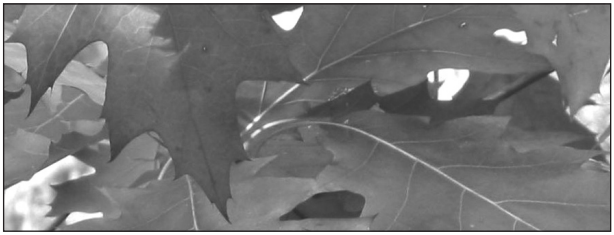
Елена Паскалева

КОМПЮТЪРНА МОРФОЛОГИЯ

РЕСУРСИ И ИНСТРУМЕНТИ



Институт за паралелна обработка на информацията, БАН
София, 2007

Описаните в тази книга разработки обхващат период от 20 години и са финансирани основно от бюджета на БАН. Подготовката и издаването на тази книга стана възможно благодарение на финансовата помощ на проекта BIS 21++ – Bulgarian IST Centre of Competence in 21 Century, Sixth Framework Programme, INCO-CT-2005-016639.

КОМПЮТЪРНА МОРФОЛОГИЯ

Ресурси и инструменти

Елена Паскалева

първо издание
ISBN 978-954-92148-1-9

Съдържание

От автора 5

Въведение 9

1 От текста към морфологичното му представяне 11

1.1 Компютърна идентификация на думата 13
1.2 Граматична идентификация на думата 14
1.3 Видове думи в компютърния текст 16
1.4 Извънезикови квазидуми 18
1.5 Извънлексикални квазидуми 19
1.6 Съкращения – идентификация и основни функции 20
1.7 Названия 22

2 Морфологични модели 35

2.1 Функции на морфологичния модел 36
2.2 Модели на морфологичен анализ и синтез – основно знание и процедури 39
2.3 Видове единици в ресурсите на модела 41
2.4 Операции върху ресурсите 42
2.5 Организация на формообразуващите елементи в парадигматична структура 43
2.6 Основен строителен материал на формообразуващия модел 46
2.7 Парадигматичната черупка – разпределителен пункт за категориални значения и буквените им изрази 48
2.8 Изчислимост и изброимост в морфологичния синтезиращ модел 56
2.9 Софтуерни приложения за морфологичен синтез 58

2.10 Морфологичен анализ vs. морфологичен синтез – условия за обратимост 62
2.11 Двупосочни морфологични модели 63
2.12 Двупосочен модел за български език на морфемно равнище 65
2.13 Преход от двупосочен към „плосък“ морфологичен модел 67
2.14 Организация на лингвистичното знание в „плоския“ морфологичен модел 69

3 Граматичен речник 71
3.1 Структура на граматичния речник 71
3.2 Съдържание на граматичния речник 74
3.3 Анотационно множество и подходи за определянето му 75
3.4 Два вида представяне на анотационното множество 77
3.5 Анотационно множество на български граматичен речник 82
3.6 Конвертиране на анотационни множества 106
3.7 Обем на анотационното множество 112

4 Пътищата след речника 109
4.1 И след речника – какво? 109
4.2 Напред към синтаксис\а 110
4.3 А без речник? 117
4.4 Без речник, но с лингвистично знание – към морфемния строеж на думата 117
4.5 Морфемният състав на думите и информационните системи 120

Заключение 135

Литература 137

От автора

Тази книга е за Думата.

За единицата – център на лингвистичното описание, която може да стане и център на един живот.

Тук не става дума за животоописание, макар че и научното описание може да има мемоарен характер – дори и за една така млада научна дисциплина като компютърната лингвистика, на чието раждане в България са бабували специалисти, живи и здрави и до ден днешен.

Автоматичната или компютърната обработка на един език, наричана днес в контекста на информационното общество със стряскащия и респектиращ термин *езикови технологии*, тръгва от Думата.

Така се започва – с обработката на първите текстове на първите изчислителни машини, минава се през завладяващата игра на подреждане и формализиране на нещата (скучно известни от нормативните граматики) в модели и експериментални компютърни програми.

Влезеш ли в царството на Думата не като случаен посетител, а като отговорен стопанин, оставаш омагьосан завинаги там.

Особено в един славянски език, дарен с високо развита морфология. За разлика от други езици тази морфология, след запознанството с нея, не ти оставя за спомен един-два признака в анотацията, за да минеш бързо към ефектния синтаксис.

Работата върху славянската, в случая българската Дума, никога не свършва.

Начините

- на нейното описание и представяне като езикова единица в строгата подредба на формализираното описание,
- на нейното събиране в различни колекции (колко вида речници съществуват само, представящи Думата от различни страни, и всичките в очакване на заслужена компютърна реализация),
- на откриването на онези нейни признаци, способни да улеснят електронните комуникации в най-естествената им форма – родния език,

предоставят широко пространство за работа, успехи и дълбоко задоволство от тях самите на отделни учени – млади и стари, както и на цели колективи.

Представените и обсъждани в тази книга решения – лингвистични и алгоритмични, натрупаните ресурси от български думи с различна функция и употреба, в различна степен формализирани и реализирани на компютър, многобройните софтуерни средства, конструирани както за описание на знанието ни за Думата, така и за подпомагане на други софтуерни реализации, са дело на научни колективи, интердисциплинарни в състава си, но обединени в общата си цел.

Тази цел в един взаимообучителен процес въведе софтуерните специалисти в екзотичната област на лингвистичните термини и възмутителната, но предизвикателна липса на формална яснота в граматическите описания. Лингвистите пък бяха внимателно и предпазливо въведени в азбуката на алгоритмичното представяне, в разликите между декларативното и процедурното знание.

Предпазливото навлизане в чуждата област бе в Началото. Днес знанието от двете области е „амалгамирано" не само в текстовете на научните статии по специалността, но и в понятийния и оперативен апарат на специалистите в тази вълнуваща дисциплина – езиковите технологии.

Взаимното „възпитание на чувствата" започна в Института по математика на БАН в началото на 80-те години, за да стане централна задача на създадената през 1987 година. Лаборатория за лингвистично моделиране в състава на Координационния център по информатика и изчислителна техника на БАН, по-късно преименуван в Централна лаборатория за паралелна обработка на информацията, а днес – Институт за паралелна обработка на информацията на БАН, Секция за лингвистично моделиране.

Само няколко години след създаването на това звено, поставеното начало в автоматичната обработка на българската Дума бе развито и продължено от многобройни научни изследвания и приложения в различни софтуерни продукти.

Нещата, описани в тази книга, са получили подкрепа и финансиране основно от бюджета на БАН и частично от международни научни проекти (твърде многобройни, за да бъдат цитирани подробно) по програмите на Европейската комисия – като се започне от Tempus и се мине през Copernicus и рамковите програми – до 6-тата включително.

Специално място заслужава да бъде отделено на програмите в INCO – Centers of excellence, по-точно на проекта BIS21++, който направи възможно издаването на тази книга[1].

[1] Bulgarian IST Centre of Competence in 21 Century, Sixth Framework Programme, INCO-CT-2005-016639

Зад всички ресурси и компютърни приложения, описани тук, стои приносът на много хора, заети в различни етапи от разработката им (от досадното „чоплене" на езиковите факти до създаването на сложни програми със специален интерфейс, наричан от благодарните потребители „linguist friendly"). Те трудно биха могли да се класират по тежестта на участието си в тази дългодишна и многостранна дейност. Всред тях имаше студенти, още преди дипломирането си пристъпили с огромно любопитство и ентусиазъм към тази екзотична и вълнуваща младия човек област, имаше колеги – математици, от дипломанти до хабилитирани лица, имаше млади хора и от двете специалности, преминали за кратко през полезната школа на дисциплиниращата и отговорна обработка на лингвистичните данни. Трудно бих ги подредила и по степента на личната ми благодарност, но една тематична и хронологична подредба извиква пред погледа ми лицата на

хората от Другата специалност:

- Кирил Симов, с който щурмувахме заедно подстъпите към сегментацията на българските думи, заедно с езика на Пролог-клаузите,
- Боянка Захариева, която беше и все още е специалист по задоволяване на всички лингвистични капризи в създадените от нея десетки програмни продукти и бази данни, подпомогнали събирането и подръжката на граматическите ресурси, описани в тази книга,
- Йожи Насвади, който овладя тънкостите на българското формообразуване и сложната процедура на морфологически синтез така, както малко лингвисти могат да я проумеят и обяснят,
- Стоян Михов, с който прекарахме много часове пред мониторите на NextStep, увлечени от задачата да преведем данните и процедурите на българския морфологичен модел в първата Intex система за български,
- Преслав Наков, чието сегашно местопребиваване обръща дневния ми режим и биоритмите ми наопъки, в името на модерните статистически методи за мулти- и монолингвистична обработка,
- Галя Ангелова, мой другар в концептуалната битка с лингвистичните структури и не само с тях,

хората с диплома от Моя факултет:

- Милена Славчева, която пое върху плещите си тежката битка със залоговото поведение на българския глагол,
- Таня Августинова, която след дългия и успешен стаж в подготовката на морфологичните данни, показа реализиран първия експеримент за построяване на синтактичен анализатор на българските аналитични времена,
- Мариана Дамова, която още в ерата преди персоналните компютри формализираше описанието на латинските флективни типове на дълги свитъци, попълвани и разчитани само на пода на просторен софийски хол,
- Орлин Чочов – от отбора на „чоплещите лингвистични данни",
- Йоана Сиракова – от същия отбор,
- Ирен Георгиева – от същия отбор,
- Илиана Гаравалова – от същия отбор.

Всички тези хора, нарочно споменати без сегашните им титли и звания, са в мо-мента или заедно с мен, но в друга работна функция, украсени с постове и звания, в софийски университети и научни учреждения, както и в престижната чужбина – в университети и софтуерни фирми, или в служба на европейската интеграция.

Има и хора, допринесли за решаването на въпросите, изложени в тази книга, с пряко или непряко, но огромно участие, които не са вече между нас, и тук с благодарност искам да спомена имената на:

- Александър Людсканов, който ме отведе, заедно с филологическите ми възторзи, в света на компютрите (пък макар и Минск-2),
- Георги Гаргов, с неговото решително участие в битката за създаване на качествено различна научна структура в областта на езиковите технологии,
- Йордан Пенчев, с критичния му поглед върху лингвистичния аспект на моделираните явления,
- Ирина Ненова, първата дипломантка по информатика, която щурмува дисциплината, въоръжена с българските окончания и машината СМ-4.

На тях – дълбок поклон, а на живите – надеждата ми за нови съвместни разработки.

Защото за българската Дума има още какво да се каже.

Елена Паскалева

Въведение

Езикът, като научно понятие (останалите му значения оставяме настрани), е определен от бащите-основатели на модерното езикознание като:

- физически обект – редица от символи или звуци, който служи за изразяване и предаване на мисли,
- системното описание на този обект.

В двата посочени смисъла (на физически обект и на конструкт) се разбират понятията *Language* и *Langue* на Сосюр [Saussure 2002], съответно *Language* и *Grammar* на Чомски. [Chomsky 1956]. Последният пък допълнително разграничава тези два прочита, като въвежда и понятията езиково изпълнение (performance) и езикова компетентност (competence). В първото си значение езикът се възприема и употребява от всички използващи го (в ролята на т.нар. наивни потребители), а във второто – от учените.

Какво представлява езикът за тези две групи хора, какви са основните му съставки, основният строителен материал и технологиите, които го изграждат, за да се превърне в най-мощното комуникативно средство?

За първата група отговорът е даден в множество изрази, някои от които – крилати, а други – идиоматични[2]. За втората група, тези, които го изучават, въпросът не стои много по-различно, макар и да е обрасъл с теоретически спорове за Единицата – основната, формално различимата, делимата и неделимата, съчетаваната и още и още...

Но колкото и далеч да се отива в тези абстрактни построения на подредени обекти, колкото и да се търси атомарният смисъл – на фонемата, морфемата, думата, фразата, изречението, дискурса, информационния поток, все пак основната единица за изследване остава Думата. Тя се разчленява на съставки, тя се обединява в конструкции и подрежда в смислови йерархии.

И подобно на живия език и живата му употреба и подредба, моделирането на тези явления в компютърния свят изисква да бъде установена връзка между реалността и правилата, между течащия по екрана на монитора поток от символи и осмислянето му. То се извършва по начин, твърде различен от дълбочината на човешкото осмисляне, но това е най-доброто в случая, което можем

[2] Най-великият наивен потребител на английския език през 17 в., а може би и за всички времена, го е определил като „words, words, words", т.е. „думи, думи, думи". Други двама потребители три века по-късно тананикат „parole, parole, parole". Така определят езика тези, които го възприемат само като поток от звуци или вериги от букви – в ежедневна употреба.

да направим заедно и едновременно с компютъра. То е предмет на съвременните софтуерни дисциплини като *natural* ***language understanding, knowledge*** *processing****/acquisition/management*** и други амбициозни наименования.

Независимо от амбициите и мечтите, а и от ярките фрази в докладите, всички тези дисциплини са построени върху реалната Дума и нейното описание в Системата.

Този основен строителен материал, производството му и подготовката за грубия строеж – дори на въздушните кули на изкуствения интелект, са описани в настоящата книга.

1 От текста към морфологичното му представяне

Като оставяме извън това изложение представянето на физическата реалност на речевия акт като поток от звукови сигнали, ще се ограничим само с неговия дигитализиран прочит като писмен текст, който можем да пренесем направо в компютъра. За това са се погрижили писмените системи, които са определили за всеки език дискретните графически порции и прехода от плавния поток на речта към тях. Очевидно само тази степен на дигитализация (звуци – букви) е необходимо, но не достатъчно условие за компютърната обработка на езиковата реалност. Буквеният запис може да съхрани един текст, но не и да ни каже нещо повече за неговите свойства – езикови, литературни, комуникационни. Тази информация можем да получим от различните видове обработка на естествения език в неговото компютърно представяне.

Първа стъпка на това представяне е да се определи основната единица на неговите структури. Очевидно тя съвпада с основната единица на лингвистичното възприятие и описание, споменати във Въведението, но за целите на формализираното описание се налага да проследим по-подробно конституирането и функционирането на тази основна единица – Думата в реалния и в компютърния свят.

За *наивния потребител* въпросът **„що е дума?" изобщо не се поставя, отговорът му е интуитивно известен**[3].

> Най обичам **думичката** *мама*, мама значи приказка любима.
>
> *Sorry* seems to be the hardest **word**.

Затова пък този въпрос е оживено разискван и в структурната лингвистика, и в езиковедски трудове с претенции за формализирано описание. Трудностите да бъде идентифициран основният обект на лингвистичното описание идат от твърде размитите критерии, приети в школските граматики: думата е основната структурна единица на езика, която притежава индивидуални фонетични, семантични и граматични признаци. Тези три признака – *неделимост* във физическото оформяне, единното изразяване на *значението* и начините на граматичната *употреба* са твърде нестабилни за формално описание. Контра-

[3] Не стои така въпросът в компютърната лингвистика, където интуитивното схващане – що е дума, се сблъсква с многообразието на всички тези групи от символи, възприемани като единичен обект – вж. следващите раздели.

примерите са безброй, особено в съпоставителен план, да напомним само следните случаи:

1. По отношение на физическата *неделимост* следните езикови единици ни навеждат на размисъл:

   - делимите префикси в немския език – понякога отделна дума, понякога в състава на думата, вж. *aussehen* (*изглеждам*), *Sie sieht gut aus* (*тя изглежда добре*);

   - определителният член – в български е част от думата, в английски и френски е отделна дума (впрочем, при *the cat* и *le chat* – за две думи ли става дума или за една?);

   - възвратната частица – в български и френски отделна дума *(мия се, se laver)*, в руски – част от думата *(умыться).* Впрочем *усмихвам се* – две думи ли са?;

   - степените за сравнение – в случаите, когато тяхната аналитична форма има синоним от една дума *(чистый* vs. *чище*, но *более чистый; moindre vs. plus petit)*, или когато са в състава на една дума, но разделени от препинателен знак – *по-чист*.

2. По отношение на единната граматична *употреба*. Тук коварно наднича и друго размито граматично понятие – думата и нейните форми, които също са думи. Изскачат въпросите:

   - ако *котка*, *котката, котките* са форми на думата *котка*, то *котенце* нова дума ли е или форма? В английските граматики, с аристократичното им пренебрежение към морфологията, често *driver* се определя като форма на *drive;*

   - отделни думи или форми на една и съща дума са видовите двойки в славянските езици (особено нелексикализираните – *читать* и *прочитать*, *скачам* и *скокна*)?

3. По отношение на единното *значение*:

   - единни ли са значенията на сложните думи, особено немските (*Freundschaftsbezeiigungen* – немска сложна дума, със страхопочитание цитирана от Марк Твен в студията му за немския език)?;

   - единни ли са значенията на аналитичните глаголни времена (понякога съдържащи до 5 думи – *щял съм да съм донесъл*)?

Ако изложените по-горе съображения тревожат структурните лингвисти[4] повече заради методологичната точност, за компютърните лингвисти определянето на Думата като обект от езиковата електронна реалност е жизнено важно, понеже представлява първият лъч в мрака, първият опит да се подреди привидния хаос на буквите така, че да бъде поет от електронната обработка.

За разлика от граматичните студии, където езиковата реалност е представена в отделни примери, които, дори и взети от реални произведения, са вчесани грижливо и подбрани главно за потвърждение на авторската теза, електронната обработка работи с *реални* текстове.

Има се предвид текстът – това, което минава пред очите ни, когато се разхождаме по web-сайтовете, в специализираните документи, този *микс* от символи – букви и цифри, всякакви съкращения, някои от които са граматично утвърдени, а други са авторска приумица, както и всички възможни особености на живия, а не препариран текст. В този микс думата трябва да бъде първо *идентифицирана* (разграничена от останалите), а после *класифицирана* (определена) за целите на по-нататъшната електронна обработка.

1.1 Компютърна идентификация на думата

Първото, най-просто изискване за идентификация на дума в граматиката е физическата ѝ неделимост, което в компютърната граматика се изразява в определението

дума = низ от буквени символи между два интервала.

Още с дефиницията започват въпросите.

Числата – редици от цифрови символи между два интервала – думи ли са? Ако за *1340* отрицателният отговор е очевиден, това не е така за *1-вият*, *1-вата*, *1-вото*, 1st, 2nd и подобни комбинации (вж. по-долу, раздел 1.4).

- *Препинателният знак* разделител на думи ли е или част от дума – т.е. буква? В примера *прочети, преди да пишеш*, символният низ между два интервала *_прочети,_* дума ли е и от колко символа се състои – 7 или 8? Същият въпрос може да се зададе и за символните низове

[4] Едно от най-добрите изложения по проблема – що е дума? вж. в [Апресян 1966].

*_(механата_* или *_„внимание"_*, първият, състоящ се от един препинателен знак и 8 букви, вторият, състоящ се от два препинателни знака и 8 букви (с „_" е означен интервалът).

- *Тирето* буква ли е? Вж. примерите: *мисля – значи съществувам* vs. *по-добър.*

Ако с редица уговорки и ограничения разрешим проблемите на *идентификацията*, остава проблемът за *класификацията* – за какви думи иде реч? Явно тези линейно подредени символни вериги не са самостоятелни, независими една от друга единици, а са групирани в система, която общо взето повтаря школската граматика, с някои необходими уговорки, които да преодолеят разнообразната *ненормативност* на компютърния текст.

1.2 Граматична идентификация на думата

Ако се върнем към школската граматика, която ни дава системното описание на езиковата реалност, думата **дума** означава много неща. Запитат ли ни за думите в стиха:

Аз съм ваше момче, момчета!
Ваше – духом. И мимоходом.

правилен ще бъде отговорът, че той се състои от девет думи. Също така е вярно, че думата *момче* се среща два пъти, както и думата *ваше* се среща два пъти. От което пък уточнение думите в стиха намаляват вече на седем. Това е прочитът на наивния потребител.

Ученикът с добър успех по граматика би казал, че думата *момче* се среща във формите си за единствено и за множествено число, а думата *ваше* е повторена два пъти във формата си за единствено число.

За компютърно описание (на веригата от компютърни думи) този либерален прочит на броя думи е безполезен, затова се налага да въведем следните досадни уточнения.

Стихът се състои от девет **словопояви** (словопояви – running **words, tokens, occurrences**) **и пет препинателни знака. От словопоявите две са различни словоформи** (wordforms) **на лемата** (lemma, **lexeme, basic form)** *момче*, а една и съща словоформа на лемата *ваш* е повторена два пъти.

В термините на онзи дял от морфологията, който се занимава с образуване на формите на една и съща дума (за разлика от образуването на нови думи), в текста са представени два члена на **парадигмата** на лемата *момче* и един член на парадигмата на лемата *ваш*. **Парадигма** се нарича съвкупността от всички форми на една дума (*момче*, *момчето*, *момчета*, *момчетата* – четири члена и *ваш*, *вашия*, *вашият*, *ваша, вашата*... и т.н. – 9 члена).

Ако сортираме деветте словопояви в това стихче, и направим техен речник, ще получим осем словоформи, а ако сведем словоформите до техните леми, ще получим седем леми. Те ни дават единиците на *граматичния речник* на този текст: *аз*, *ваш*, *духом*, *и*, *мимоходом*, *момче*, *съм*.

Граматичният анализ на стиха ще ни даде девет анализа:

> **аз** (*аз*, лич.мст. 1 л.ед.ч.) **съм** (*съм*, гл., 1 л.ед.ч. сег. вр.) **ваше** (*ваш*, прит. мст. ед.ч.) **момче** (*момче*, същ., ед.ч. нечл.), **момчета** (*момче*, същ., мн.ч. нечл.)! **ваше** (*ваш*, прит.мст. ед.ч.) – **духом** (*духом*, нареч.) **и** (*и*, съч.съюз) **мимоходом** (*мимоходом*, нареч.).

Не анализирахме препинателните знаци, но и те могат да получат значение (*пункт.).*

От тези пет препинателни знака два – *точката* и *удивителната* (в своите три появи) имат функцията на разделител на изречения (sentence boundary marker), а два – *запетаята* и *тирето* служат за разделители на думи.

Подобен граматичен анализ на всеки буквен низ в компютърния текст не е желание да се симулира човешката граматика. Целта е прагматична – да се сведат многобройните буквени низове до по-малък брой техни представители. Това изискват целите на търсенето на информация в текста (информацията за *момче* се съдържа и в словоформата *момчета*), което предполага неговото по-нататъшно структуриране. В това структуриране съществуват някои допълнителни критерии за подредба, която доближава компютърното тълкуване на текста до човешкия граматичен прочит на текста (вж. граматичните маркери за род, число, част на речта и др. в примера), колкото и дълъг път да остава да се измине до него.

1.3 Видове думи в компютърния текст

Докато достигнем до сближаването на граматичната и компютърната идентификация в термините, анализиращи дадения пример, трябва да изминем дългия път на едно прочистване на компютърния текст, в смисъл – разпознаване и класификация и на всички останали символни низове, разположени между два интервала. Очевидно компютърната класификация съдържа много по-разнообразни по вид елементи, отколкото изброените в красивия литературен пример. По екраните на компютрите много често срещаме текстове като:

> *2 . Параграф 6 от допълнителните разпоредби на Закона за уреждане на жилищните въпроси на граждани с многогодишни жилищно - спестовни влогове (обн . , ДВ , бр . 82 от 1991 г . ; изм . , бр . 62 и 94 от 1992 г . ; попр . , бр . 9 от 1993 г .) .*
>
> *21 години - за управление на моторно превозно средство от категориите C+E , D1, D, D1+E , D+E , Тмб (тролейбус) и Тмм (трамвайна мотриса).*
>
> (реален текст от информационно-правна система)

Прилагането на компютърното правило за „дума = символен низ между два интервала", дори и с уговорките за препинателните знаци и цифрите, ще ни даде списък от обособени единици, за голяма част от които трудно може да се посочи място в традиционната граматична йерархия. Тук са и съкращенията, а също и смесените низове (комбинация от букви и цифри, от букви от две азбуки), различни съчетания от главни и малки букви, съкращения от всякакъв вид, както и всички динамично образувани названия и нови думи.

Ако резервираме означението **дума** за единица от системата на езика, регистрирана (или не) в официалния му речников състав, то всичко в текста, идентифицирано като дума по горното определение, но намиращо се извън тази система, може да се определи като *квазидума*. Елементите на това множество имат важна информационна стойност, поради което съвременните методи на компютърната лингвистика отделят доста внимание за тяхното разпознаване и анализ.

Както квазидумите, така и думите трябва да бъдат обособени и разпознати в текста, за да бъдат обработени по съответния начин. Процесът на обособяване на различните текстови единици се означава с термина *tokenization,* което приблизително може да се преведе с българското *раздробяване, сегментация,*

един вид *разчленяване.* В най-общ смисъл под това действие се разбира разчленяването на нещо на съставни елементи. В това си значение терминът може да се отнася за единици от всички езикови равнища, но употребата му без уточнение, by default, в компютърната лингвистика означава – първоначална сегментация на текста на единици (каквото и да се разбира под последното). Така се осъществява и първата стъпка на прехода *реалност – система,* чрез търсене на съответствие между единиците на текста и единиците на системата.

След като сме раздробили текста на текстови единици (tokens) – с многобройни и досадни уговорки за разделителите между тях и възможните комбинации на различни символи[5], бихме могли да се опитаме да подредим и класифицираме тези сегментирани единици не толкова заради пълнотата на анализа, колкото поради информационната им стойност и мястото им в компютърното моделиране на езиковата система. С други думи, налага се едно класификационно *прочистване* на компютърния текст, за да се отделят:

- неща, които *не са* думи и няма да участват в по-нататъшния анализ, макар и да продължават да заемат мястото си в анотацията на текста (стандартите за анотация изискват регистрация за всяка текстова единица). Можем да ги наречем *извънезикови* квазидуми.

- неща, които *са* думи, но не могат да бъдат обхванати – в момента или изобщо – от езиковото описание (понеже то не ги съдържа в лексикалните си ресурси). В по-нататъшната езикова обработка на текста тези единици участват. Можем да ги наречем *извънлексикални* квазидуми.

- неща, които *са* думи, обхванати са от граматичното описание на езика и фигурират в неговите речникови ресурси. За тях остава наименованието думи, като напомняме, че става дума за истинска, граматично оформена дума.

За да достигнем до дълбочините на езиковото знание в компютърния модел, който ползва за строителен материал думите от третия вид, би трябвало да разпознаем и класифицираме чрез съответната анотация и квазидумите от първите два вида. Това се извършва с методи, които или нямат нищо общо с граматичното знание (боравят само със символи като единици на обработката), или пък използват фрагменти от това знание (формулирани обаче само на символно равнище).

[5] Не стои така въпросът в компютърната лингвистика, където интуитивното схващане – що е дума, се сблъсква с многообразието на всички тези групи от символи, възприемани като единичен обект – вж. следващите раздели.

1.4 Извънезикови квазидуми

За въведеното разделение на квазидумите – извънезикови **vs. извънлексикални** – основен критерий, освен локализацията им в/извън езиковата система и речниците на езика, остава лингвистичната им натовареност, тежестта им в лингвистичното представяне като морфологични, синтактични и прочие единици.

Извънезиковата квазидума е компютърна дума (символен низ между два интервала), която не ползва дори знаковата система на даден език (т.е. азбуката му), с други думи това са цифри, препинателни знаци, формули, знаци от друг език.

Тази единица, регистрирана чрез условно възприета анотация, изглежда на пръв поглед като досаден *шум* при извеждането на системни заключения за езика на документа – неговия лексикален състав и основни синтактични конструкции, още повече ако анализът е насочен към създаване на компютърни модели на езика с различна пълнота. Това не отменя задължителността на нейната идентификация и регистрация, защото тя участва в съчетания, които са комплексен израз на втория тип квазидуми – извънлексикалните единици (вж. 1.5).

Примери за самостоятелна и за комплексна употреба на тези квазидуми намираме най-вече в цифровите означения. Разгледани в самостоятелната си употреба, това са думи с цифров или смесен буквено-цифров състав.

Първите са чисто цифрови и представляват комбинации от цифри, евентуално с разделители, според правилата за оформяне на числата в съответния език и възприетите формати в документа. Базово означение за тези думи в стандартите за анотация е **<dig>, но в по-дълбоки анотации на следващо лингвистично** равнище целият цифров низ може да бъде причислен към част на речта - числително **<num>; да бъде компонент на единица, изразяваща време <time>** или количество **<quant>, като определя различни времеви или количествени** измерения и интервали (*1 авг. 2003*, *3.25 кг*, *2-мата*). В тази си употреба той е синтактично споен с пълнозначни думи. Независимо от колокационните си възможности, тези символни низове получават базова самостоятелна анотация като число.

Вторите са смесени цифрово-буквени низове, където буквата е елемент от числова подкласификация *(Чл. 23а)*, но може и да е морфологично споен с цифрата буквен елемент *(2-а станция, 1-ви коловоз)*. Тези два типа смесени означения принадлежат към съвсем различни равнища на езиковата система – едните са

елемент на броенето, текстов, но не езиков (граматично значещ) елемент, докато вторите са редовни лексикални единици със специфично кодиране, което позволява комплексът да се трансформира в пълнозначна дума (2-а --> втора, 1-ви --> първи и т.н.), да се анотира граматически като числително редно, да участва в синтактичната структура като атрибут на съществително и т.н.

Разгледани в синтактичната си съчетаемост, тези квазидуми в комбинация с пълноценни лексикални единици (ограничен брой с определено значение), образуват конструкции, които означават измервания на времето и на материалното количество (вж. примерите по-горе). При по-повърхностно анотиране на текста (според по-леки стандарти, като напр. в [INTERA 2003]) тези конструкции могат да не бъдат комплексно идентифицирани. Но в системи за извличане на информация разграничаването им като вид е задължително, поради очевидно високото информационно тегло на изразите, посочващи време, количество и други точни измерения – вж. 1.7.1.

1.5 Извънлексикални квазидуми

Към този раздел спадат единици, които принадлежат към знаковата система на езика, но не са включени в лексиката му (с други думи – в основния му речников фонд, в компютърната лингвистика представен чрез речници, наричани още лексикални бази данни). Причините за тяхното отсъствие в организираната лексика на езика е големият им обем, динамично и не винаги нормативно пораждане. Това се отнася за новоизковани думи, съкращения и названия. Тези думи (или изрази) са пълноправен елемент на езиковия изказ и участват в йерархията на езиковите структури в по-нататъшната обработка. Извън системно-езиковите съображения за задължителната им регистрация, допълнителен аргумент е и високата им информационна стойност.

Информационната и лингвистичната пълнота често си противоречат – елементи извън системата на описанието на езика могат да имат висока информационна стойност и обратно – пълноправни езикови единици могат да бъдат несъществени при извличането на информация от документа[6].

[6] Неслучайно всеки компютърно обработван език има своя списък на т.нар. *стоп-думи* – думи, които поради служебния си, свързващ характер не участват във формирането на информационния образ на документа. Такива са напр. формите на спомагателния глагол, всички числителни, местоимения и собствено служебните думи. За български език стоп-думите, зададени като словоформи, са между 500 и 1000 и са задължителен компонент от всяка система за информационно търсене в документи.

В квазидумите – извънлексикални единици – влизат две групи езикови единици с висока информационна и функционална стойност: съкращенията (abbreviations) и названията (named entities).

1.6 Съкращения – идентификация и основни функции

Квазидуми от втория тип, намиращи се в знаковата система на езика, но извън лексикалната, са съкращенията. На най-ниско ниво на анотиране се означават с <abbr>, а при информационно търсене могат да са елементи на названия – вж. 1.7.

Съдържателно погледнато, съкращенията са един начин на изписване в съкратен вариант на редовни лексикални единици – думи или фрази. Формалният белег на съкращението е: главна буква в начална позиция, точка в края на буквения низ, или изписване изцяло с главни букви (капитализация). Първият и вторият белег често са комбинирани. Съществуват и алтернативни варианти за изписване на едно и също съкращение – вж. напр. *Mr.* и *Mr*, *AM* и *a.m*.

Съкращението, което е уникална дума, може да е омономично с друга дума от общата лексика на езика. Тази омонимичност може да се получи от капитализацията на обикновени думи или от поставянето им в края на изречение. Такъв е случаят със съкращението *CE – Съвет на Европа* и същия буквен низ в капитализиран текст на заглавие, където обозначава възвратното местоимение *се*, или пък съкращението *ТЕ*, означаващо *Териториална единица*, както и лично местоимение в 3 л. мн.ч. (в заглавие: *ТЕ СЕ СРАЖАВАХА ЗА РОДИНАТА*). Съкращения, започващи с главна буква и намиращи се на края на изречение, имат по-лесна за снемане омонимия, но ако в същата позиция започват с малка буква, могат да съвпаднат с дума от общата лексика. Вж. напр. – *позвънете на тел. 8367689* vs. *Убоде се на ръждива тел.*

Възможно е да се направят правила за идентификация на съкращенията, които снемат поне част от тази многозначност. Такова е например елементарното съображение, че капитализирани единици, заобиколени от също такива капитализирани единици (както може да се случи в цяло заглавие), не са задължително съкращение, или че буквеният низ *тел*, следван непосредствено от цифров низ, е съкращение за *телефон*.

Идентификацията на един низ като съкращение е необходима, за да получи и анотацията си <abbr> в анотационните стандарти.

Регистрацията на съкращенията е пряко свързана с тяхното включване в организираната лексика на езика. Много от традиционните хартиени лексикографски пособия – едноезични и двуезични речници – съдържат списък на съкращенията. В този случай съкращението вече е нормална дума от речниковия фонд на езика, със своя граматика (може да образува форми, макар и в разговорен вариант – *ООН-то*, *БСП-то* и подобни, също така участва в синтактични конструкции – *генерален секретар на ООН*).

За съкращенията обаче е характерна и известна авторска оказионалност, особено в ненормативни текстове, което ги оставя извън регистрираната лексика на езика. Друг фактор за тяхното локализиране извън лексикалното пространство на езика е тяхното обновяване, навлизането на нови съкращения, често преводни или пък термини от ограничени предметни области.

Идентификацията на съкращенията е важна не само за пълнотата на анализа – да се определи мястото на всички текстови единици в лексиката на езика. Те изпълняват и една много важна функция – снемат възможното двузначие в автоматичното определяне на границите на изречението.

Много компютърни приложения се реализират върху изречението като единица на анализа (така се структурират например елементи на преводаческата памет, която съпоставя на изречение от превеждания език изречение от превеждащия език). Тъй като стандартен белег за край на изречението е точката, която следва последователност от буквени низове, възможно е съкращението да бъде определено грешно като последна дума в изречението. Само инвентаризацията на съкращенията (чрез поставянето им в общия речник или в специализиран списък) може да разреши тази омонимия.

Като текстово поведение – съчетаемост с други текстови единици, съкращенията могат да се разделят на няколко групи, в зависимост от силата на кохезията, която ги свързва със следващите ги елементи в текста (т.е. изискването след тях непосредствено да следва дума с определени формални белези). Такива групи са например:

- съкращения, предшестващи непосредствено собствени имена – обикновено титли и други съкращения за лица от вида – *проф*., *ген*., *акад., Mr*., *Mrs*. и подобни. Такива съкращения, следвани задължително от собствено име, не могат никога да са край на изречението.

- съкращения, предшестващи непосредствено числови стойности – *стр*., *фиг*., *тел*., *рис*., *вх*. и подобни. Задължително следвани от числов низ,

те не могат да са краен елемент на изречението (така се снема и омонимията на съкращението с обикновена дума, в случаите на *тел.* и *рис.*).

За съкращения, които не проявяват силна кохезия със следващ елемент, въпросът дали са в края или в средата на изречението може да се реши само въз основа на статистически наблюдения. Това е така, защото няма формални правила, които да различат следните две употреби:

Войната избухна през 1950 г. Елена беше тогава тригодишна.
През 1950 г. Елена беше тригодишна.

Безусловна употреба на съкращение в средата на изречение е случаят, когато то предшества дума с малка буква.

През 1950 г. много хора загинаха във войната.

От приведените примери се вижда, че регистрирана като съкращение или не (според пълнотата на текстовото анотиране), разглежданата единица – в отделни лексикални случаи със специфично синтактично поведение, изисква непременно да бъде включена предварително в списъци. Последните са важен инструмент за сегментацията на текста на по-едри единици от думата (като изречението), без привличане на истинско лингвистично знание (от речници и граматики), т.е. упоменатото вече използване на граматично знание на символно равнище.

1.7 Названия

Названията са текстови единици, които *именуват* различни обекти и като такива обновяват състава си постоянно – всеки нов обект в действителността влиза в текста със своето име. Така те фактически съставят едно некрайно множество. Това прави тяхната обработка изключително трудна, тъй като динамичността на формирането на състава им изключва възможността те да се въведат в речниковия фонд на езика. Друга пречка за регистрацията им в общата лексика на езика е огромният им обем – можем ли да си представим всички именувания на обектите в света, въведени в речник? Тези единици, според целта и характера на приложението, което ги използва, могат да бъдат регистрирани в отделни списъци с лични имена, географски и административни обекти, които могат да помогнат за идентификация на едно подмножество от реално именувани обекти, релевантни само за съответното изследване или приложение.

Главните категории на названията в информационно търсещите (IR и IE – *information* ***retrieval*** и ***information*** *extraction*) системи и архитектури са: *топографски названия (места)*, *лица*, *организации*, *дати*, *време*, *парични единици* и *проценти*.

Лингвистичните единици, назоваващи горните седем групи обекти, са различни преди всичко по обема на референциалното си множество (реалните обекти, назовавани от тях), по степента на вариативност на езиковото си изражение (съществуващи синонимни и родствени означения за един и същ обект), по разнообразието на лексикалните и граматични средства в границите на една група (богатство от думи и конструкции). Те са и езиково оцветени (зависят от конкретния естествен език) в различна степен. Тази им оцветеност се измерва главно с вида и дълбочината на лингвистичното знание, което използват процедурите за тяхната идентификация за различни езици.

Повече от очевидно е, че палитрата на названията – лексикална и граматична – е различна за всеки конкретен език или езиково семейство. А също е очевидно, че повече от 20 години след първите приложения на IE (осъществени за английски език) многобройните разработки за други езици позволиха на изследователите да видят конкретните измерения на езиковата специфика, особено когато прилагат готови архитектури за извличане на информация към езици, доста различаващи се от английския.

В това отношение семейството на славянските езици е сериозно предизвикателство за езикова настройка на методите за обработка на названията поради типологическата си отдалеченост от английския език особено по оста синтетичност-аналитичност, богатство на формо- и словообразователни механизми, гъвкавост vs. ригидност на синтактическите правила и много други.[7]

Настройката на една система за извличане на информация към български език се определя от типично славянските черти на езика, каквито са:

- богата морфология – проявена не само на равнището словоформа - дума, но и на равнището производни думи (много по-разнообразна от *drive-**driver***).
- като следствие от горното – лесно идентифицируеми механизми на съгласуване и синтагматично обединяване на езикови единици, само въз основа на морфологичните им свойства. Тези свойства са изразени чрез специфични буквени комбинации, обикновено на края на

[7] За българския език такава настройка бе направена в рамките на системата GATE, създадена в Университета в Шефилд [Paskaleva et al. 02].

думата (вж. *Неправителств**ена** организац**ия*** vs. *Неправителств**ени** организац**ии***) и линейното им разполагане (невъзможно: *организация неправителствена*).

Тази специфика позволи да се направи известна промяна в съотношението *лингвистични правила – списъци,* използвано в английските процедури за извличане на информация. Последните се опират в много по-голяма степен на списъци, отколкото на правила, поради слабата информативност на буквените комплекси в края на английските думи. Главното оръжие, с което се атакуват текстовите единици - названия, за да бъдат вкарани в информационния образ на документа, са списъците (gazetteers). Използването на различни списъци, помагащи да се идентифицира текстова единица, са израз на *грубата сила* в лингвистичния компютърен анализ, но при извънсистемните текстови единици това доста често е единственият подход. Списъците, които помагат за идентификацията на кандидат-названието, са от два различни функционални типа, насочващи ни към самия обект или към неговото характерно обкръжение.

Първият тип списъци съдържат текстови единици, наименования от търсения тип, идентифициращи кандидат-названието чрез пълно съвпадане (matching). Обемът на тези списъци, наричани *gazetteers*, предопределя успеха на идентификацията на названието[8]. Могат да бъдат наречени *кандидатски списъци*.

Вторият тип списъци съдържат текстови единици от различни лингвистични нива – отделни символи, морфологични единици, словоформи, леми, морфосинтактични свойства и съчетания от всички изброени дотук, които характеризират текстовото обкръжение на названието и притежават диагностична сила - вж. например наличието на думи като *г-н*, *Mr, Mme*, *prof*, *General* за идентифициране на последващата дума като име на лице; съкращенията за организации от типа на *ООД*, *Ltd*; родови понятия, назоваващи институции и организации (*съвет*, *дружество*, *council*, *organization* и др.), характерни морфеми, оформящи фамилното име и други подобни. Могат да бъдат наречени *контекстуални списъци*.

За езици със силно развита словообразователна структура и висока флективна производителност възможностите за включване на допълнителни инструменти за идентификация, освен списъците, се увеличават (вж. 1.7.3).

Подредени по възходяща тежест на използваното специфично лингвистично знание, изброените на стр. 23 названия се подреждат както следва:

[8] Във версията ANNIE на системата [GATE 2005] кандидатските списъци съдържат 60 000 единици за означаване на 80 типа названия.

- проценти,
- парични единици,
- дати и време,
- топографски названия (места),
- лица,
- организациии.

Разликата в лингвистичната тежест на процедурите, които ги идентифицират, дели тези шест вида на две основни групи.

Първите три названия образуват затворено множество от обекти – лексикални единици, които се идентифицират лесно като буквено-цифрови съчетания главно поради ограничения си обем[9]. И начините на оформяне на една дата, и лексикалните единици за изразяване на процент, и означенията на паричните единици, както и названията на части от деня се свеждат до краен брой думи. Известна разлика може да има само в граматичното оформяне и вариативност на техния контекстуален списък (например различните стандарти за означения на дати, процентите – в цифри и думи и др. подобни).

Втората група названия – имената на топографски обекти, лица и организации практически именува обектите в едно отворено множество, които се образуват динамично или пък, дори застинали в образуването си, са трудно изброими.

1.7.1 Проценти, парични единици, дати

Разпознаването на *проценти* е свързано с три вида ключови думи или символи като контекстен списък – символът за процент %; лемата *процент* във всичките й форми, както и производни от нея леми (*процентен* във всичките му форми; фразата *на сто*). При такъв кратък списък е възможно елементите му да бъдат включени директно в правилата за идентификация.

При идентификацията на *паричните единици* житейската тежест на названието определя и съществената разлика в пъстротата на наименуването им в административен (финансов, политически , информационен) текст и в разказвателен текст. В строги административни текстове трябва да се отчита единствено морфологичната вариативност на съответното съществително – слово- и формообразуваща: *щатски долар, щатски долара, доларови* (резерви).

[9] Тъй като тези единици (без времевите отрязъци) формират данни, в голяма степен официални и изискващи точно изразяване, има съществена разлика в лингвистичното им поведение в по-строг административен и в разказвателен текст (за щастие, не много чест обект на информационно търсене).

Далеч по-интересни проблеми поставят разговорните варианти на валутни единици, особено за основните световни валути. Популярните и шеговити означения на американския долар, еврото и други валутни единици, употребявани в разговорната практика, бързо преминават във вестникарските колони. За означенията на щатския долар се ползват модификатори за цвят (всяко споменаване на *зелен* в цифрова конструкция – *2000 в зелено)*, метафори като *гущери,* транскрипции на сленга на оригиналния език – руското *баксы* и др.

Тази вариативност увеличава лексикалния обем на списъците, без да променя съществено правилата. Увеличението е незначително, тъй като сленговите означения на валутите се утвърждават бавно и се проследяват лесно в развитието на езика.

Означенията за *време и дати*, особено вторите, в български са доста по-прости, отколкото в английски. И тук имаме разлика в официалното и разказвателното им изписване. Българският език, както и руският нямат разграничението *am/pm,* не използват разделителя „/" за разграничаване на годината, месеца и деня, тук стандартният разделител е интервал или точка, двоеточието пък не се използва като такъв разделител. Няма традиция да се изписва и денят от седмицата като част от датата. За разлика от английски, римските цифри се използват за месец, а арабските – за година и ден, не се използват съкращения на дните на месеците.

Тази простота и липса на вариативност е компенсирана от разнообразието на изразяването на дати в разказвателния текст, където оформянето им с пълнозначни думи влече след себе си изискването за морфологично оформяне на конструкцията и следвателно, увеличаване на вариантите.

Синтактичната структура на описателно изразената дата в български е:

- номинална фраза, състояща се от редно числително като определител и с опора – име на месец (*трети март*),
- приложение на тази фраза - друга номинална фраза в същата структура, изразяваща годината (*хиляда деветстотин и пета година*).

Тъй като на това равнище на текстов анализ – без лингвистично знание, не е възможно да се ползват граматични и синтактични белези, които да идентифицират горната конструкция, същата работа вършат списъците с изброени форми на числителни редни, плюс имената на месеците. Вариантите за комбиниране на тези елементи от списъка са лесно изброими или изчислими.

Нерешен въпрос за този етап на анализа без граматично знание остава анафорното посочване на датата или елементи от нея – *тази*, *миналата*, *същата*, *следващата година* и други подобни. Определянето на точната дата тук излиза от полето на названията и навлиза в друга, по-сложна област на информационното търсене, свързана с референцията.

За общия случай трябва да отбележим, че лингвистичният инструментариум за изразяване на темпорални факти в някои славянски езици е усложнен от морфологичната вариативност на единиците. Английското *first* и *1st* са единствените варианти, руското *1-вый* се скланя в шест падежни форми за ед.ч., а българското *1-ви* е само в една форма (българската дата не се членува, числителното се съгласува само в м.р. с името на месеца).

1.7.2 Названия-места

В групата *названия-места* са събрани множество обекти, които общо казано са топографски наименования – географски обекти, административни единици, улици, площади и др.

Наблюденията върху големи български текстови корпуси сочат, че често срещаните и общоизвестни имена на географските обекти се споменават без каквото и да било родово уточнение (*Марица* вместо *река Марица*). Това важи за високи планини и върхове, дълги реки и големи градове. Тези названия с достатъчно честа употреба влизат в кандидатския списък. Същото важи и за имената на държавите и континентите.

Невъзможността да се състави изчерпващ кандидатски списък дори само в рамките на страната определя важността на другия списък – контекстуалния. Той съдържа родови понятия като *река, море, връх, област и* др.*,* които в български текст са в пре- или постпозиция.

Богатата деривационна морфология на славянските езици позволява названията-места да станат и съставна част на названието в синонимична употреба, където названието-място е производно прилагателно – вж. *Балкан*, *Балкани* и *Балкански полуостров* (във всичките им форми). Идентификацията чрез проверка в контекстуалния списък (на единицата *полуостров*) засилва точността на търсенето, поради силната омонимия на подобни производни названия с думи от общата лексика – *Балкански полуостров* vs. *Балкански нрави*. Липсата на идентифициращ контекстуален елемент в последния пример води до нерешима омонимия, ако се намира в началото на изречението, където капи-

тализацията е задължителна, докато при позиция в средата на изречението омонимията е разрешима *(населението на Балканския полуостров* **vs. специ***фиката на балканския характер).*

Друга възможна омонимия на названието-място е тази с друг тип названия – собствените имена, които оформят фамилното име по същите морфологични правила. Нещо повече – голяма част от фамилните собствени имена в български са образувани от топографски обекти – като националния литературен герой *Ганьо Балкански*. Правилото за разпознаване на названия – собствени имена (вж. 1.7.3) оставя тези фамилии в началото на изречението многозначни, с евентуална следваща проверка в контекстуалния списък за съседната единица. Наличието на елемент от този списък, подсилено от правила за морфологичното съгласуване на двете единици (и съобразяването с парадигматични ограничения – собственото име никога не се членува!) дава превес на хипотезата *название-място* пред тази на *собственото име* (вж. – *Адриана Дунавска* vs. *Дунавската ръченица* vs. *Дунавската равнина*).

В обсега на названията-места влизат и имената на улици и площади, споменати самостоятелно или като пощенски адреси. Във втория случай има доста формални показатели за различаването им в контекста – означения за *град*, *улица*, *вход*, наличие на номер и др. (*ул. Раковски* или *Раковски 134*).

При отсъствие на съположени елементи от контекстуалния списък или специфично печатно оформяне омонимията *название-място* vs. *собствено име* е неразрешима, а в начална позиция в изречението към нея се прибавя и омонимията с обикновена лексикална единица, особено ако топографското име е синтактична конструкция – номинална група.

Когато названието не е единична дума, а синтактична група, дори при наличие на елемент от контекстуалния списък остава проблемът за определяне на дясната граница на названието на символно равнище – вж. *Trida dukelskych hrdinu* (чеш.), *Улица 26-ти бакинских комиссаров* (рус.), *Площад Велчова завера започва от Семинарията* (**бълг**.). При пълен анализ и идентификация на номинална група с други лингвистични средства комплексните названия не представляват проблем.

1.7.3 Собствени имена

Собствените имена са динамично обновяващ се езиков ресурс, където освен традиционни за езика имена (отбелязвани в речници и специални лексиког-

рафски пособия) влизат и новообразувани именни единици, а също и чужди имена, транслитерирани по различен начин. Идентификацията им се извършва с помощта на два главни ресурса – на списъците (кандидатския и контекстуален) и на специално конструирани правила.

Основният ресурс за изучаване на лингвистичната природа на собствените имена се дава от списъци от имена, събирани за други цели, главно справочни и административни. Различните именници, публикувани в мрежата, са недостатъчен ресурс за един кандидатски списък, тъй като съдържат само лични имена, с препращане към тяхната семантика, етимология и други данни. Най-представителен ресурс за собствени имена обикновено е телефонният указател. В настройката на системата GATE за български [Paskaleva et al.2002] за съставянето на българската база от названия бе използван телефонният указател на София, съдържащ 330 000 записа, комбинация от лични и фамилни имена. Експериментите върху този списък показват следната взаимовръзка между личните и фамилни имена, изразена в структурата на извлечените от него подсписъци, както следва:

- списък от лични имена – 6 500 уникални (3800 женски и 2700 мъжки). От тези два подсписъка са извлечени 250 женски и 230 мъжки многозначни имена (съвпадащи с дума от общата лексика – напр. *Бистра вода* vs. *Бистра Петрова*). Разрешаването на тази многозначност е тривиално, ако думата се среща с малка буква, но в началото на изречението въпросът не е толкова елементарен. Личните имена формират кандидатския списък, идентифициращ съответната единица.

- списък от фамилни имена – 27 500. Те не са включени в списъците за търсене, а са използвани само като експериментален материал за изработката на другия тип ресурс – правилата за разпознаване на български фамилни имена.

Подмяната на списъци с правила е възможна поради спецификата на формиране на българските фамилии – нещо, характерно само за славянските езици.

Правилата за разпознаване на фамилно име се основават на наблюдения върху морфологичната му деривационна структура, която отразява семантиката на формирането му както следва:

лично име + притежателен суфикс + маркер за род.

Така *Иван-ов* е син (потомък) на *Иван*, а *Иван-ов-а* носи значението на дъщеря (съпруга, потомка) на *Иван*. Тази базисна структура на българските фамилии

позволява върху ресурса от фамилни имена да се направят редица изследвания за честотата на срещане на различните притежателни суфикси, както и за цялостното разпространение на този вид образуване на фамилии за съвременното българско именно производство. Резултатите от изследванията са следните:

- 91% от фамилните имена в списъка могат да се разпознаят чрез техните морфологични компоненти,
- най-разпространените суфикси са както следва:
  -ов/-ова – 46,4%, *-ев/-ева* – 24,4%, *-ски/-ска* – 15,4% ,
  -ин/-ина + -йн/-йна – 3,4%, *-шки/-шка + -чки/-чка* – 1,6 %.

Извеждането на тези морфологични зависимости в правила е много по-ефективно, отколкото събирането на списъци и търсене в тях (колкото и подробен да е един списък, винаги съществува възможност в текста да се срещне ново име). Не трябва да се забравя и че има множество други думи, Завършващи с други укени низове, които често се срещат в началото на изречение, т.е. започват с главна буква и това ги прави потенциални кандидати за разпознаване от тези правила (напр. *Това – Йотова*, *Чин – Андрейчин*, *Всички – Ковачки*).

Следващ проблем е омонимията на тези единици с други названия със същата морфологична структура - имена на организации и топографски обекти, които се състоят от прилагателно и съществително и започват с главна буква (напр. *Министерски съвет*, *Перловска река*). Тази многозначност може да се разреши по три начина:

- ако думата е предшествана от лично име (разпознато в кандидатските списъци) или някой от елементите от контекстните списъци (титли, професии и др.), определя се като фамилно име;
- ако по други правила и други списъци думата е разпозната като алтернативните названия – места или организации, приема тази идентификация;
- ако думата не отговаря на никое от горните две условия, приемаме че това е фамилно име. Това решение е взето на базата на статистически наблюдения върху текстов корпус и с оглед на динамиката на образуване на фамилни имена, изпреварваща тази на местата и организациите.

Разрешаването на многозначността на личните имена, която се среща много рядко (само в началото на изреченията – *Роса роси*, *Бистра вода*, *Ела се вие, превива*) е възможно само при наличие на съседни елементи, снемащи тази многозначност. Такива са само предимно фамилните имена – капитализирани буквени низове с гореописаната морфологична структура.

Важна характеристика на българските собствени имена в официалното им изписване е, че са изградени от три части – лично, бащино и фамилно име (като последните две са морфологично определими в повечето случаи). Макар и по-рядко срещана и само в официални документи, тази употреба също трябва да бъде отчетена в правилата за разпознаване[10].

1.7.4 Названия-организации

Собствените имена и названията на организации, освен по синтагматичната си структура, се различават по скоростта на обновяването (и съответно, на остаряването) на състава си, както и по променливост на конструиращите ги правила (за собствените имена – традиционно утвърдени, за организациите – исторически променливи). Подобно на идентификацията на собствените имена, основните механизми са три: търсене в списъци (кандидатски или контекстни), морфологично изчисление, зависимости на буквено и пунктуационно равнище.

В повечето случаи имената на организации и институции са доста сложна номинална структура с произтичащи от това много варианти на морфологичното оформление (вж. например разгърнатата синтактична структура на *Българска асоциация на агенциите за развитие и бизнес центрове (БАРДА)* и простичкото *Прозрачност без граници)*. По тази причина механизмите за морфологично изчисление са трудни за формулиране и изпълнение, още повече – на буквено и пунктуационно равнище. Остава само механизмът за търсене в контекстуалните списъци, а търсенето в кандидатските списъци може да задава само конституенти на сложната номинална група – название на организация.

Търсенето на диагностични конституенти е силно затруднено от разнообразието на граматични модели за образуване на номиналната група в славянските езици – силна вариативност на отделните компоненти при различни синтактични функции. Инвариантът на тези морфологични варианти може да се открие едва в следващи етапи на обработката – лематизация и анализ (вж. част 2). Така основният инструмент за идентификация на названията на институции остава търсенето на диагностични елементи в контекстния списък. Такива елементи могат да бъдат:

- ключови думи от един сравнително малък по обем списък (вж. примерното им разпределение в названия на неправителствени организации по-долу, в Табл. 1);

[10] Още по-мощно деривационно средство, формулирано на равнището на буквения низ, за идентификация на бащини имена, намираме в друг славянски език – руския, където тези имена се образуват чрез специфични суфикси, различни от тези за фамилните имена *(Петрова* vs. *Петровна, Иванов* vs.*Иванич).*

- думи в постпозиция – еквиваленти на английското *Ltd*, *Inc* – например *ЕАД*, *ООД* и други. Срещат се по-често в официални и административни текстове, но не толкова често в разказвателни;
- придружаващи знаци. Единственият известен, макар и малко употребяван инструмент на символни равнище за означаване на комплексни названия остават кавичките.

Ключова дума	**Брой**	**Процент**
фондация	137	21,92
сдружение	136	21,76
съюз	67	10,72
център	60	9,60
дружество	57	9,12
клуб	50	8,00
асоциация	48	7,68
организация	24	3,84
агенция	18	2,88
камера	14	2,24
движение	9	1,44
без ключова дума	5	0,80

Табл. 1. Разпределение на диагностични ключови думи в названията на 625 неправителствени организации. Пример за названия без ключови думи: *Младежка толерантност*, *Европейски пространства* и др.

Кавичките обаче имат и много други значения – за означаване на цитати, а също и за логическо подчертаване, ирония и други стилистични роли. В последните години тази им функция се поема от шрифтовете – например курсив в по-модерните печатарски стандарти, но далеч не във всички видове издания.

Изборът на една от двете възможни функции на кавичките – именуване или цитиране, фактически е посочване на кандидата за названието – организация, граждан от тях. Правилата за снемане на такава многозначност могат да се съставят само чрез статистически анализ на голяма текстова база. Такъв анализ бе направен за две текстови съвкупности: български вестникарски текст – електронно издание, в обем милион и половина думи, и руски вестникарски текст – 50 пъти по-малък обем, също в електронно издание. Двата текстови корпуса са в различна степен на издателска готовност – българският е предварителен, чернови текст, преди редакция и корекции, а руският е публикуван материал.

Беше изведено едно основно правило за разрешаване на многозначността на кавичките:

- ако изразът в кавички започва с малка буква, той е цитат (срв. *И това ти наричаш „дружба"?* vs. *в „Дружба" заплатите са високи*.)

- ако изразът в кавички започва с голяма буква, той е цитат, само ако е изпълнено условието да съдържа формални разделители за край на изречението. С други думи цитатът съдържа сложна фраза, изречение или група изречения. Названията на организациите обикновено не съдържат пунктуационни знаци, тъй като са неделима синтактична група.

Това правило не е абсолютно, но почива на статистически наблюдения.

2 Морфологични модели

Простите примери в предишните раздели обясниха

- как текстът, който тече по екрана на компютъра, се разбива на основните порции – думите,
- колко е различна природата на тези единици и как се класифицират те,
- с какви операции, списъци и трикове трябва да почистим текста – да сложим настрана онези порции, които не са част от езиковата система, като ги запазим за анализа на информационното съдържание на документа,

за да стигнем до крайната цел на **сегментацията**:

- да получим чистите думи – основните тухли на лингвистичното знание, за целите на по-нататъшното му моделиране в различни компютърни приложения.

За елементите извън общата лексика, **квазидумите**, беше описана процедурата на изчисление и проверка по списъци.

За елементите на общата лексика, за думите в истинския смисъл на понятието обаче, нещата не се свеждат до просто съвпадане с елементи от списък, тъй като техните характеристики са по-сложни и взаимосвързани. Тази сложност, ако успеем да я изразим формално, позволява да се правят дълбоки анализи на текстовия материал, разкриващи неговата граматика, синтаксис и семантика, дискурсни характеристики, анафори, логическа структура, структура на знанието, комуникативна организация и много други аспекти на лингвистичното моделиране на различни езикови равнища.

Колкото и сложни да са тези равнища, всички те започват от някаква класификация на отделните думи в текста. Това означава, че трябва да разполагаме с някакъв базисен резервоар, в който държим знанието за думите – поотделно и на групи, или по-точно с някакъв склад, където е систематизирана основната езикова субстанция. Освен с този материал, трябва да разполагаме и с инструменти за обработка на думите, които свързват лингвистичния материал и процесите на обработката му.

Ролята на такива резервоари както в живата употреба на езика, така и за описанието на неговата систематика, се изпълнява от лексикографските инструменти – речниците, разнообразни по форма и характер, главен ресурс на тази

класификация (т.е. на анализа), в съчетание с граматиките.

За човешката обработка на думите – класификация и анализ – складът с лингвистичен материал и инструментите за неговата обработка са компоненти на индивидуалното лингвистично знание. Знанието **на** езика, което ни позволява да го владеем като инструмент на общуването в живата му употреба, не е предмет на аналогия и моделиране в настоящето описание. Това, което е полезно за езиковите технологии, е знанието **за** езика, за неговите процеси и основни ресурси.

Именно компонентите – ресурси и инструменти, лингвистичният материал и правилата за боравене с него правят възможно в процеса на лингвистичното моделиране да се формира текст по искано граматическо значение или да се определят граматическите значения във взет наготово текст. С други думи – да генерираме и да анализираме.

Тъй като механизмът на тези две операции в човешко изпълнение не е достатъчно добре изучен, за да твърдим точно какви ресурси, модули и процедури ги извършват, компютърното им изпълнение прави една проста симулация на този процес. То моделира само входа и изхода им, като се опира на два основни компонента – списък от всички (колкото и условно да звучи това) думи в езика, правилата за тяхното образуване и комбиниране от една страна, както и правилата за разглобяване на текста на основните му граматични съставки, изразени чрез думите, от друга.

Двата основни компонента на операциите *морфологичен анализ* и *синтез* – езиковите ресурси и правилата за боравене с тях – се съдържат в основния склад на компютърната граматика, а в термините на системите за моделиране на лингвистично знание – в морфологичния компонент на съответния модел.

2.1 Функции на морфологичния модел

Един прост пример за ролята на морфологичния модел в примера от 1.2 е: да замени думата *момчета* в текста с думата *момче* и да извлече информацията, че това е *съществително от мъжки род, единствено число, нечленувано*. За подобна дума – *мальчика* в руски текст, задачата на модела е да каже, че това е думата *мальчик*, *съществително от мъжки род, в множествено число, родителен или винителен падеж.* Тази операция моделира училищния анализ на изречението, в онази му част, където се определят морфологичните (граматични) свойства на думите.

Други училищни задачи са:

- да се определи вътрешният състав на думата – нейната *основа* и *окончание*, а при по-раздробено членение – *префикси*, *корен*, *суфикси*,
- да се изброят всички други форми, които може да приеме анализираната дума.

В термините на науката за езика тези две задачи са свързани с:

- посочване на деривационната и формообразуващата структура на думата,
- изброяване на всички членове на нейната парадигма.

С други думи, моделира се движение в две посоки: навътре (към съставните части на думата) и по вертикалата (към близките роднини по пряка линия).

Разглежданият тук фрагмент от лингвистичното моделиране има за цел да определи съставките на обекта, неговите свойства и групите, в които той влиза.

Постановката на тази задача (или група задачи) определя съдържанието на лингвистичния склад като:

- информация за морфологичните *съставки* на думата, не само като инвентар – списък, но и като функционални елементи на словообразователния (или формообразуващ) механизъм в зависимост от целите на модела,
- информация за граматичните *характеристики*, носени както от групи думи, така и от съставките на една дума,
- информация за групиране на думите в семейства от *форми*,
- правилата за връзката и операциите между елементите на горните три вида статични сведения.

Трите първи групи информация решават проблемите за представяне съответно на: *членимостта,* граматичните *свойства* и *парадигматиката* на думата. Решенията, които се взимат за всяка една от тях, са взаимосвързани. Четвъртият тип информация е повече процедурна, свързана с вида и последователността на операциите, извършвани върху различните сегменти от лингвистичния ресурс. От тяхната конфигурация зависи дали имаме анализ или синтез и на какво лингвистично равнище[11].

[11] Важно! Школската задача за определяне на ролите на думите в изречението или на техните синтактични свойства (връзките им с другите думи), т.е. движението по хоризонтала, с посочване на *съседите*, не е предмет на разглежданото моделиране. Резултатите от нашия модел участват пряко в правилата за синтактичен анализ като негови входни данни. Не случайно възприетият термин за трансфера на информацията от общия граматичен ресурс към конкретните думи в текста се нарича ***морфосинтактично*** *анотиране*, а учебниците по граматика ни напомнят, че белегът за част на речта е едновременно морфологична и синтактична категория.

Първите три вида знание в своята съвкупност очертават границите на един всеобемащ граматичен модел. Можем да го разглеждаме като изчерпващ всички аспекти на морфологичното поведение на думата, ако задачата ни е да моделираме морфологията чрез единно функционално описание.

В съвременната компютърна лингвистика обаче задачите за моделиране на фрагмент от знанието не са самоцел, а обикновено са продиктувани от нуждите на реално компютърно приложение. Целите на това приложение, характерът на обработвания материал и не на последно място мощта на изчислителния ресурс измерват степента на участие на всеки един от трите посочени компонента в модела. При това конкретните постановки и решения за всеки един от тях предопределят много често решенията в другите компоненти. Пример – при по-подробно членение на компонентите в думата се променя информативността на всеки един от тях, а оттам и обликът на парадигмата. Конкретните решения за всеки един от трите компонента имплицират и съответни решения в другите компоненти.

Независимо от богатата гама на комбинациите между порциите лингвистично знание в трите аспекта на морфологичното моделиране, основен водораздел между видовете морфологични модели си остава посоката на прехода – от текста към системата или обратно. По-ясно дефинирано – за граматичен анализ или граматичен синтез ли става дума, или за пораждащ или анализиращ граматичен механизъм. Този избор предопределя както конструкцията, така и съдържанието на съответния лингвистичен ресурс (склада, обслужващ модела).

За езици със силно развита флективност като славянските – със значителен обем на парадигмата и богатство на словообразователни и формообразуващи елементи, комбинациите от решения в трите основни плоскости – членимост, информативност и парадигматика, определят богатата вариативност на моделите. Две са обаче главните оси, по които минава водоразделът на основната им класификация:

- моделиране на словоизменително и/или словообразователно равнище,
- моделиране в посока анализ или синтез.

Това предполага да се разгледат тези модели отделно, макар че обслужващите ги ресурси се припокриват в голяма степен, а процедурите им са в основната си част обратими.

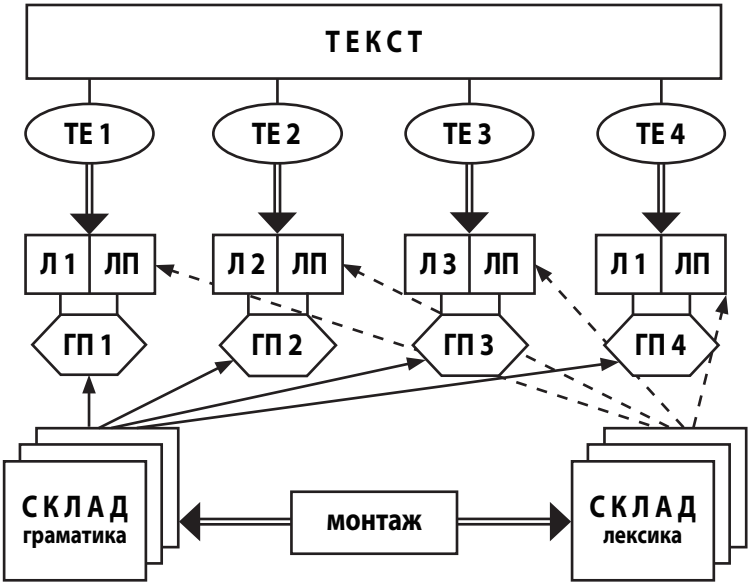
2.2 Модели на морфологичен анализ и синтез – основно знание и процедури

Аналозите на анализиращия и синтезиращ модел в компютърната морфология в училищната практика са:

- направи анализ на думите в това изречение,
- образувай формите на тази дума.

За тази цел ученикът разполага с ресурса на всички думи от езика, с които си е служил през целия си живот, и с някакви процедурни знания – правила, научени в часовете по граматика. Това, което му се задава от учителя като условие на задачата, е някакъв текст – свободно избран (за анализа) или пък съвкупност от думи и граматични значения – за синтеза.

В компютърното изпълнение на тази задача целта на морфологичния анализ е да замени последователността от зададени текстови единици с последователност от компютърно анализирани думи. *Компютърно анализираната* дума е двукомпонентна верига, състояща се от лексикален елемент (вида му ще уточним по-долу) и граматични свойства – вж. Фиг. 1.



Фиг. 1. От текста и текстовите единици (ТЕ) – думите, към лемите (Л) с техните признаци (ЛП) и граматическите признаци на словоформите (ГП) – съхранявани в складовете на лексикалното и морфологично знание.

Както се вижда от Фиг. 1, освен текста за анализ, зададен е и компютърен ресурс, който обединява:

- думи (или части от думи), разположени в лексикалната част на лингвистичния склад на системата,
- правила за монтаж, които прикачват към някакви елементи от лексикалната част на склада определени стойности,
- граматически елементи (също части от думи), които се намират в граматичната част на склада (конструкцията на склада и видът на правилата за монтаж ще бъдат разгледани по-нататък).

Целта на компютърния морфологичен синтез е движение в обратната посока – по зададен лексикален/ни елемент/и и определени граматични стойности да произведе текстови единици.

Би могло да се каже, че двата компонента разполагат с един и същ ресурс (от гледна точка на неговата лингвистична природа) и една и съща цел: да се моделира обединението на *лексикалното* и *граматичното* в Думата, само че в две противоположни посоки – от реалността към системата и обратно.

В един *идеален* двупосочен модел двете процедури би трябвало да са равноправни и симетрични, за сметка на известно усложнение на основния ресурс, който да осигури спецификата на двата симетрични прехода. В *реалното* си изпълнение двете посоки на модела се различават по достоверността на резултата (в разглеждания отрязък от лингвистични стойности).

Синтезът ни дава винаги един – правилен! резултат, а анализът може да даде повече от една вярна интерпретация. Винаги ще образуваме правилно членуваната форма ед. ч. на лемата *момче*, щом като в граматичния ресурс е зададен еднозначният преход от тази граматична стойност към окончанието – *ето*. Не винаги обаче ще познаем правилно граматичната стойност на словоформата *овч-и*, тъй като в граматичния ресурс окончанието *-и* представя и единственото число в м.р. и множественото число (*овч-и кожух* и *овч-и стада*). Правилният анализ тук може да се изчисли само от информацията от съседни думи, по-точно от техните граматични стойности – вж. в примера граматичното число на *кожух* и на *стада*, което разрешава колебанията. Както се вижда, отговорността за този избор е прехвърлена на един следващ етап на анализа и на елементи, които не влизат в нашия модел – *съседите*.

Затова трябва да направим уговорката, че граматичният анализ ни дава *всички* възможни анализи, а граматичният синтез – *единствено* правилната форма

(с изключение на периферни дублетни случаи). Тази несиметричност на резултатите се дължи на естествената многозначност на единиците в реалния текст от една страна и на известна препарираност, изкуственост на синтеза, от друга (в разглеждания отрязък, където правилните съставки са предварително подбрани).

Единиците, участващи в тези два модела, и правилата за тяхното съчетаване са събрани в лингвистичния склад. Те, макар и предназначени за производство на един и същ продукт, могат да се различават по своята граматична природа, което определя и различната конструкция на механизмите за тяхната обработка.

2.3 Видове единици в ресурсите на модела

От двата вида граматично производство – анализ и синтез, последният моделира по-добре процеса, тъй като започва от основните градивни елементи на морфологичното знание и прилага върху тях главните правила и зависимости в построяването на граматичната Дума.

Продуктът на синтеза – правилно построена дума, може да бъде поръчан да се произведе в училищно упражнение по граматика по няколко начина. За конкретно задание те звучат така:

- образувай мн.ч. членувано на думата *момче*,
- образувай мн.ч. членувана форма от основата *момч-,*
- коя е словоформата за мн.ч. членувана форма в парадигмата на лемата *момче,*
- как ще образуваш мн.ч. на словоформата *момчето*?

Първото задание звучи най-ненаучно, но може да бъде разбрано от ученика.

Второто задание предполага знание за съставките на думата, които участват в синтеза.

Третото е най-издържано по отношение на йерархията на лингвистичните единици на входа и изхода на синтеза, а четвъртото предполага и познаване на връзките между тях.

В последните две задания виждаме изразена йерархията, отношението *частно-общо*, *инвариант-вариант*, която обединява словоформите *момче*, *момчето*, *момчета*, *момчетата* в парадигмата на лемата *момче*. *Словоформи-*

те са конкретни лингвистични единици, групирани в *парадигма,* а *лемата* е абстракт, име на тяхната общност, чийто буквен състав съвпада с една от тях (за съществителните – нечленуваното съществително в ед.ч., за прилагателното – нечленуваното прилагателно в ед.ч. м.р., за глагола – формата за сег.вр. 1 л. ед.ч.). Така че формално един и същ буквен низ – лемата, назовава две различни нива на лингвистична йерархия.

От второто задание разбираме, че общото може да бъде изразено и чрез обща съставка, която притежава цялата група словоформи. От което следва, че частното, индивидуалното, се изразява с друга съставка, специфична за всеки отделен член на групата. Тези два начина на изразяване на групирането в парадигма – чрез съставките на думата и чрез отделни пълнозначни форми, ни водят до два различни типа организация на лингвистичните ресурси, участващи в синтеза.

Единият начин задава общото чрез неизменяемата буквена част на парадигмата (*момч*), а частното чрез отделни граматични формативи (*е*, *ето*, *ета*, *етата*). Другият начин представя общото и частното чрез групиране на цели думи, една от които е избрана за название на групата.

Двата начина на моделиране на формообразуването са еквивалентни, само че предполагат различна организация на лингвистичния склад от ресурси. Първият предполага складиране на основи и окончания, а вторият – на отделни думи. Съдържателната част на групирането в *варианти-инвариант*, *конкретна* група и *представител* на група, *вид и род*, *примери* и *название*, *стойности* и *променлива* остава едно и съща, тъй като отразява едни и същи граматични категории.

2.4 Операции върху ресурсите

Както беше формулирано в предишния раздел, синтезът на словоформи може да бъде организиран като конкатенация, слепване на инвариантната и вариативните съставки на думата (*момч-е*, *-ето*, *-ета*, *-етата*), а може да бъде организиран и като групиране на думи, изразяващи различни граматични форми (*момче*: *момче*, *момчето*, *момчета*, *момчетата*). Двата начина са еквивалентни, макар че първият моделира процеса чрез участващите в него съставки, а вторият – чрез групиране на резултатите, при което групирането е вторично – извършва се върху вече получените морфологични обединения на участващите съставки.

Ако тръгнем по първия път, операции върху съставките, ще формулираме заданието за получаване на формата за мн. ч. членувана форма на лемата *момче* така:

- отрязваме *е* от края на лемата и получаваме окастрен низ *момч*,
- прибавяме съответно за четирите члена на нейната парадигма редицата: *е*, *ето*, *ета* и *етата* към окастрения низ – основата на думата.

В българското формообразуване (както и в другите славянски езици) е възможна и още една операция, съпровождаща прибавянето на окончания – промяна в буквения състав на основата. По такъв модел се синтезират примерно словоформите на лемата *грък*:

- не отрязваме нищо от лемата,
- прибавяме съответно за 5-те члена на парадигмата:
  1. **нищо**,
  2. -***а*** (с трансформация *р*ъ→ ъ*р*),
  3. -***ът*** (с трансформация ръ→ ъ*р*),
  4. -***и*** (с трансформация *р*ъ → ъ*р* и *к* → *ц*),
  5. -**ите** (с трансформация ръ→ ъ*р* и *к* → *ц*).

Така трите възможни операции на прехода от лема към нейните форми са:

- отрязване (сегментация),
- залепване (конкатенация),
- промяна (алтернация) – факултативно.

За всяка първа словоформа в парадигмата (съвпадаща с лемата) операциите отрязване и залепване фактически я възстановяват в първоначалната форма, следващите операции вече построяват различаващите се от лемата словоформи.

2.5 Организация на формообразуващите елементи в парадигматична структура

Правилата за обработка на ресурсите предопределят и тяхната подредба, тъй като се реализират върху определени части от тяхната структура. Това проличава особено ясно, когато започваме да подреждаме споменатите вече участници в модела на морфологичния синтез:

- невъзможно е заданието – образувай 2 л. мн.ч. мин.св.вр. за лемата *момче,*
- невъзможно е също така заданието – образувай ед.ч. пълен чл. за лемата *момче,*

- неправилно формулирано е и заданието – образувай сег.действ.прич. м. р. за лемата *прочета*.

Конфликтът тук е между поръчаните граматични стойности и лексикално-граматичните стойности на лемата. В първия случай на лема – съществително се приписва погрешно глаголна категория, като се поръчват глаголни граматични стойности. Във втория случай на лема – съществително от среден род се приписва форма, която този подклас съществителни не образува. В третия случай се поръчва сегашна причастна форма за лема – глагол, която не я образува, понеже е от свършен вид.

Следователно стойностите на граматичните категории на лемата определят категориалните признаци на генерираните от нея словоформи. Съвкупността от тези признаци (обединени в комплекси така, както се обединяват в конкретната словоформа, а не поотделно) образува т.нар. *парадигматична черупка* (paradigmatic shape/shell), мрежата, канавата, върху която се гради постройката на генерираните словоформи (по-подробно за нея вж. 2.7)

Ако се изключат грешно формулираните задания за синтез в 3-те горни задания, вярно формулираните като задание и получени като резултат словоформи позволяват на последните да се групират около различни леми, в зависимост от приетата концепция на формообразуването. Генерираните словоформи от една лема могат да бъдат и разпределени между две леми в зависимост от това, къде сме поставили границата между форми на думата и новообразувани думи.

Като във всеки език със силно развита формообразуваща и словообразуваща морфология, в богатството от генерирани форми на българската лема има групи, за които границата между лексикалното и граматичното е размита. Това се доказва и от възможностите за лексикализация на тези междинни по природата си формообразувания. Следните примери са илюстрация за различни концепции на строеж на парадигматичната черупка.

Една типична междинна категория е глаголното *причастие*, с двойствена природа – лексикални свойства на глагол и граматични свойства на прилагателно. Това не предполага третирането му като прилагателно, но в някои решения е възможно то да се обособи като отделна част на речта (нормативните граматики не го допускат, определяйки го като отделна глаголна форма, но в едно формално описание това е съвсем възможно – с плюсовете и с минусите на такова решение)[12].

[12] Плюсовете на решението за причастието – отделна част на речта са, че така се намалява броят

Още по-далеч се измества границата между формообразуване и словообразуване при *отглаголните съществителни*, които се образуват от съответния глагол по един единствен начин, но само от глаголи от несвършен вид и „**оп**редметяват" глаголното действие. При избора на решението – отглаголното съществително *в* или *извън* глаголната парадигма – трябва да имаме предвид и голямата група от лексикализирани отглаголни съществителни със свое собствено значение, по-близко до резултата, отколкото до процеса – като *изказване*, *потвърждение* и др. Често лексикализацията се проявява и чрез алтернативна словообразователна схема – вж. *заключаване* vs. *заключение*. От друга страна езикът може да образува динамично нови отглаголни съществителни за всеки глагол, чиято форма позволява това, ако трябва да се означи едно действие в обобщен, номинализиран вид – вж. *едно плисване на вода*, *едно щипване по носа*. Тази номинализация на действието проличава и при възможните синтактични трансформации (*дразни ме неговото постоянно покашлюване –> дразни ме как постоянно покашлюва)*.

Съображенията *за* и *против* включването на отглаголните съществителни в глаголната парадигма са свързани съответно при:

- *за*: с опасността да оставим неидентифицирани отглаголни съществителни в текста, динамично образувани като нови думи, изразяващи опредметяване на действието, но не достатъчно лексикализирани, за да са включени в лексикалния ресурс като съществителни. В крайна сметка мярката за тяхната лексикализация си остава включването им в правописен или тълковен речник, които обаче не могат да догонят и регистрират образуването на нови форми за една част на речта и лексикализацията им като друга,
- *против*: с увеличаването на обема на словоформите в лексикалния ресурс и възникване на изкуствена омонимия. Това се случва, когато редовно образуваното отглаголно съществително съвпада с вече лексикализирано такова, а двете единици не се различават по поведението си в текста, за да можем да говорим за реална омонимия. В случая омонимията в ресурса е увеличена точно с толкова единици, колкото са лексикализираните отглаголни съществителни, съвпадащи с редовно образуваните в глаголната парадигма – *четене*, *писане*, *дишане* и други подобни.

на генерираните глаголни форми, а също и че новата част на речта, в съгласие с механизмите на формообразуването получава съгласувателната потенция на прилагателно, отговаряща на истинското ѝ синтактично поведение. Минусите на това решение са, че, скъсвайки с глагола, причастието губи връзката с неговите валентни свойства и възможността те да бъдат използвани в синтаксиса се изгубва. Особено драстична е тази загуба в аналитичните глаголни времена, където причастието фактически представя основното глаголно действие.

В зоната на прилагателните имена подобна алтернатива възниква при степените за сравнение, в зависимост от това дали сме приели степенуването на прилагателните като формообразуващо или словообразуващо явление. Във втория случай всяко българско прилагателно, което може да се степенува, ще се раздели на три лексикални единици – за позитивната, сравнителната и превъзходната степен. И ще приеме своите редовни парадигматични членове (така трите леми – *добър*, *по-добър* и *най-добър* ще образуват по девет словоформи). Друго решение е префиксално формообразуване чрез прилепване на *по-* и *най-* към началото на основата, така че парадигмата на качественото прилагателно да съдържа 27 форми.

2.6 Основен строителен материал на формообразуващия модел

Споменатите по-горе три компонента на формообразуването – отрежи *X*, залепи *Y*, замени *Z* с *A*, се осъществяват върху езиковите единици, участващи в този процес. При тази двукомпонентна организация на процеса – *Нещо* се прибавя към *Нещо*, единици могат да са само двата вида формативи – основната морфема, още *основата*, и флексията, още *окончанието*.

Функционалната отлика между тези два компонента е, че първият представя тази част от думата, която остава неизменяема при образуване на всичките ѝ форми (с изключение на регламентираните промени на отделни нейни елементи), а вторият дава тази изменяема нейна част, която определя уникалното значение на отделната форма.

Школската граматика определя в състава на думата съставките *префикс*, *корен*, *суфикс*, *флексия (окончание)*, а тъй като в процеса на формообразуването първите три компонента остават неизменяеми, тя ги интегрира под името *основа*, към която се прилепва съответното *окончание*.

Компютърната морфология приема същото двукомпонентно деление на двата морфокомплекса – основата и окончанието, с цялата условност на това деление, което често води до терминологични конфликти между представители на морфологичната теоретична мисъл и реализаторите на компютърни морфологични модели. При определянето на тези два комплекса за втората цел се използват дистрибутивни критерии, насочени към ефективността на процедурата – по-малко правила, приложени към по-малък брой елементи, за описание на максимален брой явления. Така например, нищо не ни пречи формално да определим за парадигмата на момче – основа *момче* и окончания *-#, -то,*

-та, -тата[13], които се прилепват към основата, но същото няма да е валидно за друг тип съществителни от среден род на *-е*, които не залепват *-та*, а заменят *е* с *я* в множествено число – *цвет-е*, *-ето*, *-я, -ята*.

Компютърното окончание, т.е. изменяемият морфокомплекс, е доста различно от традиционното окончание, ако държим на терминологичната чистота. Тъй като този комплекс трябва да бъде неделим (моделът е двукомпонентен, а критерият за обособяване на двете му части е – *променя се*/*не се променя)*, то в частта, която се променя при формообразуването, влизат различни граматични формативи. Тяхното комбиниране увеличава доста броя на стандартните окончания, но това е цената на компромиса на двукомпонентното деление.

Например критерият за промяна/непромяна при членението на причастната форма *предполагащият* (от глагола *предполаг-ам*) би определил състава на морфокомплекса-окончание като *ащият*, което определение звучи доста скандално за морфологичните норми и за техните автори. В този комплекс влизат: основна гласна, причастен суфикс, истинско окончание и постпозитивен член. Затова трябва да уговорим определението за компютърно окончание на думата като морфологичен комплекс, определен от съставящите го морфеми според приетите граници на формообразуващия модел. При промяна на модела на формообразуване, свързан с определянето на парадигматичната черупка, съставът на този комплекс (оттук нататък наричан само окончание) се променя по съответен начин.

За да реализираме пълен морфологичен модел на синтеза, трябва да разполагаме с ресурса от всички възможни основи и окончания, участващи в процеса, както и описание на възможните промени в основата. Ако предположим, че вече сме определили всички възможни парадигматични черупки в морфологичната система на даден език, в рамките на всяка една такава черупка трябва да се определи взаимната свързаност между участващите в модела основни строителни единици – основите и окончанията.

Както е невъзможно да се формулира задание – образувай 2 л. ед.ч. сег.вр. от лемата *момче,* така е невъзможно в някакъв момент от функционирането на модела да се получи залепване на окончанието *ещ* към основата *момч*. Основата *момч*, която образува всички членове на парадигмата за същ.ср.род, може да присъединява към себе си само окончанията *е*, *ето*, *ета*, *етата* (както беше показано по-горе).

[13] С # обикновено се означава нулев елемент.

Следователно, моделът трябва да разполага с правила за съчетаване на отделните формообразуващи компоненти, т.е. всяка основа трябва да е свързана с групата окончания, която формира нейната парадигма.

Групата окончания с уникална комбинация от буквени стойности, присъединявани към определена основа или група основи формира нейния формообразуващ тип или клас, наричан още *флективен тип*. Флективният тип се дефинира в границите на една парадигматична черупка – например за съществителни м.р., за глаголи от несвършен вид и т.н. Това подразделение се определя от различния обем на парадигматичните черупки – напр. пет члена за съществителните от мъжки род, четири члена за съществителните от женски и среден род и за множественото число на всички родове.

Така флективният клас *N* за съществителни (*момче*) може да съдържа: *-е*, *-ето*, *-ета*, *-етата*, а флективният клас *M* за прилагателни (*кози*): *-и*, *-ия*, *-ият*, *-я*, *-ята*, *-е, -ето*, *-и*,*- ите*. Очевидно, съществуват общи пресечни точки за някои елементи от класовете, но уникалността им обхваща пълната тяхна комбинация.

Друг компонент на флективния тип са промените в основата, които се извършват за отделни членове на парадигмата – стандартните буквени промени, известни от граматиката като прояви на палатализация, изпадането или вмъкването на *о* или *е,* премятането *ър-ръ*, *ъл-лъ* и подобни. В рамките на един флективен тип, описан чрез групата флективни елементи, могат да се обособят няколко подтипа, в зависимост от наблюдаваното редуване на основата, а в една плоска класификация (предпочитана в такъв род модели) се увеличава броят на типовете. Така клас, описан като комбинация от залепвани окончания (#, *а*, *ът*, *и*, *ите*), съответно в 5-те клетки на парадигматичната черупка на същ. м.р. неодуш. може да се раздели на няколко, в зависимост от промените в основата за отделни леми (без промяна – *квадрат, билет*, с промяна *к –> ц – белтък,* с промяна *х –> -с – кожух* и т.н.). При плоското представяне на флективните типове това не са подкласове, а отделни равноправни типове.

2.7 Парадигматичната черупка – разпределителен пункт за категориални значения и буквените им изрази

Моделът на морфологията изисква не само да се направи пълен списък на всички видове морфологични **единици, участващи в него, но да се формулират** и основните правила за комбиниране на тези единици, да се изброят изчерпателно всички възможни структурни мостри, всички възможни видове парадиг-

матични черупки, в които се подреждат споменатите единици по споменатите правила.

Тази карта на възможните конфигурации на всички генерирани думи предполага тяхната структурна подредба, определянето на йерархията на граматичните им стойности. В едно top-down (отгоре надолу) представяне се очертават различни групи и подгрупи в мрежовидно подредени стойности на граматическите категории, с отделни участъци в йерархично представяне и други – в линейна подредба.

Ако си представим граматиката (морфологията) на българския език, систематизирана по признаци (повече или по-малко общи), за които нямаме възможност да се впускаме в онтологични дефиниции, ще видим представени следните категории, в които се описва граматиката на българските думи:

- *Първо равнище.* Част на речта. Приети са десет такива категории в общия случай – глаголи, съществителни имена, прилагателни имена, местоимения, числителни, наречия, предлози, съюзи, междуметия, частици. Както ни подсказват елементарните знания по граматика, пет от тези части на речта са изменяеми думи и следователно формите им са организирани в парадигма, а за пет от тях – неизменяемите, понятието парадигма е безсмислено.

В границите на всяка една от 5-те изменяеми части на речта се проявяват различни членения на подкатегории, за които е важно да се формулират следните два подтипа: категории, валидни за цялата част на речта (един вид категории на *лемата*, още *лексикални* категории) и категории, валидни само за конкретната *словоформа*, още *граматични* категории. Те запълват следващите две равнища на граматично представяне:

- *Второ равнище.* Лексикални подкатегории, валидни за цялата лема. Пример за категории от този вид са: род и одушевеност при съществителните, вид и транзитивност при глагола, качественост/относителност при прилагателните, тип (лични, притежателни, възвратни и т.н.) за местоименията и др.

- *Трето равнище*. Граматични подкатегории. Те са валидни за всеки парадигматичен член, в уникална за него комбинация, отличаваща го от всички останали членове в парадигмата. Пример за категории от този вид са: число, членуване за съществителните, род, число и членуване при прилагателните, време, лице и число за глагола и т.н.

Подробната разработка на системата на тези категории – чрез тяхната йерархия, от една страна, чрез разновидностите им – грамеми от друга, а също и чрез начините на буквеното им изразяване, ни позволява да видим, че класификацията на синтезиращите модели по парадигматични черупки – характерните рисунки на парадигматичните значения, върви по линията на техните лексикалните характеристики на лемата, а запълването на конкретните черупки се извършва по стойностите на граматичните признаци на словоформите.

Едно кратко обръщане към термините на академичните граматики, въпреки доста нестабилната почва на използвания в тях понятиен апарат (нестабилна главно като база за компютърно моделиране и обработка поради различен адресат на обяснението), ще свърже използваното тук представяне с общоприетите граматични термини, както следва.

Част на речта е общ термин за двата вида представяне. Граматичната категория – също. Граматиката я определя като семантична, принадлежаща към плана на съдържанието. Нейните разновидности – т.нар. грамеми (напр. грамемата ж.р. за категорията род) в нашето описание ще се споменават като характеристики на категорията, нейни признаци или значения – абсолютна еднозначност не е необходима, понеже става дума за описателно изложение. Тези характеристики (признаци, значения) също принадлежат към плана на съдържанието, към семантиката на съответната категория.

Думата стойност и тук, и в едно по-точно изложение може да означава различни неща, в зависимост от това на какво равнище се проявява релацията между променлива и стойност. Затова ще уговорим, че при преминаване към плана на изразяването, към конкретния външен израз на някакво граматично значение, задължително ще използваме термина *буквена стойност*, за да означим реалния израз на това значение в езиковата система. Така категорията род в българския език има три значения (характеристики, признака) – мъжки, женски и среден, а буквената стойност на комплексната характеристика **„мъжки** род единствено число“ за прилагателни от определен флективен тип е *-и* (като в *горски*).

2.7.1 Парадигматична черупка – структура

Видът на парадигматичния рисунък (черупката, схемата) се определя от лексикалните характеристики на лемата, които запълват с буквени стойности клетките на парадигматичната мрежа на дадена част на речта. Парадигматичната мрежа не е фигура от плана на съдържанието. Тя е едно междинно производ-

но, спирка по пътя от системата към текста, мост между абстрактните граматични категории и тяхното групиране според конкретните езикови форми, изразяващи тези категории. Една клетка (слот) в парадигматичната черупка на дадена част на речта отговаря на една словоформа и съдържа комплекс от граматични значения, изразявани от конкретната словоформа. Тя ни дава в графичен вид разпределението на тези граматични комплекси в конкретните единици на формообразуването. По такъв начин тя е обърната едновременно и към плана на съдържанието, и към плана на изразяването.

Всички знаем, че българският глагол образува 52 синтетични словоформи в своето спрежение, така че неговата обща, потенциална парадигматична мрежа се състои от 52 клетки, подредени по основни граматични признаци, по принципа – една клетка за всеки комплекс от граматични значения, който образува отделна словоформа. В зависимост от лексикалните си характеристики някои глаголи запълват изцяло тази мрежа, някои – част от нея (вж. ограниченията, цитирани в 2.5). Запълнените клетки ни дават черупката, рисунъка, схемата на парадигмата.

Съществителните имат пет парадигматични черупки, които са формирани по комбинациите от лексикални характеристики за *род* и *одушевеност* – категориите, които произвеждат различни черупки (като редове на таблицата) и запълването им с онези комбинации от граматични признаци (число, членуване, форма), които получават отделна буквена стойност в склонението (колоните на таблицата). Така черупката на същ. м.р. (одушевено и неодушевено) съдържа шест члена, същ. ж.р. (одушевено и неодушевено) съдържа пет члена, а същ. ср.р. – четири члена (вж. Табл. 2 по-долу). Както се вижда от примерите, две различни черупки имат формално съвпадащ брой клетки за запълване, но съдържанието им е свързано с различни граматични признаци.

Лекс. признаци	**Комплексни граматични признаци, приписани към отделната словоформа**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
М.р. неодуш.	Ед. нечл.	Ед. кр.чл.	Ед. пъл.чл.	Бр.ф-ма		Мн. нечл.	Мн. чл.
М.р. одуш.	Ед. нечл.	Ед. кр.чл.	Ед. пъл.чл.		Зв.ф-ма	Мн. нечл.	Мн. чл.
Ж.р. неодуш.	Ед. нечл.	Ед. кр.чл.			Зв.ф-ма	Мн. нечл.	Мн. чл.
Ж.р. одуш.	Ед. нечл.	Ед. кр.чл.				Мн. нечл.	Мн. чл.
Ср.р.	Ед. нечл.	Ед. кр.чл.				Мн. нечл.	Мн. чл.

Табл. 2. Парадигматични черупки на съществителните.

Ако номерираме клетките – слотове в таблицата, от 1 до 7, излиза, че лексикалната характеристика м.р. неодушевеност запълва слотове 1,2,3,4,6,7,

м.р. одушевеност – 1,2,3,5,6,7, ж.р. неодушевеност – 1,2,5,6,7, ж.р. одушевеност – 1,2,6,7, същото запълване имаме и за ср.р.

Ако в характеристиките на черупката влизат и лексикалните характеристики, черупките са пет, а ако характеристиката на черупката включва само рисунъка на запълваните клетки, различните черупки са четири.

Прилагателните имат само 2 парадигматични черупки – ако приемем за словоформи степените на сравнение. Но ако ги приемем за нови думи, имаме по една черупка за всяка степен – позитивната, сравнителната и превъзходната. В първия случай парадигматичната черупка има 27 клетки, във втория – по 9. Деветте основни клетки са: м.р. нечл.; м.р. кр. чл.; м.р. пъл. чл.; ж.р. нечл.; ж.р. чл.; ср. р. нечл.; ср. р. чл.; мн.ч. нечл.; мн.ч. чл. Същите, умножени 3 пъти – за трите степени, дават единствената парадигматична черупка за прилагателното – 27 члена.

Парадигматичните черупки на **местоимения** и **числителни** са изключително сложни поради нееднородната им синтактична природа, която обуславя и съвсем различни типове формообразуване – номинално или адективно, с много липсващи членове в редовната парадигма, остатъци от архаични форми, главно падежни (*мене*, *ме*, *ми*), наличие на аглутативни формативи (*когото*, *комуто*, *какъвто*) и други.

Глаголната парадигматична черупка е най-сложната, тъй като синтетичните глаголни форми, които се образуват за една глаголна лема, в пълната си версия (без лексикални ограничения) са 54 на брой (към класическите 52 са прибавени и отглаголните съществителни, вж. по-горе, 2.5). Вариантите на парадигматичната глаголна черупка зависят главно от лексикалните признаци на глаголната лема, която трябва като минимум да съдържа информация за вид и преходност на глагола. Друг елемент на лексикалната класификация на глаголите е свързан със залоговото им поведение. Характеристиките на залоговото поведение на глагола, техните стойности и влиянието им върху парадигматичната черупка ще бъдат разгледани по-долу (3.5.2).

Пълната парадигматична глаголна черупка има следния вид, представен в Табл. 3.

Сег.вр.	1 ед.	2 ед.	3 ед.	1 мн.	2 мн.	3 мн.			
Мин.несв.вр.	1 ед.	2 ед.	3 ед.	1 мн.	2 мн.	3 мн.			
Сег.деят.прч.	М.р. нечл.	М.р. кр.чл.	М.р. пъл.чл.	Ж.р. нечл.	Ж.р. чл.	Ср.р. нечл.	Ср.р. чл.	Мн. нечл.	Мн. чл.
Мин.несв.прч.	М.р. нечл.	Ж.р. нечл.	Ср.р. нечл.	Мн. нечл.					
Дееприч.	Неизм.								
Пов.накл.	Ед.ч.	Мн.ч.							
Мин.св.вр.	1 ед.	2 ед.	3 ед.	1 мн.	2 мн.	3 мн.			
Мин.св.прч.	М.р. нечл.	М.р. кр.чл.	М.р. пъл.чл.	Ж.р. нечл.	Ж.р. чл.	Ср.р. нчл.	Ср.р. чл.	Мн. нечл.	Мн. чл.
Мин.стр.прч.	М.р. нечл.	М.р. кр.чл.	М.р. пъл.чл.	Ж.р. нечл.	Ж.р. чл.	Ср.р. нчл.	Ср.р. чл.	Мн. нечл.	Мн. чл.
Отгл. същ.	Ед.ч.	Мн.ч.							

Табл. 3. Пълна глаголна парадигматична черупка.

При лексикална характеристика – непреходен глагол, черупката има вида, показан на Табл. 4.

Сег.вр.	1 ед.	2 ед.	3 ед.	1 мн.	2 мн.	3 мн.			
Мин.несв.вр.	1 ед.	2 ед.	3 ед.	1 мн.	2 мн.	3 мн.			
Сег.деят.прч.	М.р. нечл.	М.р. кр.чл.	М.р. пъл.чл.	Ж.р. нечл.	Ж.р. чл.	Ср.р. нечл.	Ср.р. чл.	Мн. нечл.	Мн. чл.
Мин.несв.прч.	М.р. нечл.	Ж.р. нечл.	Ср.р. нечл.	Мн. нечл.					
Дееприч.	Неизм.								
Пов.накл.	Ед.ч.	Мн.ч.							
Мин.св.вр.	1 ед.	2 ед.	3 ед.	1 мн.	2 мн.	3 мн.			
Мин.св.прч.	М.р. нечл.	М.р. кр.чл.	М.р. пъл.чл.	Ж.р. нечл	Ж.р. чл.	Ср.р. нечл.	Ср.р. чл.	Мн. нечл.	Мн. чл.
Мин.стр.прч.									
Отгл. същ.	Ед.ч.	Мн.ч.							

Табл. 4. Парадигматична черупка на непреходни глаголи.

При лексикална характеристика – свършен вид, черупката има вида, изобразен на Табл. 5:

Сег.вр.	1 ед.	2 ед.	3 ед.	1 мн.	2 мн.	3 мн.			
Мин.несв.вр.	1 ед.	2 ед.	3 ед.	1 мн.	2 мн.	3 мн.			
Сег.деят.прч.									
Мин.несв.прч.	М.р. нечл.	Ж.р. нечл.	Ср.р. нечл.	Мн. нечл.					
Дееприч.									
Пов.накл.	Ед.ч.	Мн.ч.							
Мин.св.вр.	1 ед.	2 ед.	3 ед.	1 мн.	2 мн.	3 мн.			
Мин.св.прч.	М.р. нечл.	М.р. кр.чл.	М.р. пъл.чл.	Ж.р. нечл.	Ж.р. чл.	Ср.р. нечл.	Ср.р. чл.	Мн. нечл.	Мн. чл.
Мин.стр.прч.	М.р. нечл.	М.р. кр.чл.	М.р. пъл.чл.	Ж.р. нечл.	Ж.р. чл.	Ср.р. нечл.	Ср.р. чл.	Мн. нечл.	Мн. чл.
Отгл. същ.	Ед.ч.	Мн.ч.							

Табл. 5. Парадигматична черупка на глаголи от свършен вид.

2.7.2 Запълване на парадигматичната черупка

Няколкото показани по-горе парадигматични черупки задават конфигурацията на клетките, които предстои да бъдат запълнени с конкретни буквени стойности в процеса на формообразуването. В това е и предназначението на парадигматичната черупка – да определи местата за запълване и да забрани онези, които не могат да бъдат запълнени поради специфични лексикални ограничения. Клетката, наименувана с комплекс от значения на граматични категории, изразявани от една словоформа, предстои да бъде запълнена с буквения израз на тези категории. Запълването става според процедурата на формообразуването, описана в 2.4 – *режи*, *залепи*, *промени*.

Така посочената на Фиг. 2 парадигматична черупка на съществителните може да приеме следния вид – с посочване на целите словоформи, резултат от трите горецитирани операции – Табл. 6:

	Словоформи						
Лема	**Ед. нечл.**	**Ед. кр.чл.**	**Ед. пъл.чл.**	**Бр.ф-ма**	**Зв.ф-ма**	**Мн. нечл.**	**Мн. чл.**
дворец	дворец	двореца	двореца	двореца		дворци	дворците
син	син	сина	сина		сине	синове	синовете
сестра	сестра	сестрата			сестро	сестри	сестрите
ръка	ръка	ръката				ръце	ръцете
цвете	цвете	цветето				цветя	цветята

Табл. 6. Запълване на буквените стойности за лемите *дворец*, *син*, *сестра*, *ръка*, *цвете*.

Но може да приеме и следния вид, с посочване само на буквените вериги, извършващи конкатенацията – Табл. 7:

Лема	**Режем**	**Лепим**						
дворец	#	#	а	ът	а		и	ите
син	#	#	а	ът		е	ове	вете
сестра	а	а	ата			о	и	те
ръка	а	а	ата				е	те
цвете	е	е	ето				я	ята

Табл. 7. Буквени стойности на запълващите елементи.

А за довършване на пълната картина на формообразуването, трябва да се отчете и редуването в основата във вида, даден в Табл. 8:

Лема	**Режем**	**Променяме**						
дворец	-						е – #	е – #
син	-							
сестра	а							
ръка	а						к – ц	к – ц
цвете	е							

Табл. 8. Алтернационни промени в основата.

Табл. 6 има повече илюстративен характер – да покаже формообразуването чрез конкретните словоформи.

Табл. 7 и 8 разкриват съдържанието на въведеното по-горе понятие флективен тип, като показват явно и рисунъка на схемата (разпределението на релевантните клетки) и нейното напълване с конкретно съдържание (буквените низове, осъществяващи конкатенацията и редуването).

По такъв начин комплексната характеристика *флективен тип* обединява в себе си както сведения за формообразуващите потенции на конкретна основна морфема, така и за тяхната буквена реализация. В известен смисъл флективният тип се явява основен градивен елемент на пълния, изчерпващ модел на формообразуването – в посока синтез.

За изчерпателно морфологично описание очевидно пълният морфологичен модел на синтеза трябва да включва:

- описание на всички флективни типове, характеризиращи формообразуването на съответния език и като модели, и като конкретна буквена реализация;
- свързване на всяка изменяема лексикална единица със съответния флективен тип (по двата параметъра – конкатенация и редуване);
- изброяване на всички неизменяеми леми в езика. Последните не взимат участие във формообразуването, но представят лексикални класове, които също са уникални.

Така погледната, задачата за построяване на пълен формообразувателен модел за езика изглежда отчайващо изнурителна и досадна – не само заради изчерпващото описание на всички флективни типове, но и заради предполагаемата класификация на десетки хиляди лексикални единици в термините на приетия модел.

По отношение на първото задание – описание на възможните флективни типове, се затваря един порочен кръг, свързан с методиката на класификацията. Дали тя трябва да е умозрителна, създадена на базата на съдържащите се в обикновените нормативни граматики сведения, или пък да е емпирична – чрез изследване на всички реални словоформи на езика, за пълната инвентаризация на ресурсите и правилата на синтезиращия морфологичен модел[14].

Независимо от решението на тази дилема, което, както винаги, се намира по средата, задачата за оформяне на пълния модел на формообразуването трябва да се допълни с приписването на характеристиките на модела към всички лексикални единици на езика.

Единственият път към изпълнението на тази крайно изтощителна задача е да се създаде изчислителна процедура, съчетаваща съдържателното изчисление на лингвистични данни със специално проектиран софтуер и приятелски функционален интерфейс. Основно правило при проектирането на такава процедура трябва да бъде:

- изчисление на всичко, което е изчислимо,
- изброяване на всичко, което не може да се изчисли.

Ако въпросният модел е реализиран в компютърно приложение, организацията на съотношението между *изчислимо* и *изброимо* е поверена на генералния програмен дизайн на системата и на функционалните възможности на интерфейса.

Тези съотношения и тяхната конкретна реализация в дизайна на специален софтуер ще бъдат разгледани в следващите раздели.

2.8 Изчислимост и изброимост в морфологичния синтезиращ модел

Показаните възможни начини за обединяване и структуриране на елементите, които участват в производството на думи, т.е. в синтеза, трябва да бъдат обединени в една производствена схема за генериране на всички думи в езика.

[14] Методическото противопоставяне на построяването на описание отгоре-надолу (top-down) – от абстрактното към реалното и отдолу-нагоре (bottom-up) – от реалното към неговото обобщаване не трябва да се възприема буквално. Нито граматиките могат да ни посочат как да образуваме всички словоформи на езика, нито пък има такъв текстов корпус, който да ги съдържа (за да го ползваме като материал). Реалното решение е да се комбинират знания от граматиките и речниците, съдържащи изчерпателни таблици за формообразуването като [Зализняк 1977], [Кръстев 1990], [Попов и др. 1998].

Това е задачата-минимум за производство на словника (vocabulary) на езика, с едно малко уточнение, че традиционно словникът съдържа леми като заглавни единици, а тук става дума за производство на всички словоформи на лемата. За българския език приблизителното изчисление според различни граматични речници показва съотношение леми:словоформи приблизително 1:16. Задача-максимум е да се генерират всички възможни словоформи, заедно с граматичната информация, която те съдържат.

Очевидно синтезирането на всички словоформи в езика се базира на някакво предварително подредено лингвистично знание – както за тяхната генеративна способност, така и за буквените стойности на нейното изразяване. На базата на това знание задачата на морфологичния синтез обикновено се разглежда като производство на свързаната тройка: лема, нейните словоформи и техните граматични признаци, за всяка лексикална единица на езика. Тъй като това множество е доста различно по структура и по свойства в своите три компонента, налага се да разделим производството му на няколко етапа, които ще бъдат изложени по-долу.

1. *Определяне на лексикалните характеристики на лемите*. Първа, задължителна характеристика (която може да остане единствена в модел с минимално лингвистично знание) е тази за част на речта. Следващи характеристики са лексикалните свойства на лемите, които са важни не само за пълнотата на модела, но и за оптимизиране на следващите етапи като формирането на парадигматичната черупка. Тази операция ни дава съдържанието на парадигматичните черупки, по-точно, на техните клетки.

2. *Определяне на видовете парадигматични черупки* за всеки един от получените подкласове на лексикалните характеристики – само като брой и структура на слотовете, готови за запълване. Тази операция ни дава структурата на парадигматичните черупки – рисунъка на клетките за запълване.

3. *Определяне на буквените стойности,* запълващи слотовете на черупката – низовете, участващи в трите основни операции на прехода от лема към словоформа – отрежи, залепи, промени, т.е. определяне на всички флективни типове.

4. *Свързване* на всяка лема от езика с нейния комплект информация – лексикални характеристики, парадигматична черупка и флективен тип.

5. *Образуване* на всички словоформи в езика съгласно този комплект информация.

Операциите в тази последна фаза фактически представят истинския синтез в неговия обем, често непосилен за преодоляване в разумно време, ако си го представим извършван върху лексикална единица след лексикална единица. Дори и да имаме организирани според някаква приблизителна буквена подредба лемите в езика (например, по крайните елементи в думата – нещо, което словообразувателните възможности на българския език улесняват в известна степен – вж. напр. групите съществителни, завършващи на *-ция*, *-не*, *-ател* и техните лексикално-граматични признаци), задачата все пак се свежда до: образувай всички форми за всички леми.

Оптимизацията на тази непосилна задача предполага да се използват няколко начина за съкращаване на времето за генериране. Те са свързани:

- с организацията на първоначалните входни данни – лемите от езика,
- с организацията на граматичното знание – да се изчисли изчислимото, а неизчислимото да се изброи по удобен начин,
- и последно – по място, но не и по важност, с дизайна на използваните компютърни приложения, ако са създадени специално за целите на синтеза.

2.9 Софтуерни приложения за морфологичен синтез

Тук ще разгледаме конкретните софтуерни решения за оптимизацията на производството на български словоформи във формат:

словоформа, лема, лексикални характеристики, граматични характеристики

Те са реализирани в специализирано софтуерно приложение, получило служебното название MBL (Morphology of Bulgarian Language), описано в [Paskaleva 2005].

Софтуерът е типично поръчково изделие – създадено за български език, за задачата на морфологичния синтез, за конкретен вид граматично представяне на изходния продукт, при определено разпределение между автоматичните и ръчните процедури.

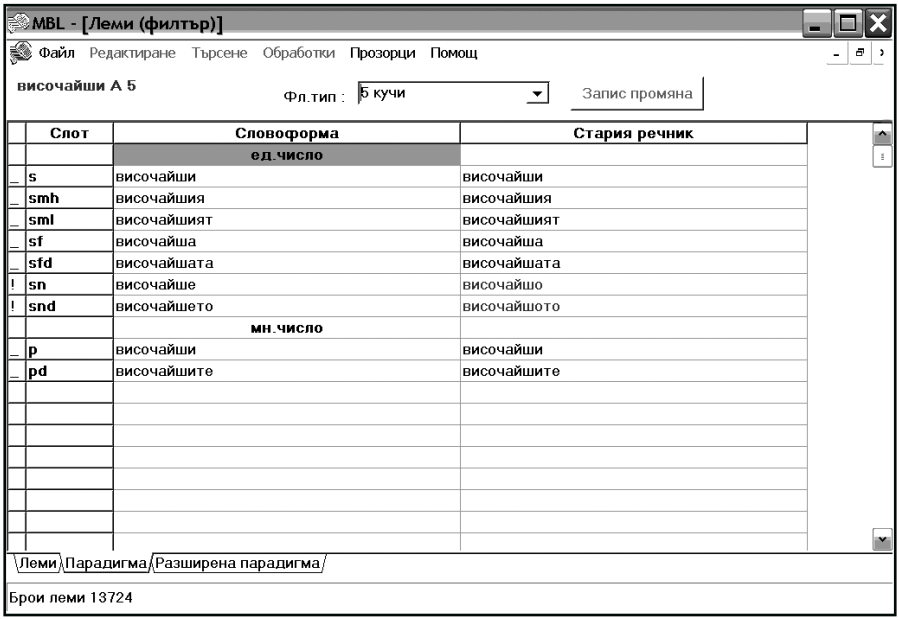
Принципите на автоматизация на някои изчисления на буквени стойности, както и възможността да се редактират основни зависимости между лексикални и граматични стойности го правят и един вид учебна илюстрация за механизма на правилата на българското склонение и спрежение, както и за проверка на правилността на образувани словоформи.

2.9.1 Входни данни и начини за въвеждането им

Основна единица, подавана на входа на системата, е лемата. Съществуват два начина за нейното въвеждане:

- поединично в меню *Нова лема*,
- във вид на речников списък във формат: словоформа, лема, лексикални и граматични характеристики, с определени разделители между тези четири типа данни и техните подтипове,
- списъци от леми без допълнителна информация. Това влече след себе си и възможности за еднотипна обработка на цели списъци, при гарантирано идентично граматично поведение.

Възможно е и въвеждането на повече от един речников списък (да ги наречем Речник 1 и Речник 2) и сравняване на техните параметри при генериране на подтипове данни – вж. Фиг. 2.



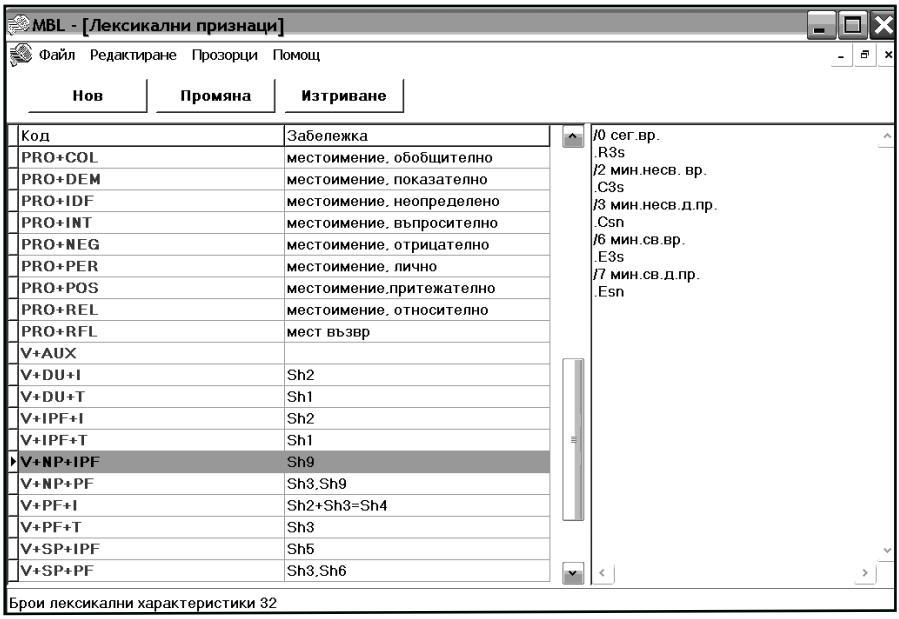
Фиг. 2. Сравнение между две версии на речника – текущ и минал, за определена лема в процеса на генериране на нейните форми, със сигнализация за различаващи се словоформи в двете версии.

Тук трябва да признаем, че българската деривационна морфология, по-точно суфиксалният й механизъм предоставя добри възможности за формиране на такива списъци. Както беше споменато, редица словообразователни суфикси

предопределят еднакво формообразуващо поведение. В достатъчно пълен речник на български словоформи, за който ще стане дума в глава 3, имаме 793 леми на същ. ср.р., завършващи на *–ние*, синтезиращи словоформи по еднакъв модел, и 1403 леми на същ. ж.р., завършващи на – *ия*, с еднакво формообразуване. Примерите могат да се продължат.

2.9.2 Организация на лингвистичното знание за максимална изчислимост

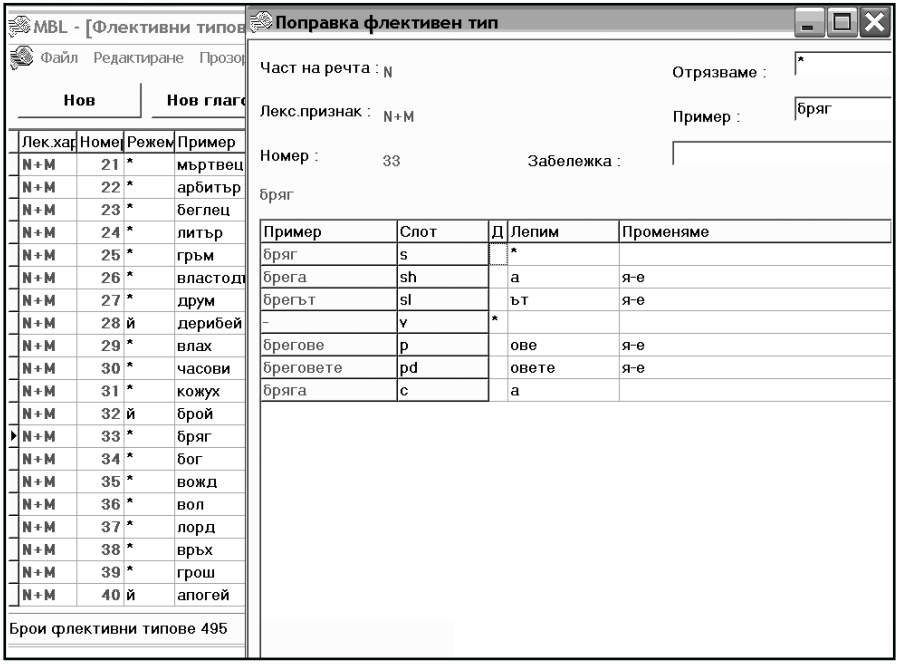
С въвеждането на нова лема в системата се пресмята и нейното място, изразено чрез йерархията на граматичните категории, като изборът на стойност на категорията на първото най-високо равнище (част на речта и лексикални признаци) предопределя възможните стойности на категориите от следващото равнище – това на граматичните категории. Тази връзка е заложена в архитектурата на системата и е специфична за конкретен език или модел. На Фиг. 3 е показана зависимостта между комбинация от признаците на лексикално равнище (за подклас безличен глагол) и броя и стойностите на граматичните категории в парадигмата на съответната лема, т.е. определянето на парадигматичния рисунък.



Фиг. 3. Комбинацията от лексикални характеристики на глагола (безличен глагол несвършен вид) обуславя парадигматична черупка от 5 члена, чиито граматични признаци са изчислени в дясно.

Този вид изчислимост – между частта на речта като основен морфосинтактичен признак и лексикално-граматичните категории, които тя имплицира, се осъществява все още в областта на системата и нейните равнища. Той е свързан със стойностите на граматичните категории като абстракти, като променливи, другояче казано само като наименования или типове.

Тази операция подготвя прехода към езиковата реалност, тъй като изборът на лексикалните категории предопределя мрежата на парадигматичната черупка. Парадигматичната черупка чрез подготвените слотове за запълване осъществява най-важния преход за синтезиращата операция – напълването на определената вече структура на черупката с конкретни буквени стойности. Тези стойности – буквени низове се попълват в разказаната вече последователност – отрежи, залепи, промени. Запълването на парадигматичната черупка на същ. м.р. *бряг*, зададено чрез флективния му тип, вж. на Фиг. 4.



Фиг. 4. Определяне на буквените стойности на даден флективен тип чрез запълване на предоставената парадигматична черупка.

Изборът на буквената стойност не винаги се свежда до един запис. С **"/"** се разделят възможни дублетни формообразувания, разрешен начин за означаване на равноправни варианти, които след това се генерират в отделни словоформи с едно и също граматично значение – вж. напр. *млян/млят*, *млях/мелих* и др.

Количеството на флективните типове, представящи формообразуването на 1 000 000 български словоформи от 65 000 български леми се разпределя по части на речта така:

- съществителни - 190 типа (мъжки род – 125, женски род – 36, среден род – 24, общ род – 2, плуралиа тантум – 3),
- глаголи – 260 типа (148 преходни, 76 непреходни, 22 безлични, 14 полулични – за класификацията на глаголите по залогово поведение вж. по-долу, 3.5.2)
- прилагателни – 20 типа,
- местоимения – 26 типа
- числителни – 11 типа.

Интерфейсът на системата ни предоставя фактически едно средство за лесно спрягане/скланяне на изменяемите леми (за 4 от частите на речта), за останалите процесът е завършил с определянето на лексикалните им характеристики, тъй като лемата и единствената словоформа в парадигмата съвпадат.

Макар и улеснен от предоставената мрежа на парадигматичната черупка, процесът на това попълване е доста бавен и еднообразен. Затова се налага да се използва още един вид изчислимост – между буквените стойности на запълнените парадигматични клетки.

Тази изчислимост не е нищо друго, освен често срещаните в нормативните граматики изрази, които свързват отделни буквени стойности на граматични значения като: ...*мекият вариант на членната форма -я/-ят се използва...,* или ...*всички съществителни имена от женски род в единствено число получават член -та,* или *окончания -я, -ят в 1 л. ед.ч. и 3 л. мн.ч. в* **I спрежение**...

В неявен вид тук се формулират константни или взаимозависими буквени стойности на граматични значения. Такава изводимост имаме за краткия и пълен член на съществителните – втората буквена стойност винаги е образувана от първата с прибавяне на едно *т*, същата зависимост имаме в постоянното равенство между буквените стойности на 2 и 3 л. ед.ч. на мин.несв.вр. (*-еше*) и за много други подобни случаи.

Тази изводимост дава възможност в слотовете (клетките) на парадигматичната черупка да се обособят т.нар. *котви* (anchors), които се задават ръчно и позволяват автоматичното запълване на други клетки, тъй като са изчислими – вж. предния пример. Не е необходимо това да става с трансформация на буквения низ, просто в системата предварително трябва да се зададат съответствията от буквени низове – *еше-еше*, *а-ът*, *я-ят*, *яхме-яхте* и др.

Функцията *Изчисли*, за която свързаните стойности на котвите и техните производни са проектирани предварително в системата, съкращава времето за ръчното запълване на парадигматичната черупка. Но докато за номиналното склонение тази икономия не е толкова съществена (спестява запълването на три-четири клетки), за глаголното спрежение тя е съществена по следните причини.

Показаната пълна парадигматична черупка на глагола съдържа 54 словоформи, чието последователно запълване очевидно е доста трудоемко. Въз основа на изследвания върху изводимостта на буквените изражения на граматичните значения бяха формулирани правила за изчисление на глаголното спрежение. Те, за разлика от *плоското* номинално изчисление на съществителните за всичките им седем възможни словоформи, например я→ят, бяха структурирани чрез обособяване на девет модула в парадигматичната черупка на глагола. Тази подкласификация позволи да се въведе два типа изчислимост – вътрешна, за стойностите в рамките на модула, и външна, използваща връзката между буквени стойности на граматически признаци в отделните модули.

2.9.3 Генерални правила за изчисляване на глаголно спрежение

Формообразувателните механизми на глаголното спрежение са разпределени в 9 основни модула, образуващи глаголната парадигма. Те са:

Сег. вр. | Сег. прч. | Мин. нсв. вр. | Мин. нсв. прч. | Деепр. |
Пов. | Мин. св. вр.| Мин. св. прч. | Страд. прч.

Първите пет се образуват от сегашната (или имперфектната основа) на глагола и са свързани в част от буквените си стойности (с някои индивидуални изключения, които са уговорени в модела). Вторите четири се образуват от аористната основа на глагола и са свързани в значенията си (също с изключения). Всеки модул има определен брой членове. Една част от тях се задават ръчно – като *котви*, а другите се изчисляват от тях. Това е *вътрешно* изчисление.

Има и изчисление между модули, което се осъществява само върху техните котви. Това е *външно* изчисление. То облекчава целия процес на попълването, като отменя необходимостта от ръчно задаване на някои котви. Прост пример за външно изчисление между аориста и неговото причастие е за класове глаголи като *мълча*, където буквената стойност на котвата 1 л.ед.ч. мин. св. вр. *ах* образува и котвата на мин.св. прич. *ал*. С доста по-сложно изчисление се задава връзката между тези две котви за глаголите с мин.несв.вр. на *ох* (*пекох-пекъл*). Интерфейсът на системата при глаголната парадигма предоставя два

типа команда *Изчисли* – изчисление на модул и генерално изчисление. В Табл. 9 по-долу се съдържа разпределението между членовете на пълната глаголна парадигма – като котви и изчислими стойности.

Модул	Брой членове	Котви			
		Задавани	Константни	Външно изчислени	Изчислими
Сег. вр.	6	2			4
Мин. несв. вр.	6	1	1		4
Сег. прич.	9			1	8
Мин. несв. прич.	4		1	1	2
Деепричастие	1			1	1
Повелително накл.	2	1			1
Мин. св.вр.	6	1			5
Мин. св. прич.	9			2	7
Мин. страд. прич.	9			2	7
Отглаг. същ.	2			1	-
Всичко	54	5	2	8	39

Табл. 9. Изчислимост в глаголното спрежение. От 54 члена в глаголната парадигма, само буквените стойности на 5 от тях се задават ръчно, 2 са константни, 8 се получават чрез външно изчисление от други модули, 39 се получават автоматично в рамките на модула.

Последователното прилагане на тази команда по веригата – вътрешно изчисление на пръв модул и последващо външно изчисление на котвите от останалите модули ни дава максималното облекчение, което може да получим в процеса на синтеза. Тази верига на изчислените стойности може да има различна дължина за различните типове спрежение на глаголите.

Най-дългата верига, която получава всички стойности само след като се зададат двете котви *м* и *ш* в сег.вр. 1 и 2 л.ед.ч., е тази на глаголите от 3 спрежение, които са 60 % от всички глаголи. След задаването на котвите всичко се изчислява автоматично.

Лингвистичният алгоритъм за глаголите от 1 и 2 спрежение е направен чрез обединяващи таблици, свързващи механизма на котви и изчисление с добре известните от нормативните граматики и ръководства по български глагол [Пашов 1966] изчерпателни описания на спрежение на глаголите, разделени на класове (наречени тук модули). Стойностите на тези таблици са отразени в алгоритъма и интерфейса на продукта, при задаването на съотношенията на изчисление. Пример за алгоритмично задаване на изчислимостта в рамките на един модул вж. в Табл. 10. Този вид таблица не фигурира в явен вид в системата, тъй като е вътрешно изчисление на софтуера.

Членове	Вид	Операция	1 мод.	2 мод.	3 мод.	4 мод.	5 мод.	6 мод.
1 ед.	Котва	1 симв (F1)	а	я	я	а	м	м
2 ед.	Котва	1 симв (F2)	еш	еш	иш	иш	ш	ш
3 ед.	Изчисл.	F2	е	е	и	и	#	#
1 мн.	Изчисл.	F2+м	ем	ем	им	им	ме	ме
2 мн.	Изчисл.	F2+те	ете	ете	ите	ите	те	те
3 мн.	Изчисл.	F1+т	ат	ят	ят	ат	т	т
Пример			*чета*	*вая*	*горя*	*мълча*	*давам*	*стрелям*

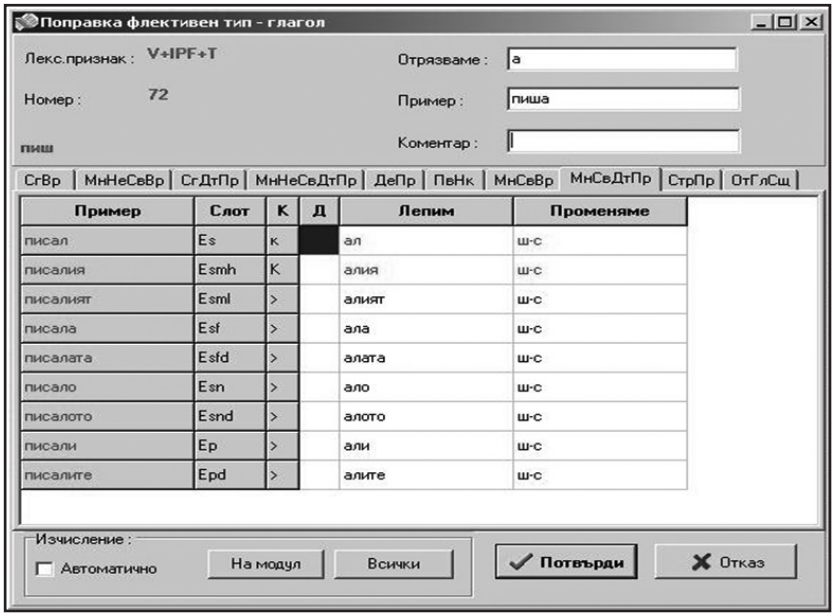
Табл. 10. Изчисление в границата на модул сегашно време за глаголи от типа, посочен в най-долния ред.

Разшифровка на операциите за тип *чета*:

F1 – единствен символ на Котва 1=*а* (*чет-А*) ; F2 – първи символ на двубуквената Котва 2=**е** (*чет-Еш*).

Изчисление след котвите: трети член – равен на **F2** (*чет-Е*), четвърти член – прибавя *м* към **F2** (*чет-еМ*), пети член – прибавя *те* към **F2** (*чет-еТЕ*) , шести член – прибавя *т* към **F1** (*чет-аТ*).

Интерфейсът на системата, осигуряващ последователно изчисление на буквените стойности в глаголната парадигматична черупка, е показан на Фиг. 5



Фиг. 5. Изчисление на буквените стойности на модула минало несвършено деателно причастие за глагола *пиша.*

2.10 Морфологичен анализ vs. морфологичен синтез – условия за обратимост

Подробно описаната процедура на морфологичния синтез, която

- получава като входен материал основните езикови единици – *лемите* (или съдържащите се в тях *основи* – чрез операция – *отрежи!*),
- по зададени граматични *категории* произвежда като изходен ресурс *словоформи* (чрез прилагане на операциите – *залепи!* и *промени!*),

би могла да бъде описана в обратния ред, а именно:

- от входния материал – *словоформи*,
- чрез *отрязване* на залепеното и евентуално *възстановяване* на промененото,
- да получи като изходен ресурс основната единица – *лемата* и граматичните категории към нея, присъщи на анализираната словоформа.

Тези операции не биха били възможни при синтеза без основния инструмент на флективните класове, които ни дават правилата за залепването, подобно на лего или пъзел играчка с посочен начин за сглобяване, т.е. към кои *X* кои *Y* може да се прилепват.

За анализа основният инструмент е аналогичен, само че обърнат, за идентификация на разглобените части – от какви *X* и какви *Y* е бил съставен сглобеният детайл.

При такава постановка процесите са обратими, участващите елементи са едни и същи, само дето заданията са формулирани в противоположна посока.

Предвид на обратимостта на морфологичния анализ и синтез, твърде подробното описание на синтезиращата процедура и нейните основни ресурси и правила, дадено в предишните раздели, може да бъде преформулирано в обратната посока, подобно на игра със сглобяване и разглобяване на лингвистичните детайли на езиковия пъзел. За разлика от пъзела, където по-лесно е да се разглоби, отколкото да се сглоби (тъй като липсва механизъм за *задължителното* комбиниране на отделните детайли, както това прави флективният тип в морфологията), в морфологичното производство е обратното. Синтезът винаги генерира правилно, а анализът не винаги познава правилно – вж. разсъжденията по този повод в 2.2.

Обратимостта на двата вида производство е позволила първите морфологични модели да се разработват едновременно за двете посоки. Основният инвентар на производството за високофлективните езици се състои от : основи, окончания и правила за тяхната съчетаемост и в двете плоскости - като текстови и като системни единици. Тази връзка на текстовото със системното представяне, на израза със съдържанието, е основният морфологичен инструмент и в двата вида производство.

2.11 Двупосочни морфологични модели

Компютърната лингвистика се е зародила много преди масовите комуникационни технологии и в първите си приложения е работила или върху изкуствено създадени данни, или върху материал, събран целево, за да обхване всички възможни варианти на предлагания изчислителен модел – най-често експериментален, за потвърждение на авторски хипотези за езика. Тежестта на нейните изследвания е била насочена към моделиране на човешкия начин на познание на езиковата реалност (моделиране на когнитивни процеси), а не към получаването на едни и същи резултати от обработката на текста от човек и машина (функционално моделиране) – както се схваща в днешни дни моделирането на езикови процеси.

Описанието на думите в езика и техните структури през 80-те години обикновено е компонент от т.нар. лексикална база данни, съдържаща богати сведения за думата на всички равнища. Естествено, задължителен компонент на тази база е описанието на процеса на образуване на нови думи. В първите сериозни модели на компютърната морфология основната обяснителна сила е поверена на генериращата процедура, а анализиращата е изводима от нея.

Движението от морфемата към думата – сглобяване на думата от съставните й части е предмет на прочутия двустепенен модел (Two level morphology), който близо две десетилетия след създаването си оказвал силно влияние върху компютърната лингвистика [Koskenniemi 1986].

Концепцията на модела е да се създаде езиково независим формален механизъм за генериране на словоформи от минимален брой елементи. Те са формулирани на морфофонетично равнище, с използване на специални символи за означаване на т.нар. *обобщена графема*, която по определени правила преминава в различни графеми – *пеК –> пек-ох* и *печ-ете*, при което осъществява прехода от дълбинното, т.нар. лексикално равнище към реализацията

на следващото, т.нар. повърхнинно равнище – lexical vs. surface level. Идеята е възникнала за финландския език, но е била аплодирана и от компютърни слависти, очаровани да видят едно стройно изложение на механизма на алтернацията в корена, съчетан с прибавянето на групи формообразуващи суфикси към последния.

В термините на двустепенния модел е разработено морфологично описание и за старобългарски език [Lindstedt1984].

Тази линия не е продължена обаче в българската компютърна лингвистика, може би поради причини, свързани с развитието на българския език, където фонетичните промени в основата не са вече така редовни и без изключения, както е било в старобългарския език (правилата за палатализацията – един от стълбовете на двустепенния модел, например, започват да не се спазват в новите думи, вж. *белег – белези* , но *лозунг – лозунги*).

Двустепенното представяне обаче, с начина на задаване на лингвистичните данни, позволява прилагането на двупосочен механизъм, който се формулира декларативно и се реализира чрез някакво изчислително устройство, достатъчно универсално, за да работи и в двете посоки.

Той е оказал влияние върху следващи изследвания главно с използването на крайните автомати за моделиране на лексикални явления.

По този начин – чрез декларативна организация на данните и изчислителните средства на езика ПРОЛОГ и насочването съответно към анализ или синтез според потребителската заявка, са организирани данните и на един от първите компютърни модели на българската морфология, чиято лексикална база съдържа 30 000 основни думи, известен като системата MORPHO-ASSISTANT [Paskaleva et al. 1990], [Simov et al. 1990].

Същата система може да се разглежда като прародител на повечето съвременни български граматични речници, без да претендира за абсолютно първенство (и преди нея са използвани инженерно-технически разработки за нуждите например на правописни коректори), но в нея за пръв път се вижда организиран – като декларативно знание и като процедури – един модел на българската морфология, опиращ се на достатъчно сериозен лексикален материал и граматични зависимости, обхващащи цялата система на българското словоизменение.

2.12 Двупосочен модел за български език на морфемно равнище

В модела MORPHO-ASSISTANT лингвистичният материал е организиран в речници на основи и на окончания. Флективният тип и типът алтернация е част от характеристиката на основата, тъй като определя еднозначно нейните формообразуващи възможности.

Като речникова единица, на повърхнинно равнище, в интерфейса за обработка на потребителската заявка обаче, MORPHO-ASSISTANT не изброява основи, а леми (основни форми). Източникът на лексикалните данни – 30 000 на брой, е правописен речник на българския език със същия обем [Андрейчин и др. 1970]. Вътрешното граматично процедурно описание към всеки флективен тип съдържа информация какъв буквен низ да се отреже от словоформата-лема, представител на парадигмата, за да се получи основата, и оттам да се премине съответно към едната от двете възможности на модела. Речниковата единица е снабдена с информация за лексикалните характеристики на лемата, което определя пътя на по-нататъшната серия от избори (сегментация, водеща към речниците на основи и окончания, след което, в зависимост от избраната посока, следва намиране на възможни граматични стойности за анализа или възможни буквени низове за синтеза).

Системата MORPHO-ASSISTANT е ползвана като генериращ софтуер предимно в демонстрационни приложения, илюстриращи възможностите на модела за компютърно обучение по българска морфология, за съставяне на тестове със същите цели и др.

Главното й достоинство и гордост за лингвистите-създатели на модела бе съчетанието между надеждно лингвистично знание и представителен лексикален материал. Формалното представяне и компютърната обработка на лингвистичните факти ползваше унификационния механизъм на features structures (структури на признаци), въведена в [Kaplan 1975], както и изчислителния механизъм на езика ПРОЛОГ.

Теоретическите съображения за чистотата на представянето от системно-лингвистична гледна точка, за единството между генериращ и анализиращ механизъм, за прозрачното представяне на механизма на формоизменението чрез участващите в него морфеми, бяха подсилени и от съображения от практически характер.

Те бяха свързани главно с изчислителната мощ на компютрите в края на 80-те години. Очевидно идеята за по-малък брой входни елементи и сложни правила за тяхната комбинация, които произвеждат голям обем изходни продукти, преследва ефективност и икономия на изчислителни ресурси.

В системата MORPHO-ASSISTANT 30 000 леми и няколко стотин окончания позволяват да се генерират или анализират около 700 000 словоформи, чийто директен запис и обработка в алтернативен, *плосък* вариант би затруднил сериозно компютрите Правец 16.

Организацията на основното лингвистично знание на системата MORPHO-ASSISTANT с решаващото и главно участие на един от нейните разработчици бяха творчески използвани и пренесени в други лексикографски традиционни и компютърни приложения, между които:

- правописни граматически речници [Попов и др. 1998],
- електронни правописни речници и коректори[15],
- мащабни компютърни разработки в областта на синтаксиса и интегрални системи за обработка на текст: BulTreeBank – българската банка на синтактични дървета [Осенова и Симов, 2007], системата за обработка на текстови корпуси CLARK[16].

Това многобройно семейство от близки и не толкова близки роднини, повечето от които благодарно отбелязват прародителя си – системата MORPHO-ASSISTANT, демонстрира една проста и изстрадана истина, а именно:

добре структурираното и формализирано знание за морфологичните процеси, където ефективен формален модел обработва изчерпателно лингвистично описание, дава стабилна основа за създаване на множество надстройки и приложения на езиковите технологии.

2.13 Преход от двупосочен към „плосък“ морфологичен модел

Времената се менят обаче и изчислителната мощ на компютрите нараства. Това позволява двупосочните морфологични модели да се развиват на друга плоскост – с надежден изчислителен ресурс, вече имплантирани в универсална система за обработка на големи текстови корпуси, анализ и синтез на текст и многобройни текстови приложения.

15 http://www.slovnik.bg
16 http://clark.space.bas.bg

Такава водеща система за обработка на текст, реализирана за много езици, е системата INTEX [Silberztein 1993]. Тя е възприемана широко и адаптирана не само заради сериозните си лингвистични основи, но и с изчислителния си апарат – системата на крайните автомати и преобразователи (finite state automata and transducers). Той продължава да се развива и до днес и дава база за приложение на много езици, последният включен в системата е македонският език [Zdravkova & Petrovski 2007]. В системата INTEX, чрез апарата на крайните автомати в изчислението и удобен графически интерфейс в представянето, са обединени данните, осигуряващи не само възможните морфологични операции в двете посоки, но и други видове обработка на текстови корпуси.

За нашето изложение интерес представлява системата от речници на INTEX, където са представени четири типа речници. Два от тях представят равнищата на лемите и словоформите, организирани по два различни начина – чрез апарата на формообразуването и чрез механизма на плоското изброяване. Това са речниците:

- DELAS, в който са изброени леми и флективни класове,
- DELAF – словоформи с техните лексикални и граматични характеристики,
- DELACF – словоформи и произведените от тях сложни думи,
- DELAC – леми и произведените от тях сложни думи.

Примери за речниковата структура на четирите типа речникова организация на INTEX[17]

DELAF **amuserions**, amuser. V3+t+4:C1p
DELAS **amuser**. V3+t+4
cousin. N32+Hum
cousin. N1+Anim
DELACF **pommes de terre**, pomme de terre. N+NDN:fp
tout de suite, tout de suite. ADV
DELAC **marche antinucleaire**. N+NA:fs/—+;une
pomme de terre. N+NDN:fs/—+;une
tout de suite. ADV

За морфологичния модел от значение са първите два речника, при което не е трудно да се види, че вторият тип (DELAS) е с аналогична структура на MORPHO-ASSISTANT, а първият тип (DELAF) е произведен от него, чрез изчерпателно генериране на всички словоформи от единиците на DELAS. Речникът DELAF съдържа всички словоформи на езика, асоциирани с лема, морфосинтактичен код, евентуално синтактико-семантични кодове и евентуално словоизменителни кодове.

17 http://www.cascadilla.com/manym/daille210-212.pdf

В примера се вижда паралела на френския морфосинтактичен код с обсъжданите по-горе (в 2.7) лексикални характеристики – част на речта, вид на глагола, а също така се виждат и граматичните характеристики на словоформата.

В сътрудничество с бащите на тази система – лабораторията LADL в Университета Париж VIII, беше разработен първият български модул на системата INTEX (версията, разработена за платформата NeXTStep и Sun workstation). Флективните класове от системата MORPHO-ASSISTANT бяха трансформирани в система от крайни автомати, представяща българското склонение. Бе генериран и българският речник на словоформите – DELAF[18], съдържащ повече от 500 000 словоформи, образувани от 30 000 леми – единиците на MORPHO-ASSISTANT (вж. по-горе, с. 69).

Безкрайната полезност на разработеното морфологично представяне в първата българска INTEX система и генерирания речник на български словоформи бе потвърдена от многобройните разширения и допълнения на данните и механизма в това представяне. Тук може да се спомена:

- уеб-приложения като Граматичния сървър на българския език, разширен с още 30 000 леми върху използваните вече принципи[19],
- граматичен речник за целите на обучението по компютърна лингвистика по програмата ERASMUS в Новия Университет на Лисабон,
- граматически речници за други езици [Паскалева 2002],
- демонстрационни версии на морфологичен анализ на български текст, както и лингвистичната анотации (в европейския проект BALRIC-LING IST-2000-26454[20],
- следващите версии на системата INTEX за български с решаващо участие на изследователи, разработвали първата версия на системата.

Бракът между организацията на лингвистичното знание на MORPHO-ASSISTANT и формата на речниците DELAF доведе до създаване на един основен граматичен ресурс – български граматичен речник, конструиран като речник на словоформите с информация за тяхната лема, лексикални и граматични признаци.

Този ресурс и неговото по-нататъшно използване в компютърни приложения заслужава отделно разглеждане като основно средство на компютърния морфологичен модел.

[18] Системата от крайни автомати, представяща българското формоизменение, бе разработена от сътрудниците на Секцията за лингвистично моделиране: Стоян Михов – формален апарат и процедури, Милена Славчева – лингвистични данни и интерпретация.

[19] http://lml.bas.bg

[20] http://www.larflast.bas.bg/balric/index/index.htm

Лексикални бази от този тип се наричат **граматични речници** и са ориентирани към практически цели, които излизат извън рамките на компютърния модел на морфологията, описван в гл. 2 на това изложение.

От времената на създаване на първите компютърни морфологични модели се промениха концепции и технологии. С нарастването на мощта на изчислителната техника и разширяването на нейните приложения се налага и един прагматичен поглед върху нещата. Той се изразява в стремежа да се разделят функционално компонентите на лингвистичното знание в системата от реалните резултати – изходни данни, които участват в индустрията на информационните технологии, в тяхното впечатляващо разнообразие.

2.14 Организация на лингвистичното знание в „плоския“ морфологичен модел

За плоския модел на представяне на морфологичното знание бе споменато при описанието на софтуера за морфологичен синтез (в 2.9), който генерира словоформи по възприет формат (този на речниците DELAF). Максималната производствена мощност на модела поражда всички словоформи на дадения език с техните лексикално-граматични характеристики. Словникът, лексикалният материал за това производство обикновено се взима от съществуващи правописни и тълковни речници, допълва се с нови думи от новинарски издания, уебсайтове и други динамично обновявани информационни източници.

Така генерираният речник фактически съдържа – вече анализирани! всички думи на езика, които (в идеалния случай) можем да срещнем в електронен текст.

Съвпадането на единица от текста с определена единица от речника фактически ни дава нейния анализ. А тъй като единицата в речника е произведена от процеса на морфологичния синтез, така се осъществява един естествен *преход от изходния продукт на синтеза към процеса на анализа.*

Тази проста концепция на търсенето – *образувай всичко възможно и търси!* заменя по-съдържателната от гледна точка на морфологичното знание операция, която беше описана като симетрична на синтеза и прилагана в двупосочните модели.

Ако искаме анализът да е симетричен на синтеза, произвел словоформите, той трябва да разделя вече обединена единица (от текста) на двете съставки – лексикална и граматична, и към нея да припише сумата от тяхната информация. Стъпките на тази операция са: сегментирай, промени (евентуално), припиши информация.

В конкретна процедура това изглежда така:

- дадена е словоформата *белязахме*,
- дадени са два списъка – на основи и на окончания, лексикален и граматичен,
- с търсене в граматичния списък се извършва операцията – отрязване на окончание, което съвпада с края на словоформата, в случая *-ахме,*
- отрязъкът *беляз-* се търси в лексикалния списък, и ако не се намери, се правят предварително зададени буквени замени, докато се установи неговата функционална идентичност с лексикалния елемент – *бележ* – основата на лемата на тази словоформа,
- обединяват се в едно лексикалните и граматичните свойства на сегментираните елементи,
- тази сума се приписва към словоформата *белязахме*.

Тази операция е красива от гледна точка на лингвистичното знание. Тя ни обяснява много неща – и за морфемната структура на думата, и за морфофонетичните промени в основата. Тя е и икономична – извършва се върху малък брой лексикални и граматични елементи (за българския език – върху няколко десетки хиляди лексикални и няколкостотин граматични).

Възниква обаче въпросът – нужна ли ни е толкова сложна процедура, при положение, че:

- нямаме нужда да пестим изчислителни ресурси при толкова ниския порядък на съотношението словоформа:лема, което е само 16:1;
- интересува ни не процесът, а резултатът от анализа на словоформите. Този резултат е само брънка от веригата на по-сложни операции, които не целят пълното моделиране на лингвистично знание, а информационни резултати, несвързани с лингвистиката.

Прагматични съображения, свързани с целта на по-нататъшната компютърна обработка, правят този въпрос реторичен, а отговорът – подразбиращ се. Плоското, линейно представяне на данни, произведени от йерархично организирано и подредено знание, не отменя лингвистичната валидност на продукта, произведен от тези данни. Главният инструмент на плоския морфологичен модел, създаден на тези принципи, е граматичният речник, който обединява единиците, произведени от елегантните процедури на морфологичния синтез.

3 Граматичен речник

Граматичният речник е централен компонент на почти всички системи за компютърна обработка на текста. Той е продукт на прагматичен подход, който не моделира нещо заради самото моделиране, тъй като се интересува само от резултата на процеса, а не от самия процес. Този прагматичен подход води до конструирането на *плосък* апарат за анализ, построен върху една единствена операция – намиране по пълно съвпадане (matching).

Естествено, тази простота на механизма не трябва да води до загуба на знание за езиковите единици, тъй като той пренебрегва само правилата на тяхната съчетаемост, използвани в построяването на речника, но не и съдържащата се в тях информация.

Тази информация може и трябва да бъде запазена и представена в други, комплексни по лингвистичната си същност ресурси на анализа. В техния обем, подредба, вътрешна структура, йерархия е отразена цялата сложност на знанието за Думата и нейното поведение.

Това знание – при подхода на пълното съвпадане на нещото, с което разполагаме, с нещото, което анализираме, е отразено пряко в структурата на граматичния речник от типа DELAF, още наричан речник на пълни словоформи (full form dictionary). Оттук нататък ще го означаваме като СФ-речник (словоформен речник).

В по-нататъшното изложение става дума за конкретен речник, с неговия формат (съставен от словоформи) и произход. Това е резултатният граматичен ресурс, обединил концепциите за моделиране на морфологичното знание в описаните дотук разработки, реализирани в Секцията за лингвистично моделиране на Института за паралелна обработка на информацията, БАН. Речникът съдържа 73 440 леми, образували 1 141 331 български словоформи. Върху конкретните лингвистични решения, отразени в данните на този речник, е построено по-нататъшното изложение.

3.1 Структура на граматичния речник

Речникът на словоформите фактически ни дава, предварително анализирани, подредени по азбучен ред, всички думи в езика.

Линейната структура на представяне на информацията в този плосък речник не отменя нуждата от обособяване в тази линейна информационна верига на отделни функционални фрагменти. На уточнение подлежат само:

- характерът на тези фрагменти и
- начините за разграничаването им.

След като уточнихме, че основна единица на този речник е словоформата, следва, че заглавието на *речниковата статия* (ако можем да наречем с това име споменатата линейна последователност от лингвистични единици и граматична информация) е *словоформа*.

За избягване на аналогията с обикновените речници – лексикографски пособия, оттук нататък ще използваме означението *речников запис* вместо речникова статия.

Речниковият запис е разделен на две зони – първата е идентификационна, а втората информационна.

Първата зона съдържа два елемента, конкретни лексикални единици:

- споменатата вече словоформа и
- лемата, от която е произведена тази словоформа.

Втората, информационната зона на речниковия запис, е също двуелементна, тъй като съдържа характеристиките на първите две единици:

- характеристиките на лемата (ще ги наричаме лексикални характеристики),
- характеристиките на словоформата (ще ги наричаме граматични характеристики)[21].

Двата вида характеристики в използвания формат се изразяват чрез комплексни буквени вериги, чиито отделни компоненти отразяват стойностите на отделни граматични категории.

[21] В нормативните граматики тези признаци се определят като лексикално-граматични, т.е. това са граматични значения, изразени на равнището на лексемата. Те се противопоставят на морфологичните, изразени на равнището на словоформата. Тук за простота оставяме означенията – лексикално и граматично, без да отричаме комплексната природа на информацията, приписана към лемата [Граматика 1998], с.306.

Четирите основни компонента на речниковата статия също са разделени със специфични разделители. Приетата структура и използвани разделители в речниковата статия на възприетия формат на СФ-речника е следната:

Словоформа, Лема.Лексика:Граматика

съкратено: Сф, Л.Лекс:Грам. Разделителите са зададени както следва: между Сф и Л – запетая, между идентификационната и информационната зона – точка, между Лекс и Грам – двоеточие.

Лекс и Грам могат да съдържат повече от един елемент, във формат

- A+B+C – за Лекс. Записът съдържа само главни букви, разделителят + (плюс) отделя съставните порции на лингвистичното знание за лемата, като една порция (признак) може да съдържа и повече от 1 символ,
- abcd – за Грам. Записът съдържа само малки букви, не повече от един символ за признак, без разделител помежду им.

Пример за речников запис:

сатириците,сатирик.N+M+H:pd,

който се разчита така: словоформата *сатириците*, произведена от лема *сатирик* (съществително м.р. одушевено), с граматични свойства – множествено число, членувано.

Както се вижда, в този запис се съдържа резултатът от процеса на синтеза, който е залепил към основата *сатирик* окончанието *ите*, като е променил *к* в *ц,* с други думи словоформата *сатириците*.

А анализът, вместо да търси в лексемния и в морфемния речник елементите *сатириц* и *ите*, да променя *ц* в *к*, за да съвпадне със *сатирик,* и да приписва на словоформата сумата от информацията към тези два елемента, може просто да потърси пълното съвпадане на *сатириците* в текста и *сатириците* в речника и след като го намери, да присъедини останалите три елемента от записа към словоформата в текста.

Предложеният формат е универсален, в известна степен езиково независим (за езици, които различават лексикалното и граматичното в думата), така че оценката на пълнотата на един речник, на степента, с която покрива лексикални и морфологични явления, е свързана повече със съдържанието на информационната част на речниковия запис.

3.2 Съдържание на граматичния речник

След като уточнихме в общи линии простата структура на речниковия запис, състояща се от четирите елемента – словоформа, лема, лексикални характеристики, граматични характеристики, остава открит въпросът – за какви точно характеристики става дума?

Отговорът на този въпрос е езиково-зависим, т.е. ориентиран е към конкретен естествен език. Обаче твърде голямото припокриване, съвпадане на категориалното описание на повечето европейски езици отдавна доведе до идеята за обединяване на принципите на речниковия запис в интернационален стандарт още през 1996 г. в проектите EAGLES[22] и MULTEXT-East[23].

И според универсалния принцип – за да обединим, трябва да разграничим, пътят към създаване на един такъв стандарт минава през грижливото изработване на принципите за определяне на информацията, приписана към словоформата като резултат от граматичния анализ.

Тази информация е базова за всички по-нататъшни етапи на компютърната обработка, каквото и да се извършва в нея – от „ювелирно" изследователско моделиране на езиково знание до сериозни информационни технологии върху огромни обеми данни. Това е първият мост, върху който се стъпва, за да се премине от реалността на текста към системата на неговото описание. Тъй като това описание в крайната си точка може да съдържа знание от най-различен вид – синтаксис, семантика, комуникационни структури, прагматика, дискурс, то очевидно е, че първият мост на прехода трябва да бъде доста солиден, за да осигури по-нататъшните пътища на обработката. Но от него не може да се очаква всичко, дори и само като замислено, а не осъществено, именно поради разнообразието в спецификата на следващите процедури. Не може да се предвидят евентуалните нужди на всички следващи обработки, а дори и за тези, които са ясни (например построяване на синтактични структури, ерго, включване на синтактични свойства в речниковата информация), не е целесъобразно информацията да се трупа предварително. В избора на това оптимално множество от признаци трябва да се приложи отново вече споменатият принцип на изчислимостта. Ако някои по-сложни, комплексни признаци, нужни за следващ етап на анализа са изводими от наличната базова съвкупност на заложените в анализа признаци, няма нужда те да фигурират в нея още от началото.

[22] http://www.ilc.cnr.it/EAGLES/home.html
[23] http://nl.ijs.si/ME/

Процесът на прехвърляне на тази базова съвкупност от признаци от територията на речника върху територията на текста – резултат от съвпадане на текстови и речникови елементи, се нарича граматично, а също и морфосинтактично **анотиране** на текста (morphosyntactic annotation). Отделните стойности на лексикалните и граматични признаци са анотационни *маркери* (tags). Последователността от маркери, приписани към една дума, ни дава нейната анотационна *верига*, а сумата от всички маркери, използвани в анотирането, ни дава *анотационното множество* (tagset) за дадения език.

Анотирането само с помощта на такъв речник може да се определи като анализ, неднозначен в много от случаите, тъй като се свежда до механично прехвърляне на маркерите от речника върху маркерите от текста при намерено съвпадение на буквената част на текстовата и речниковата единица. Поради омография на лексикални и морфологични елементи (съвпадащ буквен състав и различен граматичен прочит) това прехвърляне не е еднозначно и е възможно към една текстова единица да бъде приписан повече от един комплект анотационни маркери тагове[24].

3.3 Анотационно множество и подходи за определянето му

Стратегията за определяне на анотационното множество трябва да почива преди всичко на прагматични принципи – за какво ще ни служи тази информация в по-нататъшните фази на обработката?

Дори и да не сме много наясно какво по-точно ще правим по-нататък и как ще се развият нашите намерения, трябва да съзнаваме, че подготвяме основния строителен материал за реализирането им, а очевидно трябва да имаме представа поне за следващите няколко стъпки след анотирането.

При проектирането на един граматичен речник трябва да се избегне изкушението той да се третира като едноезичен тълковен речник, съдържащ всичко за Думата, или пък като правописен, съдържащ само думи, с откъслечни сведения за особености при формообразуването им.

---

24 Терминът *tag* в езици, твърде различни от английския, се предава често чрез калкиране – вж. руското *тег*. Процесът, означаван с термина *tagging*, е „изчистено" анотиране, без омонимия на лексикални и граматически елементи. Софтуерните средства за такова анотиране се наричат *taggers*. Въпросът за подходящ български термин, който не използва чуждото tag, остава открит. Употребява се терминът *тагер*. Що се отнася до термина *маркер*, в оскъдната българска литература по специалността се среща и терминът *етикет*. И двете находки са достатъчно небългарски, но се приемат като по-малкото зло.

Най-близка аналогия на граматичния речник са речниците, които съдържат и формообразуващи таблици, а към всяка изменяема дума има препратка към съответната таблица на спрежение или склонение. Такъв е известният граматичен речник на Зализняк, познат на всеки русист [Зализняк 1977], такъв е и цитираният речник [Попов и др. 1998], а като пособие от този тип може да се посочат и известните таблици на българското склонение и спрежение [Кръстев 1990]. Това е най-близката, но и най-бедната аналогия, тъй като граматичният речник съдържа и други показатели, освен тези на спрежението и склонението.

Каква е ролята на тези показатели?

Принципно решение, което не винаги се осъзнава от създателите на подобни речници е, че такъв речник не е обърнат към миналото, а към бъдещето, не към историята на получаването на думата, а към нейното бъдещо използване. Той не се интересува как се е образувала словоформата от дадената към нея лема. Това автоматично изключва сведения за флективен тип при всякакъв вид организация на речника. Граматичният речник дава резултата от това формообразуване, тъй като съдържа всички словоформи на дадена лема. Той ни поднася яденето и не се интересува от рецептата, нито от технологията на приготвянето му.

Кои трябва да са тези показатели?

При обсъждане на обема на анотационната верига минималният й състав очевидно трябва да включва:

- частта на речта като основен морфосинтактичен клас, който като основна характеристика на лемата предопределя и индивидуалните граматични свойства на членовете на нейната парадигма,
- стойностите на онези граматични признаци, които разграничават отделните членове на парадигмата.

Максималният състав на анотационната верига предполага отговор на въпроса – а какво още освен част на речта и разграничителните признаци за словоформата трябва да съдържа това множество?

Ако изключим от това множество признаци с определена синтактична и семантична насоченост – например глаголна валентност или семантичен клас, трябва да решим един друг въпрос. Кои от класификационните признаци, споменавани в нормативните граматики за всяка част на речта в разделите, посветени на нейните свойства и класификация, имат място в нашата верига? От тях очевидно ще ни потрябва информация, необходима поне за следващия етап

на анализа – изчистването на омонимията. Така например, маркерът *одушевеност* за съществителните от мъжки род снема омонимията на окончанието *-а*, употребявано за ед.ч. кратък член и за бройна форма (*вдигнах стола* vs. *два стола*). Граматиката отбелязва, че тази форма е характерна за неличните имена и следователно анотационната верига трябва да съдържа белега лице-нелице, за да можем да се опрем на него, когато снемаме омонимията чрез анализ на съседните думи.

Понеже не е възможно още при съставянето на речника да се предвидят всички класификационни признаци на съответната част на речта, които ще играят роля в следващото изчистване на омонимията, съставът на анотационната верига може да претърпява редакции и изменения доста време след като речникът е съставен и употребяван за други цели. Това се налага често, още повече поради изискванията за икономично анотационно множество, които се поставят обикновено от процедурите за снемане на омонимията на статистически принцип. Всяка следваща процедура, в която ще е използван този речник, може да наложи промени в анотационната верига. Така че оформянето на анотационната верига по принципа – всичко, каквото знаем за Думата, не е за препоръчване в случая.

Съставът на анотационната верига варира както между отделни речници за един и същ език, така и между версиите на един и същ речник, променян през годините, в зависимост от целите, на които е служил. Възможно е и преформатиране на анотационната верига – в означения и в структура.

За да илюстрираме тези концептуални промени, ще проследим анотационните вериги за граматичен речник на български език – както за отделни версии на един и същ речник, така и в съпоставка с други български граматични речници (там, където има материал за дискусия). Полезно е и привличането на аналогичен материал от други езици, особено родствени.

Обобщаването на взетите решения ще се осъществява поотделно за различните части на речта – така, както се процедира в общите международни анотационни стандарти.

3.4 Два вида представяне на анотационното множество

Обсъжданата дотук съдържателна информация, отразена в анотационното множество за даден език, се отнася за лингвистичната йерархия на представе-

ните в него признаци на лемата и на словоформата, локализацията им в приетите равнища на езиково описание и т.н.

Чисто формалната страна на означаването на тази верига от признаци – като подредба на съставките й, като структура на представянето им, ни предлага два типа анотиране – свободно и позиционно.

Свободното анотиране може да притежава две степени на свобода:

1. Абсолютна, при свободен плосък запис, линейна последователност от всички маркери за словоформата. Пример за такова анотиране, разглеждащо седем анотационни схеми в модела AMALGAM, вж. в Табл. 11[25].

	Brown	ICE	LLC	LOB	PARTS	POW	SEC	UPenn
select	VB	V(montr,imp)	VA+0	VB	adj	M	VB	VB
the	AT	ART(def)	TA	ATI	art	DD	ATI	DT
text	NN	N(com,sing)	NC	NN	noun	H	NN	NN
you	PPSS	PRON(pers)	RC	PP2	pron	HP	PP2	PRP
want	VB	V(montr,pres)	VA+0	VB	verb	M	VB	VBP
to	TO	PRTCL(to)	PD	TO	verb	I	TO	TO
protect	VB	V(montr,infin)	VA+0	VB	verb	M	VB	VB
.	.	PUNC(per)	.	.	.	.	.	.

Табл. 11. Анотиране на изречение според 7 анотационни схеми.

2. Относителна, при свободен структуриран запис. Той е свободен в съдържание и подредба, в рамките обаче на предварително уточнени съставни части и разделители между тях – както в цитирания запис на СФ-речника в системата INTEX.

Позиционното анотиране предполага запазване на фиксирана позиция за всеки маркер. Фиксирането на позицията може да е задължително за абсолютно всички маркери – за всички подкласове след главния клас, а може да е задължително само в рамките на един подклас (обикновено главният клас е частта на речта).

Пример за абсолютно плоско, позиционно анотиране имаме в чешката компютърна морфология, проектирана за статистически тагер[26]. В него 15 позиции определят цялото анотационно множество. Цената, която се плаща за „плоското" и фиксирано представяне, е увеличаването на стойностите в това представяне, например признакът *Род* има 11 значения. Те ни дават всички

[25] http://www.comp.leeds.ac.uk/amalgam/amalgam/amalgdesc1.html
[26] http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html

възможни комбинации от известните на всеки славянски език три граматични рода, съчетани с допълнителни признаци за одушевеност, число и др., вж. Табл. 12.

GENDER

Value	Description
-	Not applicable
F	Feminine
H	Feminine or Neuter
I	Masculine inanimate
M	Masculine animate
N	Neuter
Q	Feminine (with singular only) or Neuter (with plural only); used only with participles and nominal forms of adjectives
T	Masculine inanimate or Feminine (plural only); used only with participles and nominal forms of adjectives
X	Any of the basic four genders
Y	Masculine (either animate or inanimate)
Z	Not fenimine (i.e., Masculine animate/inanimate or Neuter); only for (some) pronoun forms and certain numerals

Табл. 12. Анотационни маркери за чешкия род.

Друг вид позиционно, плоско анотиране имаме в създадения и подържан дълги години международен стандарт за анотиране в споменатия проект MULTEXT-East[27]. Това анотиране ще означаваме оттук нататък като MTE-стандарт – европейска инициатива за унификация на морфосинтактичното анотиране, осъществено за 12 европейски езика[28].

В това анотиране в рамките на всеки подклас след главния – Част на речта, анотационната верига се конструира по следните принципи:

- категориите – атрибути са маркирани според позицията си, която е фиксирана,
- стойностите на категориите се означават с една буква,

[27] http://nl.ijs.si/ME/V3/doc/

[28] Позоваването на MTE стандарта в това изложение се нуждае от следващото уточнение: съществува един първоначален MTE стандарт за български език, който фигурира в официалните отчети на стандарта (изработен и проверен върху речника на неголям текст от 87235 думи). Впоследствие приетата в него анотационна схема е изменена, подобрена и ориентирана към по-нататъшни приложения като напр. BulTreeBank [Осенова и Симов 2007], а също и при пълното преформатиране на свободно анотирания СФ-речник, основа на това изложение, в термините и принципите на MTE.

- специални маркери (-) отбелязват нерелевантността на даден атрибут (за целия език или за дадената категория – подклас). Тук за отбелязване е различният прочит на тази нерелевантност – отнесена към част на речта, към неин подклас или към целия език. Явната нужда от уточняване на диапазона на тази нерелевантност е довела до корекции в последната версия на този стандарт за български, в преформатирана версия на СФ-речника.

Тъй като задължителната позиция на признак се реализира само в рамките на отделна част на речта (маркерът за Род е втори за съществителните, трети за прилагателните и шести за глагола), MTE-стандартът може да се определи като частично позиционен.

Позиционирането е фиксирано обаче в границите на първия подклас след Част на речта. При части на речта с неоднородно морфологично и синтактично поведение, изменящи се по различни параметри, каквото е например местоимението, броят на позициите се увеличава, съответно нарастват и празните, нерелевантни позиции, белязани с тире. Вж. например анотационното множество за българското местоимение *всички*: Pg--p----a-----,
където се запълват само следните позиции:

1 – част на речта – местоимение – P
2 – тип – обобщително – g
5 – число – множествено – p
10 – тип на референта – атрибут – a

Другите позиции: 2, 3, 4, 6-9, 11-15, остават незапълнени, тъй като част от тях са нерелевантни за този тип местоимение (като напр. падеж, число на притежател, клитика, определеност, одушевеност), а друга част – нерелевантна за българския език (последните 3 позиции).

Позиционният тип анотиране води до неудобни нотации като: Ps3----------q и Ps3----------r, маркиращи английското *whose*, първият път като въпросително, а вторият път като относително местоимение – дългата верига от нерелевантни признаци се дължи на позицията на атрибута WH-type, поставена накрая на веригата.

Подобни дълги вериги се получават обикновено при общи за няколко езика стандарти, какъвто е случаят с MTE-стандарта. Специална позиция се отделя за румънските местоимения – в силна и слаба форма, за чешките местоимения, които приемат форма на клитика, присъединяващи 's'. При тази борба за място в редицата е възможно признак, характерен само за един език, да се окаже последен. Ако го предшестват няколко нерелевантни за дадения език призна-

ка, се получава анотационна верига, подобна на тази на *whose*.

По тази причина неудобства причинява и осъвременяването на стандарта, което предполага да се прибавят нови признаци при промяна на анотационната концепция. В този случай позиционната верига не позволява разместване, а само прибавяне в края. Например, признакът *Лице на притежателя*, въведен в разглеждания стандарт само за унгарските съществителни и прилагателни, е очевидно необходим за анотация на българското притежателно местоимение. За българските притежателни местоимения в първоначалната версия на стандарта са предвидени признаците *род* и *число* на притежателя, докато *лице* на притежателя липсва. Тъй като този признак, отреден само за унгарската именна система, е поставен накрая, това прибавяне в осъвременената версия на стандарта ще доведе до запис P:s-fs-p-na-y---2 за местоимението *вашата* (разгънат прочит: местоимение, притежателно, женски род, единствено число, множествено число на притежателя, не_клитика, тип на референта – атрибут, определеност – да, второ лице на притежателя).

Подобни записи не са никак мнемонични, което затруднява прочита и оценката им от човек, но улеснява автоматичната обработка на анотирания текст, особено в междуезикови приложения, където се обработва унифицирана граматическа информация.

Заради това удобство на прочита в следващите раздели ще разгледаме подробно възможните анотации за българските словоформи, по отделни части на речта, като ползваме свободното анотиране, а сравнителната оценка на анотационни множества, предложени в други приложения, ще се извършва само по параметрите на съдържанието, а не по позиция.

Обобщено, приликите и разликите между свободното и позиционното анотиране са дадени в Табл. 13:

Параметри	**Свободно анотиране**	**Позиционно анотиране**
Единици на анотацията	Категории и техните стойности	Категории и техните стойности
Позиция на категориите	Фиксирана за наличните стойности	Фиксирана за всички стойности
Дължина на категориалната верига	Променлива – в рамките на една част на речта	Еднаква в рамките на една част на речта
Дължина на низа за една стойност	Един и повече от един символ	Един символ
Нотация на признаците	Лексикални характеристики – главни букви; граматични – малки букви	Част на речта – главна буква; останалите – малки букви
Символи извън означенията на категориите	Разделители	Маркери за нерелевантност

Табл. 8. Прилики и отлики в параметрите за свободно и позиционно анотиране.

3.5 Анотационно множество на български граматичен речник

В продължение на години разширената версия на MORPHO-ASSISTANT в DELAF формат е служила като основен модул на многобройни компютърни приложения. Това широко приложение на речника, заедно с по-ясното и стегнато представяне на лингвистичните данни, което ни дава използваният свободен анотационен формат, ни дава основание да построим по-нататъшното изложение в термините на това представяне. Ще следваме подредбата на лингвистичната анотация, започвайки след първото равнище на представянето – делението по части на речта, което е връх на дървото от граматични стойности и пръв символ в анотационната верига. Ще бъдат разгледани стойностите на анотационните маркери за всеки клас X, групирани като категории, които характеризират X, и категории, по които се изменя X – лексикално-граматичните и морфологични категории, тук обозначени като лексикални и граматични.

Поради това, че анотационното множество е обърнато към бъдещето, а не към миналото на обработката, критериите за неговата оптималност се извеждат трудно – миналото се формулира точно, докато бъдещето – не винаги. Има анотационни стойности, които очевидно ще са нужни за непосредствено следващ етап, който се очертава, без да може да се опише пълно. Има анотационни стойности, от които лингвист трудно би се отказал, не толкова заради пълнотата на системното описание, колкото заради явната им ценност за бъдещ синтактичен или семантичен анализ (колкото и да е отдалечен във времето и намеренията). Ориентир за избора между тези стойности с потенциална ценност трябва да са и нормативните граматики. Без да се преписват всички твърдения за стойностите на граматични категории в разделите по морфология, те трябва да бъдат прецизно проучени. Дори споменаването на възможни отклонения от общоприети характеристики трябва да се вземе предвид. Така в разглежданото анотационно множество за всяка част на речта се обособяват три групи:

- минимален и задължителен състав (групата на *задължителните*),
- възможни разширения, още неизпробвани, но сметнати за полезни в следващ анализ (група на *възможните*),
- разширения (споменати и отчетени), чието отхвърляне няма да е фатално на този етап, но може да се предвиди бъдещото им включване в границите на фиксирани подкласове (групата на *кандидатите*).

Тъй като границите между трите групи са малко размити, достатъчно е да формулираме най-общо техния състав за конкретната част на речта. Изборът на оптималната конфигурация на анотационната верига се определя главно от целта на анализа, както и от предоставените улеснения (концептуални и софтуерни) за промяна в речника.

Не трябва да се забравя, че минималната конфигурация отговаря на изискването на процедурите за снемане на омонимията, построени на статистически принцип (statistical taggers), където общият брой на различните анотационни вериги трябва да се сведе до минимум.

Много от разглежданите решения са претърпели промени в различните версии на граматичния речник – MORPHO-ASSISTANT, INTEX – първа версия, Граматичен сървър, актуалната версия на СФ-речника и производни от тях. Възможността за промени в зависимост от поставени цели остава и е лесно осъществима благодарение на разработения софтуер за редактиране на речници от този тип [Paskaleva 2005].

3.5.1 Съществителни имена

Лексикални признаци

Род

Българското съществително се характеризира с принадлежност към един от трите рода: мъжки, женски и среден, и това определя стойностите на първия анотационен маркер. Родът е важна категория, която участва във входните параметри на една основна синтактична връзка – тази на съгласуването.

Минималният брой маркери за тази категория се свежда очевидно до три.

Към някои размисли може да ни подтикне и наличието на съществителни, които са неутрални към рода – т.нар. съществителни от общ род – като *симпатяга*, *хаймана* и подобни, или пък съществителни, които като приложения се употребяват едновременно за женски и мъжки обекти – като *министър*, *професор* и др. Напълно възможно е и въвеждането на маркер за общ род – при свободно анотиране, или отбелязването с маркер за нерелевантност при позиционно анотиране.

Тъй като тези съществителни са малко (нормативните граматики ги изброяват към 30 [Русинов 1978]), въвеждането на още една стойност за род при налич-

на позиция в анотационната верига не представлява трудност. За по-нататъшния анализ, особено за формирането на номинални групи по съгласувателни критерии (важна операция на плоския синтаксис – вж. по-нататък, Глава 4), сигнализирането на тази нерелевантност е необходимо. Само за нуждите на морфологичния анализ приписването на стойност *м.р*. за тези съществителни не е грешно, понеже покрива част от употребите им (вж. *той е симпатяга*). При проверка на съгласуването обаче, за употребата *тя е страшна симпатяга* очевидно ще бъде сигнализирано нарушение, подобно на популярната оценка на грешките в българското съгласуване – *една гаща плява*. В някои от предишните версии на СФ-речника стойността *общ род* фигурира.

Одушевеност

Категорията *одушевеност* характеризира обекта по отношението му към *живото*. В стойностите на категорията немаркиран член е неодушевеното. За признака одушевеност, обаче, бихме могли да въведем два подкласа на живото – *лице* и *животно*.

В българските нормативни граматики опозицията живо/неживо има граматичен израз по линията *лице-нелице*, в граматиката се говори за съществителни лица и не-лица. В анотационните множества за други славянски езици – чешки и словашки, в МТЕ-стандарта се въвежда диференциацията – *одушевеност-неодушевеност*, проявяваща се обаче само за съществителните от м.р. В руската морфология, неприсъстваща в този стандарт, актуалната маркировка за одушевеност разграничава също само живото от неживото, тъй като то е значещо за формиране на падежните номинални окончания.

При тази колебливост – къде да бъде прокарана демаркационната линия на живото и неживото за съществителните, в приетото анотационно множество сме разграничили неодушевените (немаркирани) от одушевени съществителни (последните разделени на лица и животни). Това допълнително подразделяне може и да изглежда излишно, но се подкрепя от известна разлика в парадигматичните черупки за одушевените-нелица, т.е. животни, и истинските неодушевени предмети – две групи, които са от едната страна на границата лице/нелице. Например, по образуване на звателна форма, означенията на животните и предметите имат различно поведение. Съвсем граматични са формите – *свиньо*, *коньо*, *воле*, *кучко*, *кокошко*, докато за естествеността на формите *столе*, *масо*, *прозорецо* бихме могли да спорим. Очевидно тази разлика се обуславя от различната степен на позволена персонификация за двете групи имена (понякога съвсем правомерна при метафорична употреба: кого наричаме по-често *свиньо* – домашното животно или човека-събеседник?).

Броимост

Във външния си израз тази лексикално-граматична категория на лемата се проявява в потенцията ѝ да образува форми за число – единствено и множествено. Ако я изведем отделно, тази категория би могла да маркира само съществителните *сингулария* и *плуралия тантум*, т.е. употребявани само в ед. или в мн.ч. Немаркирани остават обикновените съществителни, притежаващи и двете форми за число. За отбелязване е, че двете категории (сингулария и плуралия тантум) са твърде различни по брой на лексикалните единици, които ги притежават. Тази разлика в обема им се обуславя от твърде различния диапазон на семантичните признаци, обуславящи този дефект в парадигмата – липса на форма за една от двете стойности на категорията число. Характеристиката плуралия тантум се проявява при една малобройна група съществителни – двучленни предмети *(очила, гащи)*, означения на материална маса (*трици*), както и няколко отделни съществителни (*разноски* и *финанси*). С признака сингулария тантум въпросът стои другояче. Тук спада една голяма група от абстрактни съществителни, също така означаващи материал, химични елементи и др. При това голяма част от тях може да образува множествено число при специфична употреба или метафоризация – *даденост*и в значение на телесни форми; материализиране на абстрактни свойства, напр. *все ярки индивидуалности се събрали*.

Поради разликата в стабилността на категориите сингулария vs. плуралия тантум, във версиите на речника, когато лексикалната категория броимост се въвежда в лексикалните характеристики, маркира се само плуралия тантум.

Подобни признаци, които се отразяват пряко на обема на парадигмата (в нея липсват формите за ед. или за мн.ч.), са обърнати в известна степен към миналото на обработката. Важно е такива признаци да бъдат сигнализирани при формирането на флективния тип, а оттам и при конструирането на парадигматичната черупка, за да не се получи свръхгенерация. СФ-речникът вече е отразил статуквото, зададено от парадигматичната черупка и произведено от синтеза. Ако поради пренебрегване на признаците за броимост сме свръхгенерирали някои форми, те няма да съществуват в текста, и следователно при речник, предназначен за анализ, няма да доведат до грешки или конфликти в последния.

Индивидуалност vs. всеобщност

Тук се има предвид делението на съществителните на нарицателни и собствени. Последните са твърде малко по брой в един речник на общата лексика, тъй

като са динамична и бързо попълвана категория и обикновено не намират (а и не могат да намерят) пълното си отразяване в словника на речника. Собствените имена се идентифицират преди анализа, в процеса на сегментацията (tokenization) като квазидуми в цялото си многообразие, вж. 1.7.3. В граматичен речник могат да фигурират само собствени имена, образувани според езиковата традиция, ако съставителите на речника са решили да представят извадка от именната система на езика.

Независимо от броя на включените в речника собствени имена, те са маркиран член в разглежданата опозиция и получават своята анотация в СФ-речника, като последен маркер в анотационната верига.

В MTE-стандарта това е първото равнище на подкласификация (категория *тип на съществителното*) и по законите на позиционното анотиране отделни маркери получават както собствените, така и нарицателните имена.

Граматични признаци

Стандартни граматични категории, осъществяващи формообразуването на съществителните, са: числото, определеността и падежния реликт – звателната форма, само заради която не си струва да въвеждаме за всички съществителни категория на формообразуване – падеж. Същото се отнася и за бройната форма, която би могла да се анотира отделно, като се въведе признак *форма*[29] с всичките неудобства на това решение.

Число

Това е задължителната номинална категория, свързана с броя на предметите. Тук можем да обсъдим само нейните възможни стойности, които освен традиционните единствено и множествено число включват и бройната форма на съществителното. Очевидно третирането на брйната форма като една от стойностите на числото е най-доброто решение, още повече, че и в анотацията на други славянски езици в MTE-стандарта се допуска напр. двойственото число, което е с аналогична граматична функция и синтактично поведение. Това прибиране на бройната форма в чекмеджето на категорията число е удобно и за по-нататъшна проверка на съгласуване в именната група, където информация-

[29] Няма нищо по-неприятно от решението на някаква лингвистична променлива да се припише безличния и безсъдържателен маркер *форма, тип* или *вид*, защото като лингвистична категория не ни говори нищо, докато не видим нейните стойности (докато категории като род, число и подобни са смислово натоварени). Но в някои случаи това се налага за означаване на очевидни подкласове или за класове, стърчащи извън класификацията – остатъци от стари категории, уникални форми и др. под.

та за броя на предметите, а оттам и за изразяването му се намира на една позиция и е достатъчна само проверката – единствено и не_единствено число.

Така реалните маркери за тази категория са три, със стойности – *единствено*, *множествено* и *бройна форма*.

Определеност

Категорията определеност в анотационната верига на морфемно равнище се изразява чрез определителния член. Заради специалното свойство на съществителните от м.р. да образуват пълна и кратка членна форма е предвиден и трети маркер за определеност. В анотационната верига съществителните от м.р. приемат три стойности на маркера за определеност, а всички останали, както и мн.ч. – два маркера, при което определеността се бележи различно от тази за м.р. Така се получават четири стойности за определеност: *не* – обща за всички, *да* – за съществителните от ср. и ж.р, а също и за мн.ч., *кратък* и *пълен* член (последните две – само за същ. м.р. ед.ч.).

Аналогично е съдържанието на граматичната част в анотационната верига на съществителното в други анотации за български. В позиционния MTE-стандарт за български език бройната форма е стойност на маркера *число*, звателната – на маркера *падеж*. В анотацията на BulTreeBank [Осенова и Симов 2007], която е обогатена версия на MTE-стандарта, към стойностите на звателната форма за маркера падеж са прибавени и стойностите на дателния и винителен падеж за някои архаични форми на съществителното, което формира отделен маркер за падеж на съществителното.

Това, което за нас е недостатък на позиционното анотиране, където структурата не предлага формално различие между лексикалните и граматичните признаци, е, че в този стандарт стойността *pluralia tantum* е поставена заедно със стойностите на единствено и множествено число в категорията *число*. В свободното анотиране възможността да се различи потенцията на лемата да образува определени форми от реалната проява на тази потенция в образуваните словоформи обаче съществува, както е показано по-горе (лексикална категория *броимост* vs граматична категория *число*). По същия начин е разпределена потенцията от нейната реализация при отразяването на степените на прилагателните в тяхната анотационна верига (вж. по-долу, 3.5.3).

3.5.2 Глагол

Това, което Балан нарича „слона на българската граматика“, заслужава да заеме централно място в описанието на речника, не само защото е център на фразата и нейния смисъл. Проблемите на глагола, особено в език с толкова силно развита флективност, отразена пряко и върху обема на глаголната парадигма (52 синтетични форми), са комплексно и взаимно свързани. Те се проявяват в разпределението между *лексикално* и *граматично* – в съответните характеристики, във взаимовръзките между категориите и в пряката връзка на техните стойности с обема на глаголната парадигма.

Ако поискаме да изброим категориите, които изразява българският глагол – лексикални и граматични, и се опрем на тези, които се споменават в нормативните граматики, бихме се затруднили сериозно[30].

Затова единственият начин за конструиране на анотационната верига на глагола е да определим минималното множество от маркери – първо за лемата, а после за нейните форми, с евентуално разглеждане и на групите на кандидат маркерите, и на възможните маркери.

Лексикални признаци

За да определим признаците, присъщи на цялата парадигма на глагола, т.е. на неговата лема, трябва от една страна да се насочим към традиционното лингвистично знание – какви признаци на глагола са характерни за цялата парадигма, и да решим каква част от тях ни трябва за следващи етапи на анализа, от друга. Две са категориите, които определят най-общо неговите семантични връзки, диатезата[31] му, синтактичните връзки и не на последно място обема

[30] В [Андрейчин и др. 1977], с. 190: **„За богатата морфологична система на българския глагол са** характерни следните граматични категории: 1) лице, 2) число, 3) залог, 4) време, 5) вид, 6) наклонение“ и по-долу: **„Видове глаголи по значение и по форма. Преходни и непреходни глаголи.** Обикновени и възвратни глаголи.“

В **[Бояджиев и др.1998], с. 342: „Глаголът притежава морфологичните категории число, род (при** формите, образувани с причастия), лице, време, вид на действието, таксис, наклонение, вид на изказването, залог, статус, а самостоятелно употребените причастия може и да се членуват; лексикално-граматичната категория вид на глагола; редица лексико-граматични разреди и формални класове. ...Основните лексико граматични разреди при глагола се представят чрез следните противопоставяния: лични и безлични глаголи, преходни и непреходни глаголи, глаголи, означаващи действие, и глаголи, означаващи състояние, пълнозначни и спомагателни глаголи... Основните противопоставяния на формални класове при глагола са обикновени, невъзвратни и възвратни глаголи, глаголи от първо, второ и от трето спрежение.“

[31] Диатезата на глагола задава разпределението между семантичните и синтактичните роли на глаголното действие и участниците в него, т.е. отношението между семантичните и синтактични актанти на глагола [Мельчук, Холодович 1970].

на парадигмата. Тези категории са: *вид* и *преходност*. Решаващо за общата функционалност на глаголната парадигма е и делението по оста *пълнозначни* и *спомагателни* глаголи. Особен аспект на глаголната диатеза ни дава характеристиката *лични* и *безлични* глаголи.

Форма на глагола

В свободния анотационен стандарт на СФ-речника традиционното деление на глаголите по двете оси: *лични/безлични,* както и *пълнозначни/спомагателни* е разпределено по икономичен начин, като личните глаголи са немаркирани, т.е. неанотирани, а два маркера за специална форма маркират спомагателните и безличните глаголи, макар че характеризират поведението на глаголната лема от съвсем различен ъгъл. Естествено, двете множества не се пресичат и тази икономия на записа се разчита лесно.

Тип на глагола

Тази специфична категория за български и други славянски езици (в българския език особено богата във връзките си с други лексикални и граматични категории) присъства във всяка речникова статия на български глагол – от оскъдните на информация правописни речници, до разточителните статии на тълковните речници.

Ако се замислим за ролята ѝ в модела на българския морфологичен синтез, ще видим, че пряко участие във формообразуването – в смисъл на избор на формообразуващите елементи, категорията *вид* няма. Тя има обаче решаващо участие в определяне на парадигматичната черупка на глагола, понеже нейните стойности определят броя на словоформите, образувани от глаголната лема.

Ако разглеждаме граматичния речник като резултат на вече извършено генериране на словоформи, следователно на вече установена парадигматична черупка, категорията *вид* не може да ни покаже своето пряко участие в по-нататъшен анализ, освен на доста по-късни етапи, свързани с характеристиката на глаголно действие. Към етапи, непосредствено следващи морфологичния и предшестващи синтактичния анализ, категорията също няма отношение – не притежава различителна сила в контекста на омонимични единици. Глаголният вид е повече семантична, отколкото синтактично напълнена категория. Въпреки това, дан на дългогодишната традиция за отчитането на тази категория във всички анализи на глагола е решението да я оставим в групата на минималните задължителни признаци.

Немаркираната позиция във видовото противопоставяне по отношение на образуването на словоформи е несвършеният вид на глагола, който образува всички словоформи при анализа. Маркиран член е свършеният вид, при който имаме забрана за образуване на самостоятелно сегашно време. Тази забрана не се отразява обаче на броя на формите за сегашно време в парадигмата, тъй като те участват в свързани синтактични конструкции, напр. *искам да кажеш*.

За нуждите на тези по-късни етапи категорията *вид* фигурира като лингвистична променлива в лексикалната част на глаголните маркери със своите три стойности – *свършен несвършен*, и *двувидов*.

В MTE-стандарта информация за вида на глагола в анотационните вериги на славянските езици липсва. Това очевидно се дължи на отбелязаното вече неясно участие на тази категория в бъдещите обработки. В анотационната верига на BulTreeBank категорията има две стойности – *свършен* и *несвършен* вид, като дуалното поведение на двувидовите глаголи се оставя за изчисление от контекста.

Преходност на глагола

Преходността, също като видовите характеристики на глагола, има пряко участие в определяне на обема на глаголната парадигма. Във вече генериран речник на словоформите тя присъства като резултат, но участието ѝ в бъдещи етапи на обработката е по-ясно изразено и по-близко като перспектива, отколкото използването на аспектуалността. Преходността (транзитивността) на глагола е базата, на която се гради синтаксисът на глаголните конструкции чрез основния инструмент – глаголната рамка (verbal frame, verb rection, модель управления). Тя ни задава начина на морфологичното оформяне на участниците в действието на конкретния глагол (неговите *актанти*, *аргументи*, *партиципанти*). Глаголната рамка на английския глагол *depend* задава оформянето на неговия обект (втори аргумент) чрез предлога *on*, на руския глагол *зависеть* – чрез предлога *от* и родителен падеж, на френския глагол *dependre* – чрез предлога *de* и на българския глагол *завися* – чрез предлога *от*.

Този признак с неговите две стойности – *преходен* и *непреходен глагол*, присъства във всички известни ни анотационни решения за български език, с изключение на MTE-стандарта.

Залогово поведение на глагола

В езиковедската литература залогът (каквото и да се разбира под това) е предмет на оживени дискусии: каква категория е залогът – лексикална, морфо-

логична **или синтактична, с какви граматически формативи или лексикални** средства се изразява, колко и какви залози има в съответния език, и т.н.

С какви признаци и доколко да се отрази залоговото поведение в анотационната верига на глагола в един граматичен речник е решение, което е свързано с общите съображения, изложени за категорията *вид* – релевантност на признака за бъдещи етапи на обработката. Очевидно залоговите характеристики имат значително участие в проявите на глаголната диатеза, тъй като пряко отразяват субектно-обектните характеристики на действието.

Залоговите значения се проявяват често извън рамките на една дума – пасивни конструкции, възвратни конструкции, смяна на актантите и т.н. Това бъдещо поведение, както и структурата на тези сложни единици трябва да се изчисли от информацията, приписана към главната им дума, в случая глагола. **Каква** част обаче от тези залогови характеристики трябва да бъде прикрепена към характеристиките на отделна лема? При такава постановка на въпроса **лема**та на глагола *усмихвам* трябва да съдържа информация, която да ни позволи да генерираме нейните форми, винаги придружавани от частицата *се;* лемата *боли* трябва да съдържа информация за задължителното присъединяване на клитика – местоимение във вин.пад.

На този въпрос различни анотационни концепции отговарят различно. **В на**чалните версии на системата MORPHO-ASSISTANT залоговите характеристики на глаголите се изчерпваха с посочване на безличност (личните глаголи не са маркирани), а също и с посочване на задължително присъединяваните клитики и комбинации от тях: *се*, *си*, *ме*, *ми*, *ми се*. Впоследствие тези признаци бяха разширени и обогатени в една мотивирана класификация[32], която свързва в едно цяло семантични свойства, морфологични признаци **и обем на глаголна**та парадигма. Според тази класификация основните характеристики на глаголните леми според връзката им с лицето на глаголното действие в СФ-речника се изразяват чрез класификацията на глаголите като:

- лични (немаркирани),
- безлични,
- квазилични.

По отношение на присъединяването на клитики глаголите се анотират като:

- задължително присъединяващи възвратни и местоименни клитики: *се*, *си*, *ми*, *ме*, *ми се* (*досетя се*, *отспивам си*, *докривява ми*, *счува ми се*),

32 Подробно описание на последния вариант на тази класификация, предложена и реализирана от Милена Славчева върху глаголните единици в речника, се съдържа в [Slavcheva 2003].

- глаголи, които при присъединяване на възвратните клитики *си* или *се* придобиват ново лексикално значение (*заиграя се*, *отида си*).

Характеристиката за присъединяване на клитиките се комбинира с характеристиката за *личност-безличност* на глагола в непълна комбинаторика – най-пълно всички възможности за кохезия с възвратни клитики са реализирани при личните глаголи.

В MTE-стандарта маркери за залогово поведение изобщо липсват, но в по-пълно разработения позиционен стандарт на BulTreeBank имаме опозицията *личен/безличен* глагол в категорията *форма*.

Граматични признаци

В граматичните признаци се разгръща богатството на глаголната парадигма. Тя е съставена от толкова много членове с различна морфологична природа и съответно различни характеристики, че води до значително увеличаване на дължината на маркерната верига, особено в позиционно анотиране. В граматичните признаци на глаголната словоформа виждаме поведение на вербални категории, също и адективни – за причастията, виждаме и номинални – за отглаголното съществително. Това налага да се обособи едно първо равнище на маркирането, където се диференцират личните глаголни форми (немаркирани) от причастните, деепричастните и номиналните форми (маркирани по съответен начин). Освен тази диференцираща анотация отделните класове глаголни форми приемат допълнителна анотация за основните си признаци. Така се оформят следните групи граматични характеристики.

За личните глаголни форми:

- време,
- лице,
- число.

За причастните форми:

- вид на причастието. Без да се въвежда анотация за всички темпорални (мин. vs. сег.) и залогови (деят. vs. страд.) характеристики на причастието поотделно, тук се въвеждат общо четири признака за четирите причастия в български: трите деятелни – сегашно, минало свършено и минало несвършено и едно страдателно минало причастие.
- число,
- род,
- определеност.

Деепричастието като неизменяема глаголна форма няма допълнителна анотация.

Отглаголното съществително (вж. мотивацията за мястото му в глаголната парадигма в 2.5):

- число.

Повелителното наклонение:

- число.

В позиционния MTE-стандарт позициите в глаголната анотационна верига са 15. Техният брой, както и броят на техните стойности е увеличен главно поради необходимостта да се унифицират описанията на типологично различни езици – от английски, през унгарски и естонски, до румънски и няколко славянски езика. Увеличаването на стойностите идва и от поставянето в „ един кюп“ на езикови явления от различни равнища – напр. прекалено общото понятие *тип на глагола* включва делението на личен и спомагателен, но и на глагол-връзка, в тази позиция са и модалните глаголи, а в също така универсалното понятие *форма на глагола* са включени наклоненията от една страна, заедно с личните и нелични глаголни форми от друга.

3.5.3 Прилагателни имена

Второто име в българската граматична система има по-проста анотационна верига от първото може би поради липсата на семантична стойност на самостоятелен обект. Неговите граматични признаци са съгласувателни, повтарят граматичните признаци на определяното от него съществително и следователно трябва да включат всички стойности, обусловени от характеристиките на последното.

Лексикални признаци

Способност за степенуване. Този маркер е съществувал в по-стари версии на речника, в последната е премахнат по следните причини.

Способността за степенуване (gradability) по-скоро характеризира генеративната способност на даденото прилагателно – да образува или не двете степени на сравнение освен позитивната, основна, немаркирана степен. По същество този маркер анотира семантиката на прилагателните в известното им деление на качествени и относителни.

Тъй като резултатът от генеративните способности на прилагателното е налице в речника – там, където е било възможно да се образуват степени на сравнение, те са образувани и фигурират като речникови единици, по-добро решение, осъществено в последната версия на речника, е следното. Информацията, свързана с генеративната способност на прилагателното по отношение на степените за сравнение, да бъде прехвърлена в граматичната част на маркерната верига, където образуваните степени (сравнителна и превъзходна) да бъдат анотирани като част от парадигмата на прилагателното – с два маркера, като оставят положителната степен немаркирана по този признак.

Интересни решения, илюстриращи по-скоро непоследователност в общите правила за стандартната анотация, отколкото богатство на анотирането, намираме в MTE-стандарта за трите славянски езика. Чешката анотационна схема въвежда тип на прилагателното, като разграничава качествени от притежателни прилагателни (като бълг. *хубав и майчин*). Сръбската анотационна схема освен притежателно въвежда и редно прилагателно, което по липса на пример в стандарта, не може да бъде изчислено като лексикално покритие. Авторите на словенската анотационна схема отиват по-далеч в желанията си за богата анотационна верига и освен споменатите три семантични типа – квалификативно, посесивно и редно прилагателно, обогатяват и системата на сравнителните степени. Те включват в нея и умалителността, като в придружаващ коментар (The value 'diminutive' for Degree is relevant for *derivated* adjectives – курсив мой – б.а.) признават, че разликата между словообразуване и формообразуване не е съществена за тях при описание на парадигмата на лексикалния клас.

В българската версия на MTE-стандарта липсва анотация за семантичен тип на рилагателното, но липсва и всякаква анотация за степени на сравнение.[33]

Граматични признаци

В граматичните признаци на прилагателното съзираме два вида категории.

Първите са ***съгласувателните***, или произлизащи пряко от граматичните признаци на съществителното, което прилагателното определя. Така те повтарят неговите граматични категории – като вид, количество и съдържание.

Вторите са ***собствени граматични признаци*** на формообразуването на прилагателното. Имат се предвид неговите сравнителни степени, които отразяват

[33] Това положение е коригирано в последната версия на българския MTE-стандарт, произведен чрез преформатиране на анотационните вериги на СФ-речника. В него липсва анотация за семантичен тип на прилагателното, тъй като не е много ясно на какво бъдещо анализиращо равнище ще бъде използвана тя, но степените за сравнение са анотирани като граматични признаци.

чрез специални граматични формативи степента, движението на признака, носен от прилагателното, по скалата на интензивността. Чрез това разграничаване се увеличават членовете на неговата парадигма. От това следва, че при прилагателното имаме:

- Род – мъжки, женски и среден.
- Число – единствено и множествено. За отбелязване е, че мъките при определяне на видовете число при съществителните, където се определя количеството на реални обекти (пък било то и абстрактни), тук нямат място.
- Определеност – нечленувано, членувано, пълен член, кратък член.
- Степени на сравнение – положителна (немаркирана), сравнителна и превъзходна, което умножава основната парадигматична черупка три пъти чрез формативите *по-* и *най-* в препозиция.

В парадигмата на прилагателното не може да има празни места, или различни конфигурации на парадигматичната черупка, отново поради вторичната, производна природа на неговите граматични признаци.

По този начин е третирано и прилагателното име в MTE-стандарта за другите славянски езици. Някои от добавените маркери предизвикват недоумение, като цитирания по-горе признак за словообразуващ елемент диминутив в степените за сравнение. Друг добавен маркер анотира чешките прилагателни имена като прости и сложни. Става дума за кратките форми на прилагателните с предикативна функция и пълните форми на прилагателните с чисто адективно поведение.

В MTE-стандарта липсва руската морфология, но в разработваната в момента версия на руски граматичен речник в този формат се предвиждат стойности за кратка и пълна форма (възможно означение – предикативна и адективна функция), тъй като имат морфологично изразяване и генерират нови парадигматични членове.

Въвеждането на стойност *елатив*, при степените на сравнение (абсолютен суперлатив, като бълг. *най-високчайши*), не е подкрепено достатъчно с примери, за да се оцени морфологичната природа и парадигматична проява на това явление. Нещо повече, стойността е въведена за сръбски, без примери, но затова пък липсва в хърватския език, за словенски стойността присъства, но в по-късни версии на MTE-стандарта [ИНТЕРА 2003] липсва.

3.5.4 Местоимения

Описанието на този лексикален клас в каквато и да било подредба на лингвистичното знание е задача колкото предизвикателна, толкова и главоболна. Причината може би се крие в тяхната абсолютно еклектична семантична и синтактична природа, което е естествен резултат от факта, че тази част на речта се състои само от *заместители* на нещата в реалността, но не и от самите неща. Тази *празнота* на местоименното значение се проявява и в неустойчивостта на тяхната класификация по отношение на други части на речта. Не случайно в различни описания можем да срещнем различни формулировки на хибридни класове като например местоименното наречие от почти всички местоименни видове, местоимения-числителни или числителни-местоимения и други подобни. Трудности в анотацията на местоименията, както и в определянето на тяхната парадигматична черупка създава не само синтактико-семантичната им природа, но и изобилието на архаични форми в тях, на отмрели категории, представени непълно в парадигмата, на омонимични форми, на аглутинация на морфеми (извън приетата аглутинация на постпозитивния член – вж. относителните местоимения на *-то*) и много други. Такова многообразие в изразяването – и съответно в класифицирането на лексикалните и граматичните черти на местоименията, при позиционно анотиране се проявява в дълги анотационни вериги, с много маркери за нерелевантност и трудно възприеман запис.

Свободната анотация, макар и да съкращава този запис, също се сблъсква с проблема за избор на граматичните атрибути и техните стойности, а най-вече – с грануларността на анотацията. Това особено важи за лексикалните атрибути на местоименията. Ако местоимението може да замества *всичко*, може ли да го отразим в лексикалната част на анотационната верига? При положителен отговор и довеждането на този принцип докрай бихме могли да генерираме буквения състав на местоимението само от анотационната му верига. Например – можем ли да отгатнем буквената стойност на местоимението, което има лексикални стойности: притежателно, род на притежателя – женски, число на притежателя – единствено, лице на притежателя – трето, а граматичните му стойности са мъжки род, единствено число, членувано с кратък член? Очевидно няма нужда нито от словоформа, нито от лема, за да посочим местоимението *нейния.* При такова изчерпателно задаване на стойностите на маркерите само отделни дублетни местоименни форми не могат да бъдат отгатнати еднозначно (като *туй* и *това*, *тая* и *тази*).

Многоликата природа на думите-заместители, местоименията, се проявява в броя на анотационните маркери, приписани към тази част на речта. В по-

зиционна анотационна система като MTE-стандарта анотационната верига на местоименията е най-дълга, обусловена не само от разнообразието на анотационните местоименни вериги в един език, но и от междуезиковото разнообразие в представянето на това еклектично в някакъв смисъл явление.

Разгледани в плоскостта на българския език, анотационните местоименни вериги отново впечатляват както с дължината си, така и с общия си брой. Табл. 14 показва съотношението в един граматичен речник, анотиран в два стандарта (свободния на СФ-речника и позиционния на MTE), между различни стойности на речниковите единици и техните анотационни вериги.

	СФ-речник	**MTE речник**
Анотационни вериги – общо	1866	363
Местоименни анотационни вериги	71	193
% местоименни вериги спрямо общия брой вериги	3.8 %	53%
Местоименни словоформи	308	308
% брой местоименни вериги спрямо броя на местоименните словоформи	23%	62%

Табл. 9. Местоименни словоформи, местоименни анотационни вериги и съотношения.

Числата в таблицата навеждат на някои размишления. Очевидна е неикономичността (разточителността) на позиционното анотиране. Един и същ брой словоформи – 363, се означават с 5 пъти повече различни анотационни вериги в позиционния стандарт, сравнен със свободния.

В рамките на един и същ стандарт съотношението между броя словоформи, които се анотират с една верига, също е различно. Това съотношение, което можем да определим като анотационна плътност, т.е. колко словоформи се обслужват от една верига, също може да ни говори за неикономичност на стандарта. При висока плътност на анотацията стандартът е по-икономичен – по-голям брой словоформи се обслужват от една анотационна верига. В свободния стандарт една местоименна верига анотира 4,3 словоформи, а в позиционния – 1,6 словоформи, следователно последният може да се определи като по-разточителен. Очевидно анотационната плътност не зависи само от качествата на стандарта, но и от морфосинтактичното поведение на дадената лексикална единица – по-горе стана дума за многообразието в граматичното поведение на местоименията.

Това богатство на анотациите на местоименията затруднява както тяхното определяне, така и тяхното описание.

Лексикални признаци

Описанието на лексикалните признаци на местоименията може да бъде много икономично, но може да бъде и прекалено подробно, ако се включат всички свойства на цялата местоименна верига.

Минималната и задължителна лексикална информация към лемата на едно местоимение е неговият тип (или вид) в термините на школската граматика – лични, притежателни, въпросителни и т.н. местоимения.

Такива са стойностите на атрибута *тип* в MTE-стандарта, също и в СФ-речника – на брой девет.

Видовете местоимения са следните (подредени по азбучен ред на маркерите в латинската им анотация): обобщителни, показателни, неопределителни, въпросителни, отрицателни, лични, притежателни, относителни и възвратни.

Това е минималната и задължителна информация, отразена в СФ-речника. В историческото развитие на принципите на анотирането, местоименната верига в лексикалната си част е съдържала повече информация. Тя бе съдържателно пренесена в MTE-версията на речника. Става дума за признаците, описани по-долу.

Тип на референта на местоимението

Тази информация е изключително важна, тъй като определя синтактичното и семантично поведение на лемата. Тъй като местоимението може да замества всякакви лингвистични обекти, назоваващи предмети, свойства, признаци, начини, количество и всичко останало, съдържанието на този маркер е преди всичко семантично. Тук би трябвало да възпрем желанието си за пълно описание на семантиката на заместеното при местоименията, тъй като анотационните им вериги и без това са доста на брой. Прагматичните съображения (за какво ще ни послужи това?) са основната спирачка пред това разширение и ако се вземат пред вид най-близките задачи на анализа, можем да се въздържим от тази характеристика и класификационните й дълбочини, трудни за изплуване[34]. Като възможни и реализирани в други версии на граматичния речник – вече споменатото преформатиране на СФ-речника в MTE-анотация, бихме

[34] При речници на компютърни приложения с ясна синтактико-семантична насоченост, какъвто е например морфологичния речник на системата **BulTreeBank, вж. [Осенова и Симов 2007] подобно** раздробяване е задължително и авторите са определили 13 стойности на местоименния референт, като в него са разтворили и възвратността, информацията за броя на притежателите и други подобни.

искали да отбележим значенията на синтактико-семантичната характеристика на местоименния референт, приети в него. Тези признаци са съществували в по-ранни версии на речника – напр. в демо-версията на морфологичен анализ в [BALRIC-LING 2000]. Като качества на референта са обособени следните стойности:

- предмет (OBJ),
- атрибут (ATT),
- брой (NUM),
- количество и степен (QNT),
- близост (NER),
- отдалеченост (DIS),
- количество (QLT),
- размер (SZ).

Те обхващат напълно предложените в MTE-стандарта свойства на референта – лица, притежание, атрибут, количество[35].

По такъв начин в анотационната верига на Морфоасистент и речника на BALRIC-LING се получават записи като:

- PRO+DEM+ATT+QLT – показателно местоимение с атрибутивна функция за качество (*такъв*)
- PRO+DEM+ATT+SZ – показателно местоимение с атрибутивна функция за размер (*толкав*, *толчав*)
- PRO+INT+ATT+QLT – въпросително местоимение с атрибутивна функция за качество (*какъв*)
- PRO+REL+ATT+QLT – относително местоимение с атрибутивна функция за качество (*какъвто*)
- PRO+IDF+ATT+QLT – неопределително местоимение с атрибутивна функция за качество (*някакъв*)
- PRO+NEG+ATT+QLT – отрицателно местоимение с атрибутивна функция за качество (*никакъв*)

Очевидно подобен дълъг запис в лексикалната част на местоимението може да бъде оставен само с оглед на конкретни бъдещи приложения.

В последната версия на СФ-речника в свободно анотиране същите са съкратени по прагматични причини, но са реабилитирани в неговата MTE-версия.

35 Впрочем, в първия български MTE-стандарт тези категории отсъстват.

Свойства на притежателя

Тази категория се използва само за притежателните местоимения и в първоначалната българска версия на MTE-стандарта изобщо не фигурира. В кода на BulTreeBank свойствата на притежателя са разделени семантично между типовете на референта (числото на притежателя/ите) и отделно изведена категория, отразяваща само рода му. В BALRIC-LING речника имаме 8 маркера за този признак, които характеризират притежателя едновременно по род, лице и число.

Синтактичен тип на местоимението

В MTE-стандарта това свойство е изведено отделно, като характеристика на чисто синтактичната функция, която играе референтът на местоимението, а именно – номинална, адективна и адвербиална.

Според нас тази функция може да бъде пряко изчислена от стойностите на типа на местоименния референт, които, разположени в семантичната плоскост, са много по-богати от трите посочени стойности. Затова позволяват лесно изчисляване на синтактичния тип, който по подразбиране е еднакъв за местоимението и неговия референт.

Съкровищата от склада на лексикалните категории на местоимението могат да бъдат извадени за употреба в зависимост от целите на следващи приложения.

Граматични признаци

Голямото разнообразие от семантични и синтактични типове на местоименните леми обуславя и обема на граматичните категории на техните словоформи, където се съдържат именни и адективни признаци, допълнени с някои архаични категории (падежите), които заедно образуват парадигматичната черупка на всеки вид местоимение.

Естественият подклас за местоименията е техният тип. Но и в границите на един тип парадигматичната черупка няма еднакъв рисунък, тъй като там се намират и именни (силно окастрени) типове склонение, и адективни (виж напр. склонението на *кой* и *какъв)*. Съществува и силна омонимия при отделни местоимения, съчетаващи именна и адективна функция (*кой ти каза* ***vs. кой цвят*** *избираш* – в термините на една изчерпателна и грижлива анотация това са две лексеми, едната с референт лице, а другата с референт признак).

Многообразната граматична природа на местоименията личи от списъка на категориите, приписани към словоформите им.

- Лице – стойности 1, 2 и 3 – за личните местоимения.
- Число – единствено и множествено – обща категория за местоименията обекти и признаци (*те*, но и *какви*).
- Род – обща за обекти и признаци (*тя*, но и *никаква*).
- Падеж – за обекти (*нему*, *комуто*).
- Клитика – за обекти и признаци (*дай му* vs *книгата му*).
- Определеност – за признаци (*техният* и *своите*).

Това са категориите, които фигурират в конвертирания (от СФ-речника) MTE-запис. Единствената разлика е във формулировката на *клитика* – да/не, използвана в MTE, и същата категория, формулирана като кратка/пълна форма. Това по принцип не променя нещата, тъй като двете стойности на този признак са приписани към едни и същи местоимения (личните и притежателни).

3.5.5 Числителни имена

Числителните имена са синтактично нееднородна категория, включваща лексикални единици с различни граматични значения, обединени главно от семантиката си – да изразяват количество (в брой и поредност) на предмети. Това прави морфологичното им поведение нееднородно – номинално и адективно формообразуване, изразяващо различни аспекти на количеството. За разпределение между лексикалното и граматичното може да се спори при свободното структурирано анотиране на СФ-речника. В позиционното анотиране на MTE-стандарта лексикалното не се разграничава от граматичното, но третирането на някои езикови единици като числителни – за отделни езици, а не изцяло, поставя въпроса за бъдещо обогатяване на анотационната верига, заедно с някои типологически проблеми като:

- границата между числителните и наречия и местоимения с количествено значение, като *малко, много, толкова* и др.
- границата между числителните и съществителните, които не само имат количествено значение, но и посочват точен брой – в СФ-речника например *сто* е числително, поради типичната парадигма на числително с образуване на приблизителна форма – *стотина*, а *хиляда*, *милион* и *милиард* са съществителни с типичното им формообразуване.

Лексикални признаци

Лексикалните признаци на числителните в минимален състав се свеждат до техния семантико-синтактичен *тип* – изразяване на брой със специфично

формообразуване за класа *бройни* и изразяване на поредност с адективно формообразуване – за класа *редни*. Останалите семантични подкласове биха могли да се прехвърлят към граматичните категории, тъй като образуват форми по един и същ начин за всички числителни, които имат тази способност. Такива класове (или форми) могат да бъдат свойствата приблизителност, събирателност, мъжколичен обект. В СФ-речника те са граматически признаци. В MTE-стандарта към тези признаци са прибавени, например (за чешки език) значенията на част от цяло (*третина, четвъртина, петина* и т.н.), както и изброяването на многократността на даден акт (като българското – *дваж, триж* и т.н.). В нашата концепция първите лексикални единици са съществителни, а вторите – наречия.

В BulTreeBank към тип на числителните са прибавени и наречията за количество – в отделен клас (много, малко), и наречията, означаващи приблизителен брой за предмети (мнозина, малцина), а числителните за приблизителна бройка или мъжколични не са маркирани изобщо.

Граматични признаци

Формообразуването на числителните се разполага главно в две зони – номиналната и адективната, за по-голямата част от тях. Извън тази норма остават:

- числителни като *един,* което може да бъде отнесено и към прилагателни, дори към местоименията в някои граматични концепции (в BulTreeBank), но е оставено в този клас поради съществуването му участие в структурата на сложни числителни,
- *две* и *три*, които имат форми и за род.

Останалите бройни числителни получават стойности нечленувана и членувана форма (*седем, седемте*) чрез постпозитивния член. Друга редовна форма на формообразуване е тази за приблизителен брой, която също може да се членува (*десетина, десетината*), както и формата за мъжколични обекти (*десетима*).

Числителните бройни, които образуват форми, са числителните от 1 до 20, числителните, означаващи десетки (от 20 до 90) и споменатото вече 100. Те произвеждат и редни числителни. Така се получават следните граматични стойности в анотационната верига:

- Род – за бройните *един* и *два* и всички редни.
- Число – за бройното *един* и всички редни.
- Определеност – за всички бройни и редни.
- Форма за приблизителност – за бройни числителни.
- Форма за мъжколичност – за бройни числителни.

3.5.6 Наречия

Наречието е неизменяема част на речта, където не би трябвало да се разглеждат граматичните класове, а само лексикалните. Съществуват обаче някои редовни образувания, на границата на слово- и формообразуване (като степенуването), заради които не би трябвало да се наруши така драстично природата на този традиционно неизменяем лексикален клас.

Лексикални признаци

Богатата семантична палитра на наречията, както и синтактичното им поведение, предполага наличието на твърде много лексикални класове, характеризиращи тяхната лема – фактически и единствена словоформа.

В различни анотационни системи проследяваме няколко оси на класификация на наречията.

Семантично-функционални свойства

Тук спадат всички характеристики на действието или признака, които са основна функция на наречието. Става дума за означението на време, място, начин, логическа връзка, причина и цел, количество.

Кванторни и деиктични свойства

Подобно на местоименията, всяка една от изброените горе семантични функции може да бъде обогатена от показване, отрицание, изразяване на неопределеност или всеобщност, относителност или въпросителност.

Уточненията по двете оси се комбинират – например време и отрицание ни дава *никога*, време и обобщаване ни дава *всякога*, място и неопределеност ни дава *някъде* и т.н.

Значенията по първата ос – семантичната, могат да се изразяват с различен интензитет и това води до образуване на степени на сравнение – за тази по дефиниция неизменяема част на речта.

Степени на сравнение могат да имат наречията от всички изброени семантични групи по първата ос. По втората – кванторно-деиктичната, степенуване не се допуска.

Комбинацията от два вида уточняващи подкласове плюс формите за степен водят до голям брой комбинирани маркери за наречията. В СФ-речника тези комбинации са 55.

Класификация по синтактично поведение на самото наречие липсва и е очевидно, че става дума за може би първото от възможните маркирания, което трябва да се осигури в речника в един следващ етап. Отделно от семантичната класификация наречията могат да се класифицират и по синтактична функция – нещо твърде важно за непосредствено следващи етапи на синтеза, тъй като става дума каква опорна дума поясняват в словосъчетания – глагол, прилагателно или съществително. Синтактичната функция на наречието улеснява предварителното построяване на отделни синтактични връзки и снемането на омонимията на самото наречие (вж. *хубаво дете* и *хубаво пее*).

В MTE-стандарта са отразени семантико-синтактични подгрупи за наречията в плосък позиционен запис, обединяващ класификационни резултати, разположени в различни лингвистични плоскости – модификатор vs.спецификатор, въпросително-относително vs. отрицателно, вербално vs. адективно. За българското наречие в първата MTE-версия е предвидена класификация за общо наречие и наречие, което е омонимично с краткото прилагателно в ср.р. ед.ч. (идентифицирано и като adjectival). Очевидно тази група в посочения стандарт се нуждае от унификация на предложените критерии за класификация.

3.5.7 Предлози

В тази неизменяема част на речта, възможна класификация на (само!) лексикалните свойства по понятни причини не може да включва семантиката на самия предлог, но авторите на MTE-стандарта са включили в характеристиките й управляваните от предлога падежи. Не само защото в българския език поседните липсват, но и поради това, че в едно неутрално, стегнато анотационно множество синтактични предвиждания, свързани директно с управлението изглеждат като стърчащи извън системата. Ако си помислим пък за руския език, непредставен в MTE-стандарта и богато снабден с падежи, подобна информация би била еднозначна за по-голяма част от предлозите, които често управляват различни падежи, в зависимост от семантиката на управляващата част (*положить в урну*, но *лежать в холодильнике*).

В СФ-речника предлозите нямат подкласификация на групи. Същото важи и за предлога в анотацията на BulTreeBank.

3.5.8 Съюзи

Присъщи на тази лексикална единица признаци са само съчинителен или подчинителен съюз. Това анотиране е възприето в СФ-речника, а също и в BulTreeBank, където имаме и трети тип съюз – съставката на двойния съюз.

В MTE-стандарта има въведени твърде много характеристики, свързани с управлението на съюза (какво споява и какво управлява – дума, фраза или изречение). Някои чешки съюзи, слети със спомагателния съюз, приемат и характеристика за лице, но както спецификата на българския език, така и концепцията за равнищата на информация в граматичния речник не ни поставя непосредствената задача за подоно разширяване на информацията на тази единица в речника.

3.5.9 Частици

Частицата, за разлика от съюза, има своята пълнозначна семантика и по тази причина в MTE-стандарта виждаме 11 стойности на признака *тип* (негативна, афирмативна, модална и т.н.).

В анотацията на BulTreeBank деветте български частици са класифицирани в девет семантични класа, с оглед на участието им в синтактични конструкции – разточителство, обусловено от целта на приложението. В СФ-речника частиците не са класифицирани. За нуждите на бъдещи етапи в анализа при такова равенство между броя на единиците и техните класове ни се струва, че правилото може да се зададе и чрез самия лексикален елемент, а не чрез неговия тип.

3.5.10 Междуметия

Междуметието не може да има много подкласове, макар че подобен опит за класификацията му като „изразяващо чувства“ и „други“ в MTE-стандарта не е прокаран последователно за всички езици, с изключение на румънския. Без подкласове е междуметието в СФ-речника, а също така и в анотацията на BulTreeBank.

* * *

Представените дотук общи принципи, придружени от конкретните параметри на едно анотационно множество в български граматичен речник, са базата, върху която може да разширим нашия поглед върху анотацията – и от мето-

дологична гледна точка, свързана с оценката на анотационните принципи, и от практическа гледна точка, с оглед на следващи етапи в анализа в различни посоки на обработката.

3.6 Конвертиране на анотационни множества

Както бе вече отбелязано, свободният анотационен стандарт е по-лесен за разчитане от човека-потребител, но по-труден за автоматична обработка.

Освен функционалната страна на тази оценка, съществува и оценка, свързана с целите на обрботката. Позиционният стандарт се предпочита като общ за многоезикови приложения. Когато той се прилага за група езици, както в случая с MTE, където към английски се прибавят два угро-фински езика, няколко славянски, а в следващо разширение на стандарта – и новогръцкия език, унифицираният запис се усложнява. В него количеството на нерелевантните маркери, отбелязани с тире, се увеличава, тъй като към нерелевантните за дадена категория признаци се прибавят и нерелевантните за даден език. Уговорка между потребителите на MTE-стандарта позволява да се пропуска всяка крайна верига, съставена само от маркери за нерелевантност, за да се съкрати записа.

Практически всеки свободен анотационен стандарт може да бъде преформатиран в позиционен. Особено когато последният е разработен за няколко езика и продължава да се разширява. Така се отваря пътят на съответния език за междуезикови разработки.

Така в стандарта на MTE бе преформатирано и българското свободно анотационно множество – пример за входа и изхода на тази трансформация виж на Табл. 15 по-долу.

лъчезарен,лъчезарен.A+GR:sm	лъчезарен,лъчезарен.A:-pms_n
лъчезарна,лъчезарен.A+GR:sf	лъчезарна,лъчезарен.A:-pfs_n
лъчезарната,лъчезарен.A+GR:sfd	лъчезарната,лъчезарен.A:-pfs_y
лъчезарни,лъчезарен.A+GR:p	лъчезарни,лъчезарен.A:-p-p_n
лъчезарните,лъчезарен.A+GR:pd	лъчезарните,лъчезарен.A:-p-p_y
лъчезарния,лъчезарен.A+GR:smh	лъчезарния,лъчезарен.A:-pms_s
лъчезарният,лъчезарен.A+GR:sml	лъчезарният,лъчезарен.A:-pms_f
лъчезарно,лъчезарен.A+GR:sn	лъчезарно,лъчезарен.A:-pns_n
лъчезарното,лъчезарен.A+GR:snd	лъчезарното,лъчезарен.A:-pns_nd

Табл. 15. Преформатиране на единиците на СФ-речника в MTE-стандарт.

Както всяка дейност, свързана с унификация и настройка, и тази операция бе крайно полезна, тъй като доведе до някои промени във входния и изходния стандарт. Новото множество от MTE-анотации за български добави нови характеристики в анотационната верига – например Лице на притежателя за притежателните местоимения, информация за Pluralia tantum в анотационната схема на съществителните и някои други поправки. Много от последните бяха свързани не толкова с обогатяване и задълбочаване на стандарта, а с поправяне на някои пропуски в първата българска анотационна схема.

Друг експеримент, свързан с конверсия на анотационни стандарти, е преформатирането на руски граматичен речник в DELAF формат[36] в MTE-стандарт [Паскалева 2002].

Тъй като руският речник, базиран върху известния Граматичен речник [Зализняк 1977], съдържа лексикална и граматична информация, доста по-бедна от съдържащите се както в MTE-стандарта анотационни маркери за славянски езици, така и от българското анотационно множество в СФ-речника, възниква една чисто методологическа дилема.

Ако сме обявили, че произвеждаме руски граматичен речник в MTE-формат, имаме ли право да допълваме речника на Зализняк с нови характеристики? Например местоименията в оригиналния руски речник съдържат в лексикалната си част само указание за вида склонение – адективно или номинално, без никаква диференциация на вида местоимение; наречието има единствен анотационен маркер за част на речта; числителните са маркирани само като редни и бройни. Причината за тази на пръв поглед бедност на лексикално-граматичната информация в речника на Зализняк е неговото предназначение – по-близко до това на правописните речници с пълна система на формоизменение за всяка речникова единица, отколкото за едноезичен тълковен речник.

Дилемата на конверсията в нов формат – може ли базовата информация да бъде допълнена с нови признаци, е подобна на дилемата на преводача, който има желание да промени оригинала на превода предвид на някакви други цели, или най-малкото, да го направи по-информативен (с цената на многобройни забележки под линия).

---

[36] Руският граматичен речник в DELAF формат бе произведен в съвместна инициатива между Компютърния фонд на руския език (Институт за руски език, Руска академия на науките) и Секцията за лингвистично моделиране (Институт за паралелна обработка на информацията, БАН). Данните на речника бяха предоставени като база данни, съдържаща руските лексеми с техните флективни класове, а организацията му в DELAF-формат бе извършена от програмисти от двете звена.

Соломоновско решение е да се обяви такава конверсия като *производство на речник, базиран върху словника X и анотация Y* .

Илюстрация на конверсията между двата формата е показана в Табл. 16.

вредную,вредный.**A:fsa**	вредную,вредный.Afpfsa---c
врезную,врезной.**A:fsa**	врезную,врезной.Afpfsa---c
временную,временной.A:fsa	временную,временной.Afpfsa---c
временную,временный.A:fsa	временную,временный.Afpfsa---c
вредного,вредный.A:msg:msa:nsg	вредного,вредный.Afpmsg---c,Afpmsa---c,Afpnsg---c
врезного,врезной.A:msg:msa:nsg c	врезного,врезной.Afpmsg---c,Afpmsa---c,Afpnsg---c
временного,временной.A:msg:msa:nsg	временного,временной.Afpmsg---c,Afpmsa---c,Afpnsg---c
временного,временный.A:msg:msa:nsg	временного,временный.Afpmsg---c,Afpmsa---c,Afpnsg---c
вредно,вредный.A:nsy	вредно,вредный.Afpns----n
временно,временный.A:nsy	временно,временный.Afpns----n
вреднейшими,вредный.**A:spr**	вреднейшими,вредный.Afs-p-i---c
вредов,вред.**N+NA+M:pg**	вредов,вред.Ncmpg--n
временников,временник.N+NA+M:pg	временников,временник.Ncmpg--n
вредить,вредить.V+IPF+i:AI	вредить,вредить.Vmi----------p
врезаться,врезаться.V+IPF+i:AI	врезаться,врезаться.Vmi----------e
врезываться,врезываться.V+IPF+i:AI	врезываться,врезываться.Vmi----------p
временить,временить.V+IPF+i:AI	временить,временить.Vmi----------p
вредили,вредитьV+IPF+i:APp	вредили,вредить.Vmis-p------p
врезались,врезаться.V+IPF+i:APp	врезались,врезаться.Vmis-p------p
врезались,врезаться.V+PF+i:APp	врезались,врезаться.Vmis-p------e
врезывались,врезываться.V+IPF+i:	врезывались,врезываться.Vmis-p------p
врезывались,врезываться.V+IPF+i:	врезывались,врезываться.Vmis-p------p

Табл. 16. Преформатиране на единиците на руски граматичен речник в MTE-стандарт.

Близостта на двата формата – словоформа, лема, лексико-граматични характеристики, предуславя възможността всички подобни конверсии – в рамките на един език, да се извършват със специално конструиран софтуер, който предвижда операции върху всички възможни разрези на лингвистичното знание (описание на софтуера, послужил за тази серия от конверсии вж. в [Паскалева 2005].

3.7 Обем на анотационното множество

Един важен резултат от прехода между два вида анотационно представяне на една и съща лексикална база е сравнителната им оценка им по две основни линии. Първата е концептуална, спрямо дълбочината на лингвистичното представяне, а втората е прагматична – по функционалност и икономия на представянето. Да отбележим, че ако конвертирането е само формално – да се премине от един тип структура и подредба на данните към друг (вж. Табл. 13), то би могло да се извърши дори автоматично.

Оценката по концептуална линия е свързана със съдържанието на лингвистичните данни, което трябва да се конвертира при преминаване към стандарт, който не е само прескрипция, а вече е запълнен с граматична информация за конкретен език. В такъв случай решението на задачата „конвертирай съдържанието на X в съдържанието на Y“ напомня малко древното решение на Прокруст, ако изходна точка са параметрите на X (по-богати или по-бедни от тези на Y). Конкретните измерения на това решение за всяка отделна част на речта на българския граматичен речник бяха разгледани подробно по-горе, в разделите на 3.2. В повечето случаи преходът е свързан с намаляване на лингвистичната информация, орязване на част от маркерите на входното анотационно множество. Тази разлика в обема на лингвистичното знание произтича от историята на създаване на двата продукта и от техните цели.

СФ-речникът се базира на съвкупността от лексикални и граматични признаци, участващи в българското формообразуване, а също и на основни класификационни характеристики на отделните части на речта. Всички те, взети заедно, целят да произведат един сравнително пълен граматичен анализ на българската словоформа в единиците на нормативната граматика. MTE-стандартът представлява обобщение на морфосинтактичната информация, използвана за анотиране на 12 доста различни по типология европейски езици, а целта му е да създаде обща база за тяхната по-нататъшна автоматична обработка в многоезиков режим. Това неминуемо води до стесняване на рамките на лингвистичната свобода при анотирането, а оттук и до намаляване на обема на анотационното множество.

Количествените измерения на анотационното множество могат да се отнасят до броя на маркерите в една анотационна верига, а могат да визират и броя на всички анотационни вериги в разглежданото представяне. В известен смисъл двете количества са свързани – колкото повече маркери, т.е. различни признаци са включени в една анотационна верига, толкова повече ще бъдат различните анотационни вериги, които ни дават абсолютния обем на анотационното множество. Или дължината на анотационната верига, измерена в брой на маркерите в нея, е право пропорционална на обема на анотационното множество.

Тази взаимозависимост може да бъде разгледана и в конкретните числови стойности на обема на анотационното множество, измерен чрез броя на различните анотационни вериги в двата разглеждани продукта – СФ-речника и MTE-стандарта, зададени в Табл. 17.

Част на речта	СФ-речник	МТЕ-стандарт
Съществителни	73	59
Прилагателни	36	28
Глаголи	1605	87
Числителни	22	21
Местоимения	70	193
Наречия	55	4
Съюзи	2	2
Предлози	1	1
Частици	1	1
Междуметия	1	1
Общо	**1866**	**397**

Табл. 17. Брой на анотационните вериги по части на речта в СФ-речника за еднакъв обем на лексикалната база – 1 140 000 словоформи.

Тези числови стойности се нуждаят от разяснение и навеждат на някои размисли, тъй като зад абсолютните числови стойности се крие богатството на лингвистичната информация.

Понеже става дума за един и същ брой лексикални единици, анотирани по различен начин, очебийната разлика в обема на анотационното подмножество за някоя част на речта може да се дължи на следните причини:

1. Различен обем на парадигмата на тази част на речта.

Става дума за избраното разпределение между лексикалното и граматичното в анотационното множество. Ако някои форми от състава на една парадигма са лексикализирани (т.е. станали са самостоятелни лексикални единици в друга част на речта и са приели нейната анотация), това намалява нейния обем и съответно броя на граматичните маркери. За сметка на това пък се увеличават лексикалните маркери – за новата речникова единица, но само при условие, че тя се нуждае от това – на новото си място образува свой лексикален клас. Така например включването на редовно образуваните отглаголни съществителни в глаголната парадигма (вж. 2.5 и 3.5.2) увеличава броя на граматичните маркери и съответно на различните анотационни вериги с две – за ед.ч. и мн.ч.[37]

[37] С колко члена трябва да увеличи глаголната парадигма едно отглаголно съществително е въпрос дискусионен. Става дума за случаите, когато то е редовно образувано от глагол несвършен вид и не е лексикализирано (т.е. не е преминало в разряда на съществителните като самостоятелна речникова единица , вж. *ожулване*, а се възприема само като название на глаголно действие в абстрактен смисъл, вж. *обхващане*). Тъй като става дума за абстрактно опредметяване на действие, трябва ли тази форма да получи четирите граматични варианта за същ. ср.р. или можем да се ограничим само с назоваването и в ед.ч. и евентуалното й членуване? Позволява ли този абстрактен прочит на действието образуване на множествено число? В последната версия на речника отглаголното съществително заема две клетки в парадигматичната черупка, засега определени като ед. и мн.ч, правомерността на това решение предстои да бъде проверена с експерименти върху голям текстов масив, от която да се *уловят* отглаголни съществителни, невключени в лексикалната база.

С подобно *прехвърляне* от граматично към лексикално и обратно ни сблъсква и решаването на проблема – степените за сравнение на прилагателните *в* или *извън* парадигмата? В първия случай, разглеждани като форми, те увеличават броя на граматичните маркери с 18 (по девет за всяка степен), а във втория, разглеждани като отделни леми, лексикалните маркери се увеличават с 2. В крайна сметка за прилагателните не се променя общият брой на маркерите, а само тяхното разпределение (новите два лексикални маркера ще присъединят редовните граматични девет на адективното склонение).

2. Различна дълбочина на лингвистичното знание, отразено в анотационната верига.

Илюстрация на тази разлика (в посока от по-малкото към по-голямото) ни дават обемите на анотационните вериги за местоименията и за глаголите в двата стандарта, разгледани по-долу.

Както беше споменато в 3.5.4, лексикалните маркери за **местоименията** в последната версия на СФ-речника са редуцирани главно по отношение на свойствата на местоименния референт. Комбинацията от трите маркера – за род, число и лице на притежателя при притежателните местоимения, съчетани с деветте адективни граматични маркера в парадигмите на *мой,твой, негов, неин, наш, ваш, техен* ни дава 63 нови анотационни вериги. Ако към това увеличение се прибавят и комбинациите, които ни дава маркирането на семантико-синтактичния тип на референта (обект, свойство, размер и подобни), съществували в по-ранни версии на речника, разликата от 123 анотационни вериги за тази част на речта получава обяснение. Тук трябва да отбележим, че в зависимост от целите на следващата автоматична обработка тези съкратени стойности се възстановяват лесно.

От друга страна, Табл. 14 сочи, че броят на анотационните вериги за местоименията в по-разгърнатия запис на MTE-стандарта превишава с по-малко от два пъти броя на самите лексикални единици, т.е. на една анотационна верига се падат по-малко от две словоформи, съотношение почти идентифициращо информацията с нейния носител. Това предполага лесното възстановяване на липсваща информация, още повече че става дума само за 363 речникови единици.

Огромната разлика в обема на анотационното множество за **глаголите**, което в СФ-речника е 30 пъти по-голямо от това в MTE-стандарта, се дължи на разликите в дълбочината на анотацията, разгледани подробно в 3.2.2. Броят на анотационните вериги в MTE-стандарта – 87, ни показва, че са отчетени

само граматичните стойности на словоформите, определящи ги като член на глаголната парадигма за два лексикални подкласа глаголи – пълнозначен и спомагателен. За първия тип парадигматичната черупка е в пълния си състав от 54 члена, а за втория – в намален състав от 33 члена. Признакът тип на глагола – главен (пълнозначен) или спомагателен – е единственият лексикален глаголен признак в този стандарт. Всички други лексикални свойства на глаголната лема – вид, преходност, залогово поведение – отсъстват в стандарта MTE. Затова пък наличието им в анотацията на СФ-речника увеличава общия брой анотационни глаголни вериги до 1605 различни. Редукцията на анотационното множество при преход към MTE-стандарта следователно е за сметка на изгубена лингвистична информация. Последната трудно може да бъде възстановена, тъй като броят на глаголните словоформи и в двата речника е 815 000, което дава съотношението между обема на словоформите и техните анотационни вериги съответно: 9400:1 за MTE-стандарта и 509:1 за СФ-речника.

Ако въведем единицата мярка *плътност на анотационното множество*, която се изразява чрез горното съотношение – словоформи към анотационни вериги, не е трудно да направим извода, че по-голямата плътност на анотирането за определен лексикален клас (в случая част на речта) затруднява конвертирането на анотационни множества особено в посоката от по-бедна към по-богата анотация. За сравнение – анотационната плътност за местоименията е много по-ниска – 1.6 за MTE-стандарта и 4.3 за СФ-речника. При такава ниска плътност на анотацията и малък абсолютен обем на словоформите – 308 местоименни словоформи в двете представяния, конвертирането, дори и с допълване на лингвистичната информация, не представлява трудност .

Оценката по функционална линия е свързана главно с икономичността на представянето на анотационното множество, главно чрез съкращаване както на самата анотационна верига, така и на общия брой на веригите. Това изискване е свързано главно с методите на статистическа обработка, където очевидно изборът чрез предсказване и изчисление измежду елементите на едно множество е по-лесен, когато те са по-малко на брой. Тази редукция налага подробна оценка на всеки един от параметрите на анотацията – можем ли да се лишим от него, с оглед на бъдещи операции – снемане на многозначност, синтактичен анализ, можем ли да го обединим с други и най-вече възстановима ли е тази лингвистична информация дори и с цената на допълнителни процедури, след като сме постигнали желаната ефективност на изчислението. За това ще стане дума по-нататък, в 4.2.1, при описанието на статистическите средства за анотиране.

4 От морфология към синтаксис – първи стъпки

Разнообразни формати, различно покритие на лексиката, различна дълбочина на граматическото знание – независимо от трудния избор да вземем най-подходящото за нашия език произведеният от нас речник е свършил задачата си и е преобразувал редицата от думи на текста в редица от леми и техните граматични характеристики. Няколко са посоките, в които си струва да се продължи.

Едната е – въоръжени със знанието, получено от речника, да продължим понататък, напред в обработката на текста. Като не забравяме, че речникът сам по себе си може да ни помогне да анализираме текста, но не може да отхвърли многозначните анализи – те са резултат от естествената езикова омонимия, която в човешкото възприемане и анализ на текста се снема от механизми извън лексикалното знание.

Другата е – след като сме получили всичко от този речник, да го сложим настрана, да погледнем извън него и да си представим две на пръв поглед съвсем противоположни ситуации:

- какво бихме могли да получим от неговото разширяване към един подобър или допълнителен речник на лексиката – с друга организация, с представяния от други морфологични равнища и други морфологични елементи, извън простата дихотомия лема-форма, лексикално-граматично, постоянно-изменяемо,
- а какво бихме могли да получим, ако не разполагахме с този инструмент на предварително организираната лексика, а използваме само: зависимости на буквено равнище, статистически методи и техники от машинното самообучение, неорганизирана лексика - огромни обеми текстове?

С или без речник? Подобно на кръстопътя в приказките «тръгнеш ли надясно, еди какво си, тръгнеш ли наляво, еди какво си», съществуват две посоки, условно определени като напред и назад. Те засягат повече решението – за какво ще използваме този речник в една бъдеща обработка и какво място ще му отредим, на главно или на спомагателно средство. Извън обсъжданите посоки, които засягат главно движението към бъдещи автоматични процедури,

остава главното достоинство на речника – като основен лексикален и граматичен склад на езиковия материал, който може да бъде използван в различни комуникационни технологии.

Посоката **напред,** с главно, но не единствено оръжие **– речника, ни води към** анализа на по-дълбинни езикови равнища – тези, които следват морфологията в описанията на нормативните граматики.

Посоката **назад,** с допълнителен инструментариум от други методи и средства, по нови пътища, ни води **към по-повърхнинен, но по-широкоформатен** анализ на текстовото съдържание, който ни приближава до неговия смисъл, без да ползва фино наточените (но с ограничена употреба) инструменти на езиковата семантика.

Горните две посоки с техните пространствени уточнения фактически приповтарят общоизвестната истина, че в езиковия анализ се напредва, като се копае по-дълбоко, но че не трябва да се пренебрегват и страничните пътища към целта.

За двете алтернативи изборът засяга вида на сегмента на знанието, подобно на отрязъка от тортата – дълбок и тесен или плитък и широк?

Изборът в *дълбочина* ни приближава към следващото равнище на представяне на лингвистичните зависимости – синтактичното, като ни оставя на прага му, готови да атакуваме текста с оръжията на истинския синтактичен анализ, за който вече сме подготвени.

В какво се състои нашата готовност да щурмуваме синтактичното представяне? Главно – в определянето на точните, истински граматични роли на думите в текста. Преходът от зависимостите в речника, общовалидни и системни, към зависимостите в текста, конкретни и избрани, се извършва чрез серия от свързани процедури, осъществявани върху съвкупности от:

- неанализирани думи,
- анализирани, но все още не определени еднозначно думи (поради граматична и лексикална омонимия),
- анализирани и еднозначно определени думи.

Съвкупността от трите вида текстови порции, съдържащи лингвистично знание в диапазона от нула до максимално пълно (в границите на задачата), се обработва в определена последователност чрез специализиран софтуер. Той осъществява един вид лингвистичен кастинг, избор между трите вида канди-

дати за правилната лингвистична роля. Тоест, разглеждаме всички възможни, номинираните и избраните. Процедурата на този избор ще бъде разгледана по-подробно по-долу, в 4.1.3.

Следваща стъпка по пътя напред, в дълбочина, към синтактичната структура, е сглобяването на по-сложни текстови единици, съставени от съседни анализирани думи. Получава се нещо като синтактичен полуфабрикат, който ще помогне за производството на истинското синтактично представяне, но може да се използва и преди това. За него ще стане дума по-подробно в 4.2.

Изборът на *ширина*, на по-плиткото парче от лингвистичната торта, фактически променя и основното средство за анализ, за което говорихме дотук – думите и техните граматични свойства, в комбинация със свойствата на съседните им думи. Посоката навън – към съседната дума – се сменя с посоката навътре – в самата дума, към нейните компоненти. За езици с богата морфология, която се проявява не само в изобилието на формите, а и в разнообразието на словообразователните механизми и реализиращите ги морфеми, информативността на морфемните елементи не трябва да бъде пренебрегвана. Това важи особено когато нямаме под ръка склад с подредено лингвистично знание, но затова пък разполагаме с огромен текстов материал, от който могат да се извлекат повърхнинни фрагменти от същото това знание. В тези специфични задачи не се цели нито пълен лингвистичен анализ, нито анализ на целия текст. Подобно задание може да се изпълни и в отделни модули на пълния лингвистичен анализ, натоварени с подпомагаща функция. Като средство за откриване на подобни фрагментарни зависимости се ползват статистически методи за анализ на езикови ресурси, за които граматичният речник не е нито първото оръжие на атаката, нито основната тухличка на градежа. Реалните ситуации за такава обработка са случаите, в които липсва достъпен граматичен речник (например при обработка на нови езици), или в отделни ситуации, когато речникът е безполезен поради статичността си (при голямо количество непознати думи, в ненормализиран или специализиран текст). По-подробно за тази възможност вж. 4.3.

4.1 Морфологичното представяне – от възможност към реалност

В схематичното описание на морфологичните модели като движение от текста към неговото морфологично представяне и обратно (Глава 2), както и в разглеждането на основния лексикален ресурс, база за този преход (Глава 3) съзнателно бяха пропуснати всички усложнения, произтичащи от различен

прочит на езиковите знаци. Това, което ни дават процедурите на анализа като изходен продукт в базисното просто изпълнение (думата, снабдена с информация, взета от граматичния речник), е изпълнимо само в чистите, еднозначни случаи, когато един буквен стринг е свързан с едно граматично значение. Тази връзка, еднозначна или еднозначно изчислима, между означавано и означаващо обаче е характерна само за изкуствените езици, но не и за човешкия, естествен език

Многозначният прочит на буквения низ, думата, е такъв само в изолирана употреба – подобно на речниковите статии, в които са изброени всички възможни значения на лексикалната единица. В текста прочитът е само един. Никой няма да си помисли, че в изречението *И скочиха в студените води* думата *води* значи същото, като в изреченията: *Аз също не знам – но води ме, води ме натам!* или *Скарали се кой да води бащина дружина.* Този еднозначен прочит се осъществява от механизма на човешкото разбиране, трудно подаващ се на описание и невъзможен за пълно моделиране.

Анализ, базиран на информация от граматичния речник, където горната омонимична словоформа е представена съответно от три речникови единици, ще припише към тази дума значенията на трите омонимични словоформи, представени от три речникови записа. В първия от тях фигурира лемата *вода* и признаците същ. м.р. мн.ч. нечл., във втория – лемата *водя* и признаците глагол, пов.накл. ед.ч., а в третия – лемата *водя* и признаците – глагол, сег.вр. 3 л. ед.ч. зи три прочита са проява на езикова омонимия, за случая 1 vs. 2 & 3 – лексикална, а за случая 2 vs. 3 – граматична, т.е. в първия случай имаме съвпадане на формите на различни лексикални единици, а във втория – съвпадане на различните форми на една лексикална единица.

Истинският резултат от морфологичния анализ, с който трябва да почукаме на вратата на синтактичната обработка, не позволява трите прочита едновременно, както не го позволява и човешкото разбиране на текста. Трите прочита са част от системата, по тази причина са отбелязани в речника, а единственият прочит е част от езиковия израз, тъй като един текст може да се разбира само по един начин (без да се броят случаите на специално търсена многозначност). Намирането на верния прочит е процес, в който са ангажирани съществени порции от лингвистично знание, специални методи за изчисление и на последно място, но не по значимост, човешка помощ и настройка. Операциите на специализирания софтуер, който осъществява тази текстообработка, ще бъдат разгледани по-долу, в. 4.3.

4.1.1 Граматична омонимия – методи и средства за нейното изчистване

Дадените по-горе примери за трите прочита на словоформата *води* са взети от граматичния речник. Там са регистрирани всички възможни значения на езиковите единици така, както са представени в езиковата система. Това е цялото множество, с което трябва да работи морфологичният анализ, за да определи правилната за дадения текст анотация на думата. Очевидната аналогия с човешкото разбиране, което също не би могло да реши кой от трите прочита е валиден за дадена изолирано текстова единица, ни навежда на мисълта, че само заобикалящият я контекст може да реши този проблем. Но как? Лингвистичният здрав смисъл ни подсказва, че измежду множеството контексти, сред които може да се разполага дадена многозначна дума, има някои явно диагностични – вж. *мътните води* vs. *кой води*, където лявостоящият елемент ни насочва към правилния избор между съществително и глагол. Но има и контекст, който не ни казва как да различим омонимичните глаголни форми в примера *води ме със себе си* – повелително наклонение, ед.ч., сегашно вр. 3 л. ед.ч., или аорист 2/3 л. ед.ч. Виж употребите:

[1] *Води ме със себе си и да не си гъкнал! – ми заповяда той,*

[2] *Миналото лято моят племенник ме води на много места.*

[3] *Води ме със себе си навсякъде момчето и сега.*

[4] *Спомняш ли си, води ме със себе си навсякъде, а сега отричаш.*

В [1] възклицателният знак, който може да ни подскаже повелителното значение, се намира твърде далеч от омонимичната форма. В [2] миналото време е подсказано от семантиката на отдалечената номинална група *миналото лято,* в [3] сегашното време се изчислява от темпоралното наречие *сега*, намиращо се на края на изречението, а отдалеченият подлог *момчето* определя избора на трето глаголно лице. В [4] второто лице се изчислява от глаголното лице на сказуемото в следващото съчинено изречение. Всички тези диференциращи белези, осигуряващи правилния прочит на омонимичната словоформа, се откриват само при пълно разбиране на текста, което предполага установени синтактични и семантични връзки. Такава информация не ни е достъпна в етапа на анализа, където:

- максималната лингвистична информация, с която разполагаме, са анотациите на думите, взети направо от речника,
- връзките между думите се свеждат само до означаване на съседство и брой на междинно разположените думи и

- за семантика може да говорим само в изолирани случаи, когато тя е отразена в анотационните признаци.

Невъзможността да се изведат правила за определяне на актуалния прочит в текста измежду възможните кандидати, пренесени пряко от речника, само с помощта на правила за линейно разположение и признаци на съседните думи налага да напуснем рамките на обработката, базирана на лингвистично знание, и да потърсим помощта на статистиката, на практическата езикова употреба и нейните измервания и изчисления. Те се осъществяват върху броя появи на определена анотация, в съчетание с анотациите на съседните ѝ думи. Подобно изброяване и изчисляване на честотата на характерното обкръжение на думите се извършва върху предварително анотиран текст, в който омонимията е изчистена, т.е. фигурират само реалните, а не всички възможни прочити на анотацията.

Такава порция текстов материал с еднозначна анотация, който да служи за основа за изчислителни изводи, се нарича учебен корпус (training corpus). Той е основният езиков ресурс в специализирания софтуер – *тагер*, който осъществява операцията *тагиране* (tagging). Тагирането дава като резултат анотиране с изчистване на омонимията, с други думи казано – чисто анотиране (вж. по-долу, 4.1.3)

4.1.2 Типове омонимия – в речника и в текста

Проследяването на явлението омонимия в пълен граматичен речник и в текстов корпус фактически е проследяване на проекцията на едно системно явление върху употребата му, на налагането на всички *възможни* върху *наличните*. Процедурата на това налагане е външна за тагирането – за тагера е безразлично какъв е произходът на учебният корпус, анотиран ръчно от човек или полуавтоматично – чрез граматичния речник с последващо ръчно изчистване на омонимичните вериги. Проследяването и систематизацията на омонимията в речника има повече изследователско значение и е механизъм за натрупване на данни за омонимичното поведение на езиковите единици в текста, които ще бъдат обработвани от тагиращия софтуер по съответния начин.

Омонимия в речника

В граматичния речник се съдържат всички възможни анотационни вериги, приписани към лексикалните единици на даден език.

Инвентаризацията и класификацията на типовете омонимия в речника (под тип омонимия разбираме уникално съчетание от лексикални и граматични характеристики, приписани на повече от една лема – вж. примерите на с. XXX) ни позволяват да очертаем границите на цялата възможна омонимия в езика. Ако може да се говори за някакво разпределение в това пълно множество, то засяга само броя на единиците, съвпадащи по буквен състав и различаващи се по лексикално и/или граматично значение. В зависимост от позицията и йерархията на идентичните по буквен състав единици (в парадигмата на една лема, между парадигматични членове на различни леми или в комбинации от двата случая) в работите по оценка на анотационните множества за нуждите на тагирането тези два типа омонимия (между парадигматични членове на две леми или вътре в парадигмата на една лема) се назовават различно. Те се обозначават като *лексикална* vs. *граматична* омонимия, още като *интеркатегориална* vs. *интракатегориална*. Интеркатегориалната омонимия между две различни лексикални единици се проявява на различни равнища, в зависимост от йерархията на граматическите категории. Съвпадащите единици могат да принадлежат към различни части на речта (напр. омонимия съществително – глагол), но могат да принадлежат и към различни подкатегории на частите на речта (между лично и притежателно местоимение – известната омонимия между дативното и притежателното местоимение – *казах му* vs. *пуловера му*; между отделни класове на числителни – *десетина души се отзоваха* vs. *една десетина от дозата*).

В лексикалната база на СФ-речника, състояща се от 1 141 000 словоформи, са отбелязани 118 типа омонимия, обхващащи 88 640 словоформи, към които е приписана повече от една анотационна верига (максимален брой на веригите – 5). Едно изследване на причините на тази омонимичност показва следното.

Повече от половината от тези типове (58% от речниковите записи – словоформи) се дължат на повишени изисквания за изчерпателност на лингвистичното знание, които се проявяват в прекалено раздробяване при анотацията – характерен пример са глаголите. За конкретната стойност става дума за глаголите от двойствен вид, където прочитът на миналото време е едновременно свършено и несвършено (в тази плоскост на разсъждение *автоматизирах* обединява темпоралните значения на *ходих* и *ходех* и следователно получава две анотации). В някои български анотационни стандарти, като MTE и BulTreeBank двойственият вид на глагола не се маркира самостоятелно, а приема стойността на несвършения вид. Следващи типове омонимия (по обхват на речниковите единици) са тази между наречието и краткото прилагателно в среден род (*красиво момиче* vs. *красиво казано*) – 10%, вече цитираната омонимия между

кратката членна форма и бройната форма на съществителното от мъжки род (*два стола* vs. *на стола*) – 4,4%, омонимията между нечленуваните форми на единственото и множественото число на прилагателните на *-ски* (*горски въздух* vs. *горски поляни*) – 4% и др.

Цитираните проценти на омонимичните единици се отнасят за граматическата, *интракатегориалната* омонимия. Съвпадане на буквени стойности на членовете на една парадигма имаме при доста лексикални единици с еднакъв формообразуващ механизъм (напр. всичките 5 000 неодушевени съществителни от мъжки род образуват еднакви форми за кратката членувана и за бройната форма).

Другият тип омонимия, лексикалната или *интеркатегориалната*, не е толкова редовна и се проявява при по-малък брой речникови единици – изключение прави само омонимията между формите на прилагателното в среден род и наречието за начин (*красиво момиче* vs. *красиво пее*). Обяснението за този висок процент интеркатегориална омонимия (между различни части на речта) ни дава междинната природа на тези две единици, произведени на границата между формообразуване и словообразуване – поради редовността на образуване на наречие от качественото прилагателно, което ги доближава до интракатегориалната омонимия.

Висок процент интеркатегориална омонимия имаме при съвпадането на лексикализирано отглаголно съществително в мн.ч. с кратката членувана форма на страдателно причастие, образувано от същия глагол (*литературни четения* vs. *четения вестник*). Очевидно и тук става дума за гранично явление между словообразуване и формообразуване – вж. разсъжденията за обема на парадигматичната глаголна черупка в 2.7. По тази причина речниковата омонимия от този тип е представена при 3210 речникови единици.

Интеркатегориалната омонимия от междинен тип, проявявана между подкатегории в рамките на една и съща част на речта, като споменатата вече омонимия между личното и притежателното местоимение или между различни класове числителни се проявява при малко на брой единици.

Сравнително малко са случаите и на интеркатегориална омонимия, в които участват неизменяеми думи, представени в единствената си форма, което ги лишава от богатството на комбинаториката между различните форми на изменяемите части на речта. Такива са например единичните случаи на омонимия между частицата *бе* и миналото време на спомагателния глагол *съм* (*кажи, бе!* vs. *той бе предаден*), между частицата *дали* и миналото време на глагола *дам*

(*дали са му дали достатъчно храна?),* предлога и частицата *по* (*по земята* vs. *той е по юнак*), съюза и предлога *като* (*хубав като бог* vs. *спри, като се умориш*) и други единични случаи.

Това разпределение на омонимията, наблюдавано само върху речниковите данни, ни дава по-скоро сведение за дълбочината и раздробеността на граматичното анотиране и за омонимичността на граматичните формативи в конкретната езикова система.

Омонимия в текста

Съвсем друго е разпределението на типовете омонимия в един текст, тъй като е свързано вече с честотата на лексикалните единици в реалната им употреба им.

Измервания в текстово проявените типове омонимия в български корпус от правни текстове с обем 110 000 думи показват съвсем друга картина за честотите на омонимичните типове, тъй като правилата на тяхното броене и разпределение са други. В текстовия корпус, за разлика от *потенциалните* носители на омонимията в речника, се броят *реално* съдържащите се в текста словоформи, които реализират потенциалната омонимия на речниковите единици.

В количествено изражение един омонимичен тип в *речника*, обособен като уникален по параметрите на информацията за съвпадащите низове, се проявява в X броя речникови единици. В *текста* се среща подмножеството X1 на тези речникови единици, които се срещат Y пъти (като текстови единици). Конкретен пример: омонимията от типа отглаголно съществително на *-не* или *-ние* в мн.ч, съвпадащо с кратката членувана форма на страдателно причастие (*нови назначения* vs. *назначения счетоводител*) се среща в 6808 (X) омонимични двойки в СФ-речника. В изследвания текст от 110 000 думи са представени 47 (X1) такива единици, обединяващи различни лексикални двойки, в 305 (Z) появи. Реалната граматична стойност на 10% от тези появи е страдателно причастие, а останалите 90% са съществителни имена. В този случай имаме многократно надвишаване на броя на омонимичните единици в речника в сравнение с проявите им в текста. Причината е в характера на интракатегориалната омонимия, където съвпадащите формативи за различни граматични категории се проявяват в голям брой лексикални единици с еднакъв модел на формообразуване.

Обратното съотношение между двете стойности на типа омонимия, когато текстовите прояви многократно надвишават речниковите, се проявява в интеркатегориалната омонимия. Съвпадението между някои от формите на две

леми от различни части на речта се проявява само в една речникова единица – вж. цитираните по-горе случаи на интеркатегориална омонимия *(дали*, *бе* и др.) в речника с участието на неизменяеми части на речта, повечето от които са служебни думи. Именно този служебен, свързващ характер на тези единици определя високата честота на тяхната употреба в текста. Дадената по-долу Табл. 18 показва честотата на типовете интеркатегориална омонимия (в конкретните случаи – *по, да, е, като*) и на интракатегориалната омонимия за лексикална единица с висока честота (омонимията между кратка членувана и бройна форма за съществителните *ред* и *закон*),

Инвентаризацията и подробната класификация на разгледаните типове омонимия е необходима базисна информация за конструирането на контекстни лингвистични правила, които да разрешават омонимията в етап от анотацията, предшестващ или придружаващ статистическите методи за избор на правилната анотация (вж. по-долу).

	Брой	**Процент**
Корпус – обем	110 000	100%
Неанотирани (ненамерени)	9 000	8%
Анотирани еднозначни	79 000	72%
Анотирани омонимични	22 000	20%

Табл. 18. Разпределение на неанотирани, еднозначно анотирани и омонимични думи в корпус – юридически текст.

За същия корпус от 110 000 думи Табл. 19 показва само първите най-чести типове омонимия. Корпусът е анотиран с помощта на СФ-речника, като изборът на правилното значение от няколкото омонимични е направен ръчно.

Омонимични единици – брой		**Значения**	**Първо**	**Второ**
по	1986	Предлог/частица	1986	-
да	1372	Съюз/частица	1372	-
е	885	Глагол/частица	885	-
дела	282	Същ.м.р.кр.чл. / същ. ср.р. мн.ч	-	282
като	259	Съюз/предлог	111	148
закона	247	Същ. бр.ф-ма / същ. кр.чл.	-	247
им	225	Притеж. мст. / лично мст.	177	48
му	216	Притеж. мст. / лично мст.	179	37
реда	210	Същ. бр.ф-ма / същ. кр.чл.	-	210

Табл. 19. Разпределение на значенията на някои омонимични единици в същия корпус.

Конкретните прояви на езиковата омонимия в текстов корпус, освен че ни дават сведения за честотата на явлението, са полезни за организацията на изчистване на омонимията – статистическа или основана на правила, а също и за корекция на формулираните в речника анотационни стойности. Примерите в горната таблица показват, че някои от най-често срещаните случаи на омонимия могат да престанат да бъдат такива, като се откажем от едно от значенията, поне за дадения тип текстове. Анализът на евентуалната грешка при такова елиминиране на значението „утвърдителна частица" за единицата *да* сочи, че грешката би се оказала по-сериозна при анотация на текстове от художествената проза с преобладаваща пряка реч – нещо, което за статистическата обработка на мнозинството от текстовете в големите корпуси (вж. структурата на британския национален корпус) обаче няма значение. Същото важи и за частиците *е* и *по* – в доста рядката им употреба. Не бихме могли да си позволим обаче да премахнем значението на бройната форма за съществителните от мъжки род, макар че в анализирания корпус то не се среща нито един път. За сметка на това пък тази омонимия се снема леко – чрез правила и чрез статистическа обработка (както е известно от нормативните граматики, бройната форма се употребява в ограничени случаи с типичен ляв контекст – числително или количествено наречие – *два закона*, *колко/няколко закона* и др.). Друг случай на омонимия, значението на *като* (предлог и съюз) не позволява редукция, поради сравнително равномерното разпределение на двете значения във всякакви текстове, но този тип омонимия се снема лесно поради ясната диагностика на контекста (съюзът *като* навсякъде е предшестван от запетая или от друг съюз, поне в разглеждания корпус).

Едно потвърждение на зависимостта на разпределението на типовете омонимия от характера на текста намираме при анализа на текст със същия обем, съставен от класическа и съвременна българска проза. Там единицата *да* се среща в 4 300 прояви – 4,5 пъти повече от правния текст. В 4 395 прояви тя е в състава на аналитични глаголни времена и на сложни съюзи и само в 5 прояви е утвърдителна частица в пряка реч. Омонимични лексикални единици с висока честота, произлизаща от характера на юридическия текст (като *закона* и *реда* – вж. таблицата), тук са представени съвсем скромно – *закона* – 2 прояви в значението кратка членувана форма, реда – 0 прояви.

Тъй като практиката на статистическото анотиране с помощта на учебен корпус посочва обема на последния като вариращ между 50 000 и 1 000 000 думи (вж. следващия раздел), очевидно е, че за българските случаи на омонимия, проявени в корпус от 200 000 думи (сумата от думите на споменатите два корпуса), изводите за правилния избор на значение, построени само върху наблюдава-

ните текстови прояви, са ненадеждни. Това засилва убеждението, че знанието от пълния граматичен речник може и трябва да бъде използвано и в анотиращи статистически процедури, дори и в непълен вид, с цената на известно редуциране на броя и дължината на анотационните вериги и последващото им възстановяване (вж. XXXXXX).

4.1.2 Тагери – основни методи и езикови ресурси

В съвременната компютърна лингвистика се говори за няколко вида тагери (специализиран софтуер за анотиране с изчистване на омонимията), като основната разграничителна линия минава между използването предимно на статистически методи за обработка или на методи, основани на лингвистични правила. Едните са т.нар. стохастични тагери (stochastic, statistic taggers), които са основани на вероятността единица с определена граматична стойност да бъде предшествана от специфична конфигурация от единици с определени граматични стойности (вж. примера с *гъстите води* и *той води*). Другите са т.нар. тагери, основани на правила (rule-based taggers), класическият им представител е тагерът на Брил – [Brill 1995]. От същия тип са и тагерите с памет (memory based taggers), вж. [Daelemans 1996]. Твърде често се комбинират методи от двата типа в различна последователност, която определя главния тип на тагера.

За всички видове тагери е характерно използването на основен текстов ресурс, наречен учебен (training) корпус, чийто обем варира от 50 000 до 1 000 000 думи, като решаващата сила на извлечените от този корпус правила е различна в зависимост от ролята, която е отредена на статистическата процедура в процеса на тагирането – основна, спомагателна или само настройваща. Учебният корпус представлява правилно анотиран текст, по определен езиков модел, без омонимия. Обикновено анотацията се извършва ръчно, от хора-анотатори, подпомогнати в различна степен от някакъв граматичен ресурс и специален софтуер.

На тагера, освен този *учебен* корпус, се предоставя и *тестов*, неанотиран корпус, върху който се изпробват и оценяват получените в резултат на статистическата обработка на анотирания корпус граматични характеристики на думите. Естествено, извън учебния и тестов корпус остават доста лексикални елементи и граматични форми на езика, защото дори текстов корпус от един милион думи не може да покрие цялата лексика и граматика на езика. По тази причина голямо значение има подборът на учебния корпус, характерът и съотношението на отделните текстови колекции в него.

Лингвистичните правила, които се използват за подобряване на резултатите от статистическата обработка на учебния корпус, или пък предшестват статистическата обработка, като поемат предварително част от тежестта на лингвистичното решение, са от два вида – лексикални и контекстуални. В лексикалните правила се посочват буквени низове от състава на анализираната омонимична единица, които са диагностични за определено граматично значение (например; английска дума, завършваща на *ous*, приема маркер за прилагателно). Контекстуалните правила посочват граматически признаци, а не букви и са от вида – ако отляво на разглежданата дума има дума с маркер X, то думата трябва да приеме маркер Y. Илюстрация за лингвистично правило от вида на т.нар *кръпки* (patches), което коригира грешно посочени комбинации от маркери за съседни думи, е дадена от Брил при разрешаване на омонимията съществително/глагол за думата *race*. Ако в изречението *We can race all day long* тя е маркирана погрешно като съществително, прилага се следващото коригиращо правило: *ако нещо, което може да е съществително или глагол, е маркирано като съществително, но се предшества от модален глагол, смени маркера му с глаголен.*

Тагерите от всички гореспоменати видове са разработени първоначално за английски език. Впоследствие се прилагат и за високофлективни езици. Изобилието от тагери води до своеобразен маркетинг, до изпробване на тагери от различен вид, с промяна на учебния корпус и анотационното множество. Последното може да бъде променяно в езиковия модел на анотацията или в обема си.

От 2000 г. насам, особено след въвеждането на анотационния общ стандарт в проекта MULTEXT-East, както и създаването на общ паралелен текстов корпус за езиците в проекта и стандарта, се извършиха доста експерименти, които извършват сравнителна оценка на видовете тагиращи процедури и на обема на анотационното множество, използвано в тях. В много от тях се отбелязва твърде големият обем на анотационното множество за високофлективни езици като славянските и румънския, както и за аглутинативните естонски и унгарски езици. При което става дума не за пълното анотационно множество, отразено в пълен граматичен речник за цялата езикова система, а за анотациите, съдържащи дори само в учебния корпус. Таблицата, дадена от Ян Хаич в [Hajic 2000] обобщава данните за обема на учебния корпус, неговото анотационно множество и процента на омонимични думи за някои от езиците от MULTEXT-East.

Език	Учебен корпус	Анотационни вериги	Омонимични единици
Английски	99903	139	38.65%
Чешки	87071	970	45.97%
Естонски	81383	476	40.24%
Унгарски	102992	401	21.58%
Румънски	104583	486	40.00%
Словенски	94457	1033	38.01%

Табл. 20. Езици, обем на учебния корпус на тагера, обем на анотационното му множество, процент на омонимичните думи.

Обем и структура на учебния корпус

Данните от Табл. 20 характеризират само учебния корпус и зависят както от обема му, измерван в думи, така и от неговата структура – различните типове текст. Обикновено се подбират различни текстови регистри – проза, новинарски текст, административен и официален, за по-пълно покритие на лексиката на езика, доколкото това е възможно. Невъзможността един текстов корпус, колкото и голям да е, да покрие цялата лексика на езика, съставя сърцевината на проблема за *рехавите* данни (sparse data). Тази непредставителност на данните се преодолява не само чрез увеличаване на обема на учебния корпус, но и чрез подходяща структура на включените в него текстове. Правилата за представителност на подбраната лексика важат и за големите национални текстови корпуси[38]. Много от класическите тагери за английски са експериментирани и настройвани върху големия електронен корпус Wall Street Journal corpus.

Брой на анотационните вериги

Колкото и да е голям учебният корпус, броят на анотационните вериги в него естествено е по-малък от броя на всички възможни вериги, представени в пълния речник на езика. Съотношението между тези две стойности (веригите в речника vs. веригите в корпуса) е 614:486 за румънски език, по данни от [Tufis 2000], а за словенски език е 2083:1025 [Dzeroski et al. 2000]. Пълното анотационно множество за езиците от MULTEXT-East варира от 600 до 3000 вериги (пак там).

Констатацията, че измежду по-малко на брой стойности по-лесно ще отгатнем правилната, е тривиална, но усилията за намаляване на учебното анотацион-

[38] Приблизителната структура на Британския национален корпус, състоящ се от 100 милиона думи, изглежда така. По тематика: 22% художествен текст, 8% изкуство, религия и вярвания – 3%, финанси – 8%, развлечения – 11%, естествени науки – 4%, приложни науки – 8%, социални науки – 15%, световна политика – 19%., други – 2%. По вид на изданието: книги – 59%, периодика – 31%, други публикувани източници – 4%, други непубликувани – 4%, разговорна реч – 2%.

но множество се поставят особено остро за езици с висока флективност – вж. съотношенията в Табл. 20, където чешкият и словенският език боравят с множества, съответно 7 и 8 пъти по-големи от английския. Нуждата да се намали обемът на анотационното множество е изказана ясно от Торстен Брантс, автор на един от най-използваните днес статистически тагери в [Brants 1995] :

> „От една страна, колкото повече са анотационните вериги, толкова повече информация получаваме. Но от друга – увеличаването на броя на веригите поражда проблема на „рехавите данни", който от своя страна изисква да се увеличи обемът на учебния корпус."

Съществуват и специални методики за *свиване* на анотационното множество преди анотацията и *разгъването* му след нея, за да се възстанови лингвистичната пълнота на анотирането Подобен експеримент за румънски, където 486 анотационни вериги за румънския език са свити до 92, е описан в [Tufis 2000].

Участие на граматичния речник в тагирането

Изложеното дотук ни позволява да си представим няколко възможни обема на анотационно множество в процеса на пълното анотиране с изчистване на омонимията:

1. Пълното анотационно множество на езика, взето от граматичния му речник или от друг пълен езиков модел, с постоянен обем;
2. Анотационното множество, представено в учебния корпус с обем, вариращ според обема на корпуса;
3. Редуцираното анотационно множество на учебния корпус (от т.2) с цел – улесняване на статистическите процедури, с последващо възстановяване;
4. Редуцирано анотационно множество от пълното речниково множество (от т.1) с настройка към типовете омонимия – поведението им в системата и в текста.

Определянето на работното анотационно множество се извършва в последователността на стъпките: 1, евентуално 4, 2 и евентуално 3.

Участието на един пълен граматичен речник в процеса на пълната анотация може да бъде предварително, за уточняване на вида и обема на веригите, с които ще бъде анотиран учебният корпус. Ръчната анотация на този корпус обикновено не се обсъжда като процедура, тъй като резултатът от нея е само вход на тагера. Възможни са всякакви степени на улеснение и автоматизация

на този тежък процес – изцяло ръчна анотация, подпомогната от подходящ интерфейс за избор на съответните стойности, ръчно снемане на омонимията (също подпомогнато в различна степен от автоматична процедура) от анотиран корпус, където веригите от речника са пренесени механично в текста. Така граматичният речник не участва пряко в тагирането, а само в задаването на неговите входни данни. В изходните данни информацията от граматичния речник е представена в известен смисъл фрагментарно, само в случай на съвпадане на елементите от лексикалната база на речника с тези от текста.

Това частично покритие на текстовото лингвистично знание от речниковото знание стои в сърцевината на споменатия вече проблем на «рехавите данни» (sparce data) – ненадеждността на учебния корпус, който и при най-голям обем не може да представи пълната лексика на езика. Този проблем се атакува чрез увеличаване на учебния корпус, различни конфигурации на обемите на учебния и тестовия корпус, както и в самия процес на обучение на тагера.

Ако се откажем от идеята да тренираме тагера само върху анотационното множество на учебния корпус, можем да използваме цялото граматично знание в граматичния речник, ако върху неговите пълни данни съставим процедура за снемане на омонимията. Такъв подход за използване на речника имаме в испанската версия на проекта FrameNet [7] (с крайна цел семантичната анотация на текст – задача, свързана с по-дълбинни лингвистични равнища от това на морфосинтактичния анализ). Граматичният речник може да бъде и спомагателно средство, засилващо точността на анотиращата процедура в статистически тагер.

До необходимостта да се използва информативността в един пълен граматичен речник са стигнали представители на славянски език с особено богата флективност – чешкия, вж. [Hajic et al. 2001]. Съображенията на авторите да мобилизират в битката с омонимията тежката артилерия на чешката морфология в нейното пълно снаряжение (речника) са свързани главно с типологията на езика – богата морфология и сравнително свободен словоред, характеристики, близки до тези на българския език. Тези два фактора водят до не много добри резултати от чисто статистическото тагиране, по-лоши от английските при използване на същата процедура. Богатата морфология, която обуславя големия обем на анотационното множество, прави данните за обучаване наистина рехави. Свободният словоред от друга страна утежнява голяма част от контекстните правила на статистическата обработка (сравни единствената правилна употреба *told him* с позволените вариации в български – *каза му* и *му каза*).

Подобряването на анотиращата процедура за чешки език става чрез въвеждане на лингвистични правила за обработка на по-трудните случаи на омонимия, а по-леките случаи, където изборът се определя еднозначно от точно формулиран контекст, се оставят за статистическата обработка. Двата вида обработка работят последователно върху входа на тагера. Тези входни данни фактически представляват резултата от морфологичен анализатор за чешки без снета омонимия. Върху него се прилагат първо лингвистичните ръчни правила, насочени към по-трудните за разрешаване типове омонимия, а върху резултата от тяхната дейност (останалите по-леки случаи) се пуска статистическият тагер.

За отбелязване е, че един от първите тагери с участието на ръчно съставени лингвистични правила са реализирани за един друг силно флективен език – френския, в тагера на изследователския център на XEROX [Chanod and Tapanainen 1994] . В него, както и в чешкия тагер, граматичният речник е основното средство за конструиране на ръчни правила за снемане на омонимията, които се прилагат върху анализиран с помощта на речника текст.

Само последователни експерименти с участието на различни порции лингвистично знание биха могли да ни приближат към *идеалния тагер*, настроен към спецификата на конкретния език. Настройката на тагера (като последователност от лингвистични и статистически процедури) осъществява и обратната връзка между резултатите от тагиращата процедура и състава на лингвистичната информация в речника.

Ако оставим за малко изискванията за изчерпателност на лингвистичното знание, моделиращо точно нормативните граматики, и се обърнем към по-прагматични съображения, закономерно стигаме до извода, че някои лексикални характеристики биха могли да бъдат спестени в речника – напр. определянето на някои съюзи едновременно като съчинителни и подчинителни (*че*) или като съюз и частица (*ли*); бихме могли да се откажем от частицата *бе*, която прави омонимия със спомагателния глагол, и много други практични и трезви решения.

Едва след изчистване на излишната (за практическите ни цели) омонимия, евентуално след аргументиран отказ от решаването ѝ (напр. за глаголните стойности свършен и несвършен вид, нерелевантни за синтактичната структура) и други методи за филтриране на това граматично замърсяване, може да пристъпим към създаването на тагер с помощта на правила, които допълват или улесняват статистическата обработка.

Дори и да се насочим към чисто статистическия подход, въпросът за изморителното ръчно изчистване на омонимичните случаи в текст от порядъка на

половин милион словоформи също поставя въпроса за оптимизация на анотационните вериги.

Освен това концептуално улеснение, свързано с представянето на лингвистичното знание, съществена помощ за трудоемкото анотиране на учебния корпус може да ни окажат и софтуерни средства, подпомагащи човешкия избор на омонимичните значения в един полуавтоматизиран процес. За създаването на такъв помощен софтуер неоценима помощ биха оказали предварително инвентаризираните и класифицирани типове омонимия, в процеса на пакетна обработка с подходящи интерфейсни средства за избор и решение.

4.2 Степени на кохезия между анализираните думи – текстови хапки (chunks)

Още в системата на контекстните правила, използвани в тагерите (вж. предишния раздел), се създава една сериозна лингвистична обосновка и бъдеща статистическа база за оформянето на един нов тип езикови единици, по-големи от думата. Става дума за формулиране и статистическа проверка на съществуващата текстова кохезия – непосредствено съседство на единици с определени граматични признаци. Такива обединения фактически излизат извън сферата на морфологията и бележат съществуващи синтактични връзки, в които единственото отношение е линейното непосредствено съседство, а членовете на отношението са не думи, а анотационни вериги. За пръв път това явление е разгледано и назовано в [Abney 1991]. Там се говори за обособяване на отделни текстови порции, *chunks* (възможни български преводи - парчета, хапки), което фактически е обособяване на елементарни синтактични единици. Текстовите *хапки* са определени от автора по следния начин: типичната хапка се състои от една значеща дума, обкръжена от съчетания на функционални думи и формирана по фиксиран модел. Пример: [the bald man] [was sitting] [on his suitcase]. Дефиницията и примерът са взети от [Abney 1991].

По-горе видяхме очевидни показатели за синтактични връзки между съседни текстови елементи, изразени само чрез съотношението на техните граматични признаци (*горски дух*, *горски поляни* – прилагателно и съществително, чиито стойности на род и число съвпадат). Тези обединения ни позволяват да видим текста, чийто съставни части на пръв поглед са разположени равномерно линейно, вече като неоднородна субстанция, в която отделни компоненти са слепени помежду си и картинно казано, образуват тези хапки, на които текстът се поглъща, за да бъде разбран. Не случайно тези обединения на думи се оп-

ределят и като прозодични единици, разделени от паузите на произнасянето и съответно на разбирането.

Освен по граматичните си свойства лексикалните единици могат да бъдат обединени в по-едри единици и въз основа на високия процент съвместна поява в непосредствено съседство. Такова обединение имаме при идиомите, където единиците на кохезията са конкретни лексеми (срв. *червената шапчица* vs. *червената армия* vs. *червената химикалка* – последното съчетание е свободно и с по-малка честота на съвместна поява на двата си елемента). При текстовите хапки обединението е по граматични признаци, по тази причина и основните видове кохезия са формулирани за номинална група.

За извличане на тази кохезия на чисто статистическо равнище пръв опит е направен в [Church 1988], където се за английски език се поставят задачи като: стохастично извличане на нерекурсивни номинални фрази, изчисляване на вероятността за поставяне на леви и десни скоби (в синтактичния запис). Изискването за нерекурсивност, т.е. да няма структури от типа A(xA) определя изчислените текстови отрязъци като минимални синтактични единици с проста структура, чието разширяване се допуска на ширина, в броя на елементите й, но не и в дълбочина – в синтактичната йерархия. С други думи, можем да разширяваме съчетанието *сини очи* до веригата – *красиви, разплакани, тъжни сини очи*, но разширената верига – *красиви и разплакани от тъжната вест сини очи* вече не е една текстова хапка.

Високофлективните езици като славянските позволяват тази задача да се решава на морфологично равнище, като се използва твърде силният морфосинтактичен механизъм на съгласуването, проявяващ се чрез определени конфигурации на морфологичните параметри на съседни думи.[39]

Този процес, на слепване (chunking) на текстови единици, разположени в съседство, на линейно равнище, е част от процедурите на т.нар. плитък (или частичен) синтактичен анализ. В него чрез много по-икономичен граматичен ресурс, чрез използване на правила, в която единствената връзка е тази на непосредственото следване, се постигат резултатите на нормалния синтактичен

[39] Интересен паралел, далеч в историята на езиковите технологии, ни дават първите системи за машинен превод, със сериозно лингвистично осигуряване – системата АРИАНА на Центъра по машинен превод в Гренобъл [Vauquois& Boitet 1985], някои руски системи за машинен превод [Леонтьева 1987]. На границата между изследователското и индустриалното приложение, тези системи разполагат с доста дълбок лингвистичен апарат за осъществяване на превода на синтактико-семантично равнище. И от чисто практически съображения, преди да започне истинският синтактичен анализ, се анализират (като верига от граматични признаци) и влизат (като вече оформена цяла единица) синтактични обединения като номинални групи, аналитични глаголни групи и други.

анализ, в който и връзките, и процедурите са далеч по-сложни, тъй като напускат областта на линейното представяне и влизат в областта на структурите – графовидни, дървовидни и т.н.

Методите за слепване могат да са част от сложни синтактични процедури, като ни дават предварително основната база, полуфабриката за изграждане на синтактичното дърво. Такова „каскадно" прилагане на методи на плитък и дълбок синтаксис за български имаме в [Osenova&Simov 2003].

Слепването, по-точно казано откриването на слепени части, може да има и самостоятелна стойност при информационното търсене, което обикновено не стига до синтактичен анализ, но за което е добре да бъдат идентифицирани текстови фрагменти, по-големи от дума, поради информационната им стойност (по-голямата част от специализираните термини например са номинални групи).

Каква е лингвистичната природа на текстовата хапка? Освен че са нерекурсивни, както беше казано по-горе, хапките са нещо като атомарни синтактични единици, именно в равнището на представяне на синтактичната йерархия, колкото и много члена да имат. Терминът хапка, както и английското chunk предполага неделимост, но хапката е структурирана, в нея е обособена опората (в термините на опорно-фразовите граматики – вж. [Осенова, Симов 2007]), или в по-консервативен прочит, главната дума в словосъчетанието.

Да разгледаме един алгоритъм за автоматично откриване на текстовите хапки, извършено върху анотиран текст (с маркерите на СФ-речника). Алгоритмът за откриване на номинални хапки в български анотиран текст моделира начините на композиране – по брой, по подредба, по признаци – на определенията в една номинална група. Думи-детерминанти, които могат да служат за такива определения, са прилагателните, причастията и местоименията, които се скланят като прилагателни и имат съответните адективни характеристики в анотационната верига, също и редните и бройни числителни (примери за детерминанти – *хубав, негов, онзи, пети, две).* Прилагателните могат да бъдат разширени в съчинителна конструкция (*хубави и тъжни очи*). Като определение се третират и притежателните клитики (*му, им*), които са определение на главата, но линейно следват членуваното й определение (*хубавите му книги*). При всички видове хапки се дефинира началният и крайният им компонент. Този тип номинална хапка се разделя на два подвида – определена и неопределена. Определената започва с членуван детерминант, неопределената – с нечленуван. Първият детерминант се идентифицира, разпознава в текстовата

последователност като първата срещната дума със съответна характеристика (пример: винаги ще си спомням *красивите и тъжни нейни очи*, стрелнаха го *две красиви черни очи*). Крайният елемент е съществителното – опорна дума. Като определяме състава на тази номинална хапка като състояща се от детерминант(и) и обект, трябва да отбележим, че елементи с характеристика на началния могат да се повтарят, докато елементът с характеристика на крайния е само един (много определения, но само едно определяемо). Между повтарящите се детерминанти и обекта могат да се вклинят разширения (притежателните клитики – *хубавият му пуловер* и втората част на съчинителни конструкции от прост тип – детерминант+съчинителен съюз+детерминант – *хубавият му и пухкав пуловер)*.

Без да описваме пълната процедура на тази идентификация, ще отбележим само, че за нейното автоматично изпълнение в линеен анотиран текст беше необходимо да се формулират правила за възможна съвместна поява на типове детерминанти, напр. ограничения във вида и броя местоимения, които могат да изпълняват тази функция, във вида и броя на числителните и т.н.[40]

Експерименталният софтуер за откриване на номинални хапки, приложен върху анотиран текст, дава следния изход:

Според{според.PREP} асоциацията{асоциация.N+F:sd} призивите{призив.N+M:pd} за{за.PREP}
по-широка {H=по-широк.A+GR:sf}_данъчна{E=данъчен.A:sf}_схема{T=схема.N+F:s} в{в.PREP} ЕС{} се{се.PC,себе.PRO+RFL:SA} дължат{дължа.V+IPF+T:R3p} на{на.PREP} **високата{H=висок.A+GR:sfd}_цена{T=цена.N+F:s}** на{на.PREP} **отпусканите{H=отпускам.V+IPF+T+NSE:Ppd}_субсидии{T=субсидия.N+F:p}**

където с H (head) и T(tail) са означени началният и крайният елемент на номиналната група, а с E (extension) – разширението, т.е. възможните повторени детерминанти след началния.

В [Osenova&Simov 2003] се изброяват други възможни нерекурсивни синтактични обединения във вид на хапки – с опора глагол, прилагателно, наречие, а също и предложна група. Всички те обаче се идентифицират след определяне на номиналните хапки. Очевидно богатите морфологични механизми за изразяване на синтактични връзки в български, както и в останалите славянски езици, водят до едно друго съотношение между тежестта на лингвистичните правила и вероятностните изчисления в сравнение с езиците, за които е създадена първоначално тази процедура.

[40] http://framenet.icsi.berkeley.edu/

4.3 А без морфологично знание?

Изложените по-горе методи предполагат използването на морфологично знание, прикрепено към текстовите единици, след като е предварително зададено и организирано по различен начин – събрано в речник, формулирано в правила. Очевидно получените резултати пряко зависят от качеството и обема на това знание. Въпросът какво може да получим от текста, ако се лишим от това знание, не е реторичен. Той може да възникне при две различни проекции на нашата ресурсна и методологическа готовност да атакуваме текста с налични компютърни технологии. Тези две проекции може да са разположени във времето, но може и да се отнасят за различни аспекти на поставени задачи и възможности.

Първият е – когато не са още построени лингвистичните складове с лексикален и граматичен материал, а още повече – правилата за боравене с тези материали. Това може да се случи при обработка на съвсем нов език, за който още не са създадени тези инструменти.[41]

Вторият е – когато производството на лингвистични материали и инструменти изостава от действителността на електронните комуникации – във всякаква тяхна форма не поради забавено развитие, а поради спецификата на поставените задачи. Това се случва най-вече по отношение на интернетния езиков поток, където специализирани текстове, новоизковани думи, ненормализиран текст, нови понятия и реалии нахлуват с бързина, многократно превъзхождаща тази, с която се създават съответните инструменти за обработка, ако изобщо могат да бъдат създадени[42]. В първия случай се търсят формални методи и инструменти, за да се преодолее липсата на основни езикови технологични ресурси. Във втория се търсят средства, за да се засили мощта на наличните технологии за задачи, които не са напълно решени или изобщо не могат да

41 Интересен паралел, далеч в историята на езиковите технологии, ни дават първите системи за машинен превод, със сериозно лингвистично осигуряване – системата АРИАНА на Центъра по машинен превод в Гренобъл [Vauquois& Boitet 1985], някои руски системи за машинен превод [Леонтьева 1987]. На границата между изследователското и индустриалното приложение, тези системи разполагат с доста дълбок лингвистичен апарат за осъществяване на превода на синтактико-семантично равнище. И от чисто практически съображения, преди да започне истинският синтактичен анализ, се анализират (като верига от граматични признаци) и влизат (като вече оформена цяла единица) синтактични обединения като номинални групи, аналитични глаголни групи и други.

42 Подобна ситуация в наши дни може да бъде само историческа – като спомен от първи професионални стъпки, по-скоро пионерски, във времената, когато езиковите технологии (тогава наричани другояче) са били екзотично изследователско занимание, демонстриращо само възможностите на изчислителната техника – вж. [Паскалева 1982]

се решат (тогава се гони приближен резултат, който удовлетворява конкретни практически цели).

4.3.1 Морфемният състав на думите и неговата информационна стойност

В изложеното дотук развитие на методите за морфологична обработка на текста може да се проследи известна еволюция в характера на порциите лингвистично знание, участващи в този процес. В началото на тези изследвания морфологичният анализ използва във входното и изходното си представяне лингвистични единици, които точно отразяват връзката между граматичните процеси на формообразуването и морфемната структура на думата – вж. двупосочния морфологичен модел, споменат в 2.11.

В по-нататъшното развитие на компютърното моделиране на морфологията прякото участие на формообразуващи морфеми намалява, за да останат видими най-накрая само началните данни и резултатът от процеса на обработката, по-точно текстовите единици – словоформите, и техният лексикално-граматичен образ, изразен чрез лема и граматични характеристики. Този тип моделиране – само чрез входни и изходни единици, който скрива истинския процес на анализа като в черна кутия, е резултат и от развитието на изчислителните ресурси и технологии, които позволяват складирането и използването на голям обем данни в по-плоско, разгърнато, неикономочино представяне. Така не се плаща данък на ресурсната недостатъчност, която изисква икономия на описанието и обработката – с минимален обем на зададени данни да се анализира максимален брой текстови единици.

С преодоляването на ресурсната недостатъчност компютърната морфология започва да се свежда до отношение между думи и техните характеристики, а равнището на нейното компютърно моделиране – до формообразуващите процеси, разгледани в резултата си – словоформите и граматичните им свойства.

Морфологията, обаче, като наука за думата, включва в своя предмет и изучаването на механизмите на образуване на нови думи с по-различен, макар и близък смисъл, т.е. деривационните механизми, в славянските езици твърде добре развити и изразявани чрез богата система от деривационни елементи, афикси – префикси и суфикси. В резултат от употребата на тези елементи се формира т.нар. словообразувателно гнездо на една дума, съдържащо всички думи, определяни като нейни *сродни*, *родствени*, *производни*, *деривати.*

В какъв момент и под каква форма езиковите технологии почват да се интересуват и да моделират словообразователните механизми и материалните носители на това образуване – деривационните морфеми?

Отговорът е - в момента, когато компютърното езиково моделиране започва да се интересува от значението, от смисъла на езиковите единици в малко по-друг приложен аспект. Този нов аспект е различен от моделирането на смисъла на текста като фрагмент от един пълен анализ, който предполага движение нагоре (или надолу – зависи от гледната точка) по стълбицата на езиковите равнища – от текст към морфология, синтаксис, семантика (или части от нея).

Тъй като опитът показва, че пълно моделиране на смисъла на езиковия израз не е възможно (и причините за това не са технологични), то задачите за установяване на смислова връзка между различни текстови фрагменти, от думи до цели документи (съдържащи хиляди думи) започват да се решават насочено фрагментарно,чрез установяване на смислова връзка между отделни порции информация в документа, или за ограничен брой смислови отношения.

Това фрагментарно установяване на смислови отношения между текстови съвкупности е единственият начин да се установи съдържателна връзка между документи. Това е и основната операция в информационните текстови технологии – във всичките видове търсене, реализирано в компютърни системи – за търсене на информация (Information Retrieval), за извличане на информация (Information Extraction) от документи както на един език, така и на много (cross-lingual). Формите и моделите на тази информационна обработка са различни, но в крайна сметка се свеждат до сравняване или изследване на *съдържанието* на документ (или документи) чрез налагане на някакви структури на смисъла върху неговия текст. Структурите на смисъла – модела за търсене, може да са зададени като последователност от ключови думи, може да представляват завършен смислов израз, изразен в кратък или дълъг въпрос, може да са структурирани смислови единици (концептуални графи, онтологии) и др.

Търсенето на съдържателно съответствие между въпроса и документния масив, чиито основни единици са думите, е толкова по-резултатно, колкото по-компактна и структурирана е съвкупността от думи в документа. Един вид такова подреждане – резултат от морфологичен анализ, е групирането на словоформите на една лема – формално различни, но носители на едно и също значение, което се осъществява с инструментите на морфологичния анализ. Друг вид подреждане е групирането на словоформите (по-точно, на техните представители – лемите) по родствени връзки. За информационното търсене

е добре, ако в текста думите – *вкус, вкуса, вкусът, вкусове, вкусовете* са обединени в група, но също така, ако не и по-полезно за едно дори фрагментарно представяне на смисъла на документа, е ако са обединени в група и формите на други думи – *вкус, вкусен, вкусно, вкусотия*. Това преминаване към друго равнище на морфологичните зависимости, равнището на деривационната морфология, на групите от различни думи предполага други лингвистични инструменти, а както е известно, деривационната морфология е доста трудна за формализация.

В деривационната морфология понятието парадигма е трудно да бъде формулирано, макар че обединение на продуктите на словообразуването съществува – говори се за словообразователни гнезда – съвкупността от думите, образувани от една обща дума, като прочитът на това *образуване* е твърде различен. Образувана ли е *пролет* от *лято*, каква е степента на родство между *чета* и *прочета* от една страна, и между *бера* и *пробера*, от друга и т.н. Тези примери за размисъл са български, биха могли да се дадат за всеки славянски език, но първите опити да се обединяват думи, разглеждани само като буквени вериги, в комплекси, излизащи извън простата представа за граматичната форма, са осъществявани за английски език.[43]

Не само слабите морфологични връзки, изразени формално в английската дума (откъдето идва и разликата в обема на съответните теоретични разработки и модели на морфологията и синтаксиса) карат изследователите, заети с компютърното моделиране на езикови процеси и производство на съответните технологии, да се обърнат към други методи, които не следват директно логиката на лингвистичното описание и изследване.

Развитието на информационните технологии в техния езиков сегмент не е свързано само с растящата мощ на изчислителните ресурси и новите компютърни конфигурации. Освен средството, променя се и обектът на изследване и обработка – нов не само в количествено измерение, превишаващо хиляди пъти обемите на старите текстови ресурси, често пъти подбрани или създавани специално за целта. Новите ресурси се натрупват динамично от реални източници, разпространяват се най-вече по интернет и са далеч от каноничната нор-

[43] Създателите на първите, дълго време единствени, а и досега преобладаващи основни продукти на компютърната лингвистика са англоговорящи. Оттам и известна непоследователност във формулировките, свързани с равнищата на морфологическото описание като напр. липса на критерии за разграничаване на формообразуващите от словообразувателните процеси – *driver* се определя като форма на *drive*, заедно с *driving* и др. подобни. Изглежда хипотезата на Сепир-Уорф за членението на света в зависимост от родния език е валидна и за граматиката на същия, отразена в структурата на приложно лингвистичните продукти, създавани от носителите на този език.

мативност на текста[44]. Това отклонение се проявява в наличието на много нови думи, както специализирани, така и новоизковани или просто неправилни, на специални видове организация на текста, напр. административни документи и отразява онези разрези на езика, които могат да намерят място след години в нормативните граматики – и то евентуално.

Новият материал на езиковите технологии извиква и нови методи и техники на обработка, в които организацията на езиковите факти не може да следва стройната подредба на лингвистичните равнища в теориите за езика. Както е известно, такива обекти, огромни по обем и неструктурирани, се изследват обикновено със статистически методи[45]. Така се процедира и в случая, когато не можем да спазим принципа на еднаквия обем на отрязъка от тортата на знанието – тесен и дълбок, или плитък и широк. Това, което изследват и което получават статистическите методи, е по-скоро едно неравномерно отрязано парче торта – доста широко и проникващо надълбоко само в отделни части. Тази фрагментарна дълбочина обаче е достатъчна, за да разберем вкуса на тортата, т.е. да извлечем информацията, която ни трябва.

Ако класираме статистическите методи според порциите лингвистично знание, които използват, забелязваме, че те могат да извършват пресмятанията и изводите си само върху буквите и техните линейни съчетания, могат да използват и фрагменти от организирано лингвистично знание (специално конструирани правила за съответствие между същите тези единици), могат да използват и информация от сериозни граматични ресурси като речници. Много характерна тяхна черта е елементът на самообучение, заложен в алгоритъма на изследването и архитектурата на системата. Такъв елемент може да бъде зададен предварително – чрез т.нар. учебен корпус (training corpus) или да бъде осъществяван чрез итерация на различни методи в последователността на операциите. Когато предварително зададено лингвистично знание липсва, говорим за неконтролирано, ненаправлявано самообучение, unsupervised training/learning, вж.напр [Gaussier 1999][46].

Каквито и комбинации от контролирано или неконтролирано, предварително зададено или изцяло липсващо лингвистично знание да се използват в разглежданите методи, процесите, които те моделират и извършват, се свеждат до

44 Съществува област в компютърното моделиране на езика, изследваща т.нар. грешен вход – *ill formed* ***input***.

45 Едно от най-добрите изложения на съвременните статистически методи за компютърна обработка на текст е [Manning&Schütze 1999].

46 Тук си струва да дадем пълното заглавие на този доклад - Unsupervised learning of derivational morphology from inflectional lexicons, като обърнем внимание на френския произход на автора и точно определеното равнище на морфемното членение в изследователския му продукт.

основните операции на морфологичния анализ, а именно:

- идентификация на морфологични компоненти, определяне на техните граници в състава на думата, т.е. морфемна сегментация,
- определяне на граматични свойства на думите,
- обединяване на групи думи според граматичните им свойства.

От гледна точка на дълбочината и точността на лингвистичното знание нито една от тези операции не може да достигне точността на аналогичните процедури, извършвани с помощта на граматични ресурси . Достигнатите резултати обаче подобряват резултатите на информационното търсене (вж. по-горе) и са интересни от типологична гледна точка – както за изследване на дадена морфологична система, така и за съпоставителни изследвания. Последните са улеснени от факта, че повечето от тези методи, несвързани с конкретно граматично знание, са езиково независими.

4.3.2 Определяне на морфемните граници

Тук ще бъде интересно да проследим решаването на тази задача, поставена за българския език и разрешавана само експериментално, с компютърни средства от ерата преди персоналните компютри. Нейната формулировка звучеше така: при липса на инструменти за анализ (лексикална база данни и граматика), каква граматична информация мое да се извлече от български текст, като се използва само морфологично, а не лексикално знание?

В противопоставянето на двата типа знание се има предвид, че обемът на лексикалните елементи на един език е стотици пъти по-голям от този на граматичните, ако за представители на двете множества приемаме основите/лемите и граматичните формативи. Последните, освен всичко друго, са и затворено множество – нови думи възникват непрестанно, но нови окончания – не.

Под *граматичен формативе* в задачата се разбираше комплекс от морфологични елементи, който оформя *изменяемата* част на словоформата и я прави уникален член на парадигмата. А отрязването на този комплекс от края на думата ни дава общата *основа* на словоформите в парадигмата. С изследвания върху състава на българската парадигма беше формиран ръчно списък от 104 *квазиокончания*. Те обединяваха в едно няколко вида формативи – например за причастията съдържаха в себе си някои словообразователни суфикси на причастията, тематичната гласна, истинското окончание и постпозитивния член – *алата*, *ащият*.

Всеки елемент от този списък на морфокомплекси бе наложен (отново ръчно!) върху всички словоформи, умозрително образувани от 70 000-те лексеми на Обратния речник на български език (които, според мерките на съвременните български лексикални бази генерират около 1 милион словоформи – един съвременен обем на граматичен речник). В резултат на това налагане се оказа, че голям брой словоформи формално съдържат изследвания морфокомплекс в края си, но буквената верига, получена след неговото сегментиране, не е това, което очакваме да получим, като имаме предвид граматичната стойност на морфокомплекса. Напр. при отрязване на *алата*, като информация за мин. действ. прич. ж.р. ед.ч. членувано, можем да получим *център-алата*, *з-алата*, *с-алата*, *нач-алата* – сегментация, твърде далеч от правилната и като членимост, и като граматична осмисленост на получения остатък. За всеки елемент от списъка, след налагането му върху словоформи, съдържащи го в крайния си низ, следователно се обособяват две групи словоформи – тези на правилната му сегментация, и тези на неправилната.

Съображения за икономичност на алгоритъма разделиха списъка на две части, на два типа морфокомплекса, в зависимост от съотношението между обема на правилно и грешно сегментираните словоформи. Докато за дадения по-горе пример за окончанието на минало причастие (*-алата*), валидно за голяма група глаголи, списъкът на грешните сегментации е много по-малък от този на правилните, за окончания като *-еса*, *-есата*, формиращи множественото число на малка група съществителни от ср.р.(*чудо*, *дърво*, *небе*) грешните сегментации са повече (съотношение 156:3 за *-еса* и 19:3 за *-есата* – вж. *зав-есата*, *адр-есата*, *м-есата*).

Ако към всеки елемент от списъка прикачим по-малкото подмножество от опозицията *вярно/грешно сегментиране* и му припишем един от двата типа, то за решаване на задачата *автоматично сегментиране на български текст без помощта на лексемен речник* се нуждаем от 2 списъка – този на морфокомплексите, с посочен тип, и този на съдържащите ги думи (оптимизиран според съотношението *вярно/грешно*). Тази организация на данните доведе до резултата, че 108 морфологични комплекса, разделени в 2 групи и снабдени със списъци от примери (чиято оптимизация ги сведе до 7 000) осигуряват правилната сегментация на основа и окончание за 1 000 000 български словоформи.

Практическата стойност на подобен експеримент се състои в предоставената от него възможност за намаляване на броя на лексикалните единици в текста и до възможното им обединяване по съвпадаща *начална част – основа*. Представянето на текста във вид на сегментирани (на основа и окончание) текстови

единици се стреми към една груба първоначална обработка по пътя на прехода от текста към неговото морфологично представяне. То би могло да бъде база за създаване на други граматични ресурси. Така поставената цел не можа да бъде осъществена именно по този начин, може би поради наличната изчислителна техника и доста недружелюбните възможности на използвания програмен интерфейс. Беше реализиран алгоритъм за сегментация на произволен български текст на машина СМ-4 (прототип на PDP 11-40) – вж. [Паскалева 1981], [Димитрова 1981]. Трябваше да бъде дочакана ерата на персоналните компютри (Изот и Правец), на добре проектирания потребителски интерфейс и възможностите за визуализация на ПРОЛОГ обработката, за да се пристъпи към същата задача вече с нови средства и възможности.

От гледна точка на формализацията на лингвистичното знание, за бъдещи изследвания и приложения бе направен експеримент и със сегментация на префиксалната част на словоформите по предложената методика. Методът на организация на данните бе малко по-сложен, тъй като при префиксалните образувания имаме итерация и отделянето на суфиксите трябваше да се осъществява на последователни стъпки. Подобно на алгоритъма за сегментиране на суфиксалния морфокомплекс бе обособен списък от 13 префиксални единични елемента, а съответните списъци от примери – от двата типа (*вярно/грешно сегментирано*) съдържаха 1300 единици. По разбираеми причини те са по-малко от примерите за окончанията, тъй като при последните неправилната сегментация освен на ниво лема се увеличава и от образуването на словоформи (окончанието *–ата* е грешно сегментирано само за кратката членна форма на лемата *квадрат – квадр-ата*), докато при префиксалните елементи сегментацията е еднаква за цялата парадигма. (*пред-* се сегментира грешно в цялата парадигма на глагола *пред-а*, а не в отделни негови форми).

Изходът от тази процедура като лингвистичен продукт е само посочване на правилната членимост на думата на лексикален и граматичен елемент, без по-нататъшни изводи за граматични стойности. От сегментираните окончания само 10 процента са неомонимични – т.е. могат да ни посочат и частта на реч-та, като напр. *охте, ахте.*

Горните упражнения върху формообразуващите и префиксални елементи са едно следствие от желанието да бъде използвана очевидно високата информативност на морфемната структура на българската дума за целите на автоматичната обработка. Защото ако *-я* е силно омонимично окончание, с около 8 значения, предшестващата го буква в думата започва да намалява тази омонимия. Примери за размисъл:

- от 546 словоформи на *-теля* само 2 % са глаголи – производните на *-теля* и *стеля* плюс *спрятеля,* останалите са определени не само по част на речта и род, но и по граматичен признак – същ. м.р. ед.ч. кратък член – *учителя*;
- от 724 словоформи на *-ция* 1 % са същ. м.р. членувано – *силиция*, останалите са еднозначно определени като същ.ж.р.ед.ч. нечленувано).

Това изследване, направено в зората на автоматичната обработка на текст у нас, е по-скоро една алгоритмична проверка на високата информативност на българската морфема – в различен състав и позиция в думата. Модерната проверка на информативността на българската морфема предполага качествено други условия на задачата, удовлетворяващи трите аспекта на компютърната лингвистика – формален метод, компютърна обработка и формализирано лингвистично знание. Това бе сторено 25 години след този пръв опит, в друга среда и за други цели.

Същевременно за английския език, като основен език в компютърните езикови приложения, поне в началото, опитите за автоматично извличане на морфологично знание са започнали едновременно с описаните по-горе опити за българския език на изчислителни машини почти от същото поколение – IBM 370, в Кембридж, където чрез алгоритъм за сегментация 10 000 английски думи са сведени до 6 370) – вж. [Porter 1980].

4.3.3 Стеминг и стемъри

Желанието да се обособи общата значеща част в групата думи със сходен буквен състав, които имат отношение към смислово близки понятия, води до конструирането на специални програми за автоматична идентификация на тази обща част, която се нарича *стема* – stem, а съответните програми – *стемъри* (stemmers). С известна неохота ще приемем този термин в бъдещото изложение. Лингвистичното съдържание на тази обща част на думата в основни линии покрива това, което наричаме основа на думата – общата неизменяема част в нейната парадигма при формообразуването, но покрива и термина произвеждаща основа в словообразуването. Тук лингвистичната пунктуалност трябва да отстъпи място на информационната потребност и към тази концепция се налага да се придържат и авторите на подобни продукти за славянски езици, чиято лингвистична наука много добре различава формите на една дума от образуваните от нея нови думи.

Първите стемъри датират от 80-те години на миналия век и са построени върху чисто детерминистичен подход – прилагат се правила от типа:

-ies –> -y (с изключение на случаите, когато имаме *-eies* или *–aies*)

-s –> (с изключение на случаите, когато имаме *ss* или *us*)

т.е. контекстно свързани (в отделни случаи) трансформации на крайни вериги на думата, в това число и в нулева верига (втория пример).

Примерите са взети от стемъра на Портър [Porter 1980], който използва алгоритмичния принцип, наричан suffix-stripping и е експериментиран върху системи за търсене на информация в документи по следния начин.

За всеки документ се съставя една представителна част, съдържаща неговото заглавие и текста на резюмето му, се сегментират окончанията и се получава редица от окастрени думи. Последните се налагат върху целия текст на документа и получените чрез съвпадане с тях текстови фрагменти ни дават информационния образ на целия документ. От другата страна на търсещия механизъм – заявката за търсене, се извършва същото налагане на окастрените думи от представителната част върху текста на въпроса. Търсенето се извършва чрез налагане на окастрения по същия начин въпрос върху информационните образи на документите.

Освен правилата за отрязване на определени буквени съчетания от края на думата, явно ориентирани към конкретен език, в следващите разработки на стемъри се включват по-сериозни порции от лингвистично знание – таблици с пряко съответствие между словоформи и леми, моделиращи и в известна степен достигащи почти до конкретното съотношение словоформа-лема. В крайния вариант на тази лингвистична пълнота на процедурата се появява граматичният речник, задаващ директно горното съотношение. Но задачата не може да бъде решена само с помощта на този ресурс, защото в такъв случай ще се сведе просто до автоматична лематизация. Стемата, словообразуващата основа, понякога може да е равна на лемата, но в доста случаи е по-малка от нея и представя по-голям брой лексикални единици (вж. съотношенията):

- *стол – столове*, *столът*, *стола* от една страна и *стол – столар*, *столче*, *столова* от друга, където имаме равенство между стема и лема (*стол*),
- *пазарски – пазарска*, *пазарската*, *пазарските* и *пазар – пазаря се*, *пазарувам*, *пазарен*, където стемата е част от лемата *пазарски,* но съвпада с лемата *пазар*.

Друг тип организиране на сегментацията имаме в един от най-ранните стемъри – [Hafer & Weiss 74]. Главната отлика е в посоката на сегментацията – отляво надясно, от началото на думата.

Методологическите принципи на този вид автоматично определяне на морфемния състав на думата, както се признава от авторите му, водят началото си от постулатите на американския дескриптивизъм в морфологията, развити в трудовете на З. Харис [Harris 1955]. Имат се предвид постановките за идентификация на морфологични единици без обръщане към значението им, а само чрез анализ на тяхната дистрибуция, съвместна поява и честота – принципи, изключително подходящи за компютърно прилагане върху голям текстов корпус. Именно така се извеждат и резултатите за автоматична сегментация на морфеми.

Методът се свежда до преброяване на множеството от наследници (successors variety) – това, което остава след итеративната сегментация на буквени низове от началото на думата. Сегментацията спира на мястото, в което броят различни букви в остатъка след отрязването на различни кандидат-морфеми достигне максимума.Този максимум се допълва (или корегира) статистически, чрез процентното съотношение на броя думи към броя различни букви, които оформят началото им.

В експериментите за английски език изчислението се прави върху текстов корпус. За български език то е направено върху всички словоформи от СФ-речника. Пример за прилагането на този сегментационен принцип ни дават данните за последователното изчисление на първия сегментиран префикс *пре* в словоформите, започващи с буквения низ *превключ** (на брой 61).

Задачата се свежда до сравняване на следните числа:

1. брой словоформи с *п* в начална позиция – 310 000;
2. брой на различните букви в начална позиция след сегментацията на *п* – 15;
3. брой словоформи с *пр* в начална позиция – 89 000;
4. брой на различните букви в начална позиция след сегментацията на *пр* – 7;
5. брой словоформи с *пре* в начална позиция – 41 700;
6. брой на различните букви в начална позиция след сегментацията на *пре* – 21;
7. брой словоформи с *прев* в начална позиция – 2200;
8. брой на различните букви в начална позиция след сегментацията на прев – 11;

9. брой на словоформи с *превк* в начална позиция – 61;
10. брой на различните букви в начална позиция след сегментацията на *превк* – 1;
11. брой на словоформи с *превкл* в начална позиция – 61.

Следващите стойности на словоформи и различни начални букви са същите, като в 10. и 11, до изчерпване на зададения низ – *превключ*, вж. Табл. 21.

Начален стринг	**п**	**пр**	**пре**	**прев**	**превкл**	**превклю**	**превключ**
Брой различни начални букви в остатъка	15	7	**21**	11	1	1	1
Брой думи, съдържащи началния низ	310000	89000	41700	2200	61	61	61

Табл. 21. Последователно отрязване на буквени низове, брой начални букви в остатъка и брой думи, които съдържат сегментирания низ.

Методът на изчисление на остатъците (successor variety) е известен още като peak and plateau method – т.е. методът на върхове и равнини. Както се вижда от таблицата, върхът, т.е. максимумът на стойностите за броя различни букви в остатъка, бележи върха – 21 букви за сегментацията пре-включ, което определя и избора на *пре-* като членима морфема префикс.

Една проверка на метода за определяне на границата между основата и суфикса в думата *продължавам* ни дава следните стойности в Табл. 22, където сегментацията започва от предполагаемия морфемен шев.

Начален стринг	**продълж**	**продължа**	**продължав**	**продължава**	**продължавам**
Брой различни начални букви в остатъка	3	5	1	**9**	1
Брой думи, съдържащи остатъка	140	64	54	54	2

Табл. 22. Изчисление на морфемния шев в словоформите *продължава-м*, *продължава-ме*.

Авторите на този метод прибягват до редица настройки – статистически, извършвани върху текстов корпус или речник, и специфично езикови (например забрана за разделяне на дву- и трибуквени низове, означаващи един звук – като френското *ch*, немското *sch* и други), като признават, че самостоятелното използване на метода за пълно морфемно членение води до доста шум в сегментацията. Явно принципът за минималния брой наследници (в словоформи

и различни букви в остатъка) трябва да се използва като корегиращ, но не като основен фактор.

Така се проектират и съвременните стемъри – с алгоритми, които обединяват в действието си генерални правила на буквено равнище, граматични ресурси – списъци от морфеми, граматични речници и др. Най-същественото методологично допълнение в тази процедура през последните години са наблюденията върху голям текстов корпус, който служи и за формулиране, и за проверка на хипотези, с главен източник – интернет.

Стемър за български език е построен и докладван през 2003 г. от Преслав Наков, вж. [Nakov 2003]. Той използва разглеждания тук СФ-речник, за да получи в първата стъпка списък от кандидат-окончания, произведени чрез налагане на лема върху словоформа (и двете зададени в речника). При пряко влагане на първата във втората се формира правило за отрязване на остатъка, което включва и диагностичния ляв контекст от три знака преди мястото на сегментацията (кандидат-морфемния шев).

Стемърът е трениран допълнително върху български текстови корпуси.

Пример за сегментиран текст от същия стемър:

- освен това **службат** на дел понте **започн** и **процес** за **препращан** на над 60 дела без **формулира обвин** на **национал юрисдикци** на бих, **хърват**, македони и сърбия и Черна гора.
- като **израз увереност**, че **„е създа** база за **провеждан** на **справедлив съдеб** процес“ из целия регион, дел понте **предупре**, че все пак **национал съдеб производств** трябва да бъдат **наблюдава**.

Един поглед върху горния резултат ще ни убеди в различната оценка на надеждността на тази процедура – от лингвистична и от информационна гледна точка.

Макар че стемърът използва голям граматичен речник, в него не са посочени границите на лексикалното и граматичното в словоформата. По тази причина виждаме пример за недостатъчна сегментация – *службат-* за *службат-а*. Виждаме и обратния пример за прекалена сегментация – *национал-* за *национални*. Ако в първия случай тази недостатъчност ще попречи да се идентифицират други членове на парадигмата на *служб*-а, то във втория случай прекалената сегментация ще позволи като идентични на *национални* да се определят и думи като *национализъм*, *националистически* и подобни, което вече отива към моделиране на словообразователна парадигма и повишава точността на

смисловото информационно търсене.

Същото приближаване към деривационни зависимости виждаме и при *обвин-, създа-, предупре-.* Точно в тези случаи лингвистичната коректност противоречи на информационната полезност (нещо, с което един лингвист, предоставил своето знание за статистически изследвания върху езика, трябва да се примири).

Същият стемър е използван в междуезикова процедура за информационно групиране (clusterization) на голям корпус от англо-български паралелни текстове [Rayner et al 2007]. Процедурата изисква групиране и категоризация на паралелни текстове след стеминг, който за двата езика бе осъществен от един от многобройните публично достъпни стемъри за английски и въпросният стемър за български език [Nakov 2003].

4.3.4 Отгатване на морфологични признаци

Явно само сегментацията, макар и доста полезна за информационното търсене, не е достатъчна за пълно извличане на лингвистичното знание, дори и за методи като статистическите, които не разчитат на пълната ресурсна осигуреност на модела с лексикален и граматичен материал.

Още при сегментацията на окончанията и формирането на стемите за високофлективни езици се вижда, че някои от сегментираните афикси – чисти окончания и деривационни морфокомплекси могат да ни дадат повече информация от простата идентификация на морфемния шев, както и да се разбира той. Не случайно класическият пример за отгатване на лексикални и граматични свойства много далеч преди компютърната лингвистика да се сети за това, е даден от руския академик Л.В.Щерба *Глóкая кýздра штéко будланýла бóкра и курдя́чит бокрёнка* с очевидния прочит за събитие, в което някакво животно извършва по някакъв начин действие с невръстно животно от същия вид. Би било интересно да се проследи хронологически възникването на конверсивната идея у Чомски, с неговото *coulourless idea sleep furiously*, където безсмислицата и правилното оформяне са разпределени между други лингвистични равнища.

В компютърната морфология като инструмент за информационно търсене през последните години има много *отгатващи* процедури, базирани на високата информативност на края на думата. Почти всеки тагер, снабден с лексикални правила (вж. 4.1.3 и дадения там пример за такова правило) използва зависимости между крайния буквен низ на думата и нейните граматични свойства.

Изследването на граматичната натовареност на крайния низ в словоформата за български и други славянски езици не е нищо друго, освен практическо изследване на статистическата стойност на твърдения от нормативните граматики, обвързващи значение и краен буквен низ, като например: *абстрактните съществителни, завършващи на -ство*; *отглаголните съществителни на -ние; имената, означаващи деятел, завършващи на -тел и -ар* и други подобни.

Такова изследване е направено върху граматичния СФ-речник, съдържащ близо 1 140 000 словоформи и 74 000 леми, снабдени с лексикални и граматични характеристики, би могъл да помогне съществено в съставянето на подобни отгатващи правила за български език, тъй като предоставя фактически пълната лексика на езика. Върху този материал с един прост алгоритъм за итеративна сегментация на крайните низове и количествени наблюдения върху граматичните свойства на сегментираната единица, както и близостта й с други речникови единици могат да се изведат редица зависимости между буквения състав на думите и граматичните им свойства.

Това се извършва с последователно налагане на буквени n-торки до определен праг върху всяка речникова единица. Без предварително изследване е ясно, че в това последователно отрязване ще се прояви известният принцип за обратна пропорция между дължината на езиковия елемент и неговата омонимичност (или многозначност) – по-кратките единици притежават повече значения. Така в една пълна комбинаторика на отрязаните от края на словоформата би-, три-, четири- и прочие n-грами бихме намерили n-торката, която еднозначно определя граматичните свойства на съдържащата я единица. Съвсем нормално е тази итеративна сегментация да стигне до интервала, предшестващ думата, т.е. окончанието да е изчерпило думата, без омонимията да е решена, което е и съдбата на омонимичните омонимичните думи в речника. Такова изчерпване без краен резултат ще имаме, ако чрез последователна сегментация от края на словоформата *поканите* сегментираме низовете *-е*, *-те*, *-ите*, *-ните*, *-аните*, *-каните*, *-оканите*, *поканите*. Табл 23, съдържаща омонимичните значения и съответния брой думи при последователната сегментация на елементите на петорката *-ицата* илюстрира диагностичността на различните крайни низове и тяхната омонимичност.

Окон-чание	Думи	Брой V	%	Брой A	%	Брой N	%	Брой NU	%	Брой ADV	%
а	131979	87533	66	23681	18	20475	16	89	6,7	62	4,7
та	50551	27443	54	12347	25	10372	20	59	0,1	24	0,05
ата	43522	27309	63	11927	27	4213	10	45	0,1	23	0,05
цата	424			3	0,7	419	99	2	0,3		
ицата	383					382	99,7	1	0,03		

Табл. 23. Последователна сегментация на съставките на морфокомплекс *-ицата*, омонимичност на съдържащите ги в края си думи и процентно съотношение на дадената граматична стойност към общия брой думи с това окончание.

Данните от подобни изследвания върху един пълен речник не са представителни за реалната диагностична сила и омонимията на крайните буквени низове, тъй като отразяват само потенцията, но не и реализацията на подобно *отгатване*. Освен като компонент в лексикалните правила на тагера, вж. [Cutting et al. 1992] и [Brill 1995], такава процедура има голямо значение за анализа на нови думи, не фигуриращи в речника – термини, названия и новоизковани думи от различен произход. Реалната отгатваща сила на морфемните елементи се проверява в реален голям текстов корпус, тъй като такава е и главната област на приложението на отгатващата процедура.

Комбиниране на методите на отгатване въз основа на информация от граматичен речник, статистически данни за честотата на неговите единици в реален текстов корпус, извличане на правила за реализация на механизмите на префиксация, суфиксация и оформяне чрез флексия имаме в [Mikheev 1997]. Една от базовите проверки за успешност на сегментацията е проверката дали полученият остатък след отрязването на изследвания буквен низ ни дава дума от речника. За по-голяма точност на тази проверка се въвежда и понятието *променлив суфикс*, чрез което се дефинират стъпките, които трябва да се изминат от словоформата в текста до речниковата единица (например когато отрязването на *-ied* от *specified* няма да ни доведе до речниковата единица *specify*, ако не дефинираме замяната *ied – y*).

В един експеримент върху СФ-речника, описан в [Nakov &Paskaleva 2004], илюстрираният в табл. 23 метод на пълно изчерпване на съотношението крайна буквена верига –> граматично значение е комбиниран с част от методиката на [Mikheev 1997].

Тази комбинация от двата метода плюс пълната информация от граматичния речник е приложена с процес на самообучение върху речника и върху извадка

от български текстове, като при това се постига висока степен – над 99 % отгатване на следните граматични характеристики – част на речта, род и някои граматични характеристики, които са семантично напълнени – определеността, изразена в членуването, и одушевеност/неодушевеност, взета от лексикалните характеристики на съществителното.

Честотата на отгатване е висока, правилата, свързани с изчерпателно изброяване на снемащи омонимията буквени конфигурации в края на думата, са към 1 500 на брой.

Подобни изследвания очевидно имат самостоятелна стойност като измерване на информативността на морфемните елементи на даден език (още по-интересно би било подобно проучване в съпоставително типологичен план, с прилагане на еднакви методи за различни езици, още повече, че статистическите процедури са езиково независими). Като практическо приложение, при наличие на достатъчно пълен граматичен речник и тагираща процедура те не биха могли да поемат основната тежест на определянето на граматическите свойства на думите – главна цел на морфологичния анализ. Затова основната област на използването им е новата лексика, думи извън речника, термини, неканонични и специализирани текстове (съставящи голяма част от интернетното текстово богатство).

Именно за такъв вид приложение бяха използвани статистическите методи за отгатване на граматичните свойства на българската дума с обучение върху WEB текстове. Целта на експеримента бе да изпробва тези правила за конкретна лингвистична задача – идентификация на текстови хапки, chunks (вж. 4.2), в два различни вида текстове с различно разпределение на между двата вида лексика – наличната в речника и липсващата в него, всеки един с обем от по 1000 думи. Първият текстов корпус се състоеше от медицински текстове (болнични епикризи, изобилстващи със специализирани термини), а вторият – юридически, с почти пълно покритие на лексиката от речника. Те съдържаха автоматично извлечени, с помощта на лингвистични правила на съчетаемост и обединение на словоформите според граматическите им характеристики, съответно 118 и 112 номинални текстови хапки. В първия текст 58% от номиналните хапки съдържаха словоформа с празни граматични характеристики (липсваща в речника), а във втория имаше пълно покритие на граматичната идентификация в речника и в текста.

Предложените отгатващи правила запълниха правилно над 99% от липсващите характеристики в номиналната група, само въз основа на анализ на техните

крайни буквени низове[47]. Експериментът е описан в [Nakov&Paskaleva 2005].

4.3.4 Български стемен речник

И различните видове отгатване, и стемърите, и хитроумните методи за: сегментация, проверка и отново сегментация, с или без речник, тествани и тренирани върху учебен текст или върху огромни текстове с интернетен произход, ни навеждат на следната мисъл. При толкова експерименти да се очертае приблизително или да се отгатне значещата част на българската дума (каквото и да разбираме под това), не може ли да се направи един речник на българските стеми? Той би могъл да осъществява сегментация на текстовите единици отляво надясно, без да се интересува от остатъците (деривационни и формообразуващи елементи), тъй като първоначалното смислово групиране на единиците в текста може да се извърши само чрез основните им лексикални им значения, изразявани главно от корена или основата на думата. Границите на този изразител на лексикалното значение в състава на думата ще зависят от концепцията за стемния речник – от какви морфеми се състоят неговите единици. Такъв речник би оставил за статистическите методи една допълваща и коригираща роля, като извършва основната част от сегментацията. Нима посочването на родството между думите, базирано на такъв речник, ще бъде по-малко вярно, отколкото експериментите, изчисляващи деривационната близост на буквено равнище?

Въпросът явно е реторичен, а евентуалният отговор, че задачата е много трудна, може да дойде само от теоретични терзания – но що е това *стема* и как ще я определим?

Наистина, определянето на стемата като:

- нещо, по-голямо от корена и по-малко или равно на основата на думата,
- нещо, което като буквен състав е най-близко по смисъл до термина *производяща* основа, но не винаги съвпада с нея поради алгоритмични трудности формално да се опише словопроизводственият процес,
- нещо, което като смисъл се доближава до основното значение на корена на думата, но в много случаи обхваща и сраснали с него префикси,

[47] Получените грешки се дължаха на недостатъци в самия механизъм на откриването на хапките, поради омонимия на съставките им, невъзможна за разрешаване от плитък синтактичен анализ. Вж. например:
...е установена кръвна захар, където страдателното причастие, реално – част от пасивна глаголна конструкция, е прикрепено към номиналната група, в състава на хапката установена кръвна захар. За нуждите на информационното търсене обаче двата вида групиране: [е установена] [кръвна захар] и [е] [установена кръвна захар] ще доведат до близък резултат.

може би не се подава на дефиниция в термините на назоваваните в българската морфология съставки на думата. Тогава би трябвало да приемем за стемата чисто функционална дефиниция – само по отношение на информационната ѝ стойност, обединяваща близки по буквен и смислов състав единици.

Възможно е едно оперативно определение на групата думи, произлизащи от една стема – еднокоренни и семантично близки, а степента на тази близост може да се определи от конкретната група задачи, за които ще бъде ползван речникът. Тук се сблъскваме с очевидната истина, че съвременната членимост е различна от историческата и че членимост не значи производност, но трудно бихме могли да отговорим на въпроса – каква производност гоним? Насока за търсене на отговор биха могли да ни дадат лексикографски пособия, които съдържат единици – съставки на думата и идентификацията и организацията им в по-сложни единици, но такива речници има много по-малко, отколкото речници на думите, единственият подобен речник с надеждна структура на деривационната организация за български език е [Словообразувателен речник 1999].

Така че докато чакаме българската лексикография да запълни тази ниша с повече морфемни и словообразователни речници, бихме могли, със споменатите вече алгоритмични методи – сегментация, проверка и пак сегментация, отляво надясно и отдясно наляво, да произведем един български стемен речник първо приближение.

Ясно е, че системите за търсене и извличане на информация, за въпрос-отговор и за категоризация на документи биха подобрили резултатите си, ако ползват такъв речник.

Такива бяха и съображенията ни, когато започнахме серия от експерименти върху лексикалната част на СФ-речника, т.е. върху списъка от неговите леми, на брой 74 000 (представящи 1 140 000 словоформи). Самата постановка на задачата – формирането на българската стема, общата част на думите, предполага да използваме достъпния ни вече преход от дума към лема в речника, тъй като сегментираме отляво надясно и не напускаме областта на лексикалното, тъй като ни интересува само този вид близост. Граматическите вариации, изразени чрез морфемите, разположени в десния край на думата, могат да служат като фина настройка на това търсене на близост в следващ етап след използването на стемния речник.

Простото обръщане на речниковите леми, т.е. производството на обратен речник веднага посочва тези буквени вериги, които, както се казва, плачат да

бъдат отстранени от края на думата. Можем да ги намерим и в нормативните граматики, в разделите – *образуване на X* (X – произволна част на речта), където се изброяват десетки начини за образуване на нови (по подразбиране родствени) думи от първичната, непроизводна дума. Формалният материал за това образуване, във вид на отделни буквени низове, в граматиките е класифициран по семантични стойности за отделните части на речта – напр. наставки за деятел, за означаване на процеси, на абстрактни свойства, на лица от женски пол, умалителни, увеличителни и т.н.

При опит да се осъществи автоматично еднократна сегментация на посочените от граматиките суфикси отдясно наляво от края на думата констатираме, че информативността на получените резултати не радва лингвистичния взор.

В това просто отрязване на буквени вериги, взети от нормативните граматики, лингвистичното ни знание регистрира два вида шум, който може да определим като *пресегментиране* и *недосегментиране*. Първото имаме ако отрежем *-ция* от *профанация* и получим стема *профана-*, а второто – ако отрежем същия низ от *нация* и получим *на-*.

Този шум трябва да се филтрира и премахне с проверки, най-елементарната от които е дали след отрязване на крайния низ полученото е единица, фигуриращa вече в речника. Тази проверка подобрява малко картината за случаи като *аванс*, *авансирам*, *авансов*, където сме сегментирали суфиксите *ирам* и *ов*, но я оставя непроменена за случаи като *върба, върбов, върбалак*, *върбак* – ако сме сегментирали традиционните деривационни суфикси като *-ов, -алак*, *-ак*. Повечето деривационни вериги с произвеждаща основа от същ. ж.р., а също и от глаголи (основната произвеждаща лема има реално окончание, за разлика от нулевото окончание в същ.м.р. в първия от двата примера по-горе) не могат да получат буквения състав на производящата лема чрез просто отрязване. За такива случаи в [Mikheev 1997] се въвежда понятието променлив суфикс, което свързва това, което се сегментира, и това, което го заменя, за да получим единицата, която ще търсим в речника (вж. 4.3.4) .

Следваща стъпка в изчистването на шума при сегментацията е отслабването на критерия за *остатък = речникова единица*, като заменим лемата като резултат със словоформа (което налага търсене в речника на словоформите, а не на лемите). Положението с *върбов* и *върба* няма да се подобри (тук не бихме могли да дефинираме X и Y в операцията *отрежи X – замени с Y*, като *-ов* и *-а*, тъй като отрязването на *-ов* води до друго съотношение *-ов* –> -#). Затова пък положението с *безпокой-ство* и *убий-ство* ще се подобри, ако приемем за

произвеждаща основа словоформата за повелително наклонение на глагола (отклонение, в случая простимо, тъй като в 450-те съществителни на *ство* в СФ-речника само в тези два случая имаме деривация от глагол).

Въпросът за избора на буквена верига за сегментация измежду многобройните цитирани в граматиките продуктивни словообразователни суфикси е съществен, поради голямото им количество и различната точност на формулирането и изброяването им в граматиките. Елементарни алгоритмични трикове като започване на обработката с най-дългата единица, съдържаща в себе си останалите (*-изация*, *-ация*, *-ция*) с идентична функция също могат да послужат за подобряване на правилата.

Отворен е въпросът за възможна сегментация на префиксални елементи, но префиксалното словообразуване по наше мнение е още по-трудно за описване като свързващо определен буквен състав с дадено значение, отколкото суфиксалното. Дда си представим само спектъра на значенията, които добавя префикса *пре-* към производящата основа, и далеч по-скромният списък от значения на суфикса *-ство*.

Суфикси, цитирани в нормативната граматика като редовни продуктивни суфикси, добавящи едно единствено значение (като умалителното значение на *-че*) също дават доста шум в резултатите от сегментацията. Работата е там, че човешкото знание извлича новото значение на суфикса само тогава, когато го схваща именно като суфикс – нещо прибавено към нещо, което вече има собствен смисъл – *корабче*, *ангелче* и т.н.

При производност, която е повече историческа (*врабче, менче*, *канче*), както и при формално съвпадане на този буквен низ с части от думи, в които той е неделим (*вече*, *тече*, *плаче*, *дуче*), картината се замъглява, макар и умалителното значение на *–че* да е преобладаващо в общия брой думи, съдържащи го формално в края си. В Табл. 24 се дава разпределението на единици с краен стринг *-че* по части на речта и граматичните им стойности. В Табл. 25 се разглежда разпределението на съществителните от среден род, за които стандартният прочит на значението на крайното *-че* е умалителност. Разпределението е направено по отношение на начините на присъединяване на този суфикс:

- чрез пряко залепване към стемата – речникова единица, напр. балон+че,
- чрез палатализационни промени в стемата, напр. кютук – кютуче,
- чрез други операции с различни механизми на промяна в стемата, но все пак смислово изводими като производни, напр. животно – животинче и
- неизводими от каква и да е стема, напр. дуче, *кече.*

Части на речта и граматични стойности	Брой	Процент	Примери
Съществ. ж.р. собств.	2	0,6%	Анче
Съществ. м.р. зват.ф-ма	5	1,4%	пътниче
Съществ. ср.р.	174	48,7%	балонче
Глаголи	156	43,7%	тече
Прилагателни	13	3,6%	овче
Наречия	6	1,7%	вече
Съюзи	1	0,3%	обаче
Всичко	357		

Табл. 24. Разпределение на речниковите единици с краен низ *-че* по части на речта и граматични стойности.

Начин на присъединяване	Брой	Процент	Пример
Директна конкатенация: стема + *че*	104	60%	куфар-че, букет-че
Палатализирана стема + *е*	34	20%	близнач-е, кач-е, облач-е
Смислово изводима производност по различни формални правила	25	14%	враб-че, зай-че, звън-че, кърма-че
Неизводими	10	6%	ке-че, ку-че, ду-че
Всичко	173		

Табл. 25. Разпределение на начините на присъединяване на суфикс *-че* към стемата на съществителни от среден род с резултатно умалително значение.

При цялата ненадеждност на изчислението за редовни правила на префиксалното словообразуване някои префиксални елементи все пак могат да бъдат сегментирани. Тук влиза преди всичко префикса *не-* (само при условие, че произвеждащата го дума фигурира в речника (няма да сегментираме *не-дей*, понеже няма *дей*), а също така и префиксите за степенуване *по-* и *най-* (формите за степени на сравнение – като част от парадигмата на прилагателното и като отделна лексикална единица – наречие).

Въпросните съображения, натрупани и организирани в алгоритъм за сегментация и проверка, ни дадоха следния резултат, засега междинен, тъй като подлежи на проверка и оценка както от човек, така и чрез търсещи процедури в големи текстови масиви, ръчно или полуавтоматично чрез подходящ интерфейс.

От речника на лемите, на брой 74 000, бяха произведени 28 000 стеми. Те образуват групи от производни думи, с дължина на групата от един елемент (т.е. без производни) до 64. Групите с дължина единица, т.е. без производни, са

12 000. Групите с дължина над 12 производни очевидно са признак на свръхсегментация – напр. 64 производни на *дел*, където са обединени значенията на *деля*, *дело*, *предел* и др.

В този си междинен вариант стемният речник представлява добър полуфабрикат за по-нататъшна обработка и производство на различни видове организирани списъци, които дават по-голяма семантична точност на резултатите от търсене, което цели единствено да установи тематична смислова близост. Справедливостта изисква да уточним, че понятието стемен речник в случая е това, към което се стремим, а не това, което сме произвели.

Първо, тъй като става дума за списъци от елементи без информация към тях, може да говорим по-скоро за списъци или глосарии. Глосарият в общия случай е списък от думи, с евентуално разширение – дефиниция на всяка единица, а също и посочване на връзка с други думи от списъка. В нашия случай става дума за минимална конфигурация на глосария, разглеждан като списъци от думи, обединени в групи.

Второ, докато не сме уточнили конкретно и функционално лингвистичното съдържание на понятието стема с оглед на поставена задача, бихме могли да наречем получената редица от думи – множество на квазистеми.

Независимо от названието и степента на завършеност на подобни продукти, очевидно е едно: деривационната морфология, чиито зависимости и резултати досега са използвани фрагментарно в статистически изследвания върху българския, а и други славянски езици, би трябвало да бъде систематизирана за нуждите на компютърната обработка на текстове поради неоспоримата си полезност. Има се предвид очевидната връзка между буквения състав на елементите, по-малки от дума, и изразяваните от тях значения и отношения, които много пъти са значително по-дълбоки и гранулирани, отколкото предлаганите от съвременните онтологични, концептуални и семантико-анотационни инструменти.

Тази насока, към задълбочаване на моделирането на лингвистичното знание, е още по-актуална за език като българския, с изразителни и точни словообразователни и формообразуващи инструменти на морфемно равнище.

Заключение

В това изложение, оставайки на морфологичното равнище и въоръжени с неговите средства, се опитахме да очертаем пътищата от Текста към Думата и от Думата към структурите на синтаксиса, семантиката и информационното съдържание. Повечето от тези пътища не са изминати докрай, някои са усвоени до половината, а някои са само набелязани.[48] В крайна сметка те всички водят до някакво лингвистично знание.

Труден и непосилен е пътят до това знание, ако се изминава по стария и изпитан начин – пеша, колкото и приятни изживявания да предлага той, както ги предлагат и натрупаните на бюрото езиковедски книги, чекмеджетата с листчета примери (та макар и в съвременна електронна форма – отделни текстови файлове).

Съвременният начин да се стигне по-далеч в този път е да се ползват по-модерни средства за придвижване, каквито съвременната компютърна лингвистика е създала в изобилие, подобно на модерната автомобилна индустрия.

Понякога, ползвайки тези модерни средства, можем да отминем живописна част от пътя, да не намерим най-пряката пътека и да не усетим сладостта от същината на познанието. Това е обаче жертвата, която даваме в името на скоростта и постигането на цел, която с традиционни средства е непостижима.

Затова често наблюдаваме разминаване между постановките на компютърната лингвистика и постулатите на съвременната лингвистика, дори и най-модерната, но предназначена за разбиране от носителя на езика, а не от компютъра. С това разминаване всеки лингвист, който се е посветил на компютърното моделиране на езиковите факти, би трябвало да се примири, без да го отминава. Всяко отклонение от езиковедските истини – не само като текст, а като разбиране на явленията, трябва да бъде честно регистрирано, а причините му – обяснени.

Противоположният подход – на замазване и скриване на тези отклонения, в името на модерните компютърни технологии, е царуване на еднооки в царството на слепите. А благородната цел е да се разпръсква светлината и да се сближават методите на двете науки.

---

[48] Тази констатация важи за съвременното състояние на компютърната лингвистика, но е десетократно по-вярна за изложението в тази книга, където голяма част от нещата в съвременното състояние на дисциплината са само споменати.

Защото за разлика от всички останали дисциплини, които се занимават с компютърно моделиране на различни клонове от науката, компютърната лингвистика не е само приложение на технологията, а е и средство за усъвършенстване и част от самата технология. Поради простия факт, че по-добро и качествено средство за комуникация от естествения език още не е изнамерено.

Ако това изложение е проляло лъч светлина върху поведението на думата в компютърния свят, авторът може да е само благодарен на тези, които го прочетоха докрай.

Литература

[Андрейчин и др. 1977]
Любомир Андрейчин, Константин Попов, Стоян Стоянов. Граматика на българския език, Наука и изкуство, София, 1977.

[Апресян 1966]
Апресян Ю. Д. Идеи и методы современной структурной лингвистики, Москва, 1966.

[Бояджиев и др. 1998]
Тодор Бояджиев, Иван Куцаров, Йордан Пенчев, Съвременен български език, Издателска къща „П. Берон", София, 1998.

[Димитрова и др. 1981]
Людмила Димитрова, Елена Паскалева, Ирина Ненова. Пакет програми за автоматично сегментиране на български текст. Системи и управление, № 3, 1981, с. 19-25.

[Зализняк 1977]
А. А. Зализняк. Грамматический словарь русского языка. Москва, Наука, 1977.

[Кръстев 1990]
Боримир Кръстев. Морфология на българския език в 187 типови таблици, София, 1990.

[Мельчук, Холодович 1970]
Мельчук И. А., Холодович А. А. К теории грамматического залога // Народы Азии и Африки. 1970. №4. С. 111-124.

[Осенова и Симов 2007]
Петя Осенова, Кирил Симов. Формална граматика на българския език. София, ИПОИ-БАН, 2007.

[Паскалева и др. 1981]
Людмила Димитрова, Елена Паскалева, Ирина Ненова. Автоматично сегментиране на българските словоформи. Автоматизирани системи за управление, № 4, 1981, с. 9-16.

[Паскалева 1982]
Елена Паскалева. О возможностях автоматического анализа болгарского текста без помощи словаря лексем. The Prague Bulletin of Mathematical Linguistics, 37, 1982, pp. 53-60.

[Паскалева 2002]
Елена Паскалева. Обработка русских и болгарских ресурсов в унифицированном формате. Восьмой международный симпозиум МАПРЯЛ'2002 „Теоретические и методические проблем, русского языка как иностранного в начале XXI века", В. Търново, 4-7 април 2002.

[Пашов 1966]
Петър Пашов. Българският глагол. I. Класификация. Видообразуване. Словообразуване. София, 1966.

[Попов и др.1998]
Димитър Попов, Кирил Симов, Светломира Видинска, Речник за правоговор, правопис и пунктуация, София, Атлантис, 1998.

[Правописен речник 1989]
Кратък правописен речник на българския книжовен език. София, Наука и изкуство, 1989.

[Русинов 1978]
Русин Русинов. Съществителни от общ граматичен род в съвременния български език. Помагало по българска морфология. Имена. София, Наука и изкуство, 1978.

* * *

[Alfred et al.2007]
Rayner Alfred, Elena Paskaleva, Dimitar Kazakov, Mark Bartlett. Hierarchical agglomerative clustering for cross-language Information Retrieval. International journal of translation, Vol. XX, 2007.

[BALRIC-LING 2000]
http://www.larflast.bas.bg/balric/index/index.htm

[Chomsky 1956]
Noam Chomsky. Three Models for the Description of Language. IRE Transactions on Information Theory. September, 1956.

[Church 1988]
Kenneth Church, A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, Proceedings of Second Conference on Applied Natural Language Processing, pp. 136-143, 1988.

[Cutting et al. 1992]
Doug Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A Practical Part-of-Speech Tagger. Proceedings of the Third Conference on Applied Natural Language Processing, April 1992.

[GATE 2005]
http://gate.ac.uk/ вж. също http://msh.univ-fcomte.fr/intex/downloads/Manuel.pdf

[Gaussier 1999]
E. Gaussier. Unsupervised learning of derivational morphology from inflectional lexicons. ACL workshop on Unsupervised Methods in Natural Language (ACL '99). Workshop Proceedings.

[Hafer & Weiss 74]
M. Hafer and Stephen Weiss. Word segmentation by letter successor varieties. Information Processing & Management, 10(11/12):371-385, 1974.

[INTERA 2003]
http://www.mpi.nl/INTERA/

[Kaplan 1975]
Kaplan, R. On process models for sentence comprehension. *Explorations in Cognition*, ed. by D. Norman and D. Rumelhart. San Francisco, 1975.

[Koskenniemi 1986]
Koskenniemi K. Two-level morphology: A general computational model for word-form recognition and production. Publication 11, University of Helsinki, Department of General Linguistics, Helsinki.

[Lindstedt 1984]
Lindstedt, Jouko. 1984. A two-level description of Old Church Slavonic morphology. *Scando-Slavica* 30, pp. 165–189.

[Manning&Schütze 1999]
Cristopher Manning & Heinrich Schütze. Foundations of Statistical Natural Language Processing", MIT-Press, 1999.

[Mikheev 1997]
A. Mikheev. Automatic Rule Induction for Unknown Word Guessing. Computational Linguistics, vol.23(3), pp. 405-423, 1997.

[Nakov 2003]
Preslav Nakov. BulStem: Design and Evaluation of an Inflectional Stemmer for Bulgarian. In Proceedings of Workshop on Balkan Language Resources and Tools (1st Balkan Conference in Informatics), Thessaloniki, Greece, November, 2003.

[Nakov &Paskaleva 2004]
Preslav Nakov, Elena Paskaleva. Robust Ending Guessing Rules with Application to Slavonic Languages. In Proceedings of the 3rd workshop on Robust Methods in Analysis of Natural Language Data (ROMAND), an International Workshop in Association with COLING'04, pp. 76-85, Geneva, August 29, 2004.

[Nakov&Paskaleva 2005]
Preslav Nakov and Elena Paskaleva. Dictionary, Statistical and WEB knowledge in Shallow Parsing Procedures for Inflectional Languages, In Proc. of International Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries, 2005, Borovets, Bulgaria, pp. 39-46.

[Osenova&Simov 2003]
Petya Osenova and Kiril Simov. Between Chunk Ideology and Full Parsing Needs. Proceedings of the Workshop on Shallow Processing of Large Corpora (SProLaC 2003), pp. 78-88.

[Paice 1990]
Paice, C. D. *Another Stemmer*, SIGIR Forum, 24: 56-61 (1990)

[Paskaleva et al. 1990]
Elena Paskaleva, Kiril Simov, Mariana Damova and Milena Slavcheva. The Long Journey from the Core to the Real Size of a large LDB. In: *Acquisition of Lexical Knowledge from Text*. Proceedings of a Workshop sponsored by SIGL of ACL, Columbus, Ohio, 1993, pp. 161-169.

[Paskaleva et al. 2002]
Elena Paskaleva, Galia Angelova, Mariana Yankova, K. Bontcheva, H. Cunningham, and Y. Wilks. Slavonic named entities in GATE. Technical Report CS-02-01, University of Sheffield, 2002.

[Paskaleva 2005]
Elena Paskaleva. Compilation and validation of morphological resources. In: S. Piperidis, V. Karkaletsis (Eds.), Proc. First Workshop on Balkan Language Resources and Tools, November 2003, Thessaloniki, Greece, a satellite event of the Balkan Conference on Informatics, pp. 68-74. 2005

[Porter 1980]
Porter M. F., An algorithm for suffix stripping, *Program*, **14**(3) :130-137. 1980,

[Saussure 2002]
F. de Saussure. De l'essence double du langage. Ecrits de linguistique générale, Gallimard, 2002

[Silberztein 1993]
Max Silberztein. Dictionnaires électroniques et analyse automatique de textes: le système *INTEX*. 240 p., Masson Ed., Paris

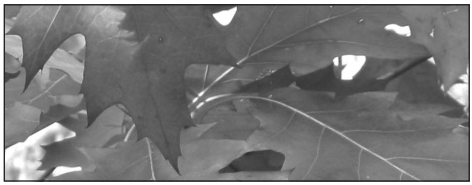
[Slavcheva 2003]
Milena Slavcheva. Some Aspects of the Morphological Processing of Bulgarian. In: Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003, Budapest, Hungary, pp. 71-77

[Simov et al. 90]
Kiril Simov, Galia Angelova and Elena Paskaleva. MORPHO-ASSISTANT: The Proper Treatment of Morphological Knowledge. Proceedings of the 13th Int. Conference COLING'90, Volume 3, pp. 455-457.

[Zdravkova & Petrovski 2007]
Katerina Zdravkova and Aleksandar Petrovski. Derivation of Macedonian Verbal Adjectives. RANLP, poster session, Borovetz 2007

КОМПЮТЪРНА МОРФОЛОГИЯ

Ресурси и инструменти



българска, първо издание

Автор: © Елена Паскалева

Графичен дизайн: © Николай Генов

Предпечат: Аспектрум ООД

ISBN: 978-954-92148-1-9

Издава: Институт за паралелна обработка на информацията, БАН

София, 2007