МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2000
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2000
Proceedings of Twenty Ninth Spring Conference of
the Union of Bulgarian Mathematicians
Lovetch, April 3–6, 2000

КОМПЮТРИЗИРАНЕ НА ЕСТЕСТВЕНИТЕ ЕЗИЦИ

Димитър Петров Шишков

Предлага се глобален проект „Компютризиране на естествените езици“ и се разглеждат основните му задачи. От тях са отбелязани трите най-важни неотложни задачи за всяка държава: въвеждането, оцифряването и разпознаването на националната реч, сканирането на цялата книжнина в държавата и създаването на универсалния компютърен речник на съответния естествен език.

Животът е кратък, изкуството и науката са вечни.

XXI век чука на вратата ни с много световни проблеми, които трудно могат да се наредят по приоритет. Всички те са изключително важни и неотложни. Един от тях е **компютризирането на естествените езици** (ЕЕ).

Тук накратко ще бъдат поставени и разгледани основните задачи на този гигантски комплекс, който нататък ще бъде наричан **Проектът** (Компютризиране на естествените езици, КЕЕ).

Засега той ще се отнася само до езици с азбучна писменост. Ако за йероглифните китайски и японски езици могат да бъдат използвани резултатите от „азбучните“ езици, възможното им включване в Проекта може да доведе и до замяната на йероглифите им с латиница. Това би било уникално постижение не само за Китай и Япония (ускорена грамотност, по-добра комуникация и др. удобства), но и за целия свят. Проектът трябва да помогне за това.

Той се дели на две основни задачи с няколко подзадачи и трета основна задача – **главно езиково средство**.

ЗАДАЧА 1. ВЪВЕЖДАНЕ И ОЦИФРЯНЕ НА ТЕКСТ (СЪЗДАВАНЕ НА КОМПЮТЪРЕН ТЕКСТ)

Ръчно въвеждане на текст

Задача 1.1. Ръчно въвеждане на текст и просто изображение от компютърна клавиатура. Задачата е напълно решена (включена е за пълнота). Въвеждането се подпомага от програмни средства – например текстов редактор Word с възможности за „просто“ рисуване. Задачата е напълно решена, като клавиатурата се използва отдавна и навсякъде в света. Нещо повече – предстои изваждането ѝ от употреба и снемане от производство.

Сканиране на текст и рисунка

Задача 1.2. Сканиране и оцифряне на печатан или по друг начин осъществен текст, произволно печатно изображение, както и на ръкопис, разглеждан като изображение. Извършва се с компютърен сканер, който може да бъде черно-бял и/или цветен с различни размери. Задачата е рутинна и е свързана единствено със средства и време (при повече средства – за по-малко време).

1.2.1. Сканиране на текст. Необходимо е да бъдат сканирани всички печатни и ръкописни материали във всички държави – ръкописи, вестници, списания, брошури, книги, афиши, ноти, географски карти и пр., които се намират в националните библиотеки, библиотеките на университетите и редица специализирани библиотеки (без дублиране), както и целият държавен архив. Изключително важна задача е да се оцифрят материалите в националните библиотеки на великите държави. Може само да се досещаме какво например ще донесе за човечеството сканирането на книжните съкровища на Ватиканската библиотека (ако това бъде разрешено).

Пред завършване е проектът за въвеждането в Интернет на 600-годишната писмена история на Отоманската империя, като са сканирани и оцифрени над 150 млн. документа от Държавния архив на Турция.

Оцифряването на световната книжнина ще доведе до това на целия човешки опит.

За съжаление засега сканирането на всяка страница е относително бавен, но все пак рутинен процес, който не изисква особена квалификация. За да се намали времетраенето му, е необходимо:

1.2.1.1. Създаване на сканер за книги. При сканирането на страница с него останалата част на книгата след страницата трябва да „виси“ под прав ъгъл извън този сканер (например една от страните на рамката на стъклото да бъде не 5 см., както е обикновено, а няколко милиметра, понеже практически всяко вътрешно бяло поле на страница е по-голямо от тях освен за изключително малки книги). Защото, в противен случай, за по-качественото разпознаване ще се наложи „дебелите“ книги да се срязват отляво, а след това да се подвързват частично отново. Това също ще изисква средства, а и красивите подвързии няма да могат да се възстановяват. Да не говорим за древните книги.

Нещо повече, за целите на Проекта трябва да се създаде робот – сканер за книги с устройство за прелистването им на вакуумен или електростатичен принцип.

1.2.1.2. Създаване на мощни комплексни OCR програми. Те трябва максимално да автоматизират процеса на сканирането на графични изображения и да могат да разчитат произволни, а значи и древни шрифтове с печатни букви. За бъдещите шрифтове трябва да се изработи световен стандарт.

Нека отбележим, че сканирането дори на една страница изисква много памет (например 300-500 Кбайта за чист текст) и практически е невъзможно изображението да се пази изцяло за по-късно разпознаване особено ако освен текст има и картина или само картина. Ето защо разпознаването на текст, получен като изображение, и превръщането му в знаков текст трябва да става незабавно след сканирането.

Паралелно с началното сканиране трябва да се извършва и

1.2.1.3. Окончателно дооформяне на оцифрен текст и рисунка. Това може да се нарече козметика. То ще струва повече и ще бъде значително по-бавно от началното сканиране. Необходимо е всеки сканиран текст да бъде разгледан незабавно (за

икономия на дискова памет) на компютърен екран и поправен (приведен в пълно съответствие с оригинала) от тесен специалист, евентуално съвместно с лице с малка компютърна грамотност – ученик, студент и др. Важно и ефективно ще бъде незабавното правене на резюме на текста и класифицирането на самия текст, за да може да се търси след това. Едва тогава може да се приеме, че той е веднъж завинаги въведен в компютърна памет (специалистът ще подписва протокол). Това е особено важно, понеже след няколко десетилетия (възможно и по-рано) хартиеният носител практически няма да се използва, и то завинаги, а старите печатни материали върху хартия физически ще загиват с времето.

В началото на 1999 г. американските фирми NuvoMedia Inc., SoftBook Press и Everybook Inc. създадоха първите устройства „електронни книги“, съответно Rocket eBook, SoftBook. и Everybook. Така „електронна книга“ стана омоним – досега беше всяка книга, която може да се прочете в Интернет, а сега произведените от трите фирми електронни устройства са за четене на всички електронни книги. Предлагам то да се нарича *универсална електронна книга* (уек; на английски ueb, юеб, което е съзвучно с web, уеб). Това е началото на предизвестения край на хартиените носители – краят на 500-годишната епоха на книгопечатането. Естествено, при сканирането, това понякога ще води до негативизъм у библиотекарите, които правилно ще почувстват, че тяхната професия е пред изчезване, макар и още много далечно.

1.2.1.4. Сканиране на ръкопис. При дотъкмяването, козметиката на текстовете и заедно с тях – и изображенията, до основна трудност ще доведат последните, а с особена трудност ще бъде преобразуването на ръкописите от изображение в текст (разчитането им). Сега почти няма разработени мощни средства за „прочитане“ на произволни, нетипографски ръкописи.. Ръкописите ще се запазват като изображения до „по-добри“ времена, макар че вече е започнало бурно развитие на компютърното разчитане на сканирани ръкописни текстове. Такива текстове, написани с печатни (т.е. отделени) букви обаче, например текстовете на средновековните ръкописи, разбира се, вече трябва да се превръщат в печатен текст. При обработването на древните ръкописи, които са написани като непрекъснат текст, те ще бъдат прочитани от специалиста пред компютъра, който ще може да поставя интервали между думите в получения компютърен печатен текст и да извършва и други полезни преобразования. Разбира се, ръкописите ще се запазват изцяло и като компютърни изображения.

1.2.1.5. Сканиране на текст върху нехартиена повърхност. Чрез сканиране трябва да се оцифрят и фотографии на текстове върху паметници (надгробни, върху колони и др.) и други материални повърхности (например вази, грънчарски произведения, златарски произведения и др.).

1.2.1.6. Сканиране на географска карта.

1.2.1.7. Сканиране на нотен текст. То е неизмеримо по-просто (дори на ръкописни нотни текстове) от сканирането на всички останали печатни текстове.

1.2.2. Сканиране на художествена картина и художествена неравнинна повърхност. Изключително важен световен проблем е и оцифряването на всички картини и предмети с художествена, историческа и етнографска стойност (независимо от това дали има текстове върху тях), които също постепенно загиват, а реставрацията на най-ценните от тях е скъпа, бавна и многократна.

Нека отбележим,че клавиатурата и сканера са единствените технически средства

за въвеждане и оцифряне на данни в масов мащаб. (За въвеждането на числови данни се използват и аналоговоцифрови преобразуватели за получаването на стойности на физически величини.)

Българската националната библиотека „Св. Св. Кирил и Методий“ и Столичната библиотека ще бъдат снабдени с френски компютри, дарени от Асоциация „Предприятия и хора“ със съдействието на българската фондация „Бъдеще за България“. Компютрите ще разполагат в паметта си с всичките произведения на 200 класици на френската литература. Предстои също така да бъдат въведени творбите на великите писатели на Англия, Италия, Германия и България. Проектът е пилотен и се провежда под патронажа на ЮНЕСКО.

През XXI век няма да се наложи сканиране на печатни материали. Всички книги и документи, които все още ще се печатат върху хартия, ще бъдат предварително оцифрени. Разбира се, трябва да се разработи специален *Закон за компютърните данни* (особен раздел в *Закон за информацията*), който ще изисква от всички издатели и чиновници (последните – за Държавния архив) депозиране във вид на компютърен текст на всички писмени материали. Макар и да звучи кощунствено, *скоро хората ще се отучат да пишат с молив, писалка и на клавиатура.* Природата не ги е създала да пишат, а да говорят (и може би да рисуват).

Превръщане на реч в компютърен текст

Задача 1.3. Разпознаването на речеви сигнали се извършва на два етапа:

1.3.1. Технически етап. Разпознават се и се оцифрят *фонемите* (*алофоните* – фонеми с обкръжението им) от речта, които за почти всички езици са около 40. Получава се непрекъснат текст без интервали и, разбира се, без пунктуация. Този текст в голямата си част е неправилен относно книжовния правопис.

В този етап трябва да се разпознават 3 езика и (ако е необходимо) да се прави идентификация на говорещия:

1.3.1.1. Основният език.

1.3.1.2. Езикът на интонацията на говорещия. Тя сигурно, но само донякъде ще послужи за отделянето на словоформите и изреченията и понякога може съществено да измени семантиката на думите (авторът често казва „бомба“, което за него означава висша оценка или похвала). Желателно е интонацията да се записва в оцифрения текст и да се изобразява на екрана с думи и знакове, подобно на това в музиката – модерато, крешчендо и пр. В текста на пиеси авторите най-често казват каква трябва да бъде интонацията на артиста, но в романите това рядко се прави. А жалко – за удобство на читателя и за изясняване на мислите на автора е добре писателите да вписват това в текстовете си.

1.3.1.3. Езикът на мимиката, жестовете и неречевите звукове (например покашляне) на говорещия. Това обаче ще се използва по-късно, когато използваните съвместно микрофон и видеокамера станат съвсем миниатюрни. Мимиката и жестовете може също да изменят семантиката на произнасяните думи.

1.3.1.4. Идентифициране на говорещия („дактилоскопия“ на речта).

1.3.2. Езиков етап. Непрекъснатият текст се разделя на думи и изречения.

Тук са възможни поне два подхода:

1.3.2.1. След получаването на не много дълъг, съответен на разпознатите фонеми, текст (за да не се получи комбинаторен взрив), *само на основата на речник да*

се търсят и отделят думите и да се поставят интервали между тях. Основна компютърноинформатична техника (виж. по-долу компютърна информатика) може да бъде бактракинг. Тук е възможно, макар и рядко, да се получи различен прочит, което е недопустимо.

1.3.2.2. При наличието на формализиран тълковен речник (вж. зад. 3, етап 2), който ще бъде в основата на семантиката на думите, е възможно процесът 1.3.2.1 да се извършва и чрез *семантичния им анализ*, което би избягнало двусмислие, но е много по-трудно в общия случай.

През 1999 г. японска национална телефонна корпорация е създала компютърна програма за разпознаване на реч, която анализира смисъла на фразата непосредствено при произнасянето ѝ. Програмата работи много по-бързо от системите, които започват да интерпретират изреченията, едва след като ги „изслушат“ докрай. Тя се използва за автоматични телефонни справки и резервация на билети. Естествено, тази програма работи със силно ограничен речник и предварително фиксирана семантика.

Техническият етап е по-лесен и може да се използват готови решения за фонетиката на основните езици. Говорният апарат на човека практически произвежда почти едни и същи фонеми (около 40 за даден език), независимо от езика. И все пак за редица държави това ще бъдат резултати за чуждоезикова фонетика. Езиковият етап обаче може да се реализира само от дадена държава за нейния национален език (езици) – едва ли някой ще го осъществи (с други думи – прецизира съответната фонетика) вместо тази държава. Ако например за българската реч се използва разпознаването на фонемите от руската фонетика, трябва да се добави фонемата на звука „Ъ“, както и да се изследват разни нюанси.

Известни са редица сериозни и скъпи разработки в САЩ, Русия (Москва), Украйна (Донецк) и другаде по въвеждането на тяхната реч.

Напоследък беше съобщено, че през последните 10 години най-мощната американска фирма Майкрософт и нейният собственик Бил Гейтс са инвестирали милиарди долари за вече близката цел – компютрите да бъдат способни да разпознават човешка реч, сами да притежават „зрение, слух и собствена реч“. Според тях това ще стане най-много до още 10 години. Мисля, че по отношение на английския език това ще стане още по-рано благодарение на техните усилия.

IBM съобщи за езика Speech Markup Language (SpeechML), с който основаните на Web приложения могат да бъдат разширени с възможности за разпознаване на реч. Този език използва спецификацията XML (Extensible Markup Language). SpeechML дава възможност на Web дизайнера да използва в своя сайт тагове (етикети) за интерактивна обработка на реч, без да има познания за тази специфична технология. Разработчикът може да маркира с таг част от съдържанието на страницата, която (част) да бъде прочетена от приложението.

1.3.3. Превръщане на звучащ музикален фрагмент в текст – ноти със съответната оркестрация (при оркестрово изпълнение) или с поръчана автоматична оркестрация на проста мелодия. Тази подзадача не е свързана с ЕЕ, но е поставена тук за пълнота и за да не бъде забравена.

ЗАДАЧА 2. ПРЕОБРАЗУВАНЕ НА КОМПЮТЪРЕН ТЕКСТ

Това е свързано с преминаване от една форма – текст, звук, изображение – в друга, а също и с преобразуването на една и съща форма.

Естествен език – естествен език

Преобразуване на компютърен текст на ЕЕ в текст на същия език

2.1.1. Проверяване на правописа на всяка дума в текст на съответния език. В този текст е възможно да има думи и от друга азбука, в т.ч. английски и латински думи и изрази, наименования на известни чужди фирми и институции и пр. Засега се прави проверяване на правописа на ограничен брой думи, като при грешка се предлагат за замяна няколко думи, подобни на сгрешената по написване, но не и по смисъл. Това често обезсмисля предлаганата замяна.

2.1.2. Предлагане на подобрения на стила. Например разместване на думи, изречения, параграфи; обръщане на внимание (дори препятстване) на въвеждането на жаргон, обидни думи, неправилни съчетания. Социално е особено важно автоматичното оценяване на грамотността и писмения стил на дадено лице. Това силно ще помогне за запазването и по-качествената употреба на ЕЕ.

Тази задача е твърде трудна и изисква продължителни научни езикови изследвания, вкл. практически пълната формализация на ЕЕ заедно с всичките му изключения. Това е необходимо, за да може да се извършва пълен синтактичен, морфологичен и семантичен анализ на всяко изречение. В тази област е сложно, но е възможно да се използват чуждоезикови постижения.

Задача 2.2. Автоматично търсене (индексиране) и реферирање на текст. Това също е трудна задача. Засега в света текстове се търсят по набор ключови думи – твърде лесен, но „първобитен“ начин. Не се търси по смисъл, по синоними и пр. Естествено е да се „губят“ текстове, въпреки че са налични, т.е. лицето ги търси, но не ги получава от търсещата програма. Отново са необходими средства и задълбочени езикови изследвания. Също е трудно, но е възможно да се използват чужди резултати.

Задача 2.3. Автоматично търсене и извличане на знания от компютърен текст. Това е основна област в изкуствения интелект (ИИ), който е най-бързо развиващият се дял на компютърната информатика (КИ; не е напълно синоним на Computer science, CSc. КИ е гигантски комплекс от науки – половината от съвременната математика, който изучава обработката на компютърни данни с компютърни системи). Сканирането на всички текстове в света е всъщност оцифряването на целия човешки опит. Това ще даде възможност за автоматичното извличане на знания и класифицирането по важност на получените предикати. Разбира се, ще се изключват всякакви повторения. Главна роля ще играе разбирането на семантиката на текстовете и знанията. Тук могат и трябва да се използват и всякакви чужди резултати. Отново са необходими средства и много големи езикови изследвания.

Преобразуване на текст на ЕЕ в текст на друг ЕЕ

Задача 2.4. Автоматичен превод от един на друг език. Вече почти 50 години в света се провеждат големи изследвания в тази област.

Напоследък се продават портативни компютри, които превеждат от английски на няколко други езика. Те съдържат около 36 000 думи и 300 най-употребявани фразеологични съчетания. Екранът е разделен на две – отляво е разположен текстът а английски, а отдясно – преводът. Компютърът пита за непознатите думи.

Известна и достатъчно качествена е руската система STYLUS за превод от английски, френски и немски на руски и обратно, която е снабдена с редица тематични речници.

Възможен е и друг подход. Преди 30 години, за транслиране на програми от m езика от високо равнище на машинните езици на n компютъра (общо $m \times n$ транслатора) бяха предложени и използвани два междинни езика АЛМО (СССР) и UNCOL. Първо програмите се превеждаха на един от междинните езици, а след това от него – на съответния машинен език. Така транслаторите се намалиха до $m + n$.

Същото може да се направи и сега, но само за превод на „по-леки“, нехудожествени текстове, например от спорта, електронната търговия, политиката и др. Може да се изгради „по-лесен“, *универсален междинен ЕЕ*, от група ЕЕ на него да се превеждат по-леки текстове, а от получения универсален текст след това да се прави превод на друг от тази група.

Отново са необходими средства, изследвания и използване на чужд опит.

Естествен език – звук (устна реч)

Преобразуване на компютърен текст на ЕЕ в устна реч на същия език

Задача 2.5. Автоматично преобразуване на компютърен текст в устна реч. Задачата е задоволително решена в някои държави. Необходимо е да се реши и за останалите ЕЕ. Текстът трябва да се „прочита“ с „глас“ от зададен пол, възраст и тембър, дори и с „гласа“ на популярни хора – артисти, диктори, политици. Последното все още е много трудно. Свързано е с идентификацията на човешкия глас. Тази задача за синтезиране и извеждане на реч трябва да се решава съвместно със зад. 1.3 – въвеждане и оцифряне на устна реч. Интересен въпрос е дали ще може да се разпознава, че именно компютър е „прочел на глас“ текст.

Решенията на зад. 1.3 и 2.5 ще се използват в системите за разпознаване и предаване на глас, в т.ч. и по Интернет, в цифровата телефония (засега само Интернет телефония), в автоматичните преводачи и другаде.

Изображение – естествен език и обратно

Преобразуване на изображение чрез пораждане на текст на ЕЕ и обратно

Задача 2.6. Синтезиране на текст въз основа на (просто) изображение (разказване на съдържанието му) и обратната

Задача 2.7. Рисуване на (просто) изображение въз основа на текст. Възможно е текстът да е получен от въвеждането на реч.

И двете задачи са много трудни и неизследвани, и двете трябва да се решават съвместно. Бързодействието и обемът на паметта на съвременните компютри дават

възможност да се започнат такива изследвания. Естествено, когато се синтезира текст по изображение, последното след това може да се „разкаже“, съгласно зад. 2.5.

Пораждане на текст на естествен език

Задача 2.8. Синтезиране на компютърен текст по сценарий (описание на съдържанието му). Това означава например създаването на проза (книга) по зададен сценарий с избор на действащи лица и избираеми разклонения на сюжета, създаването на поезия, музика и пр. Това отдавна се прави, но резултатите са примитивни, понеже проблемът се разглежда от търговската му страна и практически не се извършват сериозни научни изследвания. Най-подходящата област за това е ИИ.

Сега ще бъде разгледано накратко *основното средство за решаване на първите две основни задачи от Проекта*:

ЗАДАЧА 3. СЪЗДАВАНЕ НА УНИВЕРСАЛЕН КОМПЮТЪРЕН РЕЧНИК НА ЕСТЕСТВЕН ЕЗИК

Създаването и поддържането на Речника (УКРЕЕ) е огромна по обем държавна научноприложна задача. Тя обаче е рутинна, не изисква никакви особени нови научни изследвания, има достатъчен брой подготвени научни кадри за нея и е въпрос единствено на средства и строга организация на много хора. Лингвистиката с нейния съвременен клон математическа (компютърна) лингвистика и съвременната компютърна техника са напълно готови за решаването на тази национална задача, която стои пред всички ЕЕ.

За сравнение, вече се продава за широка употреба пълният Оксфордски речник на английския език с много милиони думи, разположени на няколко диска (сидирома). Националните абсолютни речници (отсега нататък – **Речникът**) на държавите със силно развити собствени езици също трябва да имат поне толкова. думи, понеже основните ЕЕ не са по-бедни от английския (вж. по-долу състава на Речника). Тъй като всеки език, в т.ч. ЕЕ, е средство е моделиране на света, под силно развити ЕЕ тук се разбират тези от тях, които най-пълно, най-добре и най-фино моделират обкръжаващия ни и абстрактните светове. За речевото разпознаване Речникът трябва да има още няколко пъти по толкова думи – същите, но с правилно произношение, което не е съответно на книжовния им правопис, както и неправилното произношение на всички думи. Оттук Речникът трябва да съдържа много милиони, различни по написване или произнасяне (за по-бързо търсене) словоформи, които ще бъдат обединявани с указатели (валенции) към съответната лексема. Всяка лексема именува клас от словоформи, например инфинитив на глагол – съответните му глаголни форми.

През декември 1998 г. бе завършен и издаден най-големият речник в света – „Речник на холандския език“, който се състои от 45000 страници в 40 тома и включва милиони думи. Съставян е 147 години, но до 1976 г., т.е. с излизането си вече остарял.

Известно е, че се завършва работата по създаването на пълния речник на китайския език.

След не повече от 10 години обемът на един CD-ROM ще е толкова огромен, че например на няколко диска ще се побира цялото съдържание съответно на националната библиотека, университетските и другите специализирани библиотеки

(без дублиране), държавния архив (разбира се, на една по-малка държава), както и предлагания национален Речник. Цялата книжнина и Речникът ще бъдат защитени пазарни продукти с първоначално висока цена – до няколко години търговските и въобще всякакви компютърни данни ще бъдат защитени напълно и навсякъде от присвояване, т.нар. пиратство в тази област. По-късно тази цена ще падне и книжното национално богатство ще стане достъпно за всички граждани на съответната държава, а също и за всички чужденци и изследователи, които знаят или ползват съответния ЕЕ.

Речникът може да се състои от следните подречници, включени по важност на етапи:

ЕТАП 1. Речник на: 1. Съвременните думи, вкл. жаргон и обидни думи, с написването на думите през различните периоди до наши дни. 2. Остарелите думи. 3. Диалектните думи, стари и нови. 4. Чуждите думи, вкл. и заместването им с национални синоними. 5. Чуждите думи на друга азбука, които се срещат в националните текстове, вкл. речник на крилатите латински изрази и поговорки, както на латиница, така и с изписване на съответния език; речник на латинските термини в ботаниката, зоологията и медицината; наименованията на чужди институции и фирми на националния и чужд език. 6. Терминологичен речник в хиляди подобласти на науката, техниката и занаятите, вкл. и научния жаргон. 7. Синонимен речник с оценка на близост (засега това е много трудно) на синонимите. 8. Паронимен речник. 9. Националните лични имена и фамилии. 10. Използваните в дадената държава чужди лични имена и фамилии на собствените си граждани и тези в преводната литература от чужд, а също и в националния език, при това с различно написване поради неправилно звучене. 11. Историческите имена на дадения ЕЕ и съответния чужд език. 12. Географските имена, национални и чужди, на собствен и чужд език, стари и нови. 13. Наименованията на националните институции. 14. Частичен фразеологичен речник – твърди съчетания, идиоми, метафори, поговорки и др., вкл. и в научната терминология като съчетания на национална дума и научен термин чуждица. 15. Съкращенията. 16. Некоренните морфеми – представки, наставки, окончания. 17. Сричките и др.

Тук под речник (т.е. подречник) се разбира пълен речник със стар и нов правопис. Всяка дума в него се съпътства с всичките си граматически и смислови атрибути: с разделянето си на морфеми и срички с означаването на всички възможни места за пренасяне в нея; пълно граматично описание; по възможност формализирано семантично описание; произношение – правилно, с място на ударението, неправилно, диалектно; указатели към списъка на синонимите на думата или израза; твърди фразеологични съчетания, в които участва; указатели към други списъци; други атрибути и пр. **Такъв универсален („абсолютен“) речник не е съществувал досега.**

ЕТАП 2. 18. Пълен фразеологичен речник. 19. Етимологичен речник. 20 Старият национален език с превод на новия. 21. Тълковен речник, вкл. и на научната терминология – пълна енциклопедия на съответния ЕЕ.

Този тълковен (едноезичен) речник трябва за бъде в основата на семантиката на езика. Според автора той трябва да бъде построен на основата на *аксиоматично* избрани думи с компютърноинформатичната техника буутстрапинг (саморазвиване). Аксиоматично, без определяне, се приемат например 2000 думи от езика. Чрез

конкатениране с тях вече се определят например милион думи. А от всичките аксиоматично въведени думи и определените чрез тях се определя останалата по-голяма част от лексиката на езика. И т.н. докрай, възможно чрез нови равнища на определяне. Поради важността си, създаването на тълковния речник трябва да започне незабавно с трите основни задачи.

ЕТАП 3. 22. Двуезични речници: национален – чужд език и обратните им. 23. Съкращенията в основните чужди езици. 24. Енциклопедия „Страната Х“. 25. Тематични енциклопедии. 26. Обща световна енциклопедия. 27. Други речници, в т.ч. обратен, римен и т.н..

Тъй като всеки научен или технически термин, както и всички останали думи ще имат атрибут за принадлежност към една или няколко области, във всеки момент ще може лесно да се съставя речник на съответната тематична област.

Създаването на компютърния вариант на националната книжнина и на Речника са и уникален международен, а значи и световен политически въпрос. Въпреки това, все още не се работи масирано по проблема.

ЕС като цяло засега не инвестира достатъчно в своето бъдеще, което е в развитието на технологиите и особено високите. А световното равнище изисква значителни инвестиции в изследвания и нововъведения (от доклада на Европейската комисия от средата на април 1998 г.). От 1980 до 1996 г. ЕС е вложил в разработки и проучвания само 1,8% от брутния си вътрешен продукт, докато Япония е вложила 2.8%, САЩ – 2,5%, Корея, Сингапур и Тайван – по 2,2%. Най-фрапиращото е, че Швеция води в света с 3,5%, а Франция, Германия, Великобритания и Холандия са в десятката на света по инвестиране в изследвания.

ЗАКЛЮЧЕНИЕ

Не е трудно веднага да се посочат пропуските на автора – неволни или по незнание. Много по-важно е незабавно, с по-малък бюджет да се започне във всяка държава реалната и непрекъсната работа по Проекта и преди всичко на трите му най-важни и спешни задачи: **1.** *Въвеждане, оцифряне и разпознаване на устната национална реч.* **2.** *Сканиране на цялата книжнина в държавата и* **3.** *Създаване на Речника.* Това важи и за големите държави с „големи“ езици, особено що се отнася до решаването на някои от поставените по-горе подзадачи. После ще бъде късно. Скоро клавиатурите ще изчезнат и никой в чужбина няма да се занимава с въвеждането и разпознаването на национална реч, ако тя не е на основен световен език, понеже националният пазар на малките държави (но не с малки ЕЕ) винаги ще бъде малък за чуждите компютърни компании. Създаването на Речника може да бъде само национално дело. Той ще бъде важен и за чужденците, особено при превода от националния език на чужд и обратно. След няколко десетилетия преводачите няма да са хора, а компютри – практически няма да се налага изучаването на чужди езици освен от лингвисти. Също така смятам, че рано или късно английският език ще стане основен, ако не и единствен за всички хора. Надявам се, че националните езици на малките държави ще са важни за чужденците, най-малкото докато има държави в света, тъй като и държавите са пред изчезване, може би до средата или края на XXI век.

Основна цел на Проекта е запазването на книжнината на всички езици, на цялата човешка култура и знания (човешкия опит). Предполагам, че до

100 години ще изчезнат всички езици за сметка на английския език, който ще бъде езикът на единствената Държава „Светът“. Тогава сканираната съгласно Проекта цяла човешка книжнина (освен поезията) ще може бързо и съвсем точно да бъде преведена на английски и запазена за човечеството.

Преди началото на решаването на трите основни задачи обаче трябва да се решат други три специфични задачи: **1.** Незабавно изготвяне на глобално или регионално равнище на началния вариант на *Заданието* на световния проект – дърво от задачи и подзадачи с десетична класификация **2.** Формиране на *Ръководство* на съответния национален проект. По принципа на организацията на МОК (членовете са личности, а не представители на държави), според мен, членовете на националното ръководство трябва да представят себе си, а не организациите си, т.е. важна организация може и да няма свой представител. Членовете трябва да са почтени хора и професионалисти в своята област. Иначе реализирането на Проекта може да се изроди. **3.** Създаване на *пълна национална* (собствена и чужда за дадения език) *библиография по лингвистика, в т.ч. компютърна лингвистика.* Библиографията трябва да се разпредели по подзадачите на Заданието, като дадена публикация може да фигурира в няколко или дори всички подзадачи. Лингвистичните трудовете трябва отново да се изучават и преосмислят в светлината на съвремието и Проекта. **4.** Поетапно осигуряване на *финансирането на Проекта* със средства, *получени* от държавата, чужди и национални фирми, международни и национални фондации. **5.** Използване на *армията от силни специалисти в държавата.* **6.** Регламентиране на *съвместния обмен и използването на резултатите* от работата на участниците, които представляват взаимен интерес. **7.** Разпространение на *информацията за проблемите, задачите и резултатите от изпълняването на Проекта* с цел да се получи *поддръжка от обществото.* Необичайно важно е да се преодолява възможното *съпротивление на обществеността.* Консервативността на хората съвсем сигурно ще доведе до силна съпротива срещу Проекта. Току-що в Германия бе извършена не особено голяма реформа на писмения немски език, което обаче доведе до страхотен остракизъм от немския народ, в т.ч. и масов чрез съда. Законът за полския език също среща силен отпор.

Тук трябва да се спомене, че поради липса на средства вече 9 години в България не се получават важни списания във всички области. Това ще попречи на Националния проект „Компютризиране на българския език“, понеже във всяка държава, още в началото на Проекта, трябва да се изучи световният опит в областта на Проекта.

Има много изследвания по света в областта на Проекта, но те са частични, неорганизирани и недостатъчни. Крайно време е да бъдат обединени в един световен проект. Като начало това може да бъде *регионален* проект, обхващащ например езици от една езикова група, който да прерастне в световен.

В глобален мащаб трябва да се извърши следното: **1.** Да се обединят усилията на *поне 50 държави* (в това число засега не влизат повечето африкански). **2.** Да се формира *Международно ръководство* на световния Проект. **3.** Да се приеме и непрекъснато разширява и усъвършенства *Заданието* на Проекта.

Речникът и целият Проект ще помогнат за по-бързото създаване в много държави на *Закон за опазването и развитието на съответния ЕЕ* срещу неограниченото, безсмислено и отвратително нахлуване на чужди думи и, което е по-лошо – на чужд стил в конструкциите на националния език. Франция е първата държава в света с

такъв „жесток“ закон. Според мен обаче е допуснат голям недостатък – медицината във Франция се изучава и пише на френски, а не на латински, както е в целия свят. За разлика от някои големи държави такъв Закон за борба с езиковата простащина и блюдолизничество е крайно необходим на много други по-малки с развити езици, но с малко на брой хора, които ги говорят. Той е особено нужен за запазването на националната идентификация. Напоследък в Полша се предлага драстичен Закон за полския език, който сигурно ще бъде приет. Това е похвално, но в него има и недостатъци – едва ли можем да се съгласим с това чуждите имена да се превеждат на полски, например немското фамилно име Шварц да стане Черен! Това е съвсем в стила на пуризма.

Сега нека повторим до какво трябва допълнително да доведе реализирането на Проекта:

1. Създаване на *робот за сканиране на книги*. **2.** Създаване на световен *стандарт за печатните шрифтове*. **3.** Създаване на *национални закони за депозиране в националните библиотеки на всички компютърни текстове*, които в бъдеще могат да имат национално значение. **4.** Създаване на *национални ЗАКОНИ ЗА ОПАЗВАНЕ НА ЕЗИКА*. **5.** Провеждане на *ОСНОВНА ЕЗИКОВА РЕФОРМА* на всеки език с цел възможното *фонетизиране на писмеността му, изменение на написването на думи, опростяване на пунктуацията и намаляване на изключенията му*.

Накрая нека отбележим, че след приключването му *Проектът трябва да се поддържа административно от съответните държавни институции* (например от Министерския съвет), националния и другите университети и, разбира се, от Института за националния език (ако такъв съществува).

Димитър Петров Шишков,
дом. адр.: ул. Струма 3, вх. Б, ап. 15, 1202 София,
дом. тел.: (+359 2) 83-36-16; сл. тел.: (+359 2) 46-51-91,
e-mail: dpsh@is-bg.com

COMPUTERISATION OF NATURAL LANGUAGES

Dimitar Petrov Shishkov

A project entitled “Computerisation of Natural Languages” is proposed and its main tasks are considered. Among them three most important and most urgent tasks for each country are noted: input and digitisation of national speech, scanning of all literature in the country, and creating the absolute computer dictionary of the respective natural language.