АТАНАСКА АТАНАСОВА
Институт за български език – БАН, София

МЕТОДИ И ПОДХОДИ ЗА АВТОМАТИЧНО ИЗВЛИЧАНЕ НА ЛЕКСИКАЛНИ НЕОЛОГИЗМИ[1]

Abstract: The article deals with the application of corpus approaches for extracting new words. It focuses on methods and approaches for automatic detection and automatic extraction of candidates for new words. The observations show that there are 4 methods for implementation of this task.

1. Въведение

Възникването и развитието на корпусите и корпусната лингвистика (60-те години на XX век) поставят началото на нов етап от развитието на лексикографията (80-те години на XX век), в резултат на което лексикографските решения се обективизират, лексикографската дейност се оптимизира и се създават по-качествени лексикографски ресурси. Приложението на корпусите и корпусния подход с оглед на своевременното и възможно най-пълното и прецизно отразяване на най-динамичния пласт в лексиката – новите и най-новите лексикални единици в езика, е изключително полезно, макар че с редки изключения то се свежда най-вече до автоматично идентифициране и извличане на списъци с кандидати за лексикални неологизми (нови и най-нови думи). Именно поради тази причина целта на настоящата работа е да представи приложението на корпусите и корпусния подход при автоматичното идентифициране и извличане на лексикални неологизми, като конкретно вниманието е насочено към различните методи и подходи, използвани при осъществяването на тази задача.

2. Основни корпусни подходи

В лингвистиката, в това число и в лексикографията, се говори за два основни корпусни подхода – корпусно базиран (corpus-based) и корпусно изведен (corpus-driven) подход, разграничение, направено от един от основателите на корпусната лексикография – Дж. Синклер, и доразвито от В. Уи (Уи 1998: 51 – 52).

[1] Изследването е осъществено в рамките на проект *Корпусно базирани подходи в българската неография (теоретични и научноприложни аспекти)*, финансиран по „Програма за подпомагане на младите учени в БАН“, Договор № ДФНП – 115/11.05.2016 г.

Според това разграничение при корпусно базирания подход корпусите се използват за потвърждаване, проверка и подобряване на съществуващи вече хипотези и теоретични постановки, т.е. за верифициране на информацията и за разширяване и подобряване на лингвистичното описание. При този подход изследванията се осъществяват „от горе надолу“ (top-down, Уи 1998: 52), т.е. от теорията към конкретните корпусни данни.

За разлика от него при корпусно изведения подход корпусите и корпусните данни са първоизточник за възникването на нови идеи и хипотези, въз основа на които се формулират нови теоретични постановки. Тук подходът е „от долу нагоре“ (bottom-up, Уи 1998: 52), т.е. от конкретните корпусни данни към теорията.

С оглед на лексикографията може да се каже, че днес използването на корпусни подходи е норма, а не изключение. По-широко приложение в лексикографската практика, в това число и в неографията, намира корпусно базираният подход, благодарение на който част от дейностите по изработването на даден речник се оптимизират (например автоматично извличане на списък от лексикални единици с цел създаване на словник, бързо и лесно проследяване на граматичните и синтагматичните особености на лексикалните единици, намиране на подходящи примери за илюстрация на значенията на тълкуваните единици и др.).

3. Приложение на корпусния подход в неографията

Когато се говори за приложението на корпусния подход в неографията, с редки изключения (вж. напр. Благоева, Колковска 2011, Колковска и др. 2012, Реноуф 2012) се обръща внимание най-вече на методите и подходите, използвани за автоматичното идентифициране и извличане на кандидати за нови и най-нови лексикални единици. Според съществуващите изследвания по въпроса методите са четири, а именно: 1) съпоставка с изключващи списъци; 2) използване на шаблони; 3) съпоставка с диахронни корпуси; 4) хибриден метод.

3.1. Съпоставка с изключващи списъци

При този метод за откриване на лексикални неологизми се използват списъци от лексеми, автоматично извлечени от корпус (често специално създаден) с текстове от по-нов период, които се съпоставят със списъци от лексеми и/или форми, извлечени от корпус, който съдържа текстове от фиксиран период, предшестващ изследвания период. Освен с този списък от лексеми и/или форми списъкът с кандидати за нови и най-нови лексикални единици се съпоставя и с допълнителни изключващи списъци, направени на базата на съществуващи вече лексикографски ресурси, референтни корпуси и други източници, с цел елиминиране на нерелевантните канди-

дати за лексикални неологизми. Този метод често се съчетава с използване на допълнителни филтри (невъзможни N-грами от символи за даден език, съпоставяне със списъци със собствени имена, абревиатури и под., честотност и др.) за изключване на несъществуващи лексеми или форми.

Използването на изключващи списъци с цел автоматично идентифициране и извличане на кандидати за лексикални неологизми е приложено от различни изследователи с оглед на различни езици – О'Донован, О'Нийл за английски език (О'Донован, О'Нийл 2008), Благоева, Колковска и др. за български език (Благоева 2008, Благоева 2009, Колковска и др. 2012), Янсен за испански език (Янсен 2012). Въпреки това обаче подходите, приложени от различните изследователи, са различни.

При създаването на поредното издание на *The Chambers Dictionary* Р. О'Донован и М. О'Нийл използват специално създаден за целта референтен корпус (модел на „нормалния" език (О'Донован, О'Нийл 2008: 572). Този корпус е POS тагиран и лематизиран и включва текстове от определени списания, вестници и уебсайтове, излезли за период от 12 месеца. От него автоматично е извлечен честотен списък на лексикалните единици, срещани в корпуса. След създаването на този корпус ежемесечно се създават нови (тестови) корпуси с текстове от източниците, използвани при създаването на референтния корпус, и се обработват данните в тях – определя се принадлежността на всяка дума към определена част на речта и ѝ се приписват съответните етикети; съотнасят се различните словоформи към конкретна лема. Като резултат се създават автоматично извлечени честотни списъци, които се съпоставят със списъка, извлечен от референтния корпус. По този начин се получава автоматично извлечен списък с кандидати за лексикални неологизми, който допълнително се филтрира чрез съпоставка със списък, съдържащ заглавните думи, включени в предходните издания на *The Chambers Dictionary*.

Подобен подход е приложен и по отношение на български език от Е. Пернишка, Д. Благоева и С. Колковска (Благоева 2008, Благоева 2009, Колковска и др. 2012) при изработването на *Речника на новите думи в българския език* (РНДБЕ 2010). От субкорпуса с текстове след 1990 г. на *Българския национален корпус* (http://search.dcl.bas.bg/), който също е POS тагиран и в по-голямата си част лематизиран, автоматично е извлечен азбучно-честотен списък с кандидати за лексикални неологизми. Този списък автоматично е съпоставен с изключващи азбучно-честотни списъци, извлечени от специално създадения за лексикографски цели *Електронен лексикографски корпус*, част от който към момента е интегрална част от *Българския национален корпус*. Освен с тях списъкът с кандидати за лексикални неологизми е съпоставен и с други изключващи списъци, базирани на съществуващи лексикографски ресурси, в резултат на което е изготвен редуциран списък с кандидати за лексикални неологизми.

За разлика от О'Донован и О'Нийл обаче, които сравняват само спи-

съци с леми, Колковска и др. работят и с азбучно-честотни списъци със словоформи, което предполага наличие на повече неподходящи кандидати за лексикални неологизми, включени в крайния редуциран списък.

Различен от представените до тук подход за автоматичното извличане на лексикални неологизми чрез съпоставка с изключващи списъци е приложен от М. Янсен (Янсен 2012) за испански език. Той работи с тренировъчен корпус като референтен корпус, като използва специално създаден за целта POS тагер (*NeoTag*), който е обогатен с допълнителни етикети, с помощта на които автоматично се идентифицират и извличат кандидатите за неологизми.

3.2. Използване на шаблони

При този метод за откриване на лексикални неологизми се използват предварително зададени шаблони (лексикални средства и/или пунктуационни маркери), които сигнализират за появата на нова лексикална единица в близкото им обкръжение. Лексикалните средства включват фрази като *известен като*, *наречен/наричан* (*още*), *познат като* и др., в обкръжението на които често се появяват нови лексикални единици. Към пунктуационните маркери спадат кавичките, в които много често се ограждат новите лексикални единици. Този метод е приложен от П. Паризек (Паризек 2008) по отношение на английския език, като е използван специално създаден за целта корпус с научни текстове от списание *Nature*. Корпусът съдържа 45 милиона думи.

С оглед на целта на настоящото изследване внимание заслужава фактът, че използването на шаблони за разлика от съпоставката с изключващи списъци позволява като кандидати за неологизми да се идентифицират не само отделни думи, а и цели фрази и съставни единици.

Полученият чрез приложението на този метод от Паризек списък с кандидати за лексикални неологизми на последващ етап е POS тагиран и верифициран чрез използване на изключващи списъци.

3.3. Съпоставка с диахронни корпуси

При третия метод се използват диахронни корпуси (текстове от определен фиксиран период, предходен спрямо по-нов период), на базата на които с помощта на статистически анализ и/или машинно обучение се откриват и оценяват кандидатите за лексикални неологизми. Този метод е приложен от Стенеторп при откриване на лексикални неологизми в шведски език (Стенеторп 2010), както и от Реноуф не само за идентифициране на кандидати за нови лексикални единици, но и за регистриране на нови съставни единици, както и за откриване на семантични неологизми в английски език (Реноуф 2012).

Независимо че използват един и същ метод, подходите, приложе-

ни от различните изследователи, отново са различни.

За целите на автоматичното идентифициране и извличане на лексикални неологизми в шведски език П. Стенеторп разработва и създава специална система – *Novel System for Extracting Neologisms*, която използва машинно обучение и се различава от съществуващите системи по броя етикети, които се приписват към конкретната дума, като брой срещания (в брой документи и в корпуса), първа и последна поява на думата в корпуса, „възраст“ (Стенеторп 2010: 25). Това от своя страна дава възможност за по-доброто оценяване на кандидатите за неологизми.

Системата не използва филтри за елиминиране на неподходящите кандидати. Според Стенеторп използването на филтри може да отстрани както неподходящи кандидати за нови лексикални единици, така и същински неологизми. Тя разчита на оценка, тъй като според създателя ѝ вероятността неанотираните думи и думите с ниска оценка да са неологизми, е много голяма (Стенторп 2010: 24).

За автоматичното идентифициране и извличане на неологизми (не само лексикални, но и семантични) А. Реноуф използва резултатите, получени от осъществяването на няколко предходни проекта – *AVIATOR* (Analysis of Verbal Interaction and Automated Text Retrieval, 1990 – 1993); *APRIL* (Analysis and Prediction of Innovation in the Lexicon, 1997 – 2000); *The WebCorp* (Web as Corpus, 2000 – 2003) and *WebCorpLSE* (WebCorp Linguist's Search Engine, 2003 – 2007). Също така за разлика от подхода на Стенеторп при подхода на Реноуф за разпознаването на кандидатите за нови лексикални единици се прилагат различни филтри: 1. идентифициране на неологизмите в текстове от 1989 нататък чрез сравняване с изключващи списъци, извлечени от корпус с текстове от периода 1984 – 1988 г.; 2. идентифициране на кандидатите за нови съставни единици от кандидатите за неологизми, което води до редуциране на списъка с кандидати за лексикални неологизми; 3. идентифициране на кандидатите за семантични неологизми, с което списъкът с кандидати за лексикални неологизми допълнително се редуцира; 4. проследяване на контекстуалното поведение на кандидатите за нови и на съществуващите вече лексикални единици (Реноуф 2012: 6).

3.4. Хибриден метод

При хибридния метод се съчетават два или повече от предходните методи. Такъв метод е използван от Фалк и др. за френския език (Фалк и др. 2014), от Стоянова и др. за българския език (Стоянова и др. 2016), както и от Лиу за тайванския език (Лиу 2013).

Подходът за автоматично идентифициране и извличане на лексикални неологизми, използван от Фалк и др., включва следните стъпки: 1. съпоставка с изключващи списъци; 2. филтриране на получените резултати (чрез отстраняване на собствени имена, имена на организации

и пр.; чрез изключване на думи, които съдържат правописни грешки и др.); 3. ръчно класифициране на кандидатите за лексикални неологизми като: а) plausible words (Фалк и др. 2016: 4339), чиито форми са изписани правилно, но които не присъстват в изключващите списъци, и б) неологизми; 4. обучение на специално създадената за целта система – *Logoscope*; 5. приписване на атрибути (стойности): а) формални; б) морфо-лексикални; в) тематични; 6. оценяване на двата типа кандидати за лексикални неологизми (Фалк и др. 2014: 4339 – 4340).

Стъпките за автоматично идентифициране и извличане на лексикални неологизми, приложени от Стоянова и др., са: 1. идентифициране на кандидати за неологизми и съпоставяне с изключващи списъци; 2. групиране на кандидатите (съотнасяне на словоформите към основната форма); 3. филтриране и оценяване на кандидатите (според тяхната честотност и според броя документи, в които се срещат).

За автоматично идентифициране и извличане на лексикални неологизми Лиу (Лиу 2013) използва следния алгоритъм: 1. трансформиране на всички тоукъни в биграми; 2. съпоставка с изключващи списъци; 3. филтриране на получени резултати чрез прилагане на специално създадени лингвистични правила (Лиу 2013: 253).

Както става ясно, методът, приложен от различните изследователи, отново е един и същ, но подходите отново са различни.

4. Заключение

Направените наблюдения върху методите и подходите за автоматично идентифициране и извличане на лексикални неологизми показват, че към момента съществуват четири основни метода за идентифициране и извличане на кандидати за нови лексикални единици, но подходите, чрез които става това, са значително повече.

Съществуващите изследвания по въпроса дават основание да се твърди, че независимо от факта, че дейностите по идентифицирането и извличането на кандидати за нови лексикални единици, които са изключително трудоемки и времеемки задачи, са автоматизирани, компютрите все още не могат да заместят лексикографите при изготвянето на списъци с нови лексикални единици. Изследванията показват още, че с оглед на поставената задача по-предпочитан и използван остава корпусно базираният подход.

Литература

Благоева 2008: Благоева, Д. Съвременни подходи в българската неография (проблеми и перспективи). – *Български език*, 2008, № 1, с. 5 – 14.

Благоева 2009: Blagoeva, D. Electronic Corpora and Bulgarian New-Word Lexicography. – Études Cognitives. Vol. 9. Warszawa: SOW, 2009, p. 143 – 150.

Благоева, Колковска 2011: Благоева, Д., Колковска, С. Корпусният подход в българската лексикография – практика и перспективи. – В: *Съвременни методи и подходи в лексикографската практика*. София: Авангард Прима, 2011, с. 7 – 45.

Колковска и др. 2012: Kolkovska, S., D. Blagoeva, A. Atanasova. The application of corpus-based approach in the Bulgarian new-word lexicography. – In: *Proceedings of the 15th EURALEX International Congress.* Oslo, 2012, p. 991 – 996.

О'Донован, О'Нийл 2008: O'Donovan, R., M. O'Neil. A Systematic Approach to the Selection of Neologisms for Inclusion in a Large Monolingual Dictionary. – In: *Proceedings of the 13th Euralex International Congress*. Barcelona, 2008, p. 571 – 579.

Лиу 2013: Liu, Tsun-Jui. Observing Features of PTT Neologisms: A Corpus-driven Study with N-gram Model. – In: *Proceedings of the Twenty-Fifth Conference on Computational Linguistics and Speech Processing (ROCLING)*, 2013, p. 250 – 259.

Паризек 2008: Paryzek, P. Comparison of Selected Methods for the Retrieval of Neologisms. – *Investigationes Linguisticae*, 2008, 16, p. 163 – 181.

РНДБЕ 2010: Пернишка, Е., Д. Благоева, С. Колковска. *Речник на новите думи в българския език (от края на XX и първото десетилетие на XXI в.)*. София: Наука и изкуство, 2010.

Реноуф 2012: Renouf, A. *Defining neology to meet the needs of the translator: a corpus-based perspective.* <http://rdues.bcu.ac.uk/Defining_neology_to_meet_the_needs_of_the_translator.pdf> (дата на достъп 10.11.2016).

Стенеторп 2010: Stenetorp, P. *Automated Extraction of Swedish Neologisms Using a Temporally Annotated Corpus.* Master of Science in Engineering (MSc Eng) Thesis. Royal Institute of Technology (KTH). Stockholm, 2010.

Стоянова и др. 2016: Стоянова, Ив., Св. Лесева, Св. Коева. Автоматично разпознаване на неологизми в българския език. – В: *Лексикографията в началото на XXI век. Доклади от Седмата международна конференция по лексикография и лексикология (София, 15 – 16 октомври 2015 г.).* София: Изд. на БАН „Проф. М. Дринов", 2016, с. 679 – 686.

Уи *1998*: Ooi, V. *Computer corpus lexicography*. Edinburgh University Press, 1998.

Фалк и др. 2014: Falk, I., D. Bernhard, C. Gérard. From Non-Word to New Word: Automatically Identifying Neologisms in French Newspapers. – In: *Proceedings from LREC*, 2014, p. 4337 – 4344.

Янсен 2012: Janssen, M. NeoTag: a POS Tagger for Grammatical Neologism Detection. – In: *Proceedings from LREC*, 2012, p. 2118 – 2124.