Резюмета на публикациите

на доц. д-р Петя Начева Осенова,

представени за участие в конкурса за заемане на академичната длъжност „професор”
по професионално направление 2.1. Филология (Български език – морфология, синтаксис и корпусна лингвистика),
обявен от СУ „Свети Климент Охридски” в ДВ, бр. 9/02.02.2016 г.

(Публикациите, представени за участие в конкурса, не повтарят представените за придобиване на образователната и научна степен „доктор”, както и на академичната длъжност „доцент”)

I **Монографии**

В монографията „*Граматическо моделиране на българския език (с оглед на обработката на естествен език)*”, София, изд. „Парадигма” (189 стр.) се разглеждат два начина за представяне на морфосинтактичните и синтактичните явления: (а) чрез имплементиране на ресурсна граматика и (б) чрез анотиране на синтактичен корпус спрямо определена формална лингвистична теория, който след това се премоделира в други анотационни схеми.

Основният принос на труда е акцентът върху параметрите на граматическото моделиране на нашия език с оглед на конкретни теоретични рамки, които не са били широко използвани за българския език. Показани са както езиковоспецифичните, така и универсалните граници на явленията. Предимство е, че всеки модел е свързан с реален ресурс, който е достъпен за ползване.

Двата начина на представяне на явленията въвеждат съответно и два подхода към езиковия анализ (и синтез): единият е ‘отдолу-нагоре’ (а), което означва, че се набляга върху комбинирането на данните, а другият е ‘отгоре-надолу’ (б), което означава, че се измерва обхватът и точността на цялостно изградената граматическа рамка.

Описана е подробно Българската ресурсна граматика, която е разработена в рамките на съвременна лингвистична теория − Опорната фразова граматика. В своя първи вариант тя покрива основното количество морфологични правила и синтактични явления. За целите на сравнението граматиката е синхронизирана с общ граматичен модел (матрица) и е оценена върху преведеното множество от изречения на този общ модел. Моделирането в Ресурната граматика показва най-пълно и реалистично взаимодействието между езиковите равнища (морфология, синтаксис, семантика) и между ресурсните компоненти (речник, йерархия на явленията и граматически правила).

Друг основен принос в труда е дискутирането на конституентни и депендентни анализи за българския език, което обогатява перспективите на езиковото моделиране. Всъщност, Граматика на зависимостите, или още Депендентната теория, не е прилагана широко за нашия език.

Представени са два депендентни модела за българския синтактичен корпус Бултрибанк. Единият е наречен стандартизиран, защото преносът на релации следва дословно оригиналната анотационна схема. Този пренос показва обаче каква част от лингвистичната информация е експлицирана на морфологично и на синтактично равнище.

Другият модел е наречен универсален, тъй като е свързан с общ модел, който трябва да бъде следван с идеята да се обхванат най-пълно и съпоставимо явленията във всички езици. Подобно на Ресурсната граматика тук има предварително зададена схема, затова отчетливо изпъкват разликите в моделите и в езиковоспецифичните характеристики.
И в Ресурсната граматика, и в универсалния депендентен модел се приема семантичен подход към описанието на явленията за разлика от оригиналния синтактичен ресурс Бултрибанк, при който преобладава морфосинтактичният. Принос в работата е моделирането на българския език в семантично ориентирани рамки, което е предизвикателство за флективната му граматическа природа.

Друг приносен момент е разглеждането на понятието *опора* както в различните имплементирани модели, така и в теоретичен план. Представени са тестовете за определяне на опората, видовете реализации на опората, равнището на нейното представяне, областта й на действие, като са дискутирани особеностите й в българския език.

Представените модели на лингвистично описание, както и самите езикови ресурси (граматика и корпуси), допринасят не само за обективното типологическо сравнение на българския език с други езици, но и за подобряване на автоматичната обработка на българския език в морфосинтактичен, синтактичен и семантичен план.

Статии, свързани с тематиката на монографията, от които една в съавторство с Кирил Симов (разгледани са по-нататък в текста):

Osenova and Simov 2015: Petya Osenova and Kiril Simov. Universalizing BulTreeBank: a Linguistic Tale about Glocalization. In: *Proceedings of BSNLP 2015*, Hissar, Bulgaria, pp. 81–89;
Осенова 2014**:** Петя Осенова. Славянската памет: Българският език и Граматиката на зависимостите. В: Амелия Личева, Кристина Йорданова, Милена Кирова, Надежда Стоянова, Петя Осенова (съставители) "*Езици на паметта в литературния текст", Сборник Доклади от годишната конференция на факултет „Славянски филологии“*, СУ „Св. Климент Охридски“, 2013, стр. 548–555, издателство ФАБЕР. Велико Търново. ISBN 978-619-00-0111-9.
Осенова 2012: Петя Осенова. Семантично моделиране на българските части на речта в Опорната фразова граматика. В: Бурова, Ани, Иванова, Диана, Христова, Елена, Димитрова, Славея, Аврамова, Цветанка (съст.). *Време и история в славянските езици, литератури и култури. Сборник с доклади от 11-те национални славистични четения 19–21 април, 2012. Том 1. Езикознание*. УИ „Св. Кл. Охридски”, стр. 35–41.
Osenova 2011: Petya Osenova. Localizing a Core HPSG-based Grammar for Bulgarian. In: Hanna Hedeland, Thomas Schmidt, Kai Worner (eds.) *Multilingual Resources and Multilingual Applications*, *Proceedings of GSCL 2011*, ISSN 0176-599X, Hamburg, pp. 175–180.

II **Статии**

Статиите са разделени условно на три категории, които са по тематиката на професурата: *морфология, синтаксис* и *корпусна лингвистика*, но трябва да се има предвид, че в много от случаите трите компонента се преплитат. По тази причина някои статии са разгледани в повече от една тематична област.

Допълнително са обособени още две категории: *обучение* и *общи*.

Морфология

В статиите (в съавторство с Кирил Симов) Simov and Osenova 2015: Kiril Simov and Petya Osenova. Catena Operations for Unified Dependency Analysis. In: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 320–329, Uppsala, Sweden, August 24–26 2015; Osenova and Simov 2015: Petya Osenova and Kiril Simov 2015: Modeling Lexicon-Syntax Interaction with Catenae. In: *Journal of Cognitive Science*, vol. 16/3, pp. 287–322. Seoul National University, College of Humanities. ISSN: 1598-2327; Simov and Osenova 2014: Kiril Simov and Petya Osenova. Formalizing MultiWords as Catenae in a Treebank and in a Lexicon. In: Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, Adam Przepiórkowski (eds.) *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, December 12-13, 2014, Tübingen, Germany, pp. 198–207. ISBN: 978-3-9809183-9-8; Osenova and Simov 2014: Petya Osenova and Kiril Simov. Treatment of Multiword Expressions and Compounds in Bulgarian. In: Verena Henrich and Erhard Hinrichs (eds.) *Proceedings of the Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014)*, pp. 41–46, ESSLLI, Tuebingen, Germany за първи път се въвежда представянето на т. нар. многокомпонентни думи (multiwords) в българския език чрез описание с катена. Катената служи за представяне на информация в синтаксиса и в морфологията. Катените са свързани непълни дървета (поддърво), които представят под-конституетно равнище в синтаксиса. В морфологията те се използват за анализ на многокомпонентни думи и за анализ на морфемната структура на думите.

И трите статии предлагат кодиране на многокомпонентните думи в речника и начини за свързването им със синтактичното поведение и семантиката на тези думи. Статиите нямат за цел да дискутират видовете многокомпонентни думи, макар че има различни схващания за обхвата на групата. Типичните представители са: съставни части на речта (*който и да е*; *може би; за да* и др.); идиоми (*гушнах букета*); имена (*Иван Иванов*); леки глаголни конструкции (*имам надежда*) и др.

Приносът на изследванията е в представяне на взаимодействието между поведението на тези думи в речника и текста чрез катена. Това означава, че думите се разделят на непроменящи се (при които не може да се вмъкне допълнителен материал) и променящи се (при които могат да се добавят модификатори). В речника се указват местата, където катената може да се разкъса или разшири. От морфологична гледна точка се дават основните форми на елементите на многокомпонентните думи, възможните граматически характеристики (ако има изменяеми елементи) и информация при кои части на речта са възможни разширения или промяна на граматическите характеристики.
В статиите Osenova and Simov 2015: Petya Osenova and Kiril Simov 2015: Modeling Lexicon-Syntax Interaction with Catenae. In: *Journal of Cognitive Science*, vol. 16/3, pp. 287–322. Seoul National University, College of Humanities. ISSN: 1598-2327 и Osenova and Simov 2014: Petya Osenova and Kiril Simov. Treatment of Multiword Expressions and Compounds in Bulgarian. In: Verena Henrich and Erhard Hinrichs (eds.) *Proceedings of the Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014)*, pp. 41–46, ESSLLI, Tuebingen, Germany, които използват наблюденията ми, изложени в статията Осенова 2012: Петя Осенова. Синтаксисът на думите. В: *Сб. „Магията на думите", Езиковедски изследвания в чест на проф. д.ф.н. Лилия Крумова-*

Цветкова. Академично издателство „проф. Марин Дринов", стр. 280–286, София, ISBN 978-954-322-521-7, чрез катена се представят връзките в речника между сложно съществително с два корена, съответстващ сложен глагол с два корена и синтактично съчетание. Напр. *ръкомахане − ръкомахам − махам с ръка*. Чрез механизма на катената морфите се свързват с елементите в синтактичното съчетание. Напр. *рък-* се съотнася с ръка; *маха-* се съотнася с *махам*, а в синтактичната катена се предвижда и предлогът *с*. Заключението е, че морфологичните катени, каквито са сложните думи, са стабилни откъм състав, докато при синтактичните може да се вмъква допълнителен материал. Морфемният анализ е сведен до кореновите морфеми. По-нататъшна задача е този анализ да се разшири и с участието на останалите морфеми (афикси, флексия и т.н.)

В статията Осенова 2012: Петя Осенова. Синтаксисът на думите. В: *Сб. „Магията на думите", Езиковедски изследвания в чест на проф. д.ф.н. Лилия Крумова-Цветкова.* Академично издателство „проф. Марин Дринов", стр. 280–286, София, ISBN 978-954-322-521-7 се разглежда съотнасянето на морфите в сложни (двукоренни) глаголи и съществителни към съответстващите им синтактични съчетания. Напр. *водоснабдявам = снабдявам с вода; снеговалеж = вали сняг*. Приносът на статията е в сравнението между сложни глаголи и сложни съществителни по отношение на реализация на вътрешни и външни аргументи. Оказва се, че сложните глаголи с вътрешни аргументи и адюнкти не са типични в българския език (подлогът също няма тенденция да се реализира), докато при сложните съществителни с отглаголен корен подлогът се реализира вътрешно като косвен аргумент, а също така по-често се срещат реализирани останалите аргументи и адюнктите. Това означава, че вербалността предпочита аналитичното изразяване, докато номиналността – синтетичното. В контрастивен план и по-задълбочено подобни проблеми са разисквани в книгата Verb-centered Compound Nouns in English and Bulgarian, Maria Kolarova, Sofia, St. Kl. Ohridski University Press, 2015.

Статията (в съавторство) Savkov et al. 2012: Aleksandar Savkov and Laska Laskova and Stanislava Kancheva and Petya Osenova and Kiril Simov. Linguistic Analysis Processing Line for Bulgarian. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. ELRA. 978-2-9517408-7-7 представя автоматично средство за морфосинтактичен анализ (чрез речник и статистически методи) и снемане на морфосинтактичната многозначност (чрез правила и статистически методи). Моят принос в нея е в подготовката на данните (корпус и правила) за тестване на морфологичния анотатор и също така при анализа на грешките след действието на автоматичните модули.

Статията (в съавторство) Georgiev et al. 2012: Georgi Georgiev, Valentin Zhikov, Kiril Simov, Petya Osenova and Preslav Nakov. Feature-Rich Part-of-speech Tagging for Morphologically Complex Languages: Application to Bulgarian. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, April 23–27, 2012, Avignon France pp. 492–502 описва автоматична програма за морфологичен анализ с голям набор от морфосинтактични характеристики за българския език (680 на брой като комбинации от част на речта и нейните граматически характеристики). Както е известно, българският език е морфологично богат език. Идеята е да се използват всички тези характеристики при анализа, което е трудна задача, защото са много. По-лесно би било да се направи анализ само по части на речта например. Моят принос е в описанието на типовете многозначности, честотата на срещане на най-честите думи в корпус, както и факторите, които влияят на избора на правилен морфосинтактичен етикет: локалност

срещу нелокалност на граматическите характеристики; взаимодействието между граматическите характеристики и спецификата на жанра или тематичната област, в която е текстът. Друг принос имам при анализа на грешките при морфосинтактичния анализ с оглед на лематизацията (привеждане на словоформата в основната форма). Оказва се, че ако правилно са разпознати характеристиките, свързани с лемата, лематизацията се извършва успешно дори ако има други грешни характеристики. Ако например за даден глагол са определени правилно вида и транзитивността, тогава и лемата се определя правилно, независимо че времето (аорист или имперфект) може да е разпознато грешно.

В статията Осенова 2012: Петя Осенова. Семантично моделиране на българските части на речта в Опорната фразова граматика. В: Бурова, Ани, Иванова, Диана, Христова, Елена, Димитрова, Славея, Аврамова, Цветанка (съст.). *Време и история в славянските езици, литератури и култури. Сборник с доклади от 11-те национални славистични четения 19–21 април, 2012. Том 1. Езикознание*. УИ „Св. Кл. Охридски", стр. 35–41 се разглеждат частите на речта в Опорната фразова граматика и с акцент върху семантиката. Тази статия е свързана с монографията**,** където проблемите са представени по-детайлно. В статията се описва моделирането на пълнозначните и функционалните части на речта от гледна точка на това дали са референти (съществителните и личните местоимения), или са събития (останалите части на речта). Показва се, че от семантична гледна точка представителите на една и съща част на речта се делят на различни групи. Така например, типичните съществителни не са модификатори, но има семантични групи при тях, които са (напр. за време – *вторник* ***сутринта***); прилагателните и наречията се разделят на пресичащи модификатори и модификатори с обхват; съюзите-субординатори се разглеждат като модификатори с обхват, докато съюзите-комплементизатори нямат своя собствена семантика и т.н. От друга страна, граматическите характеристики се моделират на две равнища – морфосинтактично и семантично. Това се отнася за: род, число, членуване, вид, наклонение и др. В повечето случаи двете равнища се синхронизират, но понякога има разминаване. Така например в израза *Негово величество* граматическият род е среден, но семантичният е мъжки.

В статията Осенова 2011: Петя Осенова. Видове морфологични многозначности в българския език. В: Красимира Алексова, Венче Попова и Тодор Бояджиев (съст.) *Сборник „Научни трудове в памет на Георги Герджиков"*, УИ „Св. Кл. Охридски", стр. 223–230 са описани основните типове морфологични многозначности от перспективата на отношението между лексема и словоформа на основата на морфологично анотиран корпус. Тя не е свързана с монографията. Въвежда се терминът *морфологична многозначност*, който включва конверсията, граматическата омонимия, синкретизма и др. Отделени са 4 вида многозначности: (а) в рамките на словоформите на една лексема (*2 дивана – седнах на дивана*); (б) между две или повече лексеми (*като* – предлог и *като* – съюз); (в) между лексема и словоформата на друга лексема (*лоша политика – кажете на политика*) и г) между словоформите на две или повече лексеми (*вие вървите бързо – вървите се заплетоха*). Представен е броят срещания на най-честите многозначности. Ценно е да се види каква е честотата на участниците в многозначността. Обикновено те не са равнопоставени, защото зависят от контекста и корпуса. След това е дискутирана ролята на многозначностите спрямо честотата на еднозначните думи, както и случаите на рядко срещани типове многозначности. По-късно това изследване става основа за дисертацията на Станислава Кънчева „Морфологичната многозначност в съвременния български език", защитена успешно през 2015 г.

Синтаксис

В статиите (в съавторство с Кирил Симов) Simov and Osenova 2015: Kiril Simov and Petya Osenova. Catena Operations for Unified Dependency Analysis. In: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 320–329, Uppsala, Sweden, August 24–26 2015; Osenova and Simov 2015: Petya Osenova and Kiril Simov 2015: Modeling Lexicon-Syntax Interaction with Catenae. In: *Journal of Cognitive Science*, vol. 16/3, pp. 287–322. Seoul National University, College of Humanities. ISSN: 1598-2327; Simov and Osenova 2014: Kiril Simov and Petya Osenova. Formalizing MultiWords as Catenae in a Treebank and in a Lexicon. In: Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, Adam Przepiórkowski (eds.) *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, December 12-13, 2014, Tübingen, Germany, pp. 198–207. ISBN: 978-3-9809183-9-8; Osenova and Simov 2014: Petya Osenova and Kiril Simov. Treatment of Multiword Expressions and Compounds in Bulgarian. In: Verena Henrich and Erhard Hinrichs (eds.) *Proceedings of the Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014)*, pp. 41–46, ESSLLI, Tuebingen, Germany, както беше посочено и по-горе, се разглежда представянето на езиковите елементи чрез т.нар. *катена*, или още поддърво. Дадена е формализация на явлението *катена*. От синтактична гледна точка катената е удобно средство за анализ на езикови конструкти, които не съвпадат с традиционните конституенти (някои типове многокомпонентни думи) или са проблемни за анализ (елипси, глаголни комплекси и др.).

В статията Simov and Osenova 2014: Kiril Simov and Petya Osenova. Formalizing MultiWords as Catenae in a Treebank and in a Lexicon. In: Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, Adam Przepiórkowski (eds.) *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, December 12-13, 2014, Tübingen, Germany, pp. 198–207. ISBN: 978-3-9809183-9-8 се представя моделирането на многокомпонентните думи в синтактичен ресурс (Бултрибанк) и речник. Първо са представени по-честите типове многокомпонентни думи в ресурса (AN (*вътрешен министър*); NpN (*среща на върха*) V NP (*затварям си очите* [*пред фактите*])). След това е показано моделирането им в речника и текста. Чрез катената е възможно да се осигури не само фигуративното значение на многокомпонентните думи, но и буквалното им значение. Срв. *Затварям си очите, защото ми се спи* и *Затварям си очите пред фактите*. В първия случай частите на израза са свободни, а във втория са свързани чрез катена. В лексикона се кодира катената, основната ѝ форма, валентността, семантиката, както и възможностите за модификация. По този начин се прогнозират възможните съчетаемости на катената в текста.

Статията Simov and Osenova 2015: Kiril Simov and Petya Osenova. Catena Operations for Unified Dependency Analysis. In: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 320–329, Uppsala, Sweden, August 24–26 2015 също разглежда кодирането на многокомпонентните думи в речника, като въвежда термина *лексикална катена*. Потиснатата морфосинтактична информация в речника позволява различни реализации на многокомпонентната дума (напр. съществителното *среща* може да се употреби определено – *срещата на върха* или в множествено число – *срещи/те на върха*). Допълнително се описват две операции: композиция и разширяване – които са в основата на изграждане на синтактичния анализ, базиран на катената.

Статията Osenova and Simov 2014: Petya Osenova and Kiril Simov. Treatment of Multiword Expressions and Compounds in Bulgarian. In: Verena Henrich and Erhard Hinrichs (eds.) *Proceedings of the Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014)*, pp. 41–46, ESSLLI, Tuebingen, Germany разглежда катената като подходящо средство за кодиране както на многокомпонентни думи, така и на връзката между сложни думи с техните синтактични съчетания (*билколечение = лекувам с билки*). Отбелязва се фактът, че моделирането на модификацията в този контекст не е тривиална задача.

Статията Osenova and Simov 2015: Petya Osenova and Kiril Simov 2015: Modeling Lexicon-Syntax Interaction with Catenae. In: *Journal of Cognitive Science*, vol. 16/3, pp. 287–322. Seoul National University, College of Humanities. ISSN: 1598-2327 най-пълно и подробно описва проблемите, поставени в другите статии по тематиката. Акцентът е върху взаимодействието между представянето на думите в речника и синтактичното им поведение в текста. По-прецизно е представена формализацията на катената. По-подробно са описани предишните разработки по темата за катената, както и кодирането на речниковата катена с нейните морфосинтактични характеристики, валентност и семантика.

Тематиките на горните статии не са включени в монографията.

Статиите (в съавторство) Osenova and Simov 2015: Petya Osenova and Kiril Simov. Universalizing BulTreeBank: a Linguistic Tale about Glocalization. In: *Proceedings of BSNLP 2015*, Hissar, Bulgaria, pp. 81–89; Osenova and Simov 2015: Petya Osenova and Kiril Simov 2015: Semantic Role Annotation in BulTreeBank. In: Markus Dickinson, Erhard Hinrichs Agnieszka Patejuk and Adam Przepiórkowski (eds.) *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, 11–12 December 2015 Warsaw, Poland, pp. 148–156. ISBN: 978-83-63159-18-4; Osenova et al. 2012: Petya Osenova, Kiril Simov, Laska Laskova and Stanislava Kancheva. A Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri and Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. ELRA. 978-2-9517408-7-7, pp. 2636-2640 и Osenova and Simov 2011: Petya Osenova and Kiril Simov. Syntactic-Semantic Treebank for Domain Ontology Creation. In: *Cognitive Studies*, vol. 11, Warsaw, pp. 213–225 описват въпроси, които се отнасят до синтактичния ресурс Бултрибанк.

Статиите (в съавторство) Simov and Osenova 2012: Kiril Simov and Petya Osenova. Bulgarian-English Treebank: Design and Implementation. In*: Linguistic Issues in Language Technology*, vol. 7, issue 14, 2012 и Simov et al. 2011: Kiril Simov, Petya Osenova, Laska Laskova, Aleksandar Savkov and Stanislava Kancheva. Bulgarian-English Parallel Treebank: Word and Semantic Level Alignment. In *Proceedings of the Second AEPC Workshop*, 15.09.2011, at RANLP 2011 разглеждат проблемите на българско-английския паралелен синтактичен ресурс.

Статията Осенова 2014**:** Петя Осенова. Славянската памет: Българският език и Граматиката на зависимостите. В: Амелия Личева, Кристина Йорданова, Милена Кирова, Надежда Стоянова, Петя Осенова (съставители) "*Езици на паметта в литературния текст", Сборник Доклади от годишната конференция на факултет „Славянски филологии“*, СУ „Св. Климент Охридски“, 2013, стр. 548–555, издателство ФАБЕР. Велико Търново. ISBN 978-619-00-0111-9 представя дependентната синтактична рамка на анализ, а

статията Осенова 2011: Петя Осенова. Представянето на синтаксиса в средното училище и в университета. В: *Littera et lingua* (електронно списание), есен 2011 (http://www.slav.uni-sofia.bg/naum/lilijournal/2011/8/3/osenovap) е посветена на връзката между синтактичната информация, представена в средното училище и университета.

Статията Osenova 2011: Petya Osenova. Localizing a Core HPSG-based Grammar for Bulgarian. In: Hanna Hedeland, Thomas Schmidt, Kai Worner (eds.) *Multilingual Resources and Multilingual Applications*, *Proceedings of GSCL 2011*, ISSN 0176-599X, Hamburg, pp. 175–180 описва особеностите на българската ресурсна граматика, а статията (в съавторство) Simov and Osenova 2011: Kiril Simov and Petya Osenova. Towards Minimal Recursion Semantics over Bulgarian Dependency Parsing. In *Proceedings of Recent Advances in Natural Language Processing*, pages 471–478. Hissar, Bulgaria, 12-14 September 2011 разглежда подробно създаването на семантични структури на основата на дependентен синтактичен анализ за българския език.

Статията Osenova et al. 2010: Petya Osenova, Laska Laskova, Kiril Simov. Exploring Co-Reference Chains for Concept Annotation of Domain Texts. In: *Proceedings from LREC 2010*, Malta, pp. 172–176 описва подход, при който анотирането с понятия в текста става по-плътно чрез използване на кореферентните вериги.

Следват конкретни резюмета на горните статии.

Статията (в съавторство) Osenova and Simov 2015: Petya Osenova and Kiril Simov. Universalizing BulTreeBank: a Linguistic Tale about Glocalization. In: *Proceedings of BSNLP 2015*, Hissar, Bulgaria, pp. 81–89 е пряко свързана с монографията. Статията представя основните принципи на прекодиране на езиковия модел от оригиналния корпус Бултрибанк в идеологията на инициативата за универсалните дependентности (http://universaldependencies.org/). Дискутирани са проблемите в морфологичен и синтактичен аспект при пренасяне на лингвистичното знание от една рамка на представяне (конституентна) към друга (дependентна). Новата рамка поставя Бултрибанк в среда на сравнимост с други езици и дава възможност за сравняване на автоматични морфологични и синтактични анализатори, обучени върху ресурса, с автоматичните средства за други езици. Описани са подходите при конвертирането. Те включват автоматичен модул, който използва правила, и ръчен модул, при който се налага санкция от специалист в случаите на по-трудните езикови явления (координация, апозиция и др.) или разминаване на анотационните схеми (напр. при Бултрибанк схемата е по-морфологично и формално насочена, докато при универсалните дependентности е по-синтактично и семантично изградена). Първоначалните експерименти с новия ресурс показват подобрение на морфосинтактичното тагиране, но лек спад в синтактичното, което може да се дължи на факта, че все още не целият ресурс е прехвърлен в новата схема.

Тематиката на статията (в съавторство) Osenova and Simov 2015: Petya Osenova and Kiril Simov 2015: Semantic Role Annotation in BulTreeBank. In: Markus Dickinson, Erhard Hinrichs Agnieszka Patejuk and Adam Przepiórkowski (eds.) *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, 11–12 December 2015 Warsaw, Poland, pp. 148–156. ISBN: 978-83-63159-18-4 не е включена в монографията. Приносът на статията е анотирането на синтактичния корпус Бултрибанк със семантични роли върху аргументите на предиката – подлог, пряк и непряк обект. Този процес е свързан с предварителна работа по ресурса, която се изразява в следното: анотиране на текстовете със значения от лексикалната база WordNet (този процес е описан в статията Popov et al. 2014: Alexander Popov, Stanislava Kancheva, Svetlomira Manova, Ivaylo Radev, Kiril Simov and Petya Osenova 2014: The Sense Annotation of BulTreeBank. In: Verena

Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, Adam Przepiórkowski (eds.) *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, December 12–13, 2014, Tübingen, Germany, pp. 127–136. ISBN: 978-3-9809183-9-8 в частта за корпусна лингвистика) и извличане на валентен речник (речникът е описан в статията Osenova et al. 2012: Petya Osenova, Kiril Simov, Laska Laskova and Stanislava Kancheva. A Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri and Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. ELRA. 978-2-9517408-7-7, pp. 2636-2640, която се разглежда в тази част). След анотирането на корпуса със значения за задачата се използват лексикографските класове на глаголите (когнитивни глаголи, глаголи за движение, прецепция, притежание и др.). Наборът от семантични роли е взет от лексикалния ресурс VerbNet. Стратегията на анотиране също е приносна. Тя включва следните стъпки: избиране на тип глаголен клас; оформяне на йерархия при валентните рамки (например рамката на глагола *чета* с подлог – *аз чета* – е по-обща от рамката с пряк обект – *аз чета книга*); начално по-общо анотиране със семантични роли; тъй като спецификациите във валентния речник са свързани с изреченията в ресурса, анотациите директно се пренасят в текста; преглеждане и детайлизиране или поправяне на приписаните семантични роли. Идеята беше тествана върху няколко класа глаголи (за метереологично време (verb.weather), за човешко състояние (verb.body) и за консумация/разход (verb.consumption)), като показва адекватността на подхода.

Статията (в съавторство) Osenova et al. 2012: Petya Osenova, Kiril Simov, Laska Laskova and Stanislava Kancheva. A Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri and Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. ELRA. 978-2-9517408-7-7, pp. 2636-2640 не е свързана с монографията. Тя описва процеса на извличане на валентен речник от синтактично анализирания корпус Бултрибанк. Този речник е свързан с онтология. Показани са и са коментирани най-честите валентни типове в ресурса и най-честите онтологични ограничения върху аргументите на предиката. Ограниченията са малко на брой и са изразени чрез по-абстрактни етикети като *Събитие, Лице, Обект, Артефакт, Социален обект* и др. Валентните рамки отразяват употребата си в конкретното изречение.

Статията Osenova and Simov 2011: Petya Osenova and Kiril Simov. Syntactic-Semantic Treebank for Domain Ontology Creation. In: *Cognitive Studies*, vol. 11, Warsaw, pp. 213–225 не е свързана с монографията. Статията описва създаването на синтактичен ресурс в определена тематична област (текстил и вътрешен дизайн) за целите на създаване на подходяща онтология в същата област чрез извличане на подходящи релации (*направен-от; част-от; служи-за; характеризиран-чрез* и др.). За база на синтактичния корпус служат стандарти и терминологични речници в областта. Текстовете са сегментирани, а синтактичните анализи са добавени ръчно. Поради малкия обем на корпуса и поради анотирането само на определени синтактични структури (именни групи, глаголни групи, предложни групи и подчинени изречения), за които се предполага, че имат релации, ръчната работа е подходяща заради високата точност. След това релациите са извадени по синтактични модели полуавтоматично.

Статията Simov and Osenova 2012: Kiril Simov and Petya Osenova. Bulgarian-English Treebank: Design and Implementation. In*: Linguistic Issues in Language Technology*, vol. 7, issue 14, 2012 не е свързана с монографията. Тя представя анотационна схема за българско-английски паралелен синтактичен корпус. Данните са анализирани автоматично на синтактично ниво, а след това са подравнени ръчно на равнище дума, за да се съпоставят успешно семантичните структури, изградени на основата на Минималната рекурсивна семантика. Английската част на ресурса е анализирана с ресурсна граматика, а българската част използва хибриден метод поради недостатъчното покритие на граматиката. Хибридният метод, който включва изграждане на семантични структури на основата на информацията от дependентните анализи, е приносен. След това семантичните структури са подравнени една с друга чрез правила. Описани са и основните типове грешки при това подравняване, като източникът за сравнение е английският корпус.

Статията Simov et al. 2011: Kiril Simov, Petya Osenova, Laska Laskova, Aleksandar Savkov and Stanislava Kancheva. Bulgarian-English Parallel Treebank: Word and Semantic Level Alignment. In *Proceedings of the Second AEPC Workshop*, 15.09.2011, at RANLP 2011, ISBN 978-954-452-021-2, pp. 29–38 не е свързана с монографията. Тази статия също разглежда построяването на българско-английски паралелен синтактичен корпус, но фокусът е предимно върху правилата за подравняване между думи и изрази на двата езика. Тези правила са принос към сравняването на явленията между два езика с оглед на машинния превод. Съпоставянето на семантични структури тук е направено само за малък брой изречения, които са анализирани съответно от двете ресурсни граматики. Така приносите в тази статия се явяват подготвителна фаза за резултатите в статията Simov and Osenova 2012: Kiril Simov and Petya Osenova. Bulgarian-English Treebank: Design and Implementation. In*: Linguistic Issues in Language Technology*, vol. 7, issue 14, 2012.

Статията Осенова 2014**:** Петя Осенова. Славянската памет: Българският език и Граматиката на зависимостите. В: Амелия Личева, Кристина Йорданова, Милена Кирова, Надежда Стоянова, Петя Осенова (съставители) "*Езици на паметта в литературния текст", Сборник Доклади от годишната конференция на факултет „Славянски филологии“*, СУ „Св. Климент Охридски“, 2013, стр. 548–555, издателство ФАБЕР. Велико Търново. ISBN 978-619-00-0111-9 е тясно свързана с монографията. Статията дискутира някои особености на Граматика на зависимостите, или още Дependентна граматика, в славянски контекст. След това се разглежда мястото на българския език като славянски и на българската граматическа традиция в този вид граматики. Приносните моменти са във въвеждането на нова теоретична рамка за граматическо моделиране на българския език, която не е била прилагана до момента.

Статията Осенова 2011: Петя Осенова. Представянето на синтаксиса в средното училище и в университета. В: *Littera et lingua* (електронно списание), есен 2011 (http://www.slav.uni-sofia.bg/naum/lilijournal/2011/8/3/osenovap) не е свързана с монографията. В нея се прави сравнение на дефинициите на основни синтактични термини, въведени в средното училище и в университета. Направен е извод, че няма пропаст между представянията на синтаксиса на двете места. В същото време обаче, отчитайки необходимостта от приспособяване на терминологията за средното училище, се пояснява, че е наложително да се избягват частични дефиниции (напр. че подлогът е вършител на действието) и смесване на езикови равнища (напр. приравняването на глагола със сказуемото).

Статията Osenova 2011: Petya Osenova. Localizing a Core HPSG-based Grammar for Bulgarian. In: Hanna Hedeland, Thomas Schmidt, Kai Worner (eds.) *Multilingual Resources and Multilingual Applications*, *Proceedings of GSCL 2011*, ISSN 0176-599X, Hamburg, pp. 175–180 е тясно свързана с тематиката на монографията. Тя описва параметрите на локализиране на идеята за ресурсна граматика спрямо българския език. Локализирането включва както цялостната архитектура (речник, принципи, йерархия от обекти), така и тестовото множество от преведени на български изречения за измерване на точността и покритието на граматиката.

Статията Simov and Osenova 2011: Kiril Simov and Petya Osenova. Towards Minimal Recursion Semantics over Bulgarian Dependency Parsing. In *Proceedings of Recent Advances in Natural Language Processing*, pages 471–478. Hissar, Bulgaria, 12-14 September 2011 не е свързана пряко с монографията. Тя е тематично свързана със статиите Simov and Osenova 2012: Kiril Simov and Petya Osenova. Bulgarian-English Treebank: Design and Implementation. In: Linguistic Issues in Language Technology, vol. 7, issue 14, 2012 и Simov et al. 2011: Kiril Simov, Petya Osenova, Laska Laskova, Aleksandar Savkov and Stanislava Kancheva. Bulgarian English Parallel Treebank: Word and Semantic Level Alignment. In *Proceedings of the Second AEPC Workshop*, 15.09.2011, at RANLP 2011, ISBN 978-954-452-021-2, pp. 29–38, но акцентът не е върху паралелния ресурс, а върху българския. Подробно се описва стратегията за проектиране на семантични структури в Минималната рекурсивна семантика от дependентни синтактични структури. Представени са правилата за проекция за всички основни синтактични структури, като е детайлизирано откъде идва необходимата информация – от синтактичните релации, от морфосинтактичните, или и от двете. Направена е и първоначална оценка на алгоритъма с наличните данни. Постигат се около 77 % съвместимост на семантичните структури, които са проектирани чрез правила, с тези, които са резултат от анализа на ресурсната граматика (считани за златен еталон). Оценката показва нужда от промяна на някои правила за проекция, както и от инкорпорирането на голям валентен речник.

Статията Osenova et al. 2010: Petya Osenova, Laska Laskova, Kiril Simov. Exploring Co-Reference Chains for Concept Annotation of Domain Texts. In: *Proceedings from LREC 2010*, Malta, pp. 172–176 не е свързана с монографията. Тя описва стратегия за решаване на проблема с недостатъчното покритие с понятия от онтологията в текстове от областта на информационните технологии. За целта се прави анотиране с кореференции, с помощта на които понятийната информация се разпространява в текста. Напр. между *HTML страница* и кореферирания елемент *страницата*; между *XML езика* и местоимението *той*. За целите на оценката е направен златен корпус с ръчни анотации. Експериментите показват, че подобен подход помага за постигане на по-голяма анотационна плътност с понятия, но не е достатъчен за решаването на многозначността.

Корпусна лингвистика

Нито една от статиите в тази секция не е директно свързана с тематиката в монографията . Статиите покриват най-общо следните области: семантично анотиране; различни видове автоматична обработка на български текстове; корпуси, корпусно базирани изследвания и машинен превод.

Статиите Simov, Popov and Osenova 2015: Kiril Simov, Alexander Popov and Petya Osenova. Improving Word Sense Disambiguation with Linguistic Knowledge from a Sense

Annotated Treebank. In: *Proceedings of Recent Advances in Natural Language Processing*, pp. 596–603, Hissar, Bulgaria, Sep 7–9 2015 и Popov et al. 2014: Alexander Popov, Stanislava Kancheva, Svetlomira Manova, Ivaylo Radev, Kiril Simov and Petya Osenova 2014: The Sense Annotation of BulTreeBank. In: Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, Adam Przepiórkowski (eds.) *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, December 12–13, 2014, Tübingen, Germany, pp. 127–136. ISBN: 978-3-9809183-9-8 дискутират въпроси, свързани с анотирания със значения синтактичен ресурс Булрибанк.

Първата от горните две статии използва корпуса за извличане на подходящи синтактични (синтагматични) релации, които да обогатят лексикалните (парадигматични) с повече знание в задачата за снемане на лексикална многозначност в текст. Установява се, че прибавянето на синтактични релации при граф-методите подобрява резултатите. Моят принос е в извличане на синтактични релации от Булрибанк и интерпретиране на резултатите. Втората статия проследява анотирането на Булрибанк със значения от лексикалния ресурс WordNet-BTB и от DBPedia. Тези дейности за приносни като ресурси за българския език от гледна точка на комбиниране на синтаксис, семантика и знания за света. Анотирани са пълнозначните части на речта: съществителни, глаголи, наречия и прилагателни. В статията се дискутират и причините за разминаване при решенията на анотаторите (степени на грануларност; липса на ясен контекст и др.).

Следват статии, посветени на различни модули за автоматичната обработка на българския език.

Статията Ghayoomi et al. 2014: Masood Ghayoomi, Kiril Simov and Petya Osenova. Constituency Parsing of Bulgarian: Word- vs Class-based Parsing. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 26–31 May, Reykjavik, Iceland, pp. 4056–4060 представя резултатите от два конституенти парсера, тренирани върху оригиналния ресурс Булрибанк. Поради неголемия размер на ресурса е използван метод за клъстъризация на думите, който намалява проблема с редките срещания на някои словоформи. Самите експерименти с адаптиране на станфордския парсер и парсера на Бъркли за конституентен анализ са приносни, защото в момента за българския език (а и в световен мащаб) основната линия и най-добрите резултати са на дependентните парсери. Моят принос е в подготовка на данните и оценката на резултатите.

Статията Zhikov et al. 2013: Valentin Zhikov, Georgi Georgiev, Kiril Simov and **Petya** Osenova. Combining POS tagging, Dependency Parsing and Coreferential Resolution for Bulgarian. In: Galia Angelova, Kalina Boncheva and Ruslan Mitkov (eds) *Proceedings of RANLP 2013*, pp. 755–762. ISSN 1313-8502 описва експерименти, свързани с модул, който комбинира решаването на три задачи едновременно – морфосинтактично тагиране, дependентен анализ и кореферентно свързване. Принос е не само комбинирането на тези три задачи за решаване в едно, но и фактът, че самият модел е езиково независим. Комбинираният модел постига подобни резултати на тези, които се получават от най-добрия морфологичен тагер за българския език, и по-високи резултати от тези на синтактичния парсер, включен в интегрирания модул за обработка на български текстове. Моят принос е в определянето на морфосинтактичните характеристики, които се използват в модела.

Статията Savkov et al. 2012: Aleksandar Savkov and Laska Laskova and Stanislava Kancheva and Petya Osenova and Kiril Simov. Linguistic Analysis Processing Line for Bulgarian. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. ELRA. 978-2-9517408-7-7 описва цялостния модул за автоматична обработка на български текстове. Този модул включва сегментация, разделяне на изречения, морфологичен анализ, лематизация и синтактичен анализ. Предимство е, че освен в неговата цялост като процеси модулът може да се използва и чрез отделните си модули. Моят принос е в подготовката на лингвистичните правила за снемане на многозначност като част от цялостния модул.

Статията Savkov, Laskova, Osenova, Simov and Kancheva 2011. A. Savkov, L. Laskova, P. Osenova, K. Simov and S. Kancheva. A Web-based Morphological Tagger for Bulgarian. In: D. Majchrakova and R. Garabik (eds.) *Proceedings of the 6th Interantional Conference*, Modra, Slovakia, 20-21 October, Natural Language Processing, Multilinguality. Tribun, EU, pp. 126–137 се фокусира по-конкретно върху модула за морфологична анотация, но представен като уеб услуга. Моят принос е отново в подготовката на множеството от правила и също при анализа на грешки.

Статията Osenova and Simov 2010: Petya Osenova and Kiril Simov. Using the Linguistic Knowledge in BulTreeBank for the Selection of the Correct Parses. In: M. Dickinson, K. Muurisep and M. Passarotti (eds), *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT)*, 3–4 December 2010. NEALT Proceeding Series, Vol. 9, pp. 163–174 представя модел за трансфер на информация от Бултрибанк към синтактичния ресурс, парсиран от българската ресурсна граматика. Целта е да се избере правилният анализ, тъй като ресурсните граматики по правило продуцират голям брой анализи за всяко изречение. Оказва се, че 86 % от правилните анализи от ресурсната граматика могат да се изберат с информация от Бултрибанк. Моят принос е при писане на правилата за трансфер на лингвистична информация между двата ресурса. Проблеми при преноса създават координацията и комплементите.

Статията Georgiev et al. 2009: Georgi Georgiev, Preslav Nakov, Kuzman Ganchev, Petya Osenova, and Kiril Simov. Feature-Rich Named Entity Recognition for Bulgarian Using Conditional Random Fields. In: *Proceedings of the International Conference RANLP-2009*, RANLP 2009, Borovets, pp. 113–117 представя автоматичен модул за разпознаване и категоризиране на собствени имена чрез използване на богата лингвистична информация. Моят принос е в подготовката на данните и комбинирането на различните граматически характеристики (локални, нелокални и т.н.) и ресурси (речници). Резултатите за българския език се доближават до стандартите за английски и други езици.

Следващите статии са посветени на създаването и използването на различни корпуси или други видове ресурси, каквито са речниците и т.н.

Четири статии се отнасят до корпуса с политическа и журналистическа реч (http://www.political.webclark.org/). Статията Осенова и Терзийска 2013: Петя Осенова и Надя Терзийска. Някои наблюдения върху *мога, искам* и *трябва* в публичната реч на политиците. В: *Проблеми на устната комуникация*, Книга девета, I, стр. 219–229, Велико Търново представя наблюдения за използването на модалните глаголи *мога, искам* и *трябва* в речта на политиците. За изследването са използвани подкорпусите с парламентарна реч и журналистически интервюта. Наблюденията са представени в

сравнение с общ корпус, за да се видят особеностите на специализирания ресурс. Приложен е статистически анализ и конкорданс (т.е. наблюдаване на поведението на думата в контекст).

Статиите Osenova and Simov 2012: Petya Osenova and Kiril Simov. The Political Speech Corpus of Bulgarian. In: Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Mehmet Uğur Doğan and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. ELRA. 978-2-9517408-7-7 и Осенова и Симов 2011: Петя Осенова и Кирил Симов. Корпусен поглед към политическата реч. В: Вл. Миланов и Н. Михайлова, съст., сб „*Език, морал, отговорност*". УИ „Св. Кл. Охридски", стр. 77–87 описват анотационната схема на корпуса с политическа реч. Тя включва теми, изказвания на говорещите; лингвистична информация и маркиране на отношението (положително или отрицателно). Дискутирани са и възможни сценарии за употреба на корпуса, като например статистика на думи и модели; конкорданс; подреждане според избран критерий.

Статията Осенова 2012: Петя Осенова. Как говори политикът в Парламента и по време на интервю. В: Миланов, Вл. и Сталянова-Михайлова, Н. (съст.) *Езикови портрети на български политици. Част първа*. УИ „Св. Кл. Охридски", стр. 55–61 описва особеностите на говорене на политиците в парламента и по време на интервю. За целта на изследването е използвано средството конкорданс, както и честотата на употреба на езиковите елементи. Изследвани са речевите стратегии в двете ситуации и тенденциите в разколебаването на редица езикови норми.

Статията Симов и Осенова 2012: Кирил Симов и Петя Осенова. Инфраструктура за български езикови ресурси и технологии. В: Мария Стамболиева (съст.) *Компютърна лингвистика. Проблеми и перспективи*, кн. 1, АНАБЕЛА, стр. 198–220. ISSN 1314-6823 е статия-програма. Тя представя инфраструктура за езикови ресурси и технологии. По-точно казано, описан е основният, базисният[1] пакет от ресурси и технологии за българския език. Първо е въведена основната идеология на международната инициатива за езикови ресурси и технологии CLARIN. След това са описани и самите ресурси за българския език (корпуси, речници, списъци с имена). Накрая са представени и езиковите технологии (сегментатор, средство за разпознаване на собствени имена, морфологичен анализатор, средство за снемане на морфологичната многозначност, лематизатор, чънкери (или още частични граматики) и парсери). Направен е извод, че българският език не е сред езиците, които са 'ресурсно бедни'. Затова усилията трябва да се насочат към интегриране на данни и технологии, тяхното адаптиране за различни групи потребители и задачи.

Статията Giouli, Simov and Osenova 2011: Voula Giouli, Kiril Simov and Petya Osenova. A Parallel Greek-Bulgarian Corpus: A Digital Resource of the Shared Cultural Heritage. In: Sporleder, C., Van den Bosch, A., Zervanou, K. (eds.) *Language Technology for Cultural Heritage (selected papers from the LaTeCH Workshop* Series), Springer, pp. 99–112 описва паралелен гръцко-български корпус, който съдържа литературни и фолклорни източници на двата езика като част от наследството на културната ни памет. Представени са средствата за автоматична обработка на гръцката и българската част. Моят принос е в морфосинтактичното полуавтоматично анотиране на съответните български произведения, в адаптирането на модулите за автоматична обработка за тази задача, конструирането на

[1] Basic Language Resources Kit (BLARK).

речник на имената и в ръчната поправка на грешките след автоматичното обработване на текстовете.

Статията Simov and Osenova 2010: Kiril Simov and Petya Osenova. Constructing of an Ontology based Lexicon for Bulgarian. In: *Proceedings from LREC 2010*, Malta, pp. 3840–3844 представя концепция за изработване на онтологично базиран речник за българския език. Целта е да се направи връзка между граматичен и тълковен речник с обща онтология за целите на семантичното анотиране на текстове. Използват се две основни релации: равенство и включване. Имам принос към модела за свързване на текстови единици със заглавки от речник и понятия в онтологията.

Статиите Heeringa, Nerbonne and Osenova 2010: Wilbert Heeringa, John Nerbonne and Petya Osenova. Detecting contact effects in pronunciation. In: Muriel Norde, Bob de Jonge and C. Hasselblatt (eds.), *Language Contact. New perspectives*, John Benjamins Publishing Company, pp. 131–154 и Osenova et al. 2009: Petya Osenova, Wilbert Heeringa and John Nerbonne. A Quantative Analysis of Bulgarian Dialect Pronunciation. In: *Zeitschrift fuer Slavische Philologie.* Band 66, Heft 2, 2009, Universitatetsverlag WINTER Heidelberg разглеждат българските диалекти в контакт (първата статия) или в тяхното вътрешно разнообразие (втората статия). Първата статия изследва близостта на българските диалекти във фонетично отношение със съответстващото множество думи от официалните езици на съседните ни страни – Македония, Сърбия, Гърция, Турция, Румъния. Използвани са три метода на изследване: честота на фоните, честота на характеристиките и разстояние на Левенщайн. Също така е тествана хипотезата дали до съответната граница думите от съответния съседен език и нашите гранични диалекти са по-близки в сравнение със ситуацията във вътрешността на страната. Моят принос е в подготовката и кодирането на данните за българските диалекти и другите езици, както и интерпретацията на резултатите. Хипотезата се потвърждава за Македония, Сърбия и Румъния, но не за Турция и Гърция.
Втората статия изследва подробно фонетичната близост на българските диалекти. Данните са от Диалектния атлас на Стойков и покрива множество от 36 общи думи за 490 локации. Това е един от първите пъти, когато компютърни методи се прилагат върху български диалектни данни в такива мащаби. Моят принос е в дигитализирането на данните, кодирането на произносителните варианти от лингвистична и компютърна гледна точка, анализирането на резултатите от компютърните програми. Резултатите показват интересни детайли за българските диалекти, включително и чрез визуализация с карти. Например ятовата граница не изглежда еднакво плътна по протежението си; като особени изпъкват родопските диалекти и др.

Следващите статии разискват проблеми, свързани с автоматичния машинен превод. Статията Simov et al. 2015: Kiril Simov, Iliana Simova, Velislava Todorova, Petya Osenova. Factored Models for Deep Machine Translation. In: *Proceedings of the 1st Deep Machine Translation Workshop (DMTW 2015)*, pp. 97–105, Praha, Czech Republic, 3–4 September 2015 представя модел за машинен превод от български към английски и от английски към български. Лингвистичното знание, което включва морфосинтактична и синтактична информация, както и частично семантична информация, е кодирано като фактори в системата за машинен превод Моузес. Преводът е направен върху текстове от областта на информационните технологии. Експериментите показват, че добавянето на семантично знание и голям речник със словоформи подобрява резултатите. Моят принос е в подготовката на данните и на множествата от характеристики, както и в анализа на резултатите.

Статиите Wang et al. 2012: Rui Wang, Petya Osenova and Kiril Simov. Linguistically-Enriched Models for Bulgarian-to-English Machine Translation. *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)*, ACL 2012 / SIGMT / SIGLEX Workshop, 12 July 2012, Jeju, Korea, pp. 10–19 и Wang, Osenova and Simov 2012: Rui Wang, Petya Osenova and Kiril Simov. Linguistically-Augmented Bulgarian-to-English Statistical Machine Translation Model. In: *Proceedings of Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) And Hybrid Approaches to Machine Translation (HyTra)* at EACL-2012. ACL, ISBN 978-1-937284-19-0, pp. 119–128 описват модел за машинен превод от български език към английски. Тук са представени фактор моделите с лингвистична информация, които по-късно са разширени с повече характеристики и с посоката от английски към български в статията Simov et al. 2015: Kiril Simov, Iliana Simova, Velislava Todorova, Petya Osenova. Factored Models for Deep Machine Translation. In: *Proceedings of the 1st Deep Machine Translation Workshop (DMTW 2015)*, pp. 97–105, Praha, Czech Republic, 3–4 September 2015. В статията Wang et al. 2012: Rui Wang, Petya Osenova and Kiril Simov. Linguistically-Enriched Models for Bulgarian-to-English Machine Translation. *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)*, ACL 2012 / SIGMT / SIGLEX Workshop, 12 July 2012, Jeju, Korea, pp. 10–19 допълнително към автоматичната оценка на преводите е направена и експертна оценка от специалисти. Оценява се граматичността и съдържанието. Статията Wang, Osenova and Simov 2012: Rui Wang, Petya Osenova and Kiril Simov. Linguistically-Augmented Bulgarian-to-English Statistical Machine Translation Model. In: *Proceedings of Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) And Hybrid Approaches to Machine Translation (HyTra)* at EACL-2012. ACL, ISBN 978-1-937284-19-0, pp. 119–128 описва по-подробно подготовката на данните. Моят принос в двете свързани статии е в подготовката на данните и множествата от характеристики за експериментите, както и в подготовката на ръчното оценяване на преводите.

Следващите статии не са свързани пряко с тематиката на монографията. Те имат насоченост към обучението от различни гледни точки.

Статията Осенова 2014: Петя Осенова. Електронни ресурси за електронно обучение по български език. В: Таня Ангелова (съставител), *Интегриране на електронни форми на обучение в образователния процес по български език*. Изд. Парадигма, стр. 64–72 представя описание на българския синтактичен корпус Бултрибанк в конституентен и дependентн вид; българския валентен речник и българската ресурсна граматика за целите на обучението по синтаксис. По този начин студентите могат да се запознаят с различните стратегии за изграждане на определен ресурс, но най-вече с взаимовръзките между езиковите равнища и езиковите компоненти, както и възможностите за езиково моделиране. Сред представените три ресурса ресурсната граматика е този с най-висока сложност. Той предполага разбирането на другите два.

Статията Осенова 2013**:** Петя Осенова. Нормативност и (не)отклонения в езиковата практика (с оглед на чуждоезиковото обучение). В: Панайот Карагьозов и Юлияна Стоянова (съст.) „*Минало, настояще и перспективи на чуждестранната българистика*.” УИ „Св. Кл. Охридски”, стр. 218–224 дискутира отношението между езиковата нормативност и езиковата практика. Приносът на статията е в използването на корпуси за наблюдения върху проблемни езикови норми, а също и за извличане на полезни езикови формули (напр. от интервюта). Поддържам идеята, че от дадено равнище на езикова

компетентност нататък обучаваният трябва да бъде „изложен“ на въздействието на реалните текстове, използвани в реални езикови ситуации.

Статията Osenova and Simov 2010: Petya Osenova and Kiril Simov. Semantic Annotation for Semi-Automatic Positioning of the Learner. In: Proceedings of the First Workshop on Supporting eLearning with Language Resources and Semantic Data, at LREC 2010, Malta, pp. 46–50 представя стратегия за подпомагане на преподавателя в задачата му да определи равнището на знания и компетентност на учащия. Като изходни ресурси са използвани: конспектът, който покрива съдържанието в определена тематична област, и списък с въпроси и техните отговори в същата област като златен еталон. Тематичната област в случая е информационни технологии. Ресурсите са анотирани с понятия. Когато учащият отговори на даден въпрос, понятията, които използва и са разпознати от програмата, се сравняват с тези в златния еталон. После се прави списък с термините, които съвпадат; които липсват; които са добавени от учащия. Моделът може да се подобрява, но той представлява добро помощно средство на преподавателя в оценката му на знанията на учащите.

Статията Osenova 2011: Petya Osenova. Bulgarian. In: *The Languages of the new EU Member States, Revue Belge de Phologie et D'Historie*, *Fasc. 3: Languages et Litteratures Modernes*, pp. 643–668 не е свързана с тематиката на монографията. Тази статия може да бъде определена като позиционна. Тя има интегративен характер, защото описва факти за българския език като един от езиците на Европейския съюз по схема за описания, приложена за останалите езици в книгата. Дава се информация за името на езика; типологическите му характеристики; особеностите на фонетиката, морфологията и синтаксиса; особеностите на азбуката; историческите периоди на развитието му; географски обусловените езикови контакти; особености за диалектите; информация за носителите на езика; ресурси и институции, свързани с българския език; особености на културния живот, видян през призмата на емблематични книги и периодика.