Цветомира Венкова
(София)

СЪПОСТАВИМА КОРПУСНА ИЗВАДКА НА ПОДЧИНЕНИТЕ ОБСТОЯТЕЛСТВЕНИ ИЗРЕЧЕНИЯ ЗА ВРЕМЕ В БЪЛГАРСКАТА И АНГЛИЙСКАТА РАЗГОВОРНА РЕЧ

Създаването на компютърни текстови корпуси и разработването на процедури за тяхното морфологично и синтактично анотиране е бързо развиваща се лингвистична област през последните години. Една от тенденциите в тази област е насочена към развитието на двуезични и многоезични корпуси, които, от една страна, да дадат една нова емпирична основа на съпоставителните изследвания, но най-вече да послужат за изходна и тестова база на по-добри системи за компютърен превод и за улесняване на чуждоезиковата комуникация.

В настоящото изследване се представя създаването на двуезична съпоставима корпусна извадка (СКИ), извлечена въз основа на вече съществуващи едноезични корпуси на българска и английска разговорна реч. Обединяващият критерий за съставяне на извадката е синтактичен: подчинените изречения за време в двата езика. Тези изречения, заедно с наречията за време и някои предложни групи, играят много важна роля в голям брой комуникативни ситуации, свързани с темпоралната ориентация. Тази насоченост към темпоралните отношения е тясно свързана с общата тематика на английската част на изходния корпус, както и с работата на изследователския екип в Катедрата по лингвистика (Seminar fuer Sprachwissenschaft) в Тюбингенския университет, където бе извършена съществена част от разработката.

В хода на изследването се наложи уточняване на някои съществуващи терминологични различия при назоваването на видовете компютърни двуезични корпуси и съответно на извадките, основаващи се на тях. За двуезичните корпуси, изградени от оригинални текстове на единия език и техни преводи на другия език се използват термините *преводен (translation)*, вж. Aijmer &Altenberg 1996:13, Granger 1996:38 или *паралелен (parallel)*, вж. Baker 1993:248, Erjavec et al

1996. Двуезичните корпуси, съставени от оригинални текстове от двата езика, които си съответстват по жанр или текстов тип, се наричат *паралелен (parallel)*, вж. Aijmer &Altenberg 1996:13, Granger 1996:38 или *съпоставим (comparable)*, вж. Baker 1993:248, Erjavec et al 1996, Granger 2001. Както се вижда, едно и също разграничение се определя чрез три различни терминологични двойки: *преводен/паралелен, паралелен/съпоставим и преводен/съпоставим.* По мои наблюдения разделението *паралелен* (за корпус от текстове и техните преводи на другия език) и *съпоставим* (за корпус от сходни оригинални текстове на двата езика) е получило най-широко разпространение в литературата и е най-актуално в дискусиите на научни форуми по въпросите на двуезичните корпуси. По тази причина то е възприето и в настоящото изследване, където изходните български и английски корпуси на автентична разговорна реч се разглеждат като един *съпоставим двуезичен корпус*, а съответно и корпусната извадка, извлечена от него, е наречена *съпоставима корпусна извадка.*

Като основа на изследването бе предпочетена разговорната реч, тъй като СКИ е предназначена за приложение при разработки на актуалните системи за автоматично подаване на информация в диалогов режим, а също и за автоматичен превод на специализирани диалози, а именно спонтанният диалог осигурява много важна информация в тази насока. Освен това, и от чисто теоретична гледна точка смятам, че изразяването на времеви отношения в разговорната реч има своята специфика, която е недостатъчно разработена в литературата.

Надявам се, че СКИ ще продължи традицията на постигнатото в областта на двуезичните корпусни изследвания, касаещи и българския език, напр. *Linguist Workbench* (вж. Стамболиева 1996) и *Multext-East* (вж. Erjavec et al. 1996), като обаче, от своя страна, насочи вниманието към спецификата на разговорната реч и по-точно към нейните контрастивни темпорални характеристики в български и английски.

1. Източници за съставяне на съпоставимата корпусна извадка

Като източници на СКИ са използвани два корпуса съответно на английска и на българска разговорна реч. Всеки от тях съдържа приблизително 45 000 словоформи.

Корпусът на английската разговорна реч се основава на извадка *CD8* от Тюбингенската синтактична банка (The Tuebingen VERBMOBIL

Treebank of English). Тази банка е един от продуктите на международния проект *Verbmobil* (вж. Wahlster 2001), разработен с оглед на машинния превод на три езика: английски, немски и японски. Резултатите от проекта притежават богат потенциал: разпознавател на реч, семантичен анализатор, синтезатор на реч, контекстна обработка и др. При изграждането на СКИ Тюбингенската синтактична банка се използва в два аспекта: като извадки от транскрипции на английски диалози и като синтактична банка, съответстваща на всяко изречение от транскрипцията.

Корпусът на българската разговорна реч съдържа транскрипции на диалози. Той е извадка от по-голям корпус на Цв. Николова, резултат от неин дългогодишен труд по запис на касети, транскрибиране и подреждане на материала (Николова 1987). Съществена част от този корпус (50 000) думи бе въведена от мен като компютърен текст и структурирана (Венкова 1996:264). Ползването на този корпус в Интернет на адрес: http://www.hf.uio.no/easteur-orient/bulg/mat/index.html/Nikolova е възможно благодарение на инициативата на Хетил Ро Хауге.

Въз основа на тези два корпуса след обработка по определена процедура са извлечени съответните едноезични корпусни извадки, обединени в двуезична съпоставима извадка, съдържаща подчинени обстоятелствени изречения за време.

Въпреки че по отношение на тематиката на диалозите английският корпус е в известен смисъл по-ограничен, тъй като разговорите са свързани предимно с насрочване и уточняване на срещи (appointment negotiations, scheduling appointments), това няма съществено отражение върху анализа на материала. С оглед на предимствата на двуезичните корпуси, изложени от Lauridsen1996:63, смятам, че изходните корпуси притежават необходимия потенциал за контрастивни изследвания.

2. Етапи на съставяне на съпоставимата корпусна извадка

Съставянето на СКИ се състои от няколко процедури, като стремежът е максимално да се използва компютърна помощ.

2.1. Извличане на конкорданс на съюзите за време

Най-бързият начин за отделяне на подчинените изречения за време е да се осъществи автоматично търсене на въвеждащите ги съюзи за време. Това става чрез задаване на съответния съюз като ключова дума и съставяне

на конкорданс на появите на съюза в контекста на сложното съставно изречение. Тъй като тези съюзи са много ограничен брой, а освен това са и неизменяеми части, това търсене технически се осъществява автоматично много лесно. За английските съюзи то се извършва чрез търсеща функция в заложения програмен пакет на Тюбингенската синтактична банка. Програмата намира всички появи на определен съюз и ги представя в рамките на изречението като контекст. След това всяко от тези изречения се записва като отделен файл заедно със съответстващото му дърво. За българския корпус съставянето на конкорданса става чрез разработена за целта кратка програма макрос.

2.2. Семантичен анализ на конкорданса на съюзите за време

Отделянето на група изречения само според типа на свързващия съюз, което се извършва чрез автоматично съставения конкорданс обаче не ограничава тази група достатъчно точно. Проблемът е в това, че доста съюзи са полисемантични, тъй като могат да свързват подчинени изречения от различен синтактико-семантичен тип. Така например някои съюзи могат да въвеждат подчинени изречения както за време, така и условни и определителни изречения. Освен това при неизменяемите думи се среща и полифункционалност, при която една дума може да функционира като наречие-съюз или като въпросително наречие, например англ. *when 'кога, когато'* или диалектната разновидност в бълг. *кога* на *когато* или пък като наречие и предлог, например англ. *before 'преди, преди това'*.

Ето защо като следваща стъпка се налага да се извърши семантико-граматичен анализ на всички изречения, отделени първоначално чрез съюзите, които могат да функционират като съюзи за време. При този анализ особено полезен е паралелът между речниковите статии в двата езика, а също и позоваването на някои основни теоретични постановки. В резултат на синтактико-семантичния анализ вече е възможно да се отделят само тези сложни съставни изречения, в които определен съюз въвежда подчинено обстоятелствено изречение за време. След като бъдат анализирани, тези изречения се маркират със съответния етикет (англ. *tag*), който позволява извършването на автоматични процедури с тях като например разпознаване, извличане и сортиране.

Специфични трудности при анализа възникват от това, че корпусът съдържа разговорна реч. Има много изречения, които трудно биха могли

да се причислят към структурите, извлечени на базата на художествен текст. Често се срещат явления като недовършени изречения, повторения, елипси, промяна в целта на изказването и др., които са отбелязани от редица автори (вж. напр. Ангелова 1994, Земская 1987).

Интересен проблем при извършването на анализа са и някои разлики в граматичните традиции. Сходни езикови явления в двата езика се интерпретират в литературата с различен терминологичен апарат, различна перспектива и приоритети и тези различия трябва да се разгледат внимателно и да се съпоставят. От друга страна, тази съпоставителна перспектива е много полезна за откриване на специфичните особености на изследваните структури във всеки език.

2.3. Структуриране на извадките и обединяването им в двуезична съпоставима извадка

След анализиране и подбор на подчинените обстоятелствени изречения за време от конкорданса те се отделят първоначално в самостоятелна корпусна извадка за всеки език, а след това и в съпоставима двуезична съставка.

След като се уточни съставът на всяка от едноезичните корпусни извадки, се разработва тяхната вътрешна структура, като целта е максимално да се улесни търсенето и извличането на информация от тях. Всяко изречение носи индекс, показващ точното му място в изходния корпус. За английските изречения това е номерът на изречението в него, а за българския корпус съответно – номерът на реда. Изреченията се групират по определен признак, като точната им номерация позволява и тяхното прегрупиране в съответствие с други критерии.

Извадките се обединяват в СКИ според функционалните еквиваленти на свързващите съюзни думи. Всяка функционална група подлежи на съответното подразделяне в зависимост от семантиката на подчинените изречения.

Извършва се и статистическа обработка на резултатите, като се изготвят статистически профили, които биха могли да се използват при евентуални програми за езикова обработка, основаващи се на стохастични методи.

Изреченията, които са отхвърлени при анализа, се сортират по типове като съпътстващи корпусни извадки.

3. Възможности за автоматично извличане на лингвистична информация от СКИ

3.1. Търсене на подчинени обстоятелствени изречения за време по определен съюз

Въз основа на СКИ ще е възможно да се търсят и извличат появите на подчинени обстоятелствени изречения за време според свързващия съюз. След задаване на ключова дума автоматично се съставя конкорданс, заедно със съответните статистически данни. За разлика от търсенето на съюзите от целия корпус, при търсене в СКИ ще се получават само тези появи на съюзите, в които те изпълняват темпорална функция.

3.2. Търсене на различни разговорни варианти на подчинени обстоятелствени изречения за време

Българската корпусна извадка е съставена така, че да може да се търсят и регионалните варианти на съюзите за време в разговорната реч, например *кога* или *когат'* вместо *когато*. Това ще дава възможност да се извършват анализи, отчитащи една по-широка представа за разговорния език. Ще могат също така да се извличат и връзките между тези варианти.

3.3. Паралелно търсене на подчинени обстоятелствени изречения за време в двата езика

Основното предимство на съпоставимата двуезична извадка е, че дава възможност да се извличат едновременно появите на сходни типове подчинени обстоятелствени изречения за време в двата езика.

3.4. Някои ограничения

Разбира се, извличането на информация само на нивото на неанотирания извадков текст налага своите ограничения, тъй като може да се задават за търсене и съпоставка само текстови отрязъци: една или няколко словоформи или отделни части от тях, както и препинателни знаци, но не и лексеми и граматични категории. От друга страна обаче, "чистият" текстов материал дава много добри възможности за бързи справки и за извличане на езикови примери. Освен това липсата на теоретична предпоставеност често пъти е желана при търсене на езикови данни за теоретични разработки в рамките на желана от изследователя теория.

Всъщност не е съвсем точно да се каже, че съпоставимата корпусна извадка не съдържа граматична информация. Макар и в неексплицитен вид, тази информация играе важна роля. Както бе посочено по-горе, СКИ е

извлечена по определени граматични критерии от изходните корпуси с разговорни транскрипции. Всички налични в нея изречения са били подложени на предварителен граматичен анализ и включването на всяко едно от тях е в резултат на определен избор, отнасящ се до свалянето на граматичната синонимия на съюзните връзки чрез семантично разграничаване между подчинените изречения за време и други типове подчинени изречения, въвеждани със същия съюз. Стремежът е този анализ да е в много голяма степен подчинен на общоприети и утвърдени тези. Освен това, за да бъде процесът на подбор максимално прозрачен и достъпен за потребителя, отхвърлените изречения със съответния им анализ могат да бъдат открити в съпътстваща банка към извадката.

4. Изводи

Надявам се, че двуезичната съпоставителна корпусна извадка, описана в това изследване, е една добра основа за синтактични и семантични разработки върху времевите отношения, за програмни приложения или за езиковото обучение както за всеки език поотделно, така и в съпоставителен план. Освен това, предложената процедура за изграждане би могла да се използва като прототип за създаване на съпоставима корпусна извадка и с по-широк тематичен обхват, както и да бъде доразвита в синтактична база данни, представяща комплексно проблемите на разглежданите синтактични явления.

БИБЛИОГРАФИЯ

1. **Ангелова, И.** Синтаксис на българската разговорна реч (в съпоставка с руски, чешки, полски език). С., Унив. изд. "Св. Климент Охридски", 1994.

2. **Венкова, Ц.** Компютърен конкорданс на думата *да* в разговорната реч.– В: Проблеми на социолингвистиката. Т. 5. М. Виденов, А. Ангелов, Кр. Алексова, П. Сотиров (съставители). Международно Социолингвистическо дружество. С., 1996, 263–266.

3. **Земская, Е.** Русская разговорная речь: лингвистический анализ и проблемы обучения, Москва, 1987.

4. **Николова, Ц.** Честотен речник на българската разговорна реч. С., Наука и изкуство, 1987.

5. **Aijmer, K. &Altenberg, B. (1996)** Introduction In: *Languages in Contrast*, Lundt Studies in English 88, S. Baeckman & Svartvik, J. (eds.), Lund University Press, pp.10–16.

6. **Baker, M.** Corpus Linguistics and translation studies: Implications and applications. – In: Text and technology. – In honour of John Sinclair, ed. by M. Baker, G. Francis & E. Tognini-Bonelli, 1993, pp.233–250.

7. **Erjavec, T. et al** Erjavec, T., Idle, N. Petkevic, V and Veronis, J. Multext-East: Multilingual Text tools and Corpora for Central and East European Languages. – In: Proceedings of the First European TELRI Seminar: Language Resources for Languge Technology, 1996, pp. 87–98.

8. **Granger, S.** From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. – In: *Languages in Contrast*, Lundt Studies in English 88, S. Baeckman & Svartvik, J. (eds.), Lund University Press, 1996, pp. 37–51.

9. **Granger, S. (2001)** Corpora in Contrastive Linguistics, Translation studies and Cross-linguistic NLP Applications, Paper presented at 6th TELRI Seminar, Bansko, Bulgaria, 9-11 November 2001.

10. **Lauridsen, K.** Text corpora and contrastive linguistics: which type of corpus for which type of analysis? – In: Languages in Contrast, Lundt Studies in English 88, S. Baeckman & Svartvik, J. (eds.), Lund University Press, 1996, 63–73.

11. **Stambolieva, M.** A Linguist's Workbench. In: Papers from the first conference on formal approaches to South Slavic Languages (Plovdiv October 1995), University of trondheim, Working Papres in Linguistics 28, 1996, pp. 293–301.

12. **Wahlster, W.** (ed.) Verbmobil: Foundations of Speech-to-Speech Translation. German Research center for Artificial Intelligence, (DFKI), Saarbruecken, Germany, 2001.