# Езикови ресурси и технологии за машинен превод

**Срок: 1 януари 2008 г. – 31 декември 2010 г.**

**Изпълнител от българска страна: Институт за български език “Проф. Любомир Андрейчин”, Българска академия на науките**

**Изпълнител от румънска страна: Институт за изкуствен интелект, Румънска академия на науките**

## 1. ЦЕЛ НА ПРОЕКТА

Проектът е насочен към създаването на паралелни езикови ресурси за български, гръцки, румънски, словенски и сръбски език. **Български език** се говори от около 7.5 милиона носители на езика в страната и чужбина[1]. **Сръбски език** се говори от около 17 милиона носители на езика в страната, бившите югославски републики и Америка. **Словенски език** се говори от около 2 милиона и половина души. Български, сръбски и словенски са южнославянски езици, които се характеризират с относително сродна лексика и граматична структура. **Гръцкият език** е самостоятелен дял в класификацията на индоевропейските езици и се говори от около 15 – 20 милиона души основно в Гърция и Кипър. **Румънският език** е романски език, който се говори приблизително от 28 милиона жители на Румъния и Молдова. Българският, гръцкият, сръбският и румънският език принадлежат към Балканския езиков съюз. Езиците в Балканския езиков съюз демонстрират общи черти по отношение на своята граматика, лексика и фонетичен състав. Приблизителният брой на носителите на съответните езици не е голям и вероятно това е една от причините за относително ограничения брой езикови ресурси, които съществуват за тях.

Основната цел на проекта е създаването на ***големи по обем паралелни корпуси*** за ***български, гръцки, румънски, словенски, сръбски, както и за*** английски, немски и френски, които ще се използват основно за автоматичен превод. За целите на проекта в паралелните корпуси е въведена различен тип езикова анотация: токънизация (автоматично определяне на единиците, които съставят текста - токъни (последователности от символи), изречения, тагиране (автоматично определяне на частта на речта и друга граматична информация за думите), лематизация (автоматично определяне на основната форма), съотнасяне на паралелните корпуси по изречения и думи.

Основният принос от проекта е създаването на **българо-английски, българо-гръцки, българо-румънски, българо-словенски** и **българо-сръбски** паралелни анотирани корпуси, които могат да послужат при разработването на ефективна система за ***машинен превод*** между съответните езици. Резултатите от проекта ще подпомогнат научните изследвания в областта на автоматичния превод и създаването на преводни модели за езици със сложна морфология (като южнославянските) и относително свободен словоред (каквито са балканските и южнославянските).

---

[1] Данните са от интернет.

Съществуват много области на комуникация, в които машинният превод вече се използва: например достъп до многоезикови бази от данни и информация, създаване на системи за търсене, извличане на информация и превод на документи, чуждоезиково обучение – както при традиционните форми, така и при новите форми на дистанционно или електронно обучение, в комуникациите: за превод на електронни съобщения или други документи, където бързото предаване на информация е от съществено значение, при опериране със съдържание на документи за автоматично определяне на основната тема на текста, в локализирането на описанието на продуктите за нуждите на националните или регионалните пазари чрез създаване на прилежаща документация, в професионалния превод, чрез използване на технологиите за преводна памет в системи, подпомагащи преводачите, с цел да се подобри и ускори тяхната работа и да се автоматизира основната част от преводния процес.

## 2. ОСНОВНИ РЕЗУЛТАТИ ОТ ИЗПЪЛНЕНИЕТО НА ПРОЕКТА

Електронните езикови ресурси (както и начините на описание на езиковите данни) за компютърна обработка на естествения език се различават коренно от традиционно познатите методи на работа в лингвистиката. За да могат да се използват за различни компютърни приложения, данните в електронните езикови ресурси трябва да са представени максимално пълно и непротиворечиво, а връзките между единиците, които ги съставят, да са експлицитно изразени.

Терминът езикови ресурси реферира към разнообразно множество от електронни данни, които включват както писмената, така и речевата форма на езика. В зависимост от структурата си езиковите ресурси най-общо могат да бъдат разделени на корпуси, речници (включително терминологични бази от данни, тезауруси и онтологии), лексикално-семантични мрежи, граматики и езикови модели.

Една от дефинициите за корпус е „*съвкупност от автентични езикови данни, които могат да се използват за лингвистични изследвания*" (Лиич 1997:1). Подобно е и определението за корпус като „*колекция от езикови примери, които са избрани и подредени според експлицитни лингвистични критерии, за да се използват като модел на езика*" (ЕГСКОЕ 1996a:4). С развитието на компютърната лингвистика областите на приложение на корпусната лингвистика се увеличават, затова по-скоро е приложима дефиницията: „*съвкупност от езиков материал в електронна форма, подходящ за компютърна обработка с приложение в лингвистичните изследвания или езиковите*

*технологии*" (Лиич 1997:1). Или „*компютърен корпус е корпус, кодиран по стандартизиран и хомогенен начин, даващ възможност за извличане на информация от различен характер*" (ЕГСКОЕ 1996a:5). Като компилация на многобройните дефиниции за корпус може да се предложи следното определение:

***Корпус е голяма колекция от езикови примери, представени по начин, даващ възможност за компютърна обработка, и избрани по определени (лингвистични) критерии, така че да представляват адекватен езиков модел.***

Корпусите може да съдържат текстове само от един език (или форма на съществуване на езика) или повече от един език. Това са съответно едноезичните и многоезичните корпуси. Многоезичните корпуси могат да се поделят на преводни (състоят се от преводни еквиваленти на даден оригинал или оригинали), паралелни корпуси (състоят се от преводни еквиваленти на даден оригинал или оригинали, които са съотнесени помежду си дума по дума и/или изречение по изречение – например многоезичният паралелен корпус от документи на европейския парламент JRC-ACQUIS) и съотносими корпуси (съвкупност от сходни по тематика текстове на повече от един език) – например преводът на новините в Българското национално радио Христо Ботев.

Съществуващите паралелни корпуси за български език в повечето случаи включват друг славянски, балкански или широко разпространен европейски език. Например в рамките на проекта Multext-East е създаден паралелен корпус за шест езика (български, естонски, румънски словенски, унгарски и чешки) върху основата на романа на Оруел *1984.* Корпусът е тагиран и съотнесен по изречения (Димитрова и др. 1998). Резултатите на друг проект представляват двуезичен корпус от гръцки и български текстове (Гоули и др. 2009). Корпусът се състои от около 700 000 токъна, от които 550 000 принадлежат към текстове от художествената литература, а 150 000 илюстрират фолклорни текстове или легенди. Съществуват и други проекти, насочени към създаването на паралелни корпуси, в които участва и български: RuN (Грьон и Мариянович 2010) – паралелен корпус на руски и норвежки документи, който в последно време се разширява с други европейски езици, включително български; българо-полско-литовски корпус (Димитрова и др. 2009); ParaSol (Валденфелс 2006), известен като Редесбургски паралелен корпус, включващ оригинална и преводна художествена литература както на славянски, така и на други европейски езици (включително български).

Изводът, който може да се направи, е, че паралелните корпуси, в които се включва български, не са особено големи по размери, илюстрират предимно художествена литература

(или административни документи) и са компилирани на базата на съществуващи паралелни документи в интернет, а не на планирана стратегия за балансираност и представителност на корпуса.

В рамките на проекта са създадени ***големи по обем паралелни корпуси от български, гръцки, румънски, словенски, сръбски и английски (френски, немски) текстове***. Корпусите са анотирани с информация за единиците от корпуса, частите на речта и основните форми на думите в корпуса, текстовете на различни езици са съотнесени по изречения (и думи). Върху създадените паралелни корпуси са проведени различни експерименти със съществуващи и новосъздадени методологии за машинен превод.

## 2.1. СЪЗДАВАНЕ НА ПАРАЛЕЛЕН КОРПУС ОТ БЪЛГАРСКИ, ГРЪЦКИ, РУМЪНСКИ, СЛОВЕНСКИ, СРЪБСКИ И АНГЛИЙСКИ (ФРЕНСКИ, НЕМСКИ, ЧЕШКИ) ТЕКСТОВЕ

Паралелните корпуси на два или повече езика са изключително полезен езиков ресурс особено при вероятностните подходи за обработка на езиковите данни, които се използват при задачи като извличане на информация, категоризация на документи, автоматичен превод и др. В момента съществуват паралелни корпуси на част от европейските езици, които основно се състоят от административни[2] или публицистични текстове. Тъй като компютърната обработка на езика не се ограничава до определен тип тематични области, задачата е да се създадат паралелни корпуси между български и основните европейски, южнославянски и балкански езици, които да са структурирани, така че пропорционално да съдържат различни типове документи.

В рамките на проекта са създадени два паралелни корпуса: един, съдържащ романа на Жул Верн „80 дни около света" на английски, български, сръбски и френски (тъй като за сръбски не са достъпни преводи на документите на Европейския съюз), наречен **Паралелен корпус от художествени текстове** (ПКХТ) – Фигура 1, и втори корпус, състоящ се от документи на Европейския съюз, избрани по определени критерии, на английски, български, гръцки, немски, румънски, словенски, френски и чешки, наречен **Паралелен корпус от административни текстове** (ПКАТ).

[2] http://langtech.jrc.it/JRC-Acquis.html

| | |
|---|---|
| **n1 :** En l'année 1872, la maison portant le numéro 7 de Saville-row, Burlington Gardens -- maison dans laquelle Sheridan mourut en 1814 --, était habitée par Phileas Fogg, esq., l'un des membres les plus singuliers et les plus remarqués du Reform-Club de Londres, bien qu'il semblât prendre à tâche de ne rien faire qui pût attirer l'attention. | **n1 :** През 1872 година в къщата на “Савил роу” № 7, Бърлингтън Гардънс – същата, в която през 1814 година почина Шеридан, – сега живееше Филиас Фог. Той беше един от най-странните и видни членове на Реформаторския лондонски клуб, въпреки че сякаш се стараеше да не привлича с нищо вниманието на другите. |
| **n2 :** A l'un des plus grands orateurs qui honorent l'Angleterre, succédait donc ce Phileas Fogg, personnage énigmatique, dont on ne savait rien, sinon que c'était un fort galant homme et l'un des plus beaux gentlemen de la haute société anglaise. | **n2 :** И така, един от най-големите оратори, които възхваляват Англия, бе следван от същия този Филиас Фог – загадъчна личност, за която не се знаеше нищо, освен че е изключително галантен мъж и един от най-красивите джентълмени от висшето английско общество. |
| **n3 :** On disait qu'il ressemblait à Byron -- par la tête, car il était irréprochable quant aux pieds --, mais un Byron à moustaches et à favoris, un Byron impassible, qui aurait vécu mille ans sans vieillir. | **n3 :** Говореше се, че прилича на Байрон, най-вече в лицето, защото краката му бяха безупречни. Приличаше обаче на един Байрон с мустаци и бакенбарди, на един невъзмутим Байрон, който и за хиляда години не би остарял. |
| **n4 :** Anglais, à coup sûr, Phileas Fogg n'était peut-être pas Londonner. | **n4 :** Той със сигурност беше англичанин, но вероятно не бе лондончанин. |
| **n5 :** On ne l'avait jamais vu ni à la Bourse, ni à la Banque, ni dans aucun des comptoirs de la Cité. | **n5 :** Никога не го бяха виждали нито на Борсата, нито в банката, нито дори на някой търг в Сити. |
| **n6 :** Ni les bassins ni les docks de Londres n'avaient jamais reçu un navire ayant pour armateur Phileas Fogg. | **n6 :** Нито пристанищата, нито доковете на Лондон бяха приютявали някога кораб, чийто собственик да е Филиас Фог. |
| **n7 :** Ce gentleman ne figurait dans aucun comité d'administration. | **n7 :** Този джентълмен не фигурираше в нито един управителен съвет. |

*Фигура 1.* ***Паралелен корпус от художествени текстове (френски и български)***

Текстовете са разделени коректно на изречения, преводните еквиваленти между различните езици в корпуса са съотнесени по изречения. Освен първоначално предвидените балкански и южнославянски езици – български, гръцки, румънски, сръбски и словенски, в корпуса са включени големите европейски езици – английски, немски и френски, както и чешки, тъй като за чешки е разработена лексикално-семантична мрежа (wordnet), каквато има и за останалите езици.

Новелата на Жул Верн е избрана по две основни причини: текстът е достъпен в електронна форма за повечето от европейските езици и е подходящ литературен текст за различни типове анализ, включително машинен превод. До момента са събрани текстове на 15 европейски езика: френски, английски, немски, испански, португалски, италиански, румънски, руски, сръбски, хърватски, български, македонски, полски, унгарски и гръцки. Таблица 1 илюстрира дължината на неанотираните текстове, изразена в общия брой думи, броя на уникалните форми и съотношението между тях.

| | Общо | | съотношение думи/форми |
|---|---|---|---|
| Език | думи | уникални форми | |
| френски (FR) | 71 793 | 9 433 | (7.6) |
| английски (EN) | 63 743 | 7 434 | (8.8) |

| | | | |
|---|---|---|---|
| испански (ES) | 65 064 | 9 959 | (6.5) |
| португалски (PG) | 65 037 | 10 271 | (6.3) |
| немски (DE) | 62 726 | 10 228 | (6.1) |
| италиански (IT) | 68 450 | 11 599 | (5.9) |
| гръцки (EL) | 68 615 | 11 809 | (5.8) |
| български (BG) | 58 678 | 11 217 | (5.2) |
| сръбски (SR) | 58 722 | 12 733 | (4.6) |
| руски (RU) | 56 293 | 14 708 | (3.8) |
| полски (PL) | 54 871 | 15 406 | (3.6) |

*Таблица 1* ***Разпределение на текстовите единици при различните преводни еквиваленти в ПКХТ***

Вижда се, че едно и също съдържание е изразено чрез текстове с различна дължина на различните езици: например полският текст е почти 25% по-кратък от френския оригинал, докато броят на уникалните думи е почти 60% по-голям, отколкото във френски. Този резултат се дължи частично на разликата в граматичната структура (отсъствие на определителен и неопределителен член, имплицитен подлог, изпускане на спомагателните глаголи в някои времена).

Acquis Communautaire представлява Европейското право, преведено за всички страни членки на Европейския съюз. Колекцията от законодателни документи се обогатява непрекъснато и в момента съдържа избрани текстове, създадени от 1950 година досега и преведени на езиците на държавите, членки на Европейския съюз. Следователно Acquis Communautaire се състои от паралелни текстове на следните 22 езика: български, чешки, датски, немски, гръцки, английски, испански, естонски, финландски, френски, унгарски, италиански, литовски, латвийски, малтийски, холандски, полски, португалски, румънски, словашки, словенски и шведски. Значителна част от тези текстове са компилирани от Групата за езикови технологии към Европейската комисия в паралелен корпус, наречен JRC-Acquis[3] и публикуван през 2006 г.

Този уникален езиков ресурс е сред малкото налични паралелни корпуси, които съдържат интересуващите ни езици: български, чешки, френски, гръцки, немски, румънски, словенски и английски. От същинския JRC-Acquis, в който се използват идентични идентификатори за документите (CelexNumbers), съдържащи и код за съответния език, са селектирани документите, преведени на всеки един от изброените осем езика. Резултатът е

[3] http://langtech.jrc.it/JRC-Acquis.html

1204 файла за език. Множество от осем изречения, които представят преводните еквиваленти, е структурирано като преводна единица в XML формат – Фигура 2. Големината на паралелния корпус в брой текстови единици (токъни) е илюстрирана в Таблица 2.

| език | токъни | Съотношение токъни/изречения |
|---|---|---|
| български BG | 1436925 | 23.7944824388548 |
| чешки CS | 1238981 | 20.5166669426551 |
| немски DE | 1314441 | 21.7662322608422 |
| гръцки EL | 1469642 | 24.3362532911623 |
| английски EN | 1466912 | 24.2910463826194 |
| френски FR | 1527241 | 25.2900528241898 |
| румънски RO | 1422995 | 23.5638112901356 |
| словенски SL | 1271011 | 21.0470615509447 |

*Таблица 2:* ***Статистически данни за компилирания паралелен корпус ПКАТ***

<tu id="3936">
<seg lang="bg">
<s id="31985L0337.n.83.1">
Резултатите от консултациите и информацията, събрана съгласно членове 5, 6 и 7, трябва да се вземат предвид при процедурата по издаването на разрешението.</s></seg>
<seg lang="cs">
<s id="31985L0337.n.83.1">
Informace shromážděné podle článků 5, 6 a 7 musí být brány v úvahu v povolovacím řízení.</s> </seg>
<seg lang="de">
<s id="31985L0337.n.85.1">
Die gemäß den Artikeln 5, 6 und 7 eingeholten Angaben sind im Rahmen des
Genehmigungsverfahrens zu berücksichtigen.</s></seg>
<seg lang="el">
<s id="31985L0337.n.85.1">
Οι πληροφορίες που συγκεντρώνονται δυνάμει των άρθρων 5, 6 και 7 πρέπει να λαμβάνονται υπόψη στα πλαίσια της διαδικασίας για τη χορήγηση αδείας.</s></seg>
<seg lang="en">
<s id="31985L0337.n.84.1">

Information gathered pursuant to Articles 5, 6 and 7 must be taken into consideration in the development consent procedure.</s></seg>

<seg lang="fr">

<s id="31985L0337.n.83.1">

Les informations recueillies conformément aux articles 5, 6 et 7 doivent être prises en considération dans le cadre de la procédure d'autorisation.</s></seg>

<seg lang="ro">

<s id="31985L0337.n.83.1">

Informaţiile culese conform art. 5, 6 şi 7 trebuie să fie luate în considerare în cadrul procedurii de autorizare.</s></seg>

<seg lang="sl">

<s id="31985L0337.n.83.1">

Informacije , zbrane skladno s členi 5, 6 in 7, se morajo upoštevati v postopku za pridobitev soglasja za izvedbo.</s></seg>

</tu>

*Фигура 2:* ***Преводна единица от осемезичния паралелен корпус ПКАТ***

Общата големина на корпуса зависи от целите, за които е построен - например едно лексикографско изследване се нуждае от огромен корпус с разнообразни типове текстове, за да може да се регистрират значими колокации с ниска фреквентност. В нашия случай големината на корпуса също има значение, особено за статистическите подходи при автоматичния превод или при подходите, базирани на примери. Големината на паралелния корпус ПКАТ в случая е определена от документите на Европейския парламент, които до момента са преведени на всеки един от посочените езици. И двата паралелни корпуса са отворени за допълване с нови документи и езици.

## 2.2. АНОТИРАНЕ НА ПАРАЛЕЛНИТЕ ТЕКСТОВЕ

Ползата от паралелните корпуси нараства значително, ако са анотирани, т.е. ако съдържат експлицитно представена допълнителна информация, която може да послужи за различни цели. Анотирането на даден текст представлява експлицитно представяне на езиковата интерпретация на отделните единици от текста. За анотация се използва маркиращ език, който е набор от конвенции за означенията на езиковата интерпретация.

Анотирането може да се направи автоматично или ръчно, в случая анотацията е

автоматична, тъй като става въпрос за големи по обем текстове, базира се на съществуващи езикови ресурси и програми за автоматична обработка на текста. Нелингвистичната анотация е свързана с извънлингвистични текстови характеристики и служи за отбелязване на разнообразна информация, характеризираща текста, в съответствие с разработената международна конвенция в рамките на TEI. Тя предвижда библиографски справки за период, автор, авторов период – когато става въпрос за литературен текст. Лингвистичната анотация представлява съотнасяне с граматическа или морфо-синтактична информация. Могат да се отделят различни равнища на лингвистична анотация (Лиич 1997:8-15), например: морфологично, морфо-синтактично, синтактично, семантично и дискурсно (ЕГСКОЕ 1996б: 3), като анотираните корпуси обикновено се асоциират с повече от едно равнище.

На морфологично равнище всяка лексикална единица в текста може да се асоциира със своята основна форма – лема (лематизация). Анотацията на лемата за флективни езици като български, където преходен глагол от несвършен вид има петдесет и две синтетични форми, е важна част от компютърната обработка, за да може да се осъществява връзката при употребата на различните форми на дадена дума в различни контексти. Морфо-синтактичната анотация въвежда информация за граматичния клас на дадена лексикална единица и (възможно) съответните стойности на граматичните категории, характеризиращи единицата. Морфо-синтактичната анотация отстранява граматичната многозначност на омографите и се използва в голям брой приложения за компютърна обработка на езика: за определяне на коректната фонетична стойност на дадена форма, за съотнасяне на формите към правилната парадигма, за идентификация на съставните лексикални единици, за определяне на строежа на фразите, за съотнасяне с допустимо лексикално значение и т.н. Синтактичната анотация най-общо може да представя граматичните зависимости между конституентите на словосъчетание или фразовата структура, в която конституентите се комбинират помежду си. При семантичната анотация могат да се разграничат два типа (МакЕнери и Уилсън 2001:61-62) – представяне на семантичните релации между думите в текста (например анотация на семантичните роли на аргументите, на валентността на лексикалните единици, на реализацията на семантичните фреймове) или на семантичните свойства на думите (например анотация на значението на думите). Представянето на семантичните свойства на думите може да асоциира дадена дума с нейното значение, но може да има и по-комплексен подход, при който думите се съотнасят със „семантичното поле", към което принадлежат (Уилсън и Томас 1997:54) – множество от думи с общи семантични свойства, например непосредствените хипоними. Дискурсната анотация може да

включва определяне на референцията на местоименията за отстраняване на многозначност при анафора, маркиране на изрази за изразяване на определена конвенция – извинение, поздравление и др.

Както се вижда, равнищата на анотация съответстват най-общо на равнищата на многозначност, от което следва, че една от основните области на приложение на анотираните корпуси е отстраняването на многозначността. Разбира се, анотацията е необходима и при еднозначните структури за съотнасянето им с коректните лингвистични стойности, но при многозначните структури анотацията показва не само коректните лингвистични стойности, но и избора им от множеството възможни.

### 2.2.1. АВТОМАТИЧНО ОПРЕДЕЛЯНЕ НА ЕДИНИЦИТЕ ОТ КОРПУСА (ТОКЪНИЗАЦИЯ)

Разделянето на корпуса на единиците, които го съставят, е най-ниското ниво на обработка на паралелния корпус: текстовете се разделят на единици (думи, изречения и параграфи), които се характеризират в зависимост от съставните си части. Най-общо токън се определя като поредица от символи между два отделящи символа (интервал, пунктуационен знак). Българският токънизатор разпознава последователности от букви, цифри, пунктуационни знаци, специални символи, комбинации от тях и празни символи (Фигура 3). Също така се разпознават някои изрази, които означават дати, дроби, е-мейли, интернет адреси, съкращения и др.

Интересът TOK_FUCA
към TOK_LCA
биографичните TOK_LCA
очевидности TOK_LCA
и TOK_LCA
превръщането TOK_LCA
им TOK_LCA
в TOK_LCA
текст TOK_LCA

*Фигура 3.* ***Предварителна обработка на корпусите***

Токънизаторът е базиран на регулярни правила, които дефинират начина, по който входният текст се разделя на токъни. Регулярните правила са външни за токънизатора, което

позволява лесна настройка и конфигурация спрямо специфични типове входен текст. Тези правила дефинират начина, по който входният текст се разделя на токъни (Фигура 4).

```
^([а-я][а-я\-]*[а-я\-])$
TOK_LCDA
^([А-Я][А-Я\-]*[А-Я\-])$
TOK_UCDA
^([А-Я][а-я\-]*[а-я\-])$
TOK_FUCDA
^([А-Яа-я][А-Яа-я\-]*[А-Яа-я\-])$
TOK_MCDA
```

Фигура 4. Пример за регулярни правила за определяне на единиците ан текста

Използва се и списък с регулярни шаблони, които дефинират типа на всеки токън, например състоящ се само от главни кирилски букви, от главни и малки латински букви и т.н. Типовете токъни се характеризират в зависимост от това дали се състоят от главни и малки кирилски или латински букви, цифри, пунктуация и някои специални символи.

Резултатът от работата на токънизатора се представя в два формата:

- вертикален, в който всеки токън се намира на нов ред, като информацията за типа на токъна е разделена с табулатор (Фигура 3);
- маркиращ формат, който прави асоциация с позицията и дължината на токъна във входния текст, както и с неговия тип. Това позволява входният текст да остане непроменен, както и лесна интеграция на програмата в различни системи за анотация.

Към това предварително равнище на обработка на текста принадлежи и автоматичното определяне на границите на изреченията. Определянето на границите на изреченията също е базирано на регулярни правила (Фигура 5), но се използват и списъци от лексикални единици - лексикони (например за изброяване на съкращения, след които може да има / трябва да има главна буква / цифра и т.н.). Регулярните правила и лексикона също са външни за програмата за разделяне на изречения, което позволява лесното й настройване и конфигуриране за конкретни типове входен текст.

```
$A =~ s/(\.|\!|\?|\;|\,|\/|\"|\“|\”|\„|\'\'|\`|\’|\'|\<|\>|\+|=|№|\*|\(|\)|\[|\]|\{|\}|\^|\||\&)([А-Яа-я]|[A-Za-z]|[0-9])/$1 $2/g;
```

Фигура 5. Регулярно правило за определяне на граница на изречения

Лексиконите са представени във формата на минимален ацикличен краен автомат, което позволява ефективно търсене в тях. Аналогично с токънизатора, програмата за разделяне на текста на изречения представя резултата в два формата:

- при единия специален символ, например <S>, се поставя след края на всяко изречение от входния текст;
- при другия границите на изреченията се представят като позиция и дължина във входния текст, което позволява входният текст да остане непроменен, както и лесна интеграция в различни системи за анотация.

Разпознаването и маркирането на единиците в текста е необходима предпоставка за повечето задачи, свързани с обработката на естествения език. Идентифицирането на границите на думите и на изреченията в много случаи включва отстраняване на многозначността при употребата на пунктуацията, т.е. кога даден знак означава край на изречение и кога – не.

### 2.2.2. **АВТОМАТИЧНО ОПРЕДЕЛЯНЕ НА ЧАСТИТЕ НА РЕЧТА В КОРПУСА (ТАГИРАНЕ)**

Много от интересните проблеми в областта на компютърната лингвистика, както и много от най-важните приложения при обработката на естествения език изискват автоматична система за правилно асоцииране на думите с подходящи граматични категории и техните стойности (прието е такава система да се нарича тагер). Най-общо казано тагирането (анализирането на думите по части на речта и съответните стойности на граматичните категории) включва въвеждането на многозначни граматични характеристики и отстраняването на граматичната многозначност.

Известно е, че една дума може да се използва с различно граматично значение в различните контексти. Например думата '*коси*' ще получи следното описание при морфологичен анализ (Коева 1998):

*коси – съществително, женски род, множествено число, неопределено*

*коси – глагол, сегашно време, трето лице, единствено число*

*коси - глагол, минало свършено време, второ лице, единствено число*

*коси - глагол, минало свършено време, трето лице, единствено число*

*коси - глагол, повелително наклонение, второ лице, единствено число*

Следователно някои думи при морфологичния анализ са разпознати с повече от една граматична интерпретация. Проблемът се състои в това, да се формулират тези зависимости

на непосредствения контекст, които позволяват думата да се разпознава с правилните си граматични характеристики. Автоматичното маркиране към коя част на речта принадлежи всяка дума е важна стъпка в много приложения на обработката на естествен език.

За тагирането на българските текстове се използва статистически тагер, базиран на SVM (Support Vector Machine). Използва се SVMTool[4] (Гименес и Маркес 2004), проект с отворен код за трениране на модели за тагери и прилагането им върху входен текст. SVMTool се характеризира с простота (лесно се конфигурира и тренира); гъвкавост (параметрите, с които работи, могат да се настройват, както и да се дефинират сложни параметри, включително н-грами по част на речта или класове на многозначност, също така може да се използва анализ на равнище изречение); преносимост (програмата е езиково независима); точност (в сравнение с други известни тагери има конкурираща се точност); ефикасност (времето за тагиране зависи от параметрите, които са избрани при схемата за тагиране - при едностепенно тагиране от ляво на дясно прототипът на Пърл показва скорост на тагиране от 1500 думи на секунда, а версията на C++ - над 10000 думи на секунда).

SVMTool се нуждае от корпус за трениране - в случая това е ръчно анотирания за част на речта корпус (Коева и др. 2006). Ръчно анотираният корпус за български език е конструиран от Българския Браун корпус и е с големина над 200 000 думи. От всеки един от 500-те текста в Българския Браун корпус е направена извадка от минимум 150+ думи, като извадките са разширени до край на изречение – по тази причина повечето от извадките съдържат повече от 150 думи. Преди анотацията таговете съдържат всички граматични значения, които могат да се препишат на дадена графична единица. Като резултат от автоматичната анотация 51,9% от токъните получават едно граматично значение, 46,7% - повече от едно и 1,4% - нито едно. Нетагираните думи обикновено са редки, чужди или собствени имена, които не са включени в речника. Степента на многозначност (броят на различните граматични значения) варира при различните типове токъни. Най-общо най-висока е при пунктуацията и затворените класове думи, които имат множество граматични значения за разлика от отворените класове, например запетаята има 25 различни значения според това каква функция има в изречението. Таговете имат стандартен формат, съдържащ лема, граматични характеристики на лемата, граматични характеристики на формата (ако има форми). След автоматичната токънизация и анотация следва отстраняване на граматичната многозначност, което се извършва от експерти. В резултат е получен анотиран корпус от 217 210 единици, съдържащ 172 482 думи, 42 058 пунктуационни знака и 2 670 цифри.

---

[4] http://www.lsi.upc.es/~nlp/SVMTool/

Анотираният корпус е разделен на три части - за трениране, за проверка и за тестване, което позволява трениране, настройка и проверка на резултатите.

Изборът на стратегия за трениране е от особена важност (от ляво на дясно, от дясно на ляво или и двете, в случая от ляво на дясно, макар че се твърди, че комбинацията на посоките дава най-добри резултати); дължината на контекста - колкото е по-голям, толкова са по-добри резултатите, в случая прозорец от седем елемента, дефиницията на параметрите (н-грами на думи или тагове или класове на многозначност, лексикализирани параметри като префикси, суфикси, главни букви и др.). При тагирането се използва специално редуциран тагсет.

Някои от задачите, в които тагерът е много полезен, са машинният превод, извличането на информация от текст, както и много други. В машинния превод една дума от даден език често може да се преведе по повече от един начин и това може да зависи от частта на речта. Например английската дума “fly” може да се преведе като “летя” и като “муха”. Ако се знае, че “fly” е съществително, тогава лесно може да се избере коректния превод. Разбира се това е много елементарен пример, но той ясно отразява значението на частите на речта в този вид приложения, при които частичният анализ има определяща роля за успеха на програмата за извличане на информация.

### 2.2.3. АВТОМАТИЧНО ОПРЕДЕЛЯНЕ НА ОСНОВНИТЕ ФОРМИ В КОРПУСА (ЛЕМАТИЗИРАНЕ)

Лематизацията е тясно свързана с тагирането по части на речта и включва приписването на лема, т.е. на основната форма при изменяемите думи, към всяка дума в текста след извършване на морфосинтактичния анализ, както и съответните граматични характеристики, с които се характеризира употребената форма на думата. Тоест да се разпознава, че *четеш* е глагол във второ лице, единствено число, сегашно време от глагола *чета*; *четат* - трето лице, множествено число, от глагола *чета* и т.н.

Лематизаторът е базиран на електронния Граматичен речник на българския език (Коева 1998), който в момента съдържа около 85 000 думи от основния речников фонд на българския книжовен език. Граматичният речник позволява автоматично генериране, анализиране и синтезиране на словоформите, чийто брой е около 1 140 000. Това означава, че речникът осигурява възможността да се построи парадигмата (всички форми) на произволна дума, включена в него, да се разпознае определена форма като част от парадигмата на съответната дума и да се припишат граматичните характеристики. Наред с това речникът

предоставя възможността да се решават други лингвистични задачи от различен тип – с изследователски или приложен характер, една от които е създаването на системата за корекция на правописа.

Структурата на речника е изградена с помощта на крайни преобразуватели (finate state transdusers), които имат широко приложение при създаването на съвременните електронни речници **(**Михов, 2000). Най-общо казано, крайният преобразувател е устройство, което разпознава последователности от символи и ги асоциира с други последователности от символи. Речниковите единици, представени в Граматичния речник, се състоят от основна форма, която се свързва с определена лингвистична информация, записана формално, например С+М означава съществително, мъжки род. Формалният запис на граматичната информация заедно с индекса, който го прави уникален, представлява името на крайния преобразувател, който разпознава всички кореспондиращи форми на думата:

*`а, МЕЖ, 0*

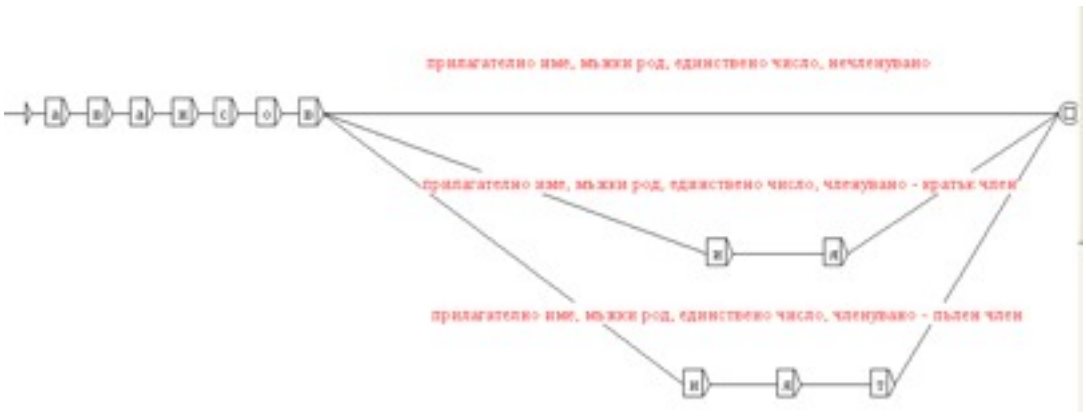
*абадж`ия, С+М+Ж, 2*

*абаж`ур, С+М, 11*

*абан`ос, С+М, 10*

*аб`ат, С+М, 6*

Всички думи в езика, които имат еднакви граматични характеристики и еднакво множество от окончания и редувания, се асоциират с един и същ краен преобразувател. Например на съществителните от мъжки род, които имат същите форми като *абаджия (чорбаджия, тютюнджия, ...),* е приписан крайният преобразувател С+М+Ж, 2, за разлика от други съществителни като *абат, дипломат,* с чието формообразуване се свързва крайният преобразувател С+М, 6. Крайният преобразувател може да се представи като граф, който илюстрира преходите от началното до крайното състояние – Фигура 6.




Фигура 6. Краен преобразувател от речника

Освен като списък от основни форми и прилежащите им крайни преобразуватели Граматичният речник съществува и в така наречения DELAF формат (Зилберщайн, 1993; 1999; Коева, 2001), който представлява списък от всички словоформи, съответната им основна форма и граматични характеристики.

*авансов, авансов. прилагателно име, мъжки род, единствено число, нечленуващо*

*авансовия, авансов. прилагателно име, мъжки род, единствено число, членувано – кратък член*

*авансовият, авансов. прилагателно име, мъжки род, единствено число, членувано – пълен член*

Речникът е представен под формата на минимален ацикличен краен автомат с етикети в крайните състояния. Етикетите показват морфологичната информация, която се асоциира с дадена дума.

Основната форма се определя, като се съотнася информацията, получена от тагера - част на речта и редуциран тагсет, с информацията в Граматичния речник за част на речта и пълен тагсет. На думите, които не се срещат в речника, се преписва редуцираният тагсет, получен от тагера. Резултатът се илюстрира от Фигура 7.

| | | |
|---|---|---|
| В | в | R--- |
| различни | различен | A---:p0 |
| части | част | NCF-:p0 |
| на | на | R--- |
| тази | този | PDB-:sf |
| книга | книга | NCF-:s0 |
| се | се | T--- |
| наложи | наложа | V2PT:R3s:E2s:E3s |
| текстовете | текст | NCM-:pd |
| на | на | R--- |
| автора | автор | NCM-:sh |
| и | и | JC-- |
| на | на | R--- |
| редактора | редактор | NCM-:sh |
| да | да | JE-- |
| бъдат | бъда | V2II:R3p |
| отбелязани | отбележа | V2PT:Qp0 |

по по R---

различен различен A---:sm0

начин начин NCM-:s0

Фигура 7. Лематизация на българските текстове

## 3. АВТОМАТИЧНО СЪОТНАСЯНЕ НА ТЕКСТОВЕТЕ ПО ИЗРЕЧЕНИЯ (АЛАЙНИРАНЕ)

За да бъде полезен един паралелен корпус, той трябва да бъде обработен със специална програма за съотнасяне по изречения и думи. Под съотнасяне (alignment) се разбира процесът на свързване на двойки от думи, фрази, термини или изречения в текстове от различни езици, които са преводни еквиваленти. Макар че съществуват паралелни корпуси, съотнесени на ръка, необходимо е автоматично съотнасяне на паралелни корпуси с висока степен на точност, тъй като се обработват големи по обем текстове. На нивото на изречението се отбелязва кое изречение S2 от текст T2 на езика L2 е преводният еквивалент на изречение S1 от текст T1 на езика L1. На нивото на думата се се изчислява вероятността кои възможни двойки от думи w1 от S1 и w2 от S2 са преводни еквиваленти.

Съотнасянето по изречения на ПКХТ е направено със системата Xalign[5]. От TEI формата, получен по този начин, са компилирани няколко версии на корпуса, съответно в TMX (Фигура 8), Vanilla и HTML формати. Съотнасянето е на равнище параграф и изречение и е установена кореспонденция 1:1 между оригиналния текст и преводния еквивалент, като е направена допълнителна проверка за коректността на автоматичното съотнасяне (Таблица 3 илюстрира количествено съотнасянето на текстовете в ПКХТ).

| | български | английски | френски | сръбски |
|---|---|---|---|---|
| Думи | 58 162 | 64 831 | 68 359 | 60 227 |
| Изречения | 4 435 | 4 435 | 4 435 | 4 435 |
| Параграфи | 1 963 | 1960 | 1963 | 1963 |

Таблица 3. Количествени характеристики на съотнасянето на паралелните текстове в ПКХТ

<tu> <tuv xml:lang="BG" creationid="n506 " creationdate="20070801T123334Z">

<seg>Сигурно добре знае, че в Индия, която е английска земя, няма да е в безопасност. </seg>

</tuv>

<tuv xml:lang="SR" creationid="n506 " creationdate="20070801T123334Z">

[5] led.loria.fr/outils.php

<seg>On dobro zna da neće biti siguran u  Indiji jer je to engleska zemlja.  </seg>

</tuv>

<tuv xml:lang="FR" creationid="n506 " creationdate="20070801T123334Z">

<seg>Il doit bien savoir qu'il ne serait pas en sûreté dans l'Inde, qui est une terre anglaise. </seg>

</tuv>

<tuv xml:lang="EN" creationid="n506 " creationdate="20070801T123334Z">

<seg>He ought to know that he would not be safe an hour in India, which is English soil.</seg>

</tuv>

Фигура 8. Съотнасяне на ПКХТ в TMX формат

Тъй като са забелязани някои грешки при съотнасянето по изречения в оригиналния корпус JRC-Acquis, 1204 файла за български, чешки, френски, гръцки, немски, румънски, сръбски, словенски и английски бяха съотнесени по изречения отново с кореспондиращите файлове на английски. От XX-EN съотнесени изречения са запазени само двойките, съотнесени 1-1, така е получен XML файл, съдържащ 60389 изречения, всяко преведено на девет езика. За съотнасянето на документите по думи е използван COWAL, за да се получи EN-XX съотнасяне по думи. От EN-X1 и EN-X2 автоматично са деривирани съотнасянията X1-X2.

## 3. ЕКСПЕРИМЕНТИ ВЪРХУ РАЗЛИЧНИТЕ МЕТОДОЛОГИИ ЗА МАШИНЕН ПРЕВОД

Съществуват редица методи за превод по аналогия и статистически методи за машинен превод, които могат да се тестват и оценят, като се използват паралелни корпуси. Използването на статистически методи позволява анализ на големи по обем паралелни корпуси и автоматично конструиране на системи за машинен превод. Експериментите показват, че показателите на съществуващите системи за машинен превод зависят в значителна степен както от изходния език, така и от паралелните корпуси, които се използват за тяхното създаване. Методът за машинен превод по аналогия (*example-based machine translation* approach) се характеризира с използването на двуезичен корпус като основна база от данни. При този подход могат да се използват различни семантични ресурси като семантични мрежи или терминологични бази от данни за постигането на по-добри резултати.

Статистическите подходи в машинния превод генерират превод на основата на двуезични корпуси. Когато такива корпуси са налице, могат да бъдат постигнати съществени резултати особено при текстове от определени тематични области.

По традиция изработването на система за сравнително коректен машинен превод се свързва с голям разход на време и усилия. Изследванията в последните години са насочени към иновативни технологии и подходи, целящи бързото имплементиране на подобна система с максимална точност. При повечето съвременни подходи в машинния превод се извлича информация от двуезичен корпус с помощта на статистически вероятностни модели или посредством метода по аналогия. Тези два подхода са фокусирани върху последователността на думите и техния превод. Това, което отличава двата подхода, е, че не са необходими синтактични или семантични правила за текстовия анализ или за избора на лексикални еквиваленти.

Наблюдава се безпрецедентно нарастване на използването на автоматичния превод (предимно на и от английски) в много области на общуването. Автоматичният превод, макар и все още да не може да замени човешкия, безспорно е в услуга на междуезиковите комуникации в областта на бизнес отношенията, образованието, културните връзки и научните изследвания. Затова един от приоритетите в съвременната интернет доминирана мултиезична епоха е развитието на системи за автоматичен превод, които са едновременно езиково прецизни и ефективни.

Машинният превод, базиран на правила, включва анализ и синтез на изречения и словосъчетания, базиращи се на подробна морфологична, синтактична и семантична информация. Статистическите методи са най-широко използваната парадигма в машинния превод през последните години. Предимствата на статистическите подходи в машинния превод над останалите известни методи са: (1) статистическите методи не са свързани с конкретни двойки езици; (2) методите, базирани на правила, изискват формулирането им, което е трудоемко и обикновено не може да се прилага при други езици. Машинният превод по аналогия (*example-based machine translation* approach), разновидност на вероятностните модели, се характеризира с използването на големи по обем паралелни корпуси, в които се измерва близостта между даден езиков фрагмент и множество от примери.

Направени са експериментални модели от румънския партньор за превод между български и английски, в който се сравняват известни стратегии и/или се прилагат нови, разработени в рамките на проекта (Туфиш и др. 2008). Експерименталният модел се базира на големи по обем паралелни корпуси на български и английски, в които е направена

предварителна лингвистична обработка: текстовете са разделени и съотнесени по думи и изречения и всяка дума е получила еднозначна граматична характеристика. За нуждите на автоматичния превод се използват и големи двуезични флективни речници на български и английски език. След съотнасянето на дума | израз към дума | израз, множеството от комбинации се проверява в корпуса и се предлагат най-вероятните резултати в зависимост от направените измервания за свързаност и близост.

## 4. ЗАКЛЮЧЕНИЕ

Многоезиковите паралелни корпуси са много богати на информация и показват как дадена езикова система си взаимодейства с друга езикова система при превод. Корпусите са маркирани с различен тип лингвистична анотация, която е въведена чрез автоматично определяне на границите и вида на единиците от текста, автоматично определяне на частите на речта и основните форми и паралелно съотнасяне на многоезиковите корпуси по думи и изречения. Върху анотираните паралелни корпуси са приложени някои от добре известните стратегии за машинен превод (статистически методи, методи, базирани на шаблони или примери, и т.н.).

## ЦИТИРАНА ЛИТЕРАТУРА


Валденфелс 2006: Waldenfels, R. Compiling a parallel corpus of slavic languages. Text strategies, tools and the question of lemmatization in alignment. in: Brehmer, B., Zdanova, V., Zimny, R. (Hrsg.); *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV*) 9. München, pp. 123--138.

Гименес и Маркес 2004: Gim´enez, Jes´us and Llu´ıs M`arquez, 2004. Svmtool: A general pos tagger generator based on support vector machines. In Proceedings of the 4th LREC Conference.

Димитрова и др. 1998: Dimitrova L., Ide, N., Petkevic, V., Erjavec, T., Kaalep, H.J., Tufiş, D. (1998) In Christian Boitet and Pete Whitelock (eds.), in: *Proceedings of the Joint 17th International Conference on Computational Linguistics*, Montreal, Canada, August 1998. pp. 315--319.

Димитрова и др. 2009: Dimitrova, L., Koseska, V., Roszko, D., Roszko, R. Bulgarian-Polish-Lithuanian Corpus – Current Development. in: *RANPL'2009 proceedings*. Borovec, Bulgaria, 17 September 2009.

Гоули и др. 2009: Ghouli V., Simov, K., Glaros, N., Osenova, P. (2009) A web-enabled and speech-enhanced parallel corpus of Greek-Bulgarian cultural texts. in: *EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pp. 35–42.

Грьон и Мариянович 2010: Grønn A., Marijanovic, I. (eds.) Russian in Contrast, in: *Oslo Studies in Language* 2 (1). pp. 1–24.

ЕГСКОЕ 1996а: *Expert Advisory Group for Language Engineering Standards Preliminary recommendations on corpus typology. EAG–TCWG–CTYP/P.* Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

ЕГСКОЕ 1996б: *EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Document EAG--TCWG—MAC/R*. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

Зилберщайн 1993: Silberztein, Max. Dictionnaires électroniques et analyse automatique de textes. Le système INTEX. Masson: Paris. 1993. (240 p.).

Зилберщайн 1999: Silberztein, M. *INTEX: a Finite State Transducer toolbox, in Theoretical Computer Science, 1999. 231:1.*

Лиич 1977: Leech, G. Introducing corpus annotation. In Garside R., Leech, G., McEnery, A. M. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Коева 1998: Коева С., Граматичен речник на българския език. Описание на концепцията за организацията на лингвистичните данни – Български език, 6, 49-58

Коева 2001: Koeva Svetla and Stoyan Mihov, INTEX 4.0 for Bulgarian, Revue Informatique et Statistique dans les Sciences humaines, Anne Dister, ed., Universite de Liege, 2001, 231-241.

Коева и др. 2006: Koeva, Sv., Sv. Leseva, I. Stoyanova, E. Tarpomanova, M. Todorova, Bulgarian Tagged Corpora. In: Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages, 18-20 October 2006, Sofia, Bulgaria, pp. 78-86.

МакЕнери и Уилсън 2001: McEnery, A. M., Wilson, A. (2001) Corpus Linguistics: An Introduction. Edinburgh: Edinburgh University

Михов 2000: Михов С., Минимални ациклични автомати. Конструкции, алгоритми, приложения, кандидатска дисертация, София, 2000.

Уилсън и Томас 1997. Wilson, A., Thomas, J. (1997) Semantic Annotation. In R. Garside, G. Leech & A. M. McEnery, (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. London: Longman. s

Туфиш и др. 2009: Tufiş, D., Koeva, Sv., Erjavec, T., Gavrilidou, M., Krstev., C. (2009) ID 10503 Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. in: *Scientific results of the SEE-ERA.NET Pilot Joint Call*, Jana Machaov, Katarina Rohsmann (eds.), Vienna, pp. 37--48.