предходните проекти Multext и EAGLES, финансирани от ЕС, и предлага ресурси за български език в стандартизиран формат със стандартно маркиране и анотация. По-късно ресурсите са разширени и допълнени по проектите ELAN (European Language Activity Network), 1998 – 1999), TELRI, фази I и II (Trans European Language Resources Infrastructure 1995 – 1998 / 1999 – 2001); и Concede (Consortium for Central European Dictionary Encoding 1998 – 2000).

Националният фонд „Научни изследвания" към Министерството на образованието, младежта и науката подкрепя научните изследвания чрез редица програми за финансиране на национално равнище. Това финансиране подкрепя редици изследователски проекти и сътрудничества с международни изследователски центрове и компании. В резултат се изгражда и основата за технологично развитие и комерсиални приложения за автоматично обработване на български.

Преди няколко години пет български академични институции основаха консорциум за създаване и развитие на интегрирана академична инфраструктура за езикови ресурси. Българските институции участват и в CLARIN. Други текущи проекти са част от EUROPEANA с цел развитие на основни технологии и стандарти, с които съдържанието в интернет да стане по-достъпно в бъдеще. Гореспоменатите проекти, както и редица по-малки проекти с по-ограничено финансиране, изграждат компетентност в областта на езиковите технологии, както и основна технологична инфраструктура за езикови ресурси за български. Публичното финансиране за проектите за езикови технологии за български е доста по-малко от онова за сравними проекти в Европа, както и спрямо инвестициите в страни като САЩ[xx].

## Достъп до инструменти и ресурси за българския език

Таблицата по-долу представя настоящата ситуация в сферата на помощните езикови технологии за български. Рейтингът на съществуващите програми и ресурси е базиран на изчисленията/прогнозите на няколко водещи експерти по следните критерии (в скалата от 0 до 6).

1. **Количество:** Налична ли е програмата/ресурсът за съответния език? Колкото повече програми/ресурси съществуват, толкова по-висока е оценката.
   - 0: няма програми/ресурси
   - 6: много програми/ресурси, твърде голяма разлика
2. **Достъпност:** Достъпни ли са програмите/ресурсите, т.е. дали са с отворен код, дали могат да се използват свободно на всяка платформа, или са достъпни при висока цена и/или при силно ограничени условия?
   - 0: практически всички програми/ресурси са достъпни само при висока цена
   - 6: голямо количество програми/ресурси са със свободен достъп под лицензите за Отворен код или Криейтив Комънс, които позволяват повторна употреба и смяна на целите
3. **Качество:** До каква степен са изпълнени съответните критерии за резултатност на програми и индикаторите за качество на ресурсите при изграждане на наличните програми, приложения и ресурси? Дали тези програми/ресурси са активни и активно поддържани?
   - 0: протопит на програмите/ресурсите
   - 6: високо качество на софтуера, с достъпна качествена анотация като ресурс
4. **Разпространение:** До каква степен най-добрите програми отговарят на съответните критерии (стилове, жанрове, видове текст, лингвистични феномени, видове входен и изходен текст, брой езици в системата за машинен превод и други)? До каква степен ресурсите са представителни за съответния език?
   - 0: ресурс или софтуер със специфична цел, отговаря на специфичен случай, слабо покритие, използва се само за специфични случаи, които не отговарят на общите контексти
   - 6: ресурс с широко покритие, стабилен софтуер, приложим в много области, поддържа много езици

5 **Развитие:** Дали софтуерът/ресурсът се смята за развит, стабилен и готов за разпространение на пазара? Дали най-добрите достъпни софтуер/ресурси могат да бъдат използвани извън конкретното приложение, или трябва да бъдат адаптирани? Дали резултатите са адекватни и готови за приложение в производството, или са само прототип? За доказателство: дали ресурсите/софтуерът са признати от обществото и се внедряват успешно в системи за езикови технологии.

- 0: предварителен прототип, доказателство за системата, примерна част от ресурсите
- 6: компонент, който може да бъде своевременно интегриран или въведен в употреба

6 **Устойчивост:** Колко добре могат софтуерът/ресурсите да бъдат поддържани/интегрирани в сегашните IT системи? Дали софтуерът/ресурсите отговорят на изискването за определено ниво на устойчивост при създаване на документация/ръководства, обяснения на примери на употреба, GUI и други? Дали използват/прилагат стандартни програмни среди (като Java EE)? Дали съществуват стандарти/квази стандарти за индустрията/изследователската област и ако е така, дали софтуерът/ресурсите могат да бъдат адаптирани (формат на данните и други)?

- 0: изцяло патентовани, ad hoc формати за данни и API
- 6: пълно покриване на стандарта, напълно документирано

7 **Приложимост:** Колко успешно най-добрите приложения или ресурси могат да бъдат адаптирани/разширени до нови задачи/области/жанрове/типове текст/примери на употреба и други?

- 0: практически невъзможно е да се адаптира софтуерът/ресурсът към друга задача, невъзможно е дори при голямо количество ресурси или човекомесеца
- 6: много висока степен на приложимост; може да се адаптира много лесно и ефективно

## Таблица на инструментите и ресурсите

| | Количество | Достъпност | Качество | Покритие | Развитие | Устойчивост | Приложимост |
|---|---|---|---|---|---|---|---|
| Езикови технологии (инструменти, технологии, приложения) | | | | | | | |
| Токанизация, Морфология (токанизация, тагиране по части на речта, морфологичен анализ/генериране) | 4 | 3 | 5 | 5 | 4 | 3 | 4 |
| Парсери (плитък и дълбок синтактичен анализ) | 2 | 2 | 4 | 4 | 3 | 3 | 3 |
| Изреченска семантика (снемане на лексикална многозначност, аргументна структура, семантични роли) | 2 | 2 | 3 | 2 | 2 | 3 | 3 |
| Семантика на текста (кореферентност, контекст, прагматика, инференция) | 1 | 1 | 2 | 2 | 1 | 1 | 2 |
| Напреднала обработка на дискурса (текстова структура, кохерентност, реторична структура/RST, аргументативно зониране, аргументация, текстови модели, типове текстове и други) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Извличане на информация (индексиране на текст, мултимедиен IR, крослингвистичен IR) | 2 | 1 | 2 | 2 | 2 | 2 | 3 |
| Извличане на информация (разпознаване на наименовании обекти, извличане на събитие/релация, разпознаване на мнение/нагласа, анализ на текст) | 2 | 1 | 3 | 3 | 2 | 2 | 3 |
| Генериране на език (генериране на изречения, резюме, доклади, текст) | 1 | 1 | 2 | 2 | 2 | 1 | 1 |
| Резюмиране, отговор на въпроси, развити технологии за достъп до информация | 2 | 2 | 2 | 2 | 2 | 1 | 2 |
| Машинен превод | 3 | 2 | 2 | 2 | 2 | 2 | 3 |
| Разпознаване на реч | 2 | 1 | 3 | 3 | 2 | 2 | 1 |
| Синтез на реч | 2 | 1 | 3 | 3 | 2 | 2 | 1 |
| Диалогов мениджмънт (диалогови способности и моделиране на потребителите) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| Езикови ресурси (ресурси, данни, познавателни основи) | | | | | | | |
|---|---|---|---|---|---|---|---|
| Представителни корпуси | 5 | 5 | 5 | 4 | 5 | 4 | 5 |
| Синтактични корпуси (фразова структура или граматики зависимости) | 2 | 1 | 3 | 2 | 3 | 2 | 2 |
| Семантични корпуси | 2 | 4 | 5 | 4 | 3 | 3 | 3 |
| Дискурсивни корпуси | 1 | 2 | 2 | 2 | 1 | 1 | 1 |
| Паралелни корпуси, Преводна памет | 3 | 1 | 4 | 2 | 2 | 2 | 3 |
| Корпуси с реч (необработени речеви данни, маркирани/анотирани речеви данни, данни с диалог) | 1 | 1 | 3 | 2 | 3 | 3 | 3 |
| Мултимедийни и мултимодални данни (текст плюс аудио и видео) | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Езикови модели | 2 | 1 | 2 | 2 | 2 | 1 | 1 |
| Лексикони, терминология | 4 | 3 | 4 | 3 | 4 | 4 | 3 |
| Граматики | 2 | 2 | 3 | 3 | 3 | 3 | 2 |
| Тезауруси, Лексикално-семантични мрежи (wordnet) | 2 | 4 | 5 | 4 | 4 | 4 | 5 |
| Онтологични ресурси за енциклопедични знания (upper модели, свързани данни) | 1 | 2 | 3 | 3 | 3 | 1 | 1 |

## Изводи

Настоящият документ за пръв път прави опит да опише ситуацията при внедряването на езиковите технологии за много европейски езици с цел сравнение и идентифициране на липсите и нуждите.

За български се налагат следните изводи за съществуващите технологии и ресурси:

- Има сравнително добър достъп до токънизатори, тагери за разпознаване части на речта и морфологични анализатори за български. Дори ако самите програми не са изцяло достъпни, ресурсите са със сравнително добро качество и покритието им е задоволително.
- Достъпът до ресурси, като представителни корпуси, речници/лексикони и лексикално-семантични мрежи (wordnet), е задоволителен, тъй като в последните години бяха разработени доста значими ресурси. Въпреки наличието на представителни като качество и количество корпуси, като например Българския

национален корпус, все още няма достатъчно големи корпуси, които да са синтактично и семантично анотирани от експерти.

- Семантиката е по-трудна за обработка от синтаксиса; текстовата семантика е по-трудна от лексикалната и изреченската семантика. Семантичните програми и ресурси са с много ниска ефективност (колкото повече семантика обработва софтуерът, толкова по-трудно е да се намерят правилните данни). Затова са нужни повече програми и инициативи в подкрепа на тази сфера както по отношение на фундаменталните изследвания, така и за създаване на адекватно анотирани корпуси.
- Съществуват и отделни продукти с ограничена функционалност в област като синтез на реч, разпознаване на реч, машинен превод и други.

Липсват достатъчно представителни паралелни корпуси, на които да се основават програмите за машинен превод. Преводът от и на български може да е много по-добър, ако съществуват повече данни.

- Има сериозни липси и при мултимедийните данни.
- Много от ресурсите не са стандартизирани, т. е. дори ако съществуват, не са достатъчно устойчиви; нужни са фокусирани програми и инициативи за стандартизиране на данни и формати за прехвърляне.

Резултатите показват, че българският е представен сравнително добре като покритие от основните езикови технологии и ресурси. Освен това са разработени и редица програми за снемане на многозначност, машинен превод, както и ресурси като паралелни и специализирани корпуси. Тези инструменти и ресурси обаче имат ограничена функционалност за определени области. Например паралелните корпуси покриват много малко езикови двойки и малко жанрове. Що се отнася до по-сложните области като семантика на текста, генериране на език и анотиране на многомодулни данни, за български липсват основни програми и ресурси, макар и някои да са в процес на разработка.

От 2000 г. насам се увеличи значително броят на проектите, финансирани от европейските фондове и от националните институции, като Фонд "Национални изследвания" към Министерството на образованието, младежта и науката.
В резултат на това през изминалото десетилетие бяха изградени редица важни електронни езикови ресурси (речници, корпуси, лексикални бази данни), както и програми за тяхната обработка (софтуер за снемане на многозначност, за корекция на правопис и други).

През последните две десетилетия разработването на езикови технологии за български не бе подкрепено от последователна национална схема за финансиране.

Процесът на създаване на приложения, програми и ресурси за български минава през комбинация от международни проекти, което разширява обсега на действие на разработките – от езиците в Западна Европа до тези в Централна и Източна Европа, на фона на процеса на разширяване на ЕС, националното финансиране за научни изследвания и ентусиазма на участниците в областта.

Очевидно са нужни повече усилия, при това насочени директно към ресурсите за български език, както и за фундаментални изследвания, иновации и развитие. Нуждата от големи масиви от езикови данни и сложността на системите за езикови технологии изискват и нови инфраструктури за споделяне и сътрудничество.

Можем да се надяваме, че участието на България в проектите CESAR и META-NET ще улесни разработването, стандартизирането и достъпа до важни езикови и технологични ресурси, като по този начин подкрепи развитието на езиковите технологии за български.

# За META-NET

META-NET е мрежа за високи постижения, финансирана от Европейската комисия. Към момента в нея членуват 47 организации от 31 страни от ЕС. META-NET поддържа развитието на Multilingual Europe Technology Alliance (META), разрастваща се общност от европейски професионалисти и организации в областта на езиковите технологии.

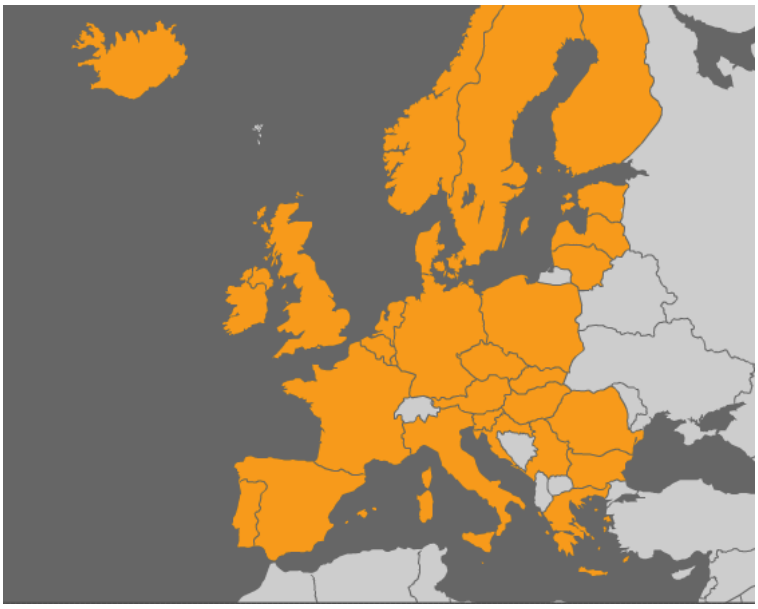


Figure 1: Countries Represented in META-NET

META-NET работи съвместно с други големи инициативи, например CLARIN, която има за задача утвърждаване на дигиталната хуманитаристика в Европа. META-NET ще подхрани технологичните основи за създаване и развитие на истинско многоезиково европейско информационно общество, което ще:

- позволи комуникацията и сътрудничеството между езиците;
- осигури равнопоставен достъп до информация и знания на потребителите с различни говорими езици;
- предложи модерни приложения за мрежово управлявани информационни технологии на европейските потребители на достъпни цени.

META-NET подкрепя и промотира мултиезиковите технологии за всички европейски езици. Технологиите ще помогнат за автоматичния превод, за създаването на съдържание и за управлението на информация и на знания за редица приложения и тематични области. Мрежата цели да подобри съществуващите подходи за подобра комуникация и сътрудничество между езиците. Европейците имат право на достъп до информация и знания независимо от езика.


*The Multilingual Europe Technology Alliance (META)*

## Посоки на развитие

Проектът META-NET стартира на 1 февруари 2010 г. с цел да подкрепи изследванията в областта на езиковите технологии. Мрежата подкрепя идеята да бъде обединена Европа от единен дигитален пазар и информационно пространство. META-NET вече организира поредица от събития. Трите посоки на развитие на META-NET са META-ВИЗИЯ, META-СПОДЕЛЯНЕ и META-ИЗСЛЕДВАНИЯ.

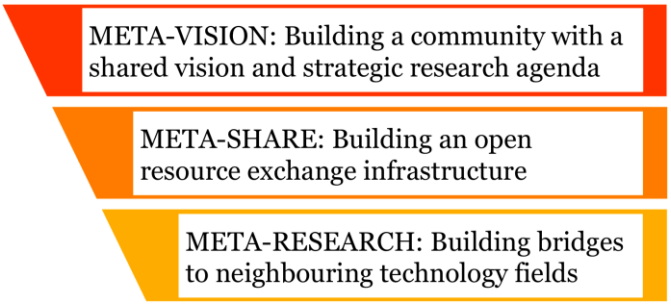



Figure 2: Three Lines of Action in META-NET

META-ВИЗИЯ работи за създаването на динамично и влиятелно активно общество, обединено от споделена визия и обща стратегическа изследователска програма (SRA).
Основната задача е изграждането на единна езиково-технологична общност в Европа чрез свързване на представители от фрагментираната и разнообразна група на участниците в нея. През първата година проектът META-NET беше представен на FLaReNet Forum (Испания), Language Technology Days (Люксембург), JIAMCATT 2010 (Люксембург), LREC 2010 (Малта), EAMT 2010 (Франция) и ICT 2010 (Белгия). Според предварителните изчисления META-NET вече се е свързала с повече от 2500 професионалисти в областта на езиковите технологии, с които да работи по целите и визията си. На META-FORUM 2010 в Брюксел META-NET представи предварителните резултати от процеса на изграждане на визията си пред над 250 участници. В серия от интерактивни сесии участниците в META-ВИЗИЯ споделиха своето мнение и предложения за визията на мрежата.
META-СПОДЕЛЯНЕ създава отворена среда за обмен и споделяне на ресурси. Мрежовата архитектура ще съдържа езикови данни, инструменти и уеб услуги с висококачествени метаданни, организирани в стандартизирани категории. Достъпните ресурси включват свободни материали с отворен код, както и продукти с ограничен достъп, комерсиално разпространение или получаване срещу такса. META-СПОДЕЛЯНЕ включва както вече съществуващи езикови бази данни, инструменти и системи, така и нови продукти или продукти, които са в процес на разработване и които участват при изграждане и оценяване на новите технологии, продукти и услуги.

Повторната им употреба, комбинация, смяната на целите, както и обновяването на езиковите данни и инструменти са много важни. META-СПОДЕЛЯНЕ ще стане важна част от пазара на езикови технологии, чиито клиенти ще са създатели на софтуер, експерти по локализация, изследователи, преводачи и езикови професионалисти от малки, средни и големи предприятия. META-СПОДЕЛЯНЕ е съсредоточено към цялостното развитие на езиково-технологичния цикъл – от извършването на изследвания до създаването на иновативни продукти и услуги. Инициативата META-СПОДЕЛЯНЕ трябва да се превърне във важна част от европейската и глобалната инфраструктура за езиково-технологичното общество.

Инициативата META-ИЗСЛЕДВАНЕ изгражда контакти между свързаните технологични области. Тя се стреми да се възползва от напредъка в други области и да получи достъп до иновативни изследвания за езиковите технологии. Една от по-конкретните цели е увеличаване на ролята на семантиката при машинния превод, оптимизиране на разделението на труда при хибридните системи за машинен превод, използване на контексти, в които автоматичните преводи да послужат за емпирична основа на машинния превод. META-ИЗСЛЕДВАНЕ си сътрудничи с други области и дисциплини като машинното обучение и семантичната мрежа. По META-ИЗСЛЕДВАНЕ се събират данни, подготвят се пакети от данни и се организират езикови ресурси за оценка; компилират се списъци с инструменти и методи; организират се работни семинари и семинари за обучение за членовете на общността. Вече са идентифицирани аспекти от машинния превод, в които семантиката може да повлияе за подобряване на най-добрите практики. Създадени са и ръководства за интегриране на семантичната информация в системите за машинен превод. META-ИЗСЛЕДВАНЕ завършва изграждането на нов езиков ресурс за машинен превод – Анотирания хибриден примерен корпус за машинен превод, в който са включени данни за езиковите двойки английски-немски, английски-испански и английски-чешки. META-ИЗСЛЕДВАНЕ разработва и софтуер, който събира многоезични корпуси, скрити в уеб пространството.

## Организации членки

Приложената по-долу таблица съдържа организациите и техните представители, които участват в META-NET.

Състав на мрежата за високи постижения META-NET

| Страна | Член (Институция) | За контакт |
|---|---|---|
| Австрия | Университет на Виена | Герхард Будин |
| Белгия | Университет на Антверпен | Валтер Делеманс |
| | Университет на Льовен | Дирк Ван Комперноле |
| България | Българска академия на науките | Светла Коева |
| Хърватия | Универитет на Загреб | Марко Тадич |
| Кипър | Университет на Кипър | Джак Бърстън |
| Чешка република | Карлови университет в Прага* | Ян Хаджич |
| Дания | Университет на Копенхаген | Бент Мегаард, Болеете Санфорд Педерсен |
| Естония | Университет на Тарту | Тиит Роосмаа |
| Финландия | Университет Аалто* | Тимо Хонкела |
| | Университет на Хелзинки | Кимо Коскениеми, Кристер Линден |
| Франция | CNRS, LIMSI* | Жозеф Мариани |
| | ELDA* | Халид Чукри |
| Германия | DFKI* | Ханс Уцкорайт, Георг Рем |
| | RWTH Аахен* | Херман Ней |
| Гърция | ILSP, R.C. “Атина”* | Стелес Пиперидис |
| Унгария | Унгарска академия на науките | Тамаш Варади |
| | Технически университет в Будапеща | Геза Немет, Габор Оласи |
| Исландия | Университет на Исландия | Ейрикюр Рьонвалдсон |
| Ирландия | Дъблин Сити Юнивърсити* | Джоузеф Ван Генабит |
| Италия | Национален изследователски съвет* | Николета Калцолари |
| | Фондация “Бруно Кеслер”* | Бернардо Манини |
| Латвия | Tilde | Андрейс Васильевс |
| | Университет на Латвия | Ингуна Скадина |
| Литва | Институт за литовски език | Йоланта Забарскайте |
| Люксембург | Arax Ltd. | Варткес Гьотчериан |

| Страна | Член (Институция) | За контакт |
|---|---|---|
| Малта | Университет на Малта | Майк Рознър |
| Нидерландия | Университет на Утрехт * | Ян Одик |
| Норвегия | Университет на Берген | Коенраад Де Смет |
| Полша | Полска академия на науките | Адам Пжепорковски |
| | Университет на Лодз | Пьотр Пезик |
| Португалия | Университет на Лиссабон | Антонио Бранко |
| | Институт за системи, инженеринг и компютри | Изабел Транкозо |
| Румъния | Румънска академия на науките | Дан Туфиш |
| | Университет “Александру Йоан Куза” | Дан Кристеа |
| Сърбия | Университет в Белград | Душко Виташ, Цветана Крстев, Иван Обрадович |
| | Институт “Михайло Пупин” | Саня Вранеш |
| Словакия | Словашка академия на науките | Радован Гарабик |
| Словения | Институт “Йозеф Стефан”* | Марко Гробелник |
| Испания | Barcelona Media* | Тони Бадиа |
| | Технически университет на Каталуня | Асунсион Морено |
| | Университет Помпеу Фабра | Нуриа Бел |
| Швеция | Университет на Гьотеборг | Ларш Борин |
| Великобритания | Университет на Манчестър | София Ананиаду |

*: институции – основатели

**Как да участваме?**

META-NET и META предлагат много възможности за участие и сътрудничество. Моля, проверявайте на сайта ни www.meta-net.eu за информация за бъдещи срещи и дейности

# Референции


[i] European Commission, *Multilingualism: an asset for Europe and a shared commitment*, Brussels, 2008 (http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf).

[ii] UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf).

[iii] European Commission Directorate-General for Translation, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies).

[iv] http://www.ethnologue.com/show_language.asp?code=bul

[v] http://www.aba.government.bg/?show=english

[vi] http://www.nsi.bg/census2011/pagebg2.php?p2=36&sp2=37

[vii] Dictionary of New Words in Bulgarian (2010)

[viii] За сравнение – новите заемки от англ. превъзхождат многократно новите заемки от другите западноевропейски езици, в речника са включени 61 думи с произход от френски, 13 от немски, 27 от италиански и 4 от исп. Характерно е освен това, че заемки от други езици, с които БЕ няма пряк контакт, навлизат в БЕ през английски (или други западноевроп. езици) – напр. японски по произход, но навлезли с посредничеството на англ., са думи като караоке, аниме, суши.

[ix] http://bg.wikipedia.org/wiki/Шльокавица

[x] http://www.oecd.org/document/61/0,3746,en_32252351_46584327_46567613_1_1_1_1,00.html

[xi] http://nces.ed.gov/surveys/pirls/

[xii] http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

[xiii] http://www.gemius.com

[xiv] http://www.internetcee.com

[xv] http://www.internetworldstats.com

[xvi] Wikipedia metadata: http://meta.wikimedia.org/wiki/List_of_Wikipedias.

[xvii] http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html

[xviii] http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html

[xix] The higher the score, the better the translation, a human translator would get around 80. K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of ACL, Philadelphia, PA.

[xx] Gianni Lazzari: „Sprachtechnologien für Europa“, 2006: http://tcstar.org/pubblicazioni/D17_HLT_DE.pdf