Технологии за лексикални изследвания на славянски езици (с различна степен на близост) чрез паралелни корпуси

Елена Паскалева, ИПОИ, БАН, Секция за лингвистично моделиране

URL= http://lml.bas.bg/~hellen/ , E-mail: hellen@lml.bas.bg

1 Увод.

В езиковите технологии, важна област от комуникационните технологии, чийто главен обект (също инструмент на изследването) е най-старата комуникационна технология – **езика**[1], компаративните изследвания заемат важно място. Не случайно началото на началата за компютърната лингвистика и всички компютърни приложения е машинният превод.

Във всички съвременни компютърни приложения *многоезичието* има два аспекта и два прочита – говори се за многоезикови (multilingual) и междуезикови (crosslingual) изследвания и приложения. Разликата между тях е в характера и последователността на операциите – за многоезикови приложения говорим , когато създаденият за един език метод или инструмент за компютърна обработка се прилага върху друг езиков материал, а за междуезикови приложения говорим, когато се обработва с един метод материал от два езика едновременно. Класическо междуезиково приложение е автоматичният превод, но извън неговите амбициозни цели се развиват и много други дейности, свързани с получаване на информация в многоезикова среда – класически пример – междуезиковото информационно търсене – в база от документи на разни езици на въпрос, зададен на един език, се получава отговор за документи на други езици. Перспективно приложение е и създаването на двуезични речници (автоматично извличане на преводни съответствия) върху материали на големи многоезикови текстови корпуси.

Независимо от инструментите за лингвистичен анализ и равнищата на езиковото представяне във всички тези разнообразни приложения, главното условие за тяхната разработка е наличието на голяма текстова **многоезична** база, чието създаване и структурни характеристики ще проследим в това изложение.

[1] В света на компютърната лингвистика има място за понятието **език** така, както е определяно от бащите-основатели на модерната лингвистика – parole на Сосюр и language на Чомски: в реалността на дигитализирания текст или реч, и langue на Сосюр и grammar на Чомски – в постулатите на компютърната лингвистика.

2 Паралелни корпуси и тяхното „подравняване”

Както всички компаративни изследвания, и компютърните междуезикови приложения се проектират и извършват върху двойка текстови съвкупности – на оригинала и превода, оформящи т.нар. **паралелни** текстове.

Операцията – поставяне в съответствие, позната още като **alignment (**възприет български превод – **подравняване)**, представлява автоматично разбиване на двата текста по единиците на подравняване и образуване на смесен текст, съдържащ последователност от подравнени двойки (т.е. трансформацията Ta – Ea1, Ea2, Ea3… vs. Tb – Eb1, Eb2, Eb3… -→ {Ea1-Eb1), {Ea2 - Eb2, Eb3) и т.н. (T – език, E – единица).

На този принцип *единици на подравняването* на Ta и Tb могат да бъдат всички единици на текстовата линейна структура, което ни дава следните степени на членение на линейните единици на текста, а именно: *файлове* (по електронна идентификация), *параграфи*, *изречения*, *фрази* и *думи*.

Първото съответствие не е релевантно за междуезикови изследвания, освен за идентификацията на обработваните обекти. Второто е лесно за установяване, но е съпровождано от информационен шум, поради обема на параграфите, мерен в думи и изречения, който може да бъде много голям и да обезсмисли понятието съответствие.

Равнището на изреченията е най-използваният вид подравняване за създаване на този междуезиков ресурс. Единицата изречение е достатъчно голяма, и линейната й структура е сравнително проста по състав, за да бъде лесна нейната идентификация. Тя се осъществява чрез анализ на веригите от символи: букви – главни и малки, цифри, препинателни знаци. Тази идентификация е езиково независима, т.е. универсална, с изключение на участващите в нея списъци от съкращения, характерни за всеки език.

Един от най-използваните алгоритми за подравняване по изречения е статистическият алгоритъм на Гейл-Чърч [*Gale&Church,1993*]. Подравняването по изречения не означава просто разделяне на параграфа на изречения и смесване на двойките входни и изходни изречения, поради това, че едно изречение може да бъде преведено с две и обратно. Затова се извършва статистически анализ на разпределението на символите в двата текста и се установява т.нар. *сегмент на подравняването*, където 2 изречения могат да са подравнени към 1, 3 към 2 и обратно. Споменатият алгоритъм е основното формално средство за подравняване на големи текстови корпуси, използвано в компютърни приложения - *aligners*, български термин – *подравнители*.

Такъв подравнител, снабден с богат интерфейс за редактиране и с възможности за търсене в подравнени единици е програмата *Mark Alister*, разработен през 1997 г. [*Paskaleva&Mihov,1997*], с който е извършен описаният по-долу експеримент (Фиг.1).

Последният създаден в ИПОИ-БАН подравнител - *LORA* работи в пакетен режим (групи от хиляди файлове), върху текстове, извличани пряко от web сайтове [*Genov 2007*].

Използван е за създаване на голям корпус от текстове на балканските езици, подравнени към английски [*Paskaleva 2007*].

Подравняването *по фрази* – извличане на фразови съответствия, се използва в статистическия машинен превод, както и подравняването по *думи* (IBM метод).

Подравняване по думи не може да бъде осъществено за всички текстови единици - буквалният прочит на този израз би ни довел до буквалния превод *дума по дума*, заклеймяван и в човешкия превод.[2] Подравняванията по *фрази* и *думи* дават повече материал за междуезикови съответствия на различни равнища, докато подравняването по *изречения* създава само базата за тези наблюдения. Първите определят съответствията между смислово значими единици от двата езика, използват се за производство на речници, в междуезиковото информационно търсене и в компонентите на т.нар. *преводаческа памет* – компютърно подпомогнат превод.

Извличанията на съответствията между фрази и думи задължително се извършват върху вече *подравнени по изречения* текстови корпуси.

3 Създаване на база от подравнени корпуси за славянските езици

Засилващият се интерес към междуезиковите изследвания с многобройните им приложения поставя като първа изследователска задача намирането на съответната ресурсна база от подравнени текстови корпуси.

Неслучайно в последните години в европейски мащаб се работи върху създаването на големи многоезикови масиви от подравнени текстове, за всички езици на общността. Най старите традиции, извън Европа, но за европейски езици, датират от времената на създаването на англо-френския корпус на текстовете на Канадския парламент, [*HANSARD 1995]*, за да се стигне до създадения преди няколко месеца корпус от подравнени текстове на европейското право и документи за 22 европейски езика [*JRC-Acquis 2007*].

3.1 Източници за събиране на паралелни корпуси

Събирането на паралелни корпуси, за целите на машинния превод и преводаческата памет, започва в началото на 90-те години - идеята е изказана първо в [*Harris 1988*] . Оттогава нещата са се променили по отношение на обема, технологиите на събиране на корпусите и стандартите за тяхното представяне. Но богатството на тази текстова подръжка за отделни двойки езици е различно. Причините за това са геополитически и технологични.

Неслучайно първите големи паралелни текстови сбирки са правени за основните европейски езици – естествен резултат от дългите години евроинтеграция. Немалка роля

2 Опити за съответствие между всяка дума от едното изречение и дума от второто води до христоматийни безсмислици като намерения от мен пример в една стара руска разработка за подобен метод като помощник на преводача – вж. превода дума по дума на английския израз with a little trouble с руското без много труда, водещ до преводите: with- без, little – много, trouble – труд

изиграват и компютърните технологии, навлезли в книгоиздаването – с възможността за пряко използване на електронни текстове, а с масовото навлизане на web услугите в информационното пространство нещата се ускориха съществено, но за съжаление, не равномерно за различните езици.

Първите електронни сбирки на паралелни корпуси, включващи български, бяха създадени през 1997 г. за проекта GLOSSER [*Nerbonne et al.1997*] , с помощта на цитирания подравнител *Mark Alister*. Повечето от тях бяха електронни файлове на преводи, получени от български издателства и от представителства на международни институции, а трети бяха въвеждани ръчно в компютъра или сканирани. Web източниците бяха незначителна част от сбирката. [3]

Понастоящем намирането на такива текстове е улеснено поради почти задължителното web представяне на политически, административни и правни документи, а също и на електронните новинарски сайтове.

Как стои въпросът със славянските езици?

3.2 Паралелни корпуси за славянските езици

Както във всички големи семейства, и тук не всички са равни – по отношение на богатството от електронни ресурси, едноезикови и многоезикови. Причините са отново технологични и геополитически. Първите засягат наличието на методи за автоматична обработка – следствие от национална езикова политика, достъп до високи технологии и пр. Специално по отношение на многоезиковите електронни ресурси определящи са и геополитическите предпоставки – присъединяването на страната към европейските структури и към документалното им обслужване. Така езиците на първенците - славяни, приобщени към европейските структури – полски, словашки и словенски, получиха възможност за представяне в общоевропейските web портали. Отскоро и българският език е получил тази възможност и е включен в сбирката *JRC-Acquis,* заедно с още 21 европейски езика [*JRC-Acquis 2007*]. Останалите южнославянски езици са в *листата на чакащите*, а западнославянските руски, белоруски и украински не са допуснати още и до опашката.

Освен споменатите общоевропейски официални сайтове, съществуват и много други интернетни източници, с преводи и оригинали на славянски езици, но те са ориентирани към една двойка езици (най-често вторият член в двойката е английският).

От гледна точка на езиковите технологии, ако гоним обем (нещо важно за статистическите изследвания), споменатите общоевропейски административни сайтове са солиден ресурс. Но ако се интересуваме от езиковедската стойност на компаративните

[3] Когато се събират паралелни корпуси за двойка езици, наличието на добър превод на даден текст и оригинал - и двата в електронна форма, е рядко и щастливо събитие, ако сме се насочили към конкретен жанр или, още по-неосъществимо, към конкретен автор или произведение.

изследвания, материалът е твърде беден за наблюдения и изводи – юридическите и административни текстове ползват ограничена лексика и стандартен опростен синтаксис.

Доста по-голям диапазон от езикови явления са представени в новинарските сайтове, тъй като вестникарският информационен текст, макар и с доста *успокоен* синтаксис, съдържа богата лексика, описвайки различни събития и страни от действителността.

Една щастлива находка в търсенето на многоезикови електронни ресурси, създадени специално за южнославянските и балкански езици (за някои от които славистичната наука има доста резерви в определянето им като език), е използваният от нас през последната година в рамките на европейския проект **BIS-21++, Center of competence** (FP6 INCO-CT-2005-016639) новинарски сайт http://www.setimes.com , наричан от балканските лингвисти *Balkan Times.* Той се подържа от Министерството на отбраната на САЩ и геополитическите възгледи на авторите му са определили следните езици-участници в него: английски, албански, български, босненски, гръцки, македонски, румънски, турски, хърватски (до миналата година и сръбски - кирилица). Понеже количеството думи, натрупани от 2002 г. досега за всеки език, възлиза на около 3 млн думи, предоставя се добра възможност за електронна обработка на двойки езици, за която преди няколко години компютърният лингвист - славист е могъл само да мечтае. [4]

Това текстово богатство, съчетано с актуалните методи за статистически междуезикови изследвания, принадлежащи към инвентара на т.нар. статистически машинен превод, даде възможност да се изследват междуезикови зависимости, свързани най-вече с лексикалната близост между езиците. Възможността върху обща текстова база да се *измерва* тази близост и да се съпоставят нейните параметри, е едно типологическо предизвикателство за езиковеда-славист – вж. например формалните параметри на лексикалната близост между тройката *езици* – български, сръбски и македонски. Върху езиков материал от посочения сайт са направени експерименти за автоматично извличане на фразови съответствия между български и македонски език [*Paskaleva et al.2007*]. Изследвания по измерване на лексикална близост между руски и български (намиране на т.нар. *когнати*, включително и фалшиви) са правени със същите методи върху друг корпус (вж. [*Nakov et al.2007-1*] и [*Nakov et al.2007-2*], както и следващия раздел).

Други междуезикови изследвания върху материал от този корпус, свързани с методите за намиране на близост между цели документи, са правени върху българо-английския материал в него [*Alfred 2007*] .

За отбелязване е обаче печалният факт, че докато в европейското интернет-пространство в последните години се намират доритекстове, с които да се конструират паралелни корпуси за *език* като босненския, такива материали за междуезикови изследвания

[4] Сайтът www.setimes.org ни бе посочен от гръцкия партньор (ILSP, Greece – Института за обработка на език и реч, Атина) в цитирания европейски проект, за създаване на основните компоненти на българо-гръцка преводаческа памет.

с участието на руски език липсват изобщо. Това е случаят, когато геополитическите и технологическите причини за липсата на ресурси не са свързани, а си противоречат. Тази липса на паралелни текстови ресурси за някои езици налага индивидуално решение за тяхното събиране извън многоезиковите колекции, вж. раздел 4.

4 Други източници за събиране и обработка на паралелни корпуси – неудобства и компромиси.

При липсата на многоезикови web източници за отделни двойки езици се налага да се връщаме в зората на тяхното събиране и обработка и да търсим отделни електронни ресурси на оригинален текст и неговия превод.

4.1 Оригинали и преводи в електронните библиотеки

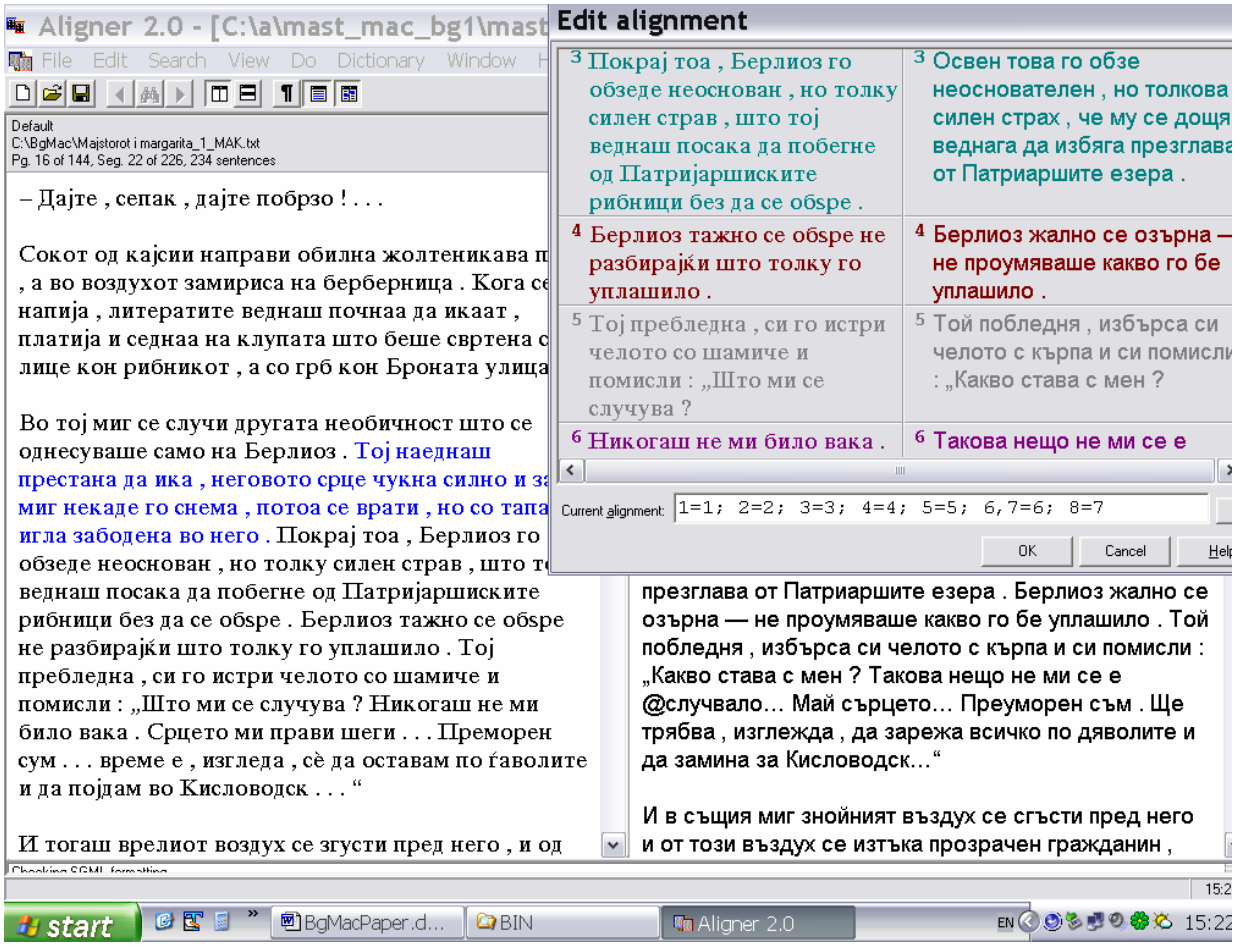
Такъв е вече цитираният случай с руския език, където като достъпен електронен ресурс могат да се използват широко известните руски електронни библиотеки, в онова малко тяхно подмножество, за което съществуват и електронни преводи на български. Така беше събрана и ресурсната база за статистически изследвания на лексикалната близост между руски и български в [*Nakov et al.2007-1*] и [*Nakov et al.2007-2*] – върху електронния оригинал и българските преводи на романа на В.Беляев – *Властелин мира* и книгата на И.Бунич – *Золото партии*. Подобен начин за събиране на електронни ресурси, освен че не може да достигне необходимия за статистически изследвания обем, не позволява разширяването на текстовата база с други езици – трудно бихме могли да намерим например, сръбския превод на същите произведения в електронна форма. Развиващите се в последните години електронни библиотеки от друга страна, се попълват предимно с оригинални, а не с преводни произведения. Сериозна пречка за разширяване на обема на електронните публикации са и все по-строгите правила на копирайта, за щастие не толкова строги в руското интернет пространство.

При компаративни изследвания с руски език в двойката текстове липсата на езиков материал засяга единствено електронната форма, а не превода като такъв, тъй като руският език е широко и отдавна превеждан. Съвсем другояче стои въпросът при славянски езици, за които обемът и на преводите изобщо, и на електронните им версии е значително по-малък.

4.2 Подравняване на преводи – начин за преодоляване на бедността на текстовия ресурс.

За двойки славянски езици, извън руския, намирането на електронни версии на оригинал и превод е почти невъзможно. Например, в българските електронни библиотеки, чийто брой непрекъснато намалява (поради мерките за борба с пиратството), в електронната библиотека http://www.e-bookbg.com/ от 1429 произведения на чужди автори само 3 са преводи от славянски езици извън руския – един на Карел Чапек и два на Иво Андрич.

Поради слабия дебит на междуславянските преводи едно алтернативно решение е да се подравняват славянски *преводи* на известни английски или руски оригинали (най-превежданите сега и в миналото езици). Така бе осъществен и експериментът с подравняването на българския и македонски преводи на *Майсторът и Маргарита* на М.Булгаков, първия взет от електронна библиотека, а втория – любезно предоставен от македонско издателство. Обработката бе извършена с подравнителя *Mark Alister,* който има дружелюбен интерфейс, позволяващ не само да се редактира подравняването в междинните му фази, но и да се извършват операции за търсене на определена дума от единия език и разположението на нейния превод в текстов сегмент на другия език (вж. Фиг.1).



Фиг.1. Mark Alister – интерфейс с постредакция на подравняването.

Търсенето на преводи на отделни лексикални елементи доведе до заключението, че:

Подравнените преводи на оригинал от трети език не могат да се разглеждат като преводи в истинския смисъл на думата и, строго погледнато, може да ги възприемем като вид *съпоставими корпуси,* колкото и далеч да са от истинските такива (за определението на т.нар. *comparable corpora* вж. [*Teubert 1996*]) . Не бихме могли да ги използваме за дълбоко изследване на механизма на превода за конкретни езици, но за сравнително изследване на лексикалния състав на двата езика – да.

Един пример за тази невъзможност :

Македонската частица **пак** се появява 12 пъти в глава 1 на изследвания роман, и в 10 от тях няма еквивалент в преведения български сегмент, тъй като очевидно е употребена като усилителна паразитна частица (подобно на българското **пък,** което обаче не е фигурира като превод в примерите). Само в 2 случая тя е преведена като темпорално наречие – **отново**, очевидно във второто си значение. Тази липса на лексикално съответствие в македонско-български план се обяснява, ако погледнем оригинала на подравняваните преводи. В руско-македонския корпус тази частица, в 7 от 10-те паразитни употреби е превод на руското **же**, което пък няма български еквивалент. Очевидно може да говорим по-скоро за преводачески решения в двата езика за предаване на значението на усилителна паразитна частица.

Рус.: *Берлиоз **же** хотел доказать поэту , что*
Мак.: *Berlioz , **pak** , sakace da mu dokafe na poetot deka*
***Ълг**.: А Берлиоз искаше да докаже на поета , че*

***Рус** . И **опять** крайне удивились и редактор и поэт*
Мак. *I **pak** krajno se poyudija i urednikot* i *poetot*
Ълг *И **отново** безкрайно се учудиха и редакторът , и поетът*

Интересни съпоставки от този вид могат да се правят и при подравняване на различни преводи на едно и също произведение (за анализ на преводачески подход).

4.3 Превод и оригинал? В подравнени корпуси от Интернет? Забравете!

Няма спор, че големите текстови сбирки, извлечени от Интернет от един и същ източник, са прекрасна база за междуезикови изследвания, статистически обоснована и проверена. Ако желаем да използваме тази база за компаративни езиковедски изследвания на по-дълбинни равнища, трябва да се откажем от характеристиката – посока на превода, оригинал и превод, превеждан и превеждащ език. В много случаи езикът на оригинала (или езикът на преводача, който винаги е в дълбока анонимност [5]) очевидно е английският. Това важи и за огромните текстови бази от документи на Европейския Съюз. В глобализираните комуникации е трудно да се отговори на въпроса за езика на оригинала или превода, както и за езика на съставителите на документите.

Затова желаещите да използват подобни текстови сбирки за изследване на преводния механизъм трябва да разчитат само на отделни версии на превеждани литературни произведения.

[5] В заглавията на файловете от сайта www.setimes.com се срещат славянски имена, без да е уточнена тяхната роля в превода (наличието на едно и също име във всички преводи на един материал сочи по-скоро към редактор, а не преводач).
http://www.setimes.com/cocoon/setimes/xhtml/sq/features/setimes/newsbriefs/2003/03/030302-**IVAN**-001.

Което от своя страна не намалява значимостта на основния строителен материал за електронни съпоставителни изследвания – подравнените корпуси с обем милиони думи. Особено за езиците от южнославянското семейство и балканския езиков съюз.

ЛИТЕРАТУРА

[*Harris 1988*] **Brian Harris**. Bi-text : A New Concept in Translation Theory, *Language Monthly* Issue 54, March 1988; *Are You Bi-textual?*, Language Technology, Issue 7, p. 41.

[*Gale&Church,1993*] **W. Gale A., K. Church** (1993), A Program for Aligning Sentences in Bilingual Corpora , *Computational Linguistics* 19 (1): 75-102. (http://acl.ldc.upenn.edu/J/J93/J93-1004.pdf)

[*HANSARD 1995]* http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20

[*Teubert 1996*] **Teubert, W**. Comparable or parallel corpora? *International Journal of Lexicography* 9(3), 238-264 вж. също **Anthony Mcenery & Zhonghua Xiao**. Parallel and comparable corpora: What are they up to? http://www.lancs.ac.uk/postgrad/xiaoz/papers/corpora%20and%20translation.doc

[*Nerbonne et al.1997*] **J.Nerbonne, L.Karttunen, E.Paskaleva, G.Proczeky, T.Roosmaa**. Reading more in Foreign Languages*, Fifth Applied Natural Language Processing Conference*, April 1997, Washington, ACL, вж. http://ucrel.lancs.ac.uk/acl/A/A97/A97-1020.pdf

[*Paskaleva&Mihov,1997*] **Paskaleva, E. and St. Mihov**: Second Language Acquisition from Aligned Corpora. Proc. Int. Conference *"Language Technology and Language Learning"*, Groningen, April 1997. (http://citeseer.ist.psu.edu/15845.html)

[*Vanilla 1997*] http://nl.ijs.si/telri/Vanilla/doc/ljubljana/

[*Hunalign 2006*] http://mokk.bme.hu/resources/hunalign

[*JRC-Acquis 2007*] http://langtech.jrc.it/JRC-Acquis.html

[*Alfred 2007*] **R. Alfred, E. Paskaleva , D. Kazakov , M. Bartlett.** Hierarchical agglomerative clustering for cross-language Information Retrieval. *International journal of translation*, Vol. XX, предадена за печат.

[*Genov 2007*] **N. Genov.** LORA - a basic tool for creation and primary processing of multilingual Balkan text corpora aligned to English. Presentation on *BIS21++ Information days*, maj 2007, Hissar.

[*Nakov et al. 2007 -1*] **P. Nakov, S. Nakov and E. Paskaleva**. Improved word alignments using the Web as a corpus, *RANLP-2007*, September, 2007, Borovetz, приета за печат.

[*Nakov et al.2007-2*] **P. Nakov, S. Nakov and E. Paskaleva.** Cognate or False Friend? Ask the Web! *Acquisition and management of multilingual lexicons*. Workshop, RANLP-2007, September, 2007, Borovetz, приета за печат.

[*Paskaleva 2007*] **E. Paskaleva**. Balkan – South East corpora aligned to English. *A common natural language processing paradigm for balkan languages*. Workshop, RANLP-2007, September, 2007, Borovetz, приета за печат.

[*Paskaleva et al.2007*] **E.Paskaleva, P.Nakov, V.Pacovski**. Extracting Translation Lexicons from Bilingual Corpora: Application to South Slavonic Languages. *A common natural language processing paradigm for Balkan languages*. Workshop, RANLP-2007, September, 2007, Borovetz, приета за печат.